Seamless Multimedia over Heterogeneous Wireless Networks

by

Abolfazl Nazari

Submitted in total fulfillment of the requirement for the degree of

Doctor of Philosophy

Centre for Advanced Internet Architectures
Swinburne University of Technology
Australia
December 2013
Abstract

In recent years, we have observed a rapid proliferation of mobile devices. This trend and the explosion of mobile data traffic are expected to continue. For example, Cisco predicts that smart phones will generate 50 times more data traffic than they do today. Network operators are striving to meet the ever-growing demand for ubiquitous, cost-effective, and high-speed data services. They are integrating various radio technologies such as LTE, WiMAX, and WiFi into their existing network infrastructure. This integration enables mobile operators to combine the benefits of different radio technologies including higher data rates, cheaper service, better mobility support and better radio coverage.

Next Generation Mobile Network (NGMN) is a term used to express a vision for the future of wireless communication. NGMN is a standard IP-based platform which seamlessly integrates various radio technologies and provides users with ubiquitous real-time multimedia services as well as data applications. In this network environment, the Mobile Station (MS) is presented with a variety of connectivity options, each provided using a specific wireless technology. The MS continuously searches for available radio cells and seamlessly performs inter-technology handovers to keep its users Always Best Connected (ABC).

Despite many promises these multi-radio wireless networks, also referred to as heterogeneous networks, present many challenges including ensuring strong security, consistent Quality of Service (QoS) across different radio technologies, and efficient seamless mobility management. In this research we address the latter challenge.

Seamless mobility in multi-radio networks has proved to be particularly challenging due to the inherent limitations of Internet Protocol, non-cooperative wireless technologies, and fluctuations in radio channels. The original Internet specifications were developed for fixed hosts and the usage of mobile devices was not envisaged. As such, terminal mobility often causes connection teardown due to changes of the MS’s IP address. Although protocols such as Mobile IP and Session Initiation Protocol (SIP) may be used to safeguard data session continuity, high delay associated with renewing IP addresses and registering with candidate radio technology combined with high packet loss rates caused by randomly varying radio conditions often cause service interruptions.

To provide high-quality real-time multimedia services over heterogeneous networks, an efficient mobility management framework is required which minimises handover delay and packet loss. Many research solutions have been developed in the literature to improve the efficiency of seamless mobility management frameworks; however most have not been constructed to be compatible with the NGMN
framework. In this research, we contribute to these efforts by developing an NGMN-compatible mobility management framework and propose two novel mechanisms for reducing the delay and packet loss of inter-technology handovers. We provide the following contributions. First, an NGMN compliant seamless mobility management framework, called the Uninterrupted Proactive connection Transfer for IMS Mobility Enhancement (UPTIME) framework, is proposed. UPTIME follows NGMN requirements for an efficient mobility management solution including the separation of service and transport strata and signal and media function. This framework is based on the IP Multimedia Subsystem (IMS) which means the existing IMS architecture and mechanisms are used to deploy mobility management mechanisms. Second, the Pre-Registration for IMS Mobility Enhancement (PRIME) is developed for the purpose of handover delay optimisation. The PRIME handover preparation mechanism significantly optimises the handover preparation process and reduces the associated delay. Third, a Conservative Soft Handover (cSHO) scheme is proposed for reducing packet loss while conserving radio resource and battery power. A conventional solution for packet loss minimisation is using a Soft Handover (SHO). The proposed cSHO scheme uses the same approach but reduces the resource consumption of a conventional SHO by nearly 50%.
I wish to begin by expressing my gratitude to my principal supervisor Dr Philip Branch for his great support, useful guidance and constructive critiques of this research work. His efforts during the planning and development of this research have been much appreciated. I would also like to thank Dr Jason But and Professor Hai Vu for their engagement, technical advice and assistance in writing this thesis.

Special thanks should be given to Professor Grenville Armitage for creating a friendly research environment, the Centre for Advanced Internet Architectures (CAIA), filled with enthusiasm and joy. Advice given by Associate Professor Lachlan Andrew is also much appreciated. I would like to thank my other friends and colleagues in CAIA.

I also wish to express my deepest gratitude to my parents for their unconditional love and support which have always inspired and encouraged me to do my best. I would like to thank my sister and brothers as well. I am eternally grateful.

Last, but not least, my special thanks are extended to all my friends in Melbourne, Tehran, and around the world for their supports and encouragement which made this thesis possible.

Thank you!
Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma at any university. To the best of my knowledge this thesis contains no materials previously published or written by another person except where a reference is made.

Abolfazl Nazari
December 2013
# Contents

1 Introduction ........................................................................... 1  
  1.1 Research motivations and objectives .................................. 2  
  1.2 Contributions ................................................................ 6  
  1.3 Thesis organisation ........................................................ 7  

2 Next Generation Mobile Networks ............................................... 9  
  2.1 NGMN Overview ............................................................. 10  
    2.1.1 NGMN Characteristic and Requirements ......................... 10  
    2.1.2 NGMN Architecture .................................................. 12  
  2.2 WiMAX technology ......................................................... 16  
    2.2.1 History and Characteristics ........................................... 16  
    2.2.2 WiMAX Architecture ................................................ 18  
    2.2.3 WiMAX Network Entry Overview ................................. 20  
    2.2.4 Synchronisation, Ranging, and Capability Exchange ........... 21  
    2.2.5 Authentication and Security Key Exchange ..................... 22  
    2.2.6 Base Station Registration .......................................... 26  
    2.2.7 IP Configuration ....................................................... 26  
    2.2.8 QoS and Session Creation .......................................... 26  
    2.2.9 Mobility Support in WiMAX ........................................ 29  
  2.3 LTE Technology ............................................................... 30  
    2.3.1 LTE History and Characteristics .................................... 30  
    2.3.2 LTE Network Architecture ......................................... 31  
    2.3.3 LTE Network Entry Overview ...................................... 32  
    2.3.4 Cell Search and Selection ........................................... 32  
    2.3.5 Synchronisation and Random Access ............................ 33  
    2.3.6 Authentication Procedure in LTE .................................. 34  
    2.3.7 Location Update ...................................................... 35  
    2.3.8 Default Bearer ........................................................ 36  
    2.3.9 QoS Support and Dedicated Bearer Establishment ............ 37  
    2.3.10 Mobility Support in LTE .......................................... 40  
  2.4 IP Multimedia Subsystem (IMS) .......................................... 40
2.4.1 IMS Architecture ........................................... 41
2.4.2 IMS Registration Process ................................. 43
2.5 Mobility Management in NGMN ............................. 45
   2.5.1 Mobility Types, Scenarios, and Definitions .......... 45
   2.5.2 Mobility Management Procedure Overview ............ 48
   2.5.3 Network Discovery ..................................... 49
   2.5.4 Handover Preparation .................................. 52
   2.5.5 Handover Execution .................................... 53

3 Seamless Handover Control .................................. 57
   3.1 Introduction ............................................. 57
   3.2 The Challenge of Seamless Mobility ....................... 58
   3.3 Seamless Mobility Frameworks and Architectures .......... 62
   3.4 Handover Delay Optimisation: Cross-layer Approaches .... 68
       3.4.1 Network Discovery Optimisation ....................... 68
       3.4.2 Proactive Preparation Optimisation .................... 70
       3.4.3 Proactive Handover Execution ......................... 74
   3.5 Packet Loss Reduction: SHO Approaches ................... 75
       3.5.1 SHO for Network Discovery And Preparation .......... 75
       3.5.2 SHO for Handover Execution ........................ 77

4 UPTIME mobility framework ................................. 83
   4.1 Introduction ............................................. 83
   4.2 Design Considerations .................................... 84
   4.3 UPTIME System Architecture ................................ 88
       4.3.1 Handover Server ..................................... 90
       4.3.2 Media Duplication Function .......................... 91
       4.3.3 PoA Handover Agent .................................. 93
       4.3.4 Handover Adaptation Layer and MS protocol stack ... 95
   4.4 UPTIME Procedures Overview ............................... 100
   4.5 Handover Preparation with PRIME Mechanism ............... 103
       4.5.1 RAN Discovery Procedure .............................. 104
       4.5.2 CN Pre-Registration .................................. 106
       4.5.3 IMS Pre-Registration ................................ 109
       4.5.4 PoA Discovery and Selection .......................... 110
       4.5.5 PoA Pre-Registration ................................ 111
   4.6 PRIME Signaling Flow for LTE and WiMAX ................. 113
       4.6.1 LTE Handover Preparation ............................. 114
       4.6.2 Handover to WiMAX .................................. 120
5 Conservative Soft handover

5.1 Introduction .................................................. 125
5.2 Motivations, Rationale, and Design Principles ................ 126
5.3 cSHO Procedure and Signaling Flow .......................... 130
5.4 Interface Selection in cSHO ................................. 135
5.5 Extensions to SIP and MIH Protocols ......................... 139
  5.5.1 Proposed SIP Handover Header ......................... 139
  5.5.2 Proposed MIH Link Actions ............................ 142

6 Performance Evaluation ........................................... 145

6.1 Introduction .................................................. 145
6.2 Handover Delay Analysis ..................................... 146
  6.2.1 Introduction and definitions ............................ 147
  6.2.2 Handover Delay of the HHO Scheme .................... 149
  6.2.3 Handover Delay of Pre-Registration scheme ............ 150
  6.2.4 Handover Delay of the SHO Mechanism ................ 153
  6.2.5 UPTIME Handover Delay Analysis ....................... 154
6.3 Packet Delivery Analysis ..................................... 156
  6.3.1 Radio Channel Modeling ................................ 157
  6.3.2 Finite State Markov Chain Model ....................... 159
  6.3.3 Packet Error Rate (PER) of HHO ....................... 162
  6.3.4 Packet Error Rate (PER) of cSHO ...................... 163
  6.3.5 Packet Error Rate (PER) of SHO ....................... 165
  6.3.6 Proof for Given Formulae ................................ 165
6.4 Simulation Environment ........................................ 166
  6.4.1 OPNET Modeler Overview ............................... 167
  6.4.2 Network Level Simulation Architecture ................. 167
  6.4.3 UPTIME Node Models .................................. 170
  6.4.4 MS mobility pattern .................................... 174
  6.4.5 Shadow fading model ................................... 175
  6.4.6 LTE and WiMAX Model Assumptions and Parameters ..... 176
    6.4.6.1 WiMAX ............................................. 176
    6.4.6.2 LTE ............................................... 177
  6.4.7 Network Traffic Model .................................. 178
    6.4.7.1 Network and Handover Control Signaling ........ 178
6.5 Simulation Results ............................................ 179
  6.5.1 PRIME Handover Preparation Delay ..................... 179
  6.5.2 Accuracy of Developed Analytical Model ............... 181
  6.5.3 Packet Delivery Evaluation ............................ 183
7 Conclusion

7.1 Research Summary ................................................. 188
7.2 Contributions ...................................................... 190
   7.2.1 UPTIME Mobility Framework .............................. 190
   7.2.2 PRIME Handover Preparation Mechanism ............... 191
   7.2.3 cSHO Handover Execution Mechanism .................... 192
7.3 Limitations and Future Directions .............................. 193
## List of Figures

1.1 Ubiquitous network connectivity using hybrid (heterogeneous) wireless networks [1] ........................................... 3

2.1 Priorities for NGMN system characteristics [2] .................................................. 12
2.2 Separation of services from the transport network in NGMN as presented by ITU-T in [3] ................................ 13
2.3 A more detailed NGMN logical architecture presented by Knightson et al. in [4] ........................................... 14
2.4 NGMN physical architecture ................................................................. 15
2.5 WiMAX network reference model [5] ...................................................... 18
2.6 Detailed presentation of WiMAX logical architecture ............................. 20
2.7 WiMAX network entry process .......................................................... 21
2.8 Initial base station association in WiMAX technology ......................... 22
2.9 WiMAX authentication protocols[5] ...................................................... 23
2.10 WiMAX authentication procedure [5] .................................................. 25
2.11 Registration to the WiMAX base station ............................................. 26
2.12 DHCP procedure for IP address allocation in WIMAX ....................... 27
2.13 DHCP procedure for IP address allocation in WIMAX ....................... 28
2.14 LTE system architecture ................................................................. 31
2.15 LTE procedure for network attachment and session creation ............ 33
2.16 Contention-based Random Access procedure for initial access to the LTE channel ............................................... 34
2.17 LTE authentication procedure .......................................................... 35
2.18 Default bearer creation procedure ...................................................... 38
2.19 Dedicated EPS bearer creation procedure ........................................ 39
2.20 IMS architecture ............................................................................. 42
2.21 IMS registration process ................................................................. 44
2.22 Mobility scenarios based on coverage: 1) complementary coverage, 2) overlay networks, 3) hotspot ....................... 48
2.23 SIP handover execution using re-INVITE method ................................ 55
2.24 The SCC handover execution procedure ........................................... 56
3.1 Radio signal hysteresis solution to avoid the ping-pong effect . . . . . . 61
3.2 Scenarios for system architecture in interworking of wireless networks 63

4.1 UPTIME system architecture ........................................ 89
4.2 MDF architecture .................................................... 93
4.3 PHA architecture ..................................................... 95
4.4 MS protocol stack ................................................... 96
4.5 UPTIME procedure .................................................. 102
4.6 RAN Pre-Registration process ..................................... 112
4.7 UPTIME process overview for LTE handover preparation ............ 115
4.8 CN-PR process for LTE handover preparation ........................ 116
4.9 P-PR process for LTE handover preparation ........................ 119
4.10 CN-PR process for WiMAX handover preparation .................. 121
4.11 WiMAX P-PR signaling flow ..................................... 123

5.1 A vertical handover scenario to improve radio coverage ............. 127
5.2 Inter-technology handover in hotspots ................................ 128
5.3 cSHO procedure ..................................................... 132
5.4 cSHO handover completion or temporary halting .................... 135
5.5 Sampling SNR of radio interfaces .................................. 136
5.6 Three components of the received signal [6] ........................ 137
5.7 PER vs SNR curves for WiMAX [7] and LTE [8]. .................... 139
5.8 SIP handover header ................................................. 141
5.9 An example of SIP re-INVITE message with the handover header . 141
5.10 The $MIH\_Link\_Actions$ request message format .................. 143

6.1 Handover delay in traditional (a) HHO, (b) SHO, (c) Pre-Registration,
and (d) PRIME mechanisms ............................................. 148
6.2 Quantisation levels for FSMC model of IF1 .......................... 160
6.3 The Markov chain model for the combination of IF1 and IF2 states . 161
6.4 OPNET top-down approach in network simulation .................... 168
6.5 Dual-mode LTE/WiMAX MS model with the HAL module .......... 169
6.6 Dual-mode LTE/WiMAX MS model with the HAL module .......... 171
6.7 Configuration of HAL operation modes and required parameters .... 172
6.8 Configuration of HAL operation modes and required parameters .... 173
6.9 Configuration of HAL operation modes and required parameters .... 174
6.10 Mobility pattern of MS in simulation scenarios ..................... 175
6.11 PER vs SNR curves for WiMAX and LTE [7] and [8] ............... 182
6.12 PER of different handover schemes ................................ 183
6.13 The gain of SHO and HHO methods ................................ 184
6.14 Received VoIP packets for different handover schemes ............ 184
6.15 PER for HHO, SHO, and cSHO ........................................ 185
6.16 CDF for length of packet loss bursts .............................. 186
6.17 The radio transmission energy usage of all MSs .................. 187
6.18 The bandwidth usage of different schemes .......................... 187
**List of Tables**

2.1 The evolution of WiMAX standard ........................................... 17  
2.2 Comparison of handover control solutions ................................. 54  
4.1 MIH_SAP primitives used in UPTIME ........................................ 100  
4.2 Examples of the MIH Information Elements used in UPTIME ............ 106  
5.1 The proposed link actions for the MIH protocol ........................... 143  
6.1 Number of control messages for WiMAX handover ....................... 151  
6.2 Number of control message for LTE handover ............................ 151  
6.3 PR mechanism and the number of control messages for WiMAX hand-  
    over .................................................................................. 152  
6.4 PR mechanism and the number of control messages for LTE handover 153  
6.5 The number of control messages in an UPTIME handover to the  
    WiMAX technology .................................................................. 155  
6.6 The number of control messages in an UPTIME handover to the  
    WiMAX technology .................................................................. 156  
6.7 Network level parameters of simulation scenarios .......................... 170  
6.8 Parameters of mobility pattern .................................................. 175  
6.9 Attributes of the voice stream in the multimedia application .......... 178  
6.10 Attributes of the video component of the multimedia application ... 178  
6.11 Attributes of the video component of the multimedia application ... 179  
6.12 Handover preparation delay for UPTIME and pre-registration method 181  
6.13 Simulation and analytical results comparison for PER ................... 182  
7.1 Number of control messages required to prepare a radio connection .. 192  
7.2 Comparison of HHO, SHO, and cSHO performance and their resource  
    consumptions ....................................................................... 193
## Glossary

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3GPP</td>
<td>3rd Generation Partnership Project</td>
</tr>
<tr>
<td>4G</td>
<td>Fourth Generation</td>
</tr>
<tr>
<td>AAA</td>
<td>Authentication, Authorization, and Accounting</td>
</tr>
<tr>
<td>ABC</td>
<td>Always Best Connected</td>
</tr>
<tr>
<td>ACK</td>
<td>Acknowledgement</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interfaces</td>
</tr>
<tr>
<td>ARQ</td>
<td>Automatic Repeat Request</td>
</tr>
<tr>
<td>AS</td>
<td>Application Servers</td>
</tr>
<tr>
<td>ASN</td>
<td>Access Service Network</td>
</tr>
<tr>
<td>ASN-GW</td>
<td>Access Service Network Gateway</td>
</tr>
<tr>
<td>BER</td>
<td>Bit Error Rate</td>
</tr>
<tr>
<td>BS</td>
<td>Base Stations</td>
</tr>
<tr>
<td>CCDF</td>
<td>Complementary Cumulative Distribution Function</td>
</tr>
<tr>
<td>CH</td>
<td>Corresponding Host</td>
</tr>
<tr>
<td>CID</td>
<td>Connection ID</td>
</tr>
<tr>
<td>CN</td>
<td>Core Network</td>
</tr>
<tr>
<td>CN-PR</td>
<td>Core Network Pre-Registration</td>
</tr>
<tr>
<td>CSCF</td>
<td>Call Session Control Functions</td>
</tr>
<tr>
<td>cSHO</td>
<td>Conservative Soft Handover</td>
</tr>
<tr>
<td>CSN</td>
<td>Connectivity Service Network</td>
</tr>
</tbody>
</table>
DHCP  Dynamic Host Configuration Protocol
DNS  Domain Name System
E-UTRAN  Evolved Universal Radio Access Network
EAP  Extensible Authentication Protocol
EAP-AKA  EAP Authentication and Key Authorization
eNB  E-UTRAN NodeB
EPC  Evolved Packet Core
ETSI  European Telecommunications Standards Institute
FSMC  Finite State Markov Chain Model
GPRS  General Packet Radio Service
GUI  Graphical User Interface
HAL  Handover Adaptation Layer
HARQ  Hybrid Automatic Repeat Request
HHO  Hard Handover
HOS  Handover Server
HSPA  High Speed Packet Access
HSS  Home Subscriber Server
I-CSCF  Interrogating Call Session Control Functions
I-PR  IMS Pre-Registration
IE  Information Elements
IEEE 802.16  IEEE Standard Specification for WiMAX
IETF  Internet Engineering Task Force
IF  Interface
IK  Integrity Key
IMS  IP Multimedia Subsystem
ISC  IMS Service Control
ITU-T  International Telecommunication Union Telecommunication Standardization Sector
LTE  Long Term Evolution
MCS  Modulation and Coding Schemes
MDF  Media Duplication Function
MICS  Media Independent Handover Command Service
MIES  Media Independent Handover Event Service
MIH  Media Independent Handover
MIHF  Media Independent Handover Function
MIIS  Media Independent Handover Information Service
MIP  Mobile IP
MIPv4  Mobile IP Version 4
MIPv6  Mobile IP Version 6
MME  Mobility Management Entity
MRFP  Media Resource Function Processor
MS  Mobile Station
mSCTP  mobile Stream Control Transmission Protocol
NACF  Network Attachment Control Functions
NAI  Network Access Identity
NGMN  Next Generation Mobile Network
NGN  Next Generation Network
NSP  Network Service Provider
OFDM  Orthogonal Frequency Division Multiplexing
OFDMA  Orthogonal Frequency Division Multiple Access
P-CSCF  Proxy Call Session Control Functions
P-GW  Packet Data Network Gateway
P-PR  Point of Attachment Pre-Registration
PCC  Policy and Charging Control
PCRF  Policy and Charging Rules Function
PDF  Probability Density Function
PER  Packet Error Rate
PHA  Point of Attachment Handover Agent
PHY  Physical Layer of the OSI Model
PoA  Point of Attachment
PRIME  Pre-Registration for IMS Mobility Enhancement
PSTN  Public Switched Telephone Network
QoS  Quality of Service
RACF  Resource and Admission Control Function
RAN  Radio Access Network
RFC  Request for Comments
RSS  Received Signal Strength
S-CSCF  Serving Call Session Control Functions
S-GW  Serving Gateway
SA  Security Association
SCC  Service Continuity and Centralization
SCTP  Stream Control Transmission Protocol
SDP  Session Description Protocol
SFA  Service Flow Authorization
SHO  Soft Handover
SIM  Subscriber Identity Module
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIP</td>
<td>Session Initiation Protocol</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to Noise Ratio</td>
</tr>
<tr>
<td>TEK</td>
<td>Traffic Encryption Key</td>
</tr>
<tr>
<td>UMTS</td>
<td>Universal Mobile Telecommunications System</td>
</tr>
<tr>
<td>UPTIME</td>
<td>Uninterrupted Proactive connection Transfer for IMS Mobility Enhancement</td>
</tr>
<tr>
<td>URI</td>
<td>Uniform Resource Identifier</td>
</tr>
<tr>
<td>VoIP</td>
<td>Voice over IP</td>
</tr>
<tr>
<td>WiFi</td>
<td>Wireless Fidelity</td>
</tr>
<tr>
<td>WiMAX</td>
<td>Worldwide Interoperability for Microwave Access</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

In recent years, we have observed a rapid proliferation of mobile devices. It is predicted that in 2016, there will exist more mobile devices on earth than people, and that smart phones will generate fifty times more data traffic than they do today [9]. This explosion of data demand is pressuring mobile operators to enhance their network capacity by deploying faster wireless technologies, more base stations and wider radio channels. Currently, common radio technologies include Wireless Fidelity (WiFi), Worldwide Interoperability for Microwave Access (WiMAX), High Speed Packet Access (HSPA), and Long Term Evolution (LTE). However, the evolution of wireless technologies is still ongoing and the arrival of newer technologies such as the LTE-Advanced and WiMAX Release 2 is imminent.

To protect their capital investments, mobile operators often integrate new technologies into their existing network infrastructure resulting in a heterogeneous network environment. Heterogeneous networks are also deployed to combine the benefits of different radio technologies including higher data rates, cheaper service, better mobility support, and better radio coverage. Heterogeneity is also created in multi-operator mobile markets where operators implement different technologies, but rather than isolating their networks they allow inter-technology roaming. Despite many advantages, heterogeneity causes complexity, especially in providing seamless multimedia services.

Multimedia services are becoming increasingly important for mobile operators. The advent of relatively powerful smart phones with high-resolution cameras and large screens combined with high-speed radio connections is contributing to the rapid adoption of multimedia communication. While traditional cellular networks were capable of circuit-switched voice and low-rate video calls, modern wireless networks are expected to be able to provide rich and real-time multimedia services. Recognising this requirement, the 3rd Generation Partnership Project (3GPP) developed the IP Multimedia Subsystem (IMS) [10] as a standard platform for delivering multimedia
services over all IP wireless technologies. The IMS served as the basis for Next Generation Networks (NGN) developed by ITU-T to project its vision of wired and wireless networks integration. With their focus on mobile networks, the terms Next Generation Mobile Network (NGMN) [11] and Next Generation Wireless Systems (NGWS) [12] are used to express the same vision: a standard IP-based platform which seamlessly integrates various radio technologies and provides ubiquitous Always Best Connected (ABC) [13] services to mobile users. Despite promises of such a network, seamless mobility in multi-radio networks has proved to be particularly challenging due to the inherent limitations of Internet protocols, non-cooperative wireless technologies, and fluctuations in radio channels.

The fluctuation of radio channels affects the availability of the Radio Access Network (RAN). As a Mobile Station (MS) moves among base stations, new RANs appear and disappear. To remain ABC, the MS discovers the best available RAN and performs an inter-technology handover if necessary. However, the MS may not be able to react to sudden radio changes because the preparation of the target RAN and the migration of the IP-based multimedia session is too slow. The associated delay and packet loss is often beyond the QoS requirements of multimedia applications and this leads to service interruption.

Many research solutions have been developed in the literature to improve the seamless provision of multimedia service over heterogeneous networks. In this research, we contribute to these efforts by developing an NGMN-compatible mobility management framework and propose two novel mechanisms for reducing the delay and packet loss of inter-technology handovers. The proposed seamless mobility solution makes inter-technology mobility transparent to network users. As such, it enables mobile operators to integrate different wireless technologies into their network infrastructure without causing service interruption.

The rest of this chapter is organised as follows. In Section 1.1 we characterise heterogeneous networks and describe challenges that motivate us to develop a novel mobility management solution. In Section 1.2, we present our contributions in more detail. Section 1.3 explains the organisation of this thesis.

1.1 Research motivations and objectives

As untethered and ubiquitous access to the Internet becomes a daily need for users, mobile operators are finding new opportunities and facing new challenges. Network operators are employing modern IP-based wireless technologies to serve users and generate more revenue. The challenge is how to integrate these technologies seamlessly to satisfy various user requirements including high transmission rates, pervasive radio coverage, and seamless mobility when users are traveling at high
Providing pervasive Internet access using heterogeneous networks which combine different wireless technologies was considered in the late 1990s and early 2000s. Pahlavan at al. in [1] demonstrate an Internet access scenario through hybrid (heterogeneous) networks. In this scenario, as depicted in Figure 1.1, several WiFi, cellular, and satellite networks provide overlay radio coverage. The MS tries to connect to the wireless network with the shorter radio range to obtain higher data rates. As the MS travels and moves out of the coverage of the serving RAN, an inter-technology handover to a wider-range wireless network is triggered. This results in an upward vertical handover. Conversely, a downward vertical handover is initiated whenever the MS approaches a short-range Point of Attachment (PoA) such as a WiFi access point. Within a single Radio Access Network (RAN), horizontal handovers are carried out to manage terminal mobility.

This hybrid network scenario provides pervasive high-speed network connectivity. However, the optimality of radio connections was a further advancement proposed by Gustafsson and Johnson in the Always Best Connected (ABC) scenario [13]. In the ABC scenario, the MS constantly considers parameters such as service costs, security, QoS, and reliability to determine the user’s personal best option for network connectivity. As network conditions vary and the user switches between different applications, inter-technology handovers are required to keep users ABC.

Although vertical handovers are necessary to ensure consistent and optimal network connectivity, the original Internet protocol stack exhibits limited capabilities for mobility management. This protocol stack was designed for fixed hosts and the
usage of Internet-connected mobile devices was not envisaged \[14\]. Most notably, the IP address functions as both the host identifier and the location (serving network) identifier. When the MS changes its RAN, it usually obtains a new IP address identifying the new serving network. However, in the original Internet protocol, the task of mobility management was not assigned to any protocol layer and a change in IP address often results in connection teardown \[15\].

For managing a mid-session change of IP address, different protocols have been proposed. The Mobile IP protocol \[16\] operates in the IP (network) layer and allows the MS to register its newly obtained IP address with its home network. The home network is then able to tunnel data packets to the visited network. Taking a rather different approach, the Session Initiation Protocol (SIP) \[17\] functions in the application layer and informs the Corresponding Host (CH) that the MS’s IP address has been changed. The CH then directly communicates with the MS. The Mobile IP and SIP protocols are considered the leading mobility management solutions \[18\]. However, they both suffer from high handover delays.

In vertical handovers, the MS is required to discover the optimum RAN, prepare a radio connection, and execute a handover using protocols such as SIP and Mobile IP. As the number of candidate RANs increases and more handover decision criteria are considered, the network discovery delay increases. The handover preparation process is also time-consuming and commonly includes tasks such as time and frequency synchronisation with the target PoA, basic capability exchange, mutual authentication between the MS and the target network, IP address configuration, and bearer creation for QoS purposes. Similarly, the handover execution process requires several message exchanges between the MS and the target network resulting in excessive overall handover delay. This delay is well beyond the QoS requirements of real-time multimedia applications such as Voice over IP (VoIP) and video calls which can only tolerate a few hundreds of milliseconds of packet latency \[19\].

Real-time multimedia applications are important services provided in Next Generation Mobile Networks (NGMN). Since the access network in an NGMN is heterogeneous and consists of different RANs, a seamless mobility management solution is required to ensure interruption-free roaming between these RANs. However, achieving seamless handovers is specifically challenging in NGMN. In NGMN, the service stratum is considered to be independent from the underlying transport stratum \[20\]. As an independent network, the service stratum contains its own subscriber information and separately authenticates NGMN clients. Before initiating a multimedia session, an NGMN client registers with both the transport and service strata. This independent authentication process imposes an additional delay to an already significantly high handover delay.

High inter-technology handover delays combined with varying characteristics of
radio channels result in excessive packet loss and service interruption. Due to effects such as shadow fading, radio channels change rapidly and as a consequence the received signal strength varies. In this condition, the MS often does not have enough time to prepare the new radio link and observed link disconnection on its current interface. In cases where the MS is able to connect to the target network in a timely fashion, the radio condition over the target link may deteriorate and packet loss may occur. However, due to the high handover delay and signaling costs, the MS is not able to return to the old RAN. Therefore, frequent handover or the notorious ping-pong effect must be avoided [21].

A body of research in recent years has resulted in many proposed solutions and approaches to the problem of seamless mobility. The approaches used in these solutions include reducing the number of exchanged messages, cross-layer optimisation of the handover process, and Soft Handover (SHO) preparation and execution. However, most of previously adopted solutions are not suitable for NGMN or are not optimal in delay and packet loss reduction.

For example, combining the authentication processes of the transport and service strata has been proposed to reduce the number of exchanged messages and minimise handover delay [22]. However, the NGMN requirement of independence of transport and service strata is not considered.

Moreover, solutions based on proactive handover preparation by predicting an imminent handover have been suggested [23, 24, 25, 26], but their performance is not optimum and can be significantly improved.

Finally, Soft Handover (SHO) schemes have been adopted to minimise packet loss and remove service interruption by simultaneously using several radio interfaces to send and receive data packets [27, 28, 29, 30]. It has been shown that receiving duplicated media streams significantly improves the radio connection reliability and reduces the Packet Error Rate (PER) [31, 32]. However, the drawback is the excessive consumption of radio resources and battery power required to send duplicated packets.

Considering the importance of seamless mobility in the delivery of interruption-free multimedia services in NGMN, an efficient seamless mobility management framework is required which satisfies the following requirements:

- NGMN compliance: The mobility framework should be developed to comply with the NGMN architecture and its design principles. The relevant architectural principles include the independence of the service stratum from the underlying transport network, the separation of control functions from media forwarding entities, the division of the transport network into the core network and radio access network, the utilisation of open standard protocols, and the protection of users' security and privacy [33]. NGMN-compliance also requires
compatibility with the IP Multimedia Subsystem (IMS) which serves as the basis for the service stratum of the NGMN.

- Optimum performance for real-time multimedia applications: These applications have strict QoS requirements including low packet latency and PER. The seamless handover solution should combine various techniques to reduce handover delay as much as possible and avoid packet loss.

- Minimum resource consumption: The excessive resource consumption of the SHO mechanism limits its applications in wireless networks. The SHO scheme can only be applied in cases where the MS has abundant battery power and radio cells have more than enough radio resources, a scenario which is not common. A seamless mobility management solution is required which is more efficient in consuming radio resources and battery power.

1.2 Contributions

In this research, we focus on seamless mobility management for real-time multimedia application in heterogeneous network environments. We develop an NGMN-compliant mobility management framework with two novel mechanisms to reduce handover delay and packet loss. In particular, we provide the following contributions:

- An NGMN compliant seamless mobility management framework called the Uninterrupted Proactive connection Transfer for IMS Mobility Enhancement (UPTIME) framework: The UPTIME framework is a mobility management architecture which follows the NGMN requirements including the separation of service and transport strata. It is IMS-compliant which means the IMS architecture and mechanisms are used to deploy the mobility management framework. We introduce a Handover Server (an IMS application server) which assists the MS in performing inter-technology handovers. The UPTIME framework includes many features including interaction with authentication servers, coordination with QoS provisioning frameworks such as the Policy and Charging Control (PCC) framework, and interaction with the underlying transport network for implementing certain handover execution functions such as packet duplication and buffering.

- The Pre-Registration for IMS Mobility Enhancement (PRIME) for handover delay optimisation: The PRIME handover preparation mechanism optimises the handover process based on the concept of separating radio transmission functionalities from network-connectivity tasks. It takes advantage of the
NGMN architecture in which the network is divided into the Core Network (CN) and the Radio Access Network (RAN). In the PRIME mechanism, the MS is able to pre-register with the CN and perform radio-independent handover preparation tasks long before the actual handover takes place. The result is a significant reduction in handover delay compared with previous pre-registration schemes.

- A Conservative Soft Handover (cSHO) scheme for reducing resource consumption: The cSHO scheme reduces the resource consumption of a conventional SHO by nearly 50%. This resource conservation is achieved by a process of packet duplication in the network and buffering at the current and target Points of Attachment (PoA). Since both PoAs receive the duplicated media stream, the MS can quickly activate the network interface with the better radio conditions. The result is a significant Packet Error Rate (PER) reduction when compared with a conventional hard handover. Because only one interface is active at each given time in the cSHO scheme, battery power and radio resources are conserved.

By utilising the PRIME and cSHO mechanisms in the UPTIME mobility framework, seamless handovers over heterogeneous wireless networks become possible. The handover delay is significantly reduced by the PRIME mechanism while the PER is minimised using the cSHO method. Through simulation and mathematical analysis, we confirm the viability of the proposed solutions.

1.3 Thesis organisation

This thesis is organised into seven chapters. In Chapter 2, we provide a brief introduction to modern wireless technologies. We introduce NGMN concepts and its network architecture approach. We then overview two mobile technologies, LTE and WiMAX, which are likely to be used in the transport network of the NGMN. We describe the signaling flows for basic tasks such as user authentication and IP address assignment. These signaling flows are used later to demonstrate the effectiveness of the PRIME mechanism in reducing the handover delay. In introducing these mobile technologies, we also point out the architectural paradigm of dividing wireless networks into the CN and the RAN. This network architecture is used in the design of the PRIME mechanism and the reduction of the handover delay. In Chapter 2, we also introduce the IMS network architecture and basic procedures. Finally, we overview the mobility management procedure in NGMN.

Chapter 3 contains our literature review on seamless mobility management in heterogeneous networks. We overview the challenge of mobility management in het-
heterogeneous networks, discuss the existing solutions for seamless mobility, and identify their shortcomings. It is revealed that previously adopted approaches are either not NGMN-compliant, not optimal in delay reduction, or extravagant in resource consumption.

In Chapter 4, we present our UPTIME mobility framework and explain the overall system architecture. We then describe the proposed PRIME mechanism for reducing handover preparation delay. As two examples for UPTIME operation, we demonstrate the signaling flow for typical handovers to LTE and WiMAX technologies.

In Chapter 5, the cSHO mechanism is proposed and its procedure is described. Since cSHO is independent of the underlying wireless technology, the signaling flow is the same for both the LTE and WiMAX technology. In Chapter 5, we also discuss the required extensions to the standard Media Independent Handover (MIH) and SIP protocols to realise the PRIME and cSHO mechanisms.

In Chapter 6, we analyse the effectiveness of the proposed mechanisms. The PER performance of the cSHO mechanism is compared analytically with each of the SHO and HHO schemes. The accuracy of the analytical model and the results obtained are confirmed in simulation scenarios developed using the OPNET Modeler [34]. Chapter 6 also includes our analysis and simulation results for handover delay evaluation.

Finally, in Chapter 7 we conclude the thesis and present some possible future work.
Chapter 2

Next Generation Mobile Networks

In this chapter, we provide an overview of Next Generation Mobile Networks (NGMN) which is expected to provide real-time multimedia services as well as data applications over heterogeneous wireless networks. We begin by outlining NGMN characteristics, requirements, and architecture. In the NGMN architecture, a semi-independent service delivery platform, based on IP Multimedia Subsystem (IMS), runs over a transport network which may consist of different radio technologies. The LTE and WiMAX technologies are expected to form the foundation of NGMN transport network.

The NGMN outlines an architecture but leaves the issue of mobility management largely undefined. Different mobility management frameworks and mechanisms can be employed to provide seamless inter-technology handovers, one of the fundamental requirements for NGMN. The mobility management framework must be optimised based on NGMN characteristics and its radio technologies. In this thesis we develop such a framework which makes use of existing protocols and functionalities within the NGMN service platform and radio networks to deal with mobility management. In this chapter we describe IMS, LTE, and WiMAX procedures that are important in our seamless mobility management framework.

This chapter is organised as follows. In Section 2.1, we provide a background on NGMN, its characteristics and requirements, and its functional architecture. In Section 2.2 and 2.3 we overview the WiMAX and LTE technologies. We present various procedures including network entry, authentication, and QoS setup. These procedures are used later to optimise handover delay by reducing the number of signaling messages and analysing the effectiveness of our proposed mobility management mechanisms.

In Section 2.4, we describe the IMS architecture and procedures which have important implications for mobility management in NGMN. Finally, in Section 2.5 we provide a background on different aspects of mobility management in NGMN and we describe common handover control solutions such as Mobile IP and Session
2.1 NGMN Overview

The International Telecommunication Union - Telecommunication standardization (ITU-T) sector in Recommendation Y.2001 [20] defines the Next Generation Network (NGN) as a packet based network capable of utilising multiple broadband access networks to provide different telecommunication services. In this network, service provisioning is independent of the underlying transport technologies. Multiple wired and wireless technologies may be utilised to provide users with ubiquitous and unfettered access to different services and service providers. These concepts which are included in the NGN definition are also referred to as “generalised mobility”, one of the most important aspects of this network.

The above definition presented in ITU-T recommendations was published in 2004. However, the concept of NGN was developed in the late 1990s [35]. Motivations for advancing towards NGN include the need for market deregulation and operator competition, a surge in data traffic, strong demand for multimedia applications, and the need for general mobility. In a response to these motivations, in 2002 the ITU-T formed a study group to prepare recommendations on NGN.

The term Next Generation Mobile Network (NGMN) was originally proposed by a group of international mobile operators and vendors who formed the NGMN Alliance in 2006 in order to influence standards organisations and technology providers [36]. An NGN with a focus on mobile communications is commonly referred to as a Next Generation Mobile Networks (NGMN) [11]. NGMN is envisaged as a mobile network capable of integrating different radio technologies under a common IP-based platform to provide both data and multimedia services. Other terms such as the Fourth Generation (4G) mobile networks, Next Generation Wireless Networks (NGWN) [37], and Beyond 3G (B3G) are also used in the literature to refer to a similar vision [11].

2.1.1 NGMN Characteristic and Requirements

ITU-T in the Recommendation Y.2001 [20] outlines the main requirements and characteristics of NGN. Some of the important characteristics of NGN are:

- Packet-based transport network: The NGN is a purely IP-centric network. The circuit switch part of legacy networks such as Public Switched Telephone Network (PSTN) and GSM is replaced with a packet switched network.

- Separation of services and transport: This is one of the fundamental characteristics of NGN which ensures service and transport strata can be offered
separately and evolve independently. As such, NGN architecture requires a clear separation between service control functionalities and media handling nodes such as routers and media gateways.

- Support for a wide range of services including real-time and streaming multimedia applications.
- Support of high data rates access networks with better QoS support.
- Support of multiple access technologies: The access network is envisaged to be heterogeneous comprising different radio access technologies.

The NGMN imposes a number of requirements on the service platform [20]. The service platform should be extended to support telephony and multimedia services. It should provide open Application Programming Interfaces (API) and required proxy servers to enable third parties to create and develop services. These APIs are important because it is expected that many network operators will cooperate with third-party developers to create new services and remain competitive [38]. The service platform is also required to support the provisioning of services over different access networks and roaming. Moreover, it must contain mechanisms to support user presence and service customisation based on user availability. As we discuss later in this chapter, the IMS supports all of these requirements.

In the transport network, radio technologies are required which provide higher data rates, lower latencies, and better support for QoS. Currently, data rates of around 100 Mbps and latencies of less than 30 milliseconds are what mobile operators are expecting from the NGMN [36]. To satisfy the growing user demand for faster network access services, new radio technologies may be introduced. The transport network is required to include mechanisms for integrating these technologies into existing network infrastructure.

The NGMN Alliance in [2] outlines mobile operators expectations from NGMN and their priorities, depicted in Figure 2.1. In this figure, the inter-working term refers to the coexistence with legacy networks while seamless mobility refers to the service continuity without interruption when roaming within a wireless network and between different radio technologies. The high priority of seamless mobility stresses the importance of an efficient mobility management framework.

In traditional mobile networks, seamless mobility is limited to one technology. The MS is able to handover from one radio cell to another without experiencing service degradation. NGMN however requires generalised mobility. ITU-T in [20] defines generalised mobility as the ability to use different access technologies at different locations to access and manage the same set of services. NGMN does not necessitate seamless handovers in generalised mobility. However, NGMN is
envisaged to support seamless mobility to allow the integration of disparate radio technologies [11]. In this thesis we focus on developing a mobility management framework which supports seamless handovers while following NGMN requirements.

### 2.1.2 NGMN Architecture

In designing a seamless mobility management framework and optimising handover mechanisms, the NGMN functional and physical architectures must be considered. The NGMN architecture has been designed to satisfy various requirements such as the provision of rich multimedia services over multiple access networks, the separation of service and transport functionalities, and consistent end-to-end QoS. By following NGMN architectural principles, we ensure the applicability of our mobility management framework and its interoperability with other NGMN components. In our review of the architecture we start with a high-level view of the NGMN architecture which represents the main functionalities and we continue towards a more detailed architecture which contains more logical functions and physical entities.

To emphasise the separation of service provision from the underlying transport network, NGMN functionalities are divided into two strata, the transport stratum and the service stratum [3]. This logical view of NGMN is depicted in Figure 2.2. The transport functions are only concerned with the transmission of data between different entities. The Internet Protocol (IP) is the main protocol used for this purpose. On the other hand, the service stratum includes functions for controlling and managing a variety of services including voice telephony, multimedia services, and data applications. It contains mechanisms to authorise and control user ac-

Figure 2.1: Priorities for NGMN system characteristics [2].

![Figure 2.1: Priorities for NGMN system characteristics [2].](image-url)
cess to services provided, and interfaces with application developers to create new applications.

Knights et al. in [4] present a more detailed logical architecture of NGMN which is depicted in Figure 2.3. Besides including a more detailed view of logical functions, this architecture also highlights interactions between the service and transport strata. In the transport stratum, Access functions represents the radio technologies which connect the MS to the network. The data traffic from access networks is aggregated and transmitted to other networks (including the Internet) using Core transport functions.

Aside from transport functions, the transport stratum includes logical entities for controlling network access and managing resources. The Network Attachment Control Functions (NACF) authenticate the MS and create security associations with the MS when it initially attaches to the network or when it performs an intertechnology handover. Meanwhile, the Resource and Admission Control Function (RACF) is responsible for reserving adequate resources and ensuring a consistent QoS in the transport network. In 3GPP terminology the RACF is called the Policy and Charging Rules Function (PCRF).

The RACF is one of the interaction points between the service stratum and the transport stratum. During the establishment of a data session, the service control functions determine the required QoS parameters. The RACF is then requested to authorise the required resources and instruct other network nodes to create bearers. To make policy decisions, the RACF may use user profiles stored in transport stratum databases.

The transport stratum also includes Gateway functions which enable an NGMN network to connect to legacy circuit-switch networks such as PSTN and non-compatible data networks. These gateways can be controlled by the service stratum. For example, to create a voice call session between an NGMN user and a PSTN subscriber, the service stratum may invite a PSTN gateway node into the session.

In the service stratum, the Service Control Functions initiate data and multi-

Figure 2.2: Separation of services from the transport network in NGMN as presented by ITU-T in [3]
Figure 2.3: A more detailed NGMN logical architecture presented by Knightson et al. in [4]

media sessions based on requests received from the MS or Application Functions (application servers). Some of the application servers are controlled by the network operator and usually provide standard services such as the Multimedia Messaging Service (MMS) and the Short Messaging Service (SMS). However, in NGMN, network operators are not the only service providers. Third-party application developers may create new services and implement them in third-party controlled application servers. The servers connect to the service control functions using the standard NGMN APIs.

The service stratum also includes a Service Authentication and Authorization Function (not shown in Figure 2.3) which authenticates users in the service stratum [39]. In order to use multimedia services, MSs are required to be authenticated in both the transport and service strata. In the general scenario, the authentication process in the service stratum is independent from transport network authentication. This independence causes more flexibility in network implementation and service provisioning. It enables the transport and service networks to be implemented and operated by different operators. Each operator can implement a specific authentication method and maintain its own user profiles. The implication of two independent authentication processes on mobility management is discussed in Chapter 3.

The independence of the service stratum from the underlying transport stratum is not limited to the authentication procedure. The service control platform is also
transport-ignorant meaning multiple transport networks can be used to access the offered services. As a result, the transport network can evolve with newer radio technologies without altering the service stratum. Likewise, the service stratum can evolve to include new service features and capabilities without affecting the underlying transport stratum.

Figure 2.4 shows a high-level view of the NGMN physical architecture. In the transport stratum, multiple wireless networks may be implemented and each network may utilise different radio technologies such as LTE, WiMAX, and WiFi. In the general case, each wireless network is divided into the Radio Access Network (RAN) and Core Network (CN) parts. The RAN section of the network contains radio transmission functionalities while the CN section includes other mechanisms such as network access control and IP configuration. This architectural paradigm is usually followed in modern wireless technologies such as LTE and WiMAX.

In some cases, one CN may be implemented to support several RANs. For example, the LTE CN supports non-3GPP access networks such as WiMAX and WiFi [40]. This situation is common when one network operator implements more than one radio technology to exploit the benefits of each technology. For example, a cellular network operator may choose to implement WiFi access points to offload data traffic from its main LTE network. In this case, the operator can implement only one CN to reuse its authentication and resource management servers for the complementary RAN (the WiFi network).

In the service stratum, the IP Multimedia Subsystem (IMS) is usually deployed as the service control platform. The IMS is the standard framework for providing
multimedia services over wireless networks. Different wireless technologies can be
used to connect users to the IMS network and IMS is hence access ignorant. In
subsequent sections we describe LTE and WiMAX as modern wireless technologies
which are expected to be used in NGMN transport stratum. We then discuss the
IMS.

2.2 WiMAX Technology

In this section, we first provide an overview of the WiMAX network architecture.
The WiMAX architecture follows the typical NGMN paradigm where network func-
tionalities are separated into a RAN and a CN. We then describe WiMAX procedures
which are performed during the initial network entry or when the MS executes a
handover to a WiMAX network. In Chapter 4, we use the WiMAX network archi-
tecture and its procedures to develop our mobility management framework and
analyse its performance.

2.2.1 History and Characteristics

Worldwide Interoperability for Microwave Access (WiMAX) is a wireless technology
designed by the IEEE. The original WiMAX standard (IEEE 802.16) first appeared
in 2001 [41]. The aim was to define a technology for long distance and high frequency
radio links. The original IEEE 802.16 was designed for radio frequencies of 10 to
66 GHz with a maximum data rate of 120 Mbps [42]. However, in these frequencies
the standard only supported radio links with Line of Sight (LOS). This limitation
made the standard unsuitable for use as an access network of ISPs in urban areas.
To address this shortcoming, the standard was revised in 2003 to support non-LOS
transmissions.

The modified standard is referred to as the 802.16a standard. This new standard
utilised the Orthogonal Frequency Division Multiplexing (OFDM) mechanism for
non-LOS radio links. It also extended the supported radio frequencies to the 2GHz
to 11GHz range. In order to make the IEEE 802.16a more suitable for RAN ap-
lications, further modifications were introduced and IEEE 802.16-2004 (also called
the 802.16d standard) was created [43].

Despite performance improvements, two major shortcomings hindered the adop-
tion of the 802.16 standard by vendors and service providers. First, only fixed
access scenarios were supported; mobile applications were not envisaged. Second,
the standard specified too many radio profiles (a set of parameters such as channel
bandwidth, modulation, and coding) which was a major concern for implementa-
tion and interoperability of products [44].
Table 2.1: The evolution of WiMAX standard

<table>
<thead>
<tr>
<th>Standard</th>
<th>Published date</th>
<th>Main features</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE 802.16</td>
<td>2001</td>
<td>Fixed access, LOS links, 10 to 66 GHz range</td>
</tr>
<tr>
<td>IEEE 802.16a</td>
<td>2003</td>
<td>Fixed access, non-LOS links, 2 to 11 GHz range</td>
</tr>
<tr>
<td>IEEE 802.16d</td>
<td>2004</td>
<td>Fixed access, superseding previous standards</td>
</tr>
<tr>
<td>IEEE 802.16e</td>
<td>2005</td>
<td>Mobile access, MIMO, 2 to 11 GHz range</td>
</tr>
<tr>
<td>IEEE 802.16m</td>
<td>2011</td>
<td>Up to 1 Gbps for fixed access and 100 Mbps for mobile users</td>
</tr>
</tbody>
</table>

To overcome the first shortcoming, the IEEE improved the 802.16 standard in 2005 and published the IEEE 802.16e specification, also known as 802.16-2005 or Mobile WiMAX. The new standard improves the non-LOS penetration by using radio techniques such as Hybrid Automatic Repeat Request (HARQ), Adaptive Antenna System (AAS), and Multiple Input Multiple Output (MIMO) [42]. These enhancements improved network coverage, bandwidth usage, and frequency reuse, which are important factors in mobile applications.

The mobile WiMAX technology has the following characteristics:

- Data rates of 63 Mbps in download and 28 Mbps in uplink directions [42],
- Support for mobility speeds of up to 125 km/h [6],
- Support of five QoS classes for real-time VoIP, multimedia, and data applications,
- Flexible frequency spectrum assignment and radio channels width of 1.25 MHz to 20 MHz,
- Efficient management of power consumption by introducing sleep and idle operation modes.

The enhancement of the WiMAX technology continued with the IEEE 802.16m [45] published in 2011. This standard is also referred to as WiMAX Release 2. Some of the new features of IEEE 802.16m include lower transmission delays through the introduction of subframes, better support of MIMO, support for radio channels wider than 20 MHz, support for femtocells, increased VoIP capacity, and improved support for handover schemes [45]. The IEEE 802.16m standard provides download data rates up to 100 Mbps for mobile users and 1 Gbps for fixed and nomadic applications. These parameters satisfy the requirements set by ITU for IMT-Advanced technologies. Table 2.1 summarises WiMAX evolution in recent years.

In this work, we use the IEEE 802.16e specifications because at the time of conducting the research the newer standard was not published yet. However, the devised handover control mechanisms can still be applied to 802.16m networks.

To ensure the interoperability of 802.16 products and promote standard, the WiMAX Forum was formed in 2001 [46]. the WiMAX Forum certifies products
which implement selected profiles of the IEEE 802.16 standards. The first products based on 802.16d standard were certified in 2006 and can be used in fixed wireless access applications [6].

2.2.2 WiMAX Architecture

Beside ensuring product interoperability, the WiMAX Forum also provides specifications for the WiMAX network architecture. In a similar manner to the 3GPP, the WiMAX Forum publishes its specifications in different releases. WiMAX Release 1 was developed based on IEEE 802.16e. At the time of writing, the latest published specification is WiMAX release 2 which is based on the newest standard, IEEE 802.16m. In this research, we use the specifications of the WiMAX release 1.5 which enhanced the original WiMAX release.

Figure 2.5 shows the WiMAX network reference model as defined by the WiMAX Forum [5]. This model is the logical view of the WiMAX architecture and defines functional entities and reference points which determine the way WiMAX entities communicate with each other. As can be seen, the WiMAX network is divided into the Access Service Network (ASN) and the Connectivity Service Network (CSN). The ASN only contains functionalities related to the provision of reliable and efficient radio connectivity. These functionalities include the following:

- Implementation of the PHY and MAC layers of the WiMAX technology.
- Radio resource management.
- Intra-ASN mobility management.
- Relaying signaling between the MS and the CSN for Authentication, Authorisation, and Accounting (AAA) or IP configuration purposes.
The ASN is generally independent of the CSN and is owned and managed by an entity called the Network Access Provider (NAP). The NAP might implement the ASN to provide a connectivity service to more than one NSP. This means subscribers of more than one operator can connect to the Base Stations (BS) of an ASN.

The CSN, on the other hand, is owned and managed by an entity called the Network Service Provider (NSP). The CSN mostly contains functionalities required to provide IP connectivity to the data networks such as the Internet, or to service platforms such as the IMS. The CSN has the following responsibilities:

- Authenticating subscribers and authorising different services (AAA).
- IP address assignment and providing other IP configurations.
- Policy and Charging Control (PCC) and QoS implementation in the core network.
- Supporting inter-ASN mobility and inter-CSN roaming.
- Providing IP connectivity to the Internet and other service providers.

The separation of the CSN from the ASN is an architectural paradigm which is also observed in other modern technologies such as LTE. This paradigm reduces the network complexity because the complexity of the radio connectivity is contained within the RAN and the effect on the CN is minimised. It also allows radio technologies to evolve without changing the CN.

A more detailed logical architecture of the WiMAX technology can be observed in Figure 2.6 [5]. As can be seen, radio resource management functions are limited to the ASN while functions for authentication, QoS and policy control, inter-ASN mobility, and IP configuration processes involve both the ASN and the CSN.

In WiMAX Release 1.5, the ASN functions are located either in the Base Station (BS) or the ASN Gateway (ASN-GW) [5]. This architecture was previously referred to as Profile C of WiMAX Release 1. The BS is the network node providing WiMAX link layer (L2) connectivity to the MS through the implementation of the WiMAX PHY and MAC layers. The ASN-GW is the aggregation node for control signaling in the ASN domain (among BSs), between the ASN and the CSN, or among different ASNs.

In the WiMAX security architecture, the BS functions as a relay between the MS and the AAA server in the CSN. The ASN-GW functions as the authenticator and key distributor and receives security keys from the AAA server. For QoS and policy control, the BS contains the Service Flow Management (SFM) which creates the radio bearer based on instructions received from the Service Flow Authorization (SFA) module located in the ASN-GW. If the dynamic Policy and Charging Control
(PCC) framework is implemented, the SFA contacts the PCRF in the CSN for QoS decisions. In the radio resource management framework, the BS includes the Radio Resource Agent (RRA) which measures local radio resource indicators. It then reports to the Radio Resource Controller (RRC) which is usually located in the ASN-GW and holds a regional view of radio resources.

The intra-ASN mobility management process involves Data Path, Handover, and Context functions which are responsible for data transmission, handover control, and context storing and transferring.

In the following sections we describe various WiMAX procedures which are relevant in developing and analysing the UPTIME mobility framework.

### 2.2.3 WiMAX Network Entry Overview

Figure 2.7 depicts the process performed by the MS when entering a WiMAX network. The network entry procedure may include other tasks such as receiving the time of the day and other operational parameters. The black solid circles in this figure indicate the involvement of different nodes in each network entry phase.

In order to join a WiMAX network, the MS is required to perform the following tasks:

- **Synchronisation**: The MS performs time and frequency synchronisation with the selected BS.

- **Ranging**: This phase is required for transmission power and timing adjustment.

- **Basic capability exchange**: The MS and BS negotiate basic operational parameters such as supported modulation schemes.
Authentication and key exchange: The MS and the WiMAX network mutually authenticate each other.

BS registration: The MS registers with the BS to receive a management connection identity and exchange additional capabilities.

IP address allocation using the standard Dynamic Host Configuration Protocol (DHCP) mechanism.

Service flow creation: The MS request the creation of service flows with desirable QoS parameters.

2.2.4 Synchronisation, Ranging, and Capability Exchange

When a WiMAX MS powers up, it first scans radio channels to determine if it is within the coverage area of a WiMAX network and synchronised with a radio channel. Since WiMAX uses an Orthogonal Frequency Division Multiple Access (OFDMA) radio transmission technique, both time and frequency synchronisations are required. By listening to preamble signals transmitted from the BS, the MS is able to perform the synchronisation process. The MS then listens to the broadcast
PHY and MAC parameters and obtains information on uplink channels and how to send initial control messages.

After synchronisation, the MS performs the initial ranging operation which allows it to adjust its relative timing with the BS and its power-level [6]. In the ranging operation the MS sends a ranging request (RNG-REQ) message to the BS. Since a connection has not yet been established with the BS, the MS has to use a contention-based mechanism to obtain a transmission opportunity. If the MS does not receive a ranging response (RNG-RES) message, it tries again with a higher transmission power. If the ranging process is successful, the BS sends a RNG-RES message which contains a primary Connection ID (CID) and assigns a time slot for the MS to send additional RNG-REQ messages for fine tuning purposes.

After a successful ranging, the MS and BS exchange their capabilities such as supported PHY and bandwidth allocation parameters, supported power saving methods, and supported security parameters. For this purpose, the MS sends a Subscriber Station Basic Capability Request (SBC-REQ) message to the BS. The BS responds with a SBC-RES message to let the MS know what parameters should be used [6]. Figure 2.8 shows the procedure for synchronisation and initial association between the MS and the BS.

### 2.2.5 Authentication and Security Key Exchange

A secure wireless technologies should provide mechanisms for MS authentication, user authorisation, data encryption, and integrity check of transmitted packets. To satisfy these requirements, the WiMAX technologies use a flexible security architecture and a mixture of protocols designed by the IEEE and the IETF.

The security framework of the WiMAX technology consists of five network functional (and not necessarily physical) entities. The CSN contains the AAA server. The ASN includes the Authentication Relay, the Authenticator, the Key Receiver,
and Key Distributor. In a physical implementation, the first two ANS components are implemented in the ASN-GW, while the BS contains the Key Receiver and the Authentication Relay.

The protocol stack for WiMAX security is depicted in Figure 2.9. The establishment of the Security Association (SA) between the MS and the BS is based on the Privacy Key Management Protocol version 2 (PKMv2) procedure which is defined in the IEEE 802.16e standard. Over PKMv2, the Extensible Authentication Protocol (EAP) [47] is run to allow the AAA server in the CSN and the MS to mutually authenticate each other.

As the name suggests, the EAP protocol is extensible and supports different authentication methods. For example, the EAP Authentication and Key Authorization (EAP-AKA') scheme is based on the LTE authentication method which uses the Universal Subscriber Identity Module (USIM) to store the user information. The EAP messages are carried between the ASN-GW (as the Authenticator) and the AAA server by an AAA protocol such as the Remote Authentication Dial In User Service (RADIUS) protocol [48], or alternatively the Diameter protocol [49].

The WiMAX user authentication procedure using the EAP-AKA method can be seen in Figure 2.10. The process starts when the ASN-GW sends an EAP Identity Request (EAP-REQ/Identity) message to the MS. The MS replies with the
EAP-RES/Identity message which includes the MS’s identity in the Network Access Identity (NAI) format, for example, username@domain. The Message is forwarded to the AAA server using the RADIUS or Diameter protocols. If the MS is in a roaming scenario, the AAA proxy in the visited CSN uses the MS’s ID to route EAP messages to the AAA server in the home CSN.

The AAA server uses the Authentication Vector (AV) to construct an authentication challenge and sends it to the MS. This message contains an encrypted random number which can be decrypted only by security keys located in the USIM card of the MS. The MS authenticates the WiMAX network by decrypting the challenge and replies with another encrypted message which allows the AAA server to authenticate the MS.

Upon successful authentication, the MS and AAA server generate a 512-bit Master Session Key (MSK). The AAA server, at step (10), sends the MSK to the Authenticator (the ASN-GW). Using this key, the MS and the authenticator are able to produce the Pairwise Master Key (PMK) by truncating the MSK to 160 bits; the Authentication Key (AK) is then generated from PMK. The algorithm to generate the AK is called the Dot16KDF algorithm and uses the Base Station Identity (BSID) and the MS’s MAC address to derive the AK [50]:

\[
AK = \text{Dot16KDF}(\text{PMK}, \text{MS MAC Address}|\text{BSID}|"AK", 160) \tag{2.1}
\]

where the | sign denotes a concatenation.

The dependence of the AK on the BSID has important implications for the novel pre-registration method we propose in Chapter 4.

In step (11), the Key Distributor in the Authenticator transfers the AK to the Key Receiver in the BS. In next steps the MS and the BS verify the SA using a three-way handshake. After the handshake, or during this process for the case of WiMAX re-entry, the BS delivers a Traffic Encryption Key (TEK) to the MS. The handshake starts with the BS sending a SA-TEK Challenge message to the MS. This message identifies the AK key. The MS authenticates the BS and replies with SA-TEK Request which indicates the MS possess the AK.

The BS then sends the SA-TEK Response which completes the SA verification handshake. In the case of initial entry, the MS requests the BS to send a new TEK, as depicted in steps (15) and (16). In the case of WiMAX re-entry where the MS performs a handover to a new BS, the SA-TEK Response message includes a field called the SA-TEK-Update which contains a fresh TEK and the MS does not need to request a security key. For the case of first authentication within the WiMAX network, the MS is required to request a fresh TEK.

After this stage, the MS and the BS are mutually authenticated and all subse-
Figure 2.10: WiMAX authentication procedure [5]
quent messages sent over the radio link are securely encrypted.

### 2.2.6 Base Station Registration

After the mutual authentication process, the MS registers with the BS by sending a Registration Request (REG-REQ) message and receiving a Registration Response (REG-RES). By this process, the MS receives a secondary Connection ID (CID) which is used to transfer management messages which are delay tolerant. For example, the MS uses this CID to exchange DHCP messages with the WiMAX network to obtain an IP address configuration.

In the REG-REQ message, the MS also includes the secondary capabilities not covered in the SBC-REQ message. For example, the MS and the BS can exchange information such as the IP version, mobility parameters, handover support, and support for Automatic Repeat Request (ARQ) mechanism. Figure 2.11 depicts the BS registration mechanism.

### 2.2.7 IP Configuration

In WiMAX, the primary IP address allocation mechanism is DHCP [6]. In this mechanism, the ASN contains a DHCP relay function which transfers the messages between the MS and the DHCP server which is located in the CSN. Alternatively, the CSN can allocate the IP address in conjunction with the AAA procedure. In this case, the ASN contains a DHCP proxy which retrieves the IP address from the AAA server. The DHCP proxy then uses the DHCP mechanism to deliver the IP address to the MS.

Figure 2.12 shows the signaling flow for IP address allocation using the DHCP procedure.

### 2.2.8 QoS and Session Creation

The MAC layer of the WiMAX technology is designed to carry packets of both real-time multimeida applications and non real-time applications. In WiMAX, being a connection-oriented wireless technology, the MS is required to set up a logical
connection with the BS before transferring data packets. Each unidirectional connection between the MS and the BS is identified by a Connection Identifier (CID). CIDs are put in MAC packets to identify the source and destination nodes. As such, CIDs can be regarded as temporary identifiers for WiMAX nodes. The basic and primary CIDs are allocated to the MS during the initial ranging process and are used to carry signaling packets. The MS may also receive a secondary management CID in the BS registration process. Additional secondary CIDs are allocated to the MS during session creation process and are used for data transmission.

To create a data session, service flows must be established first. A service flow is a unidirectional connection (from MS to BS or vice versa) which has certain QoS parameters and is identified with a Service Flow Identity (SFID). The BS assigns SFIDs and maps them to unique CIDs. SFIDs are not changed during a handover to a new cell, but CIDs are temporary and refreshed. In the backbone side, service flows can be mapped to DiffServ codes or MPLS labels to provide end-to-end QoS [6].

Some of the QoS parameters of a service flow are traffic priority, minimum reserved rate, maximum latency and jitter tolerance, and scheduling type. WiMAX specifies five scheduling types which include the Unsolicited Grant Services (UGS), the Real-time polling services (rtPS), the Extended real-time variable rate (ERT-VR) service, the Non-real-time polling service (nrtPS), and the Best-effort (BE) services.

Service flow creation can be initiated by both the MS and the WiMAX network. Figure 2.13, based on the WiMAX Forum specifications in [5] and [51], shows the signaling flow for an MS-initiated creation of a dedicated service flow. To start the process, the MS sends a Dynamic Service Allocation Request (DSA-REQ) message with the required QoS parameters. The SFM function receives the request and forwards it to the SFA function in the ASN Gateway. If dynamic Policy and Charging
Control (PCC) framework is deployed, the SFA asks the PCRF to decide on QoS parameters and service flow creation. Otherwise, the SFA uses local policy and information and either accepts or rejects the service creation request.

The PCC framework of WiMAX is based on 3GPP specifications for PCC [51]. In this framework, when the PCRF receives an indication of service establishment from the ASN Gateway, it queries the Subscription Profile Repository (SPR) to obtain information related to the service flow creation. Based on this information, the PCRF makes policy decisions. If the service flow creation is authorised, the PCRF informs the ASN Gateway which enforces the decision. The ASN Gateway then requests the BS to create the service flow by sending a Dynamic Service Allocation Request (DSA-RES) message to the MS. To complete the process, the MS acknowledges the creation of the service flow.

Network-initiated creation of a dedicated service flow happens in cases where a downlink packet for the MS is received, or when an IMS application server requests
the PCRF to create a bearer with suitable attributes. The process is similar to the one depicted in Figure 2.13 and is started by an application server sending an Application/Service Info message to the PCRF. The PCRF then makes a policy decision and informs the ASN Gateway which in turn requests the BS to send a DSA-REQ message to the MS.

2.2.9 Mobility Support in WiMAX

WiMAX supports seamless handovers between BSs for movement speeds up to 120 km/h [6]. In WiMAX network infrastructure both ASN-anchored mobility (in which the MS’s IP address is not changed) and the CSN-anchored mobility (where the MS roams between different subnetworks) are supported. The first mobility scenario is supported by deploying the Data Path, Handover, and Context functions. To manage the second scenario, the WiMAX Forum recommends the implementation of the Mobile IP (MIP) protocol.

In both ASN-anchored and CSN-anchored mobility scenarios, three handover methods can be used:

- Hard Handover (HHO).
  In this method, the original BS allocates a scanning interval to the MS. During this interval, the MS collects information on the target BS and optionally performs the initial ranging process. When a handover decision is made (by the MS, BS, or a handover control network entity), the MS detaches from the old BS and joins the target BS.

- Fast Base Station Switching (FBSS).
  The support for this method is optional. In FBSS, the MS maintains radio connections with several BS and changes the anchor BS when a handover is needed.

- Macro diversity handover.
  This method is another optional handover scheme. The MS simultaneously communicates with several BSs and combines received packets using diversity-combining techniques. This method and the FBSS scheme are more complicated than the HHO method and require BSs to be synchronised, use the same radio frequency channel, and exchange information on the backbone network.

The IEEE 802.16 standard specifies methods to minimise handover delay by reducing signaling messages and increasing cooperation between involved BSs. Since this research is not focused on handover within the WiMAX network, we do not cover these handover procedures in detail.
2.3 LTE Technology

In this section we describe the LTE technology, its network architecture, and procedures. We highlight the LTE architecture paradigm of separating radio-related tasks and non-radio mechanisms into a RAN and a CN. Later in this thesis, we use this paradigm along with the LTE procedures to develop and analyse a mobility management procedure.

2.3.1 LTE History and Characteristics

The Long Term Evolution (LTE) is an evolution of the Universal Mobile Telecommunications System (UMTS) technology (or 3G). 3GPP started the LTE project in 2004 [52] to address the growing demand for higher data rates, faster channel access, and improved network architecture. The LTE specifications were completed in December 2008 and the first commercial LTE network was launched in 2009.

The LTE network consists of the Evolved Universal Terrestrial Radio Access Network (E-UTRAN) and an Evolved Packet Core (EPC). The E-UTRAN is the radio section of the LTE technology and employs a new modulation technique; it is not backward compatible with previous 3GPP radio technologies such as the High Speed Packet Access (HSPA). The E-UTRAN was designed to provide higher throughput and more flexibility in supporting radio channels wider than 5 MHz.

The EPC network is designed to replace the General Packet Radio Service (GPRS) core network. The EPC has a flat IP-based architecture which is more suitable for converged voice, video, and data communications.

The LTE technology provides the following features [53]:

- Peak data rates of 100 Mbps in downlink direction and 50 Mbps in uplink direction.
- Access delay of less than 5 milliseconds where access delay is defined as packet transmission latency between the MS and the edge of access network.
- “Always on” terminals requiring less than 100 milliseconds for transition between idle to active states.
- Full mobility support for speeds up to 350 km/h.
- Support for different channel bandwidth from 1.25 MHz to 20 MHz.
- Pure packet switched core network and packet-optimised radio interface.
- Supporting of different QoS classes for a variety of applications.
2.3.2 LTE Network Architecture

The LTE system architecture is depicted in Figure 2.14. The LTE network is divided into an IP-based Core Network (CN), called the Enhanced Packet Core (EPC), and a modern Radio Access Network (RAN), called the Evolved Universal RAN (E-UTRAN) [52]. The RAN includes functionalities for the radio transmission of data packets, while the CN manages network access and IP-connectivity issues such as authentication, policy control and QoS management, and IP address allocation.

The E-UTRAN consists of E-UTRAN NodeB (eNB) radio equipment which provides the LTE radio connection to MSs. The eNB is in charge of radio transmission and reception in one or more LTE cells. Some of the important eNB functionalities include [54]:

- Radio resource management and scheduling.
- Routing users data packets to the EPC.
- Encryption of data streams.
- IP header compression.

eNBs of an E-UTRAN are connected to each other via a network interface, denoted by X2, which enable them to cooperate to improve the performance of handovers between radio cells. The eNB is connected to the EPC by a link denoted by S1. This interface is split into S1-U, which carries users data packets, and S1-MME, which carries signaling.

In the EPC, the Mobility Management Entity (MME) is responsible for controlling subscribers and data sessions. The MME authenticates users by obtaining subscriber information from the Home Subscriber Server (HSS). The MME registers
the MS in the LTE network and updates the location information. The MME is also involved in creating the MSs default and dedicated bearers which are used for signaling and data transmission.

In contrast to the MME which only manages signaling, the Serving Gateway (S-GW) in the EPC is responsible for terminating the data path towards the E-UTRAN and routing data packets towards the Packet Data Network Gateway (P-GW). The S-GW also functions as an anchor for mobility between eNBs.

The P-GW is the LTE gateway to the IMS network and the rest of the Internet. The P-GW assigns IP addresses to MSs and is also used in lawful interception. The role of LTE nodes is explained in more detail in the next sections where we describe the different LTE procedures.

2.3.3 LTE Network Entry Overview

When the MS powers up, it performs the initial network attachment procedure to connect to an LTE network. It then establishes dedicated network bearers in order to initiate a multimedia session. The LTE initial network entry process, as explained by Lescuyer and Lucidarme in [53], is summarised in Figure 2.15. The procedure starts with the MS scanning radio channels to discover and select a target eNB. The MS then synchronises with the eNB and receives a temporary identity for radio transmission.

Following radio synchronisation, the MS and the LTE network mutually authenticate each other. Upon successful authentication, the MS location is updated in the HSS. The next step is to create the default bearer which is used for the transmission of control messages and occasional data traffic. The MS then requests the establishment of dedicated network bearers with desirable QoS parameters to carry the data traffic of the multimedia application. In the following sections we describe the LTE attachment procedure in more detail.

2.3.4 Cell Search and Selection

In order to enable MS operations every LTE network needs to broadcast certain information. The most important broadcast parameters, such as the downlink bandwidth and reference transmit power, are transmitted in the Broadcast Channel (BCH) and are refreshed every 40 milliseconds [53]. Less important information is sent over the Downlink synchronisation Channel (DL-SCH) transport channel which has more flexibility. This information includes parameters such as the cell ID, the ID of the mobile networks utilising this cell (in RAN sharing scenarios), the cell barring information (preventing certain MSs selecting this cell), and radio channel parameters.
Based on the received broadcast information, the MS selects an eNB as the target. The selection criteria include parameters such as network operator, cell status, signal quality, and radio transmission technology. The MS then performs the initial access process to associate with the target eNB.

2.3.5 Synchronisation and Random Access

If the MS determines that the cell is suitable for a radio connection and the MS is not barred from accessing the cell, it tries to connect to the eNB using the Random Access mechanism. In LTE technology, the random access mechanism is used by MSs which do not have a dedicated radio bearer for data transmission. There are two types of random access schemes [54]:

1. Contention-based random access. In this type, the eNB does not have any record of the MS. As such no dedicated radio resource can be allocated to the MS for the transmission of initial control message. The MS selects a five-bit random identity and competes with other MSs to transmit a Random Access Preamble message to the eNB. The eNB responds with the Random Access Response message which contains a temporary identity for the MS, timing alignment information, and a scheduled transmission opportunity. The MS can then transmit the initial control messages (such as an attach request) in the allocated slot without contending with other stations. If multiple MSs
request access to a specific radio resource, the eNB will send a *Contention Resolution* message to determine the successful station. The procedure is depicted in Figure 2.16.

2. Non contention-based random access. In cases such as the handover or reception of downlink traffic, the eNB assigns a random access preamble to the MS. The assignment can be sent to the MS via the old eNB (in handover cases) or a downlink control channel. The MS then is able to send the *Random Access Preamble* message without contention. Consequently, the eNB replies with the *Random Access Response* message. This results in reduced signaling delay.

### 2.3.6 Authentication Procedure in LTE

Before using any network services (aside from emergency calls) the MS must first perform the authentication process. In this process, the MS and the LTE network mutually authenticate each other and generate the security keys required for data encryption and integrity checking.

The LTE authentication mechanism is the same method that was used in the UMTS technology and is called the EPS Authentication and Key Agreement (EPS AKA) mechanism. The EPA AKA procedure is depicted in Figure 2.17. The procedure is initiated by the MME when it receives an unencrypted Attach Request message from the MS. In the initial network attachment there is no record of the MS in the LTE network. As such, the MME sends an *Identity Request* message to the MS. The MS replies with its International Mobile Subscriber Identity (IMSI) number.

The MME then retrieves the user’s information from the HSS. User authentication in the EPS AKA mechanism is based on the main security key (K) which is
stored both in the USIM inside the terminal and the HSS in the network. This key never leaves its secure locations. When the MME requests the HSS for user information the HSS generates and sends an Authentication Vector (AV) which contains five elements: the Integrity Key (IK), the Ciphering Key (CK) both these derived from K, the authentication token (AUTN), a random number (RAND), and an Expected Result (XRES).

The MME stores the AV and sends the AUTN and RND to the MS. The MS uses the AUTN to authenticate the network. It then encrypts the RAND and returns the result to the MME. If the encrypted RAND and XRES are equal, the user is successfully authenticated.

Upon successful authentication, the Access Security Management Entity (ASME) key, the $K_{ASME}$, is derived using the CK and IK keys. $K_{ASME}$ is an intermediate security key generated in both the MS and the MME. From the $K_{ASME}$, all the keys required to secure the MS communication are generated. For example, encryption and integrity check keys for Non-Access-Stratum (NAS) control messages, denoted by the $K_{NAS\text{enc}}$ and $K_{NAS\text{int}}$, are derived from $K_{ASME}$ and used secure communication between the MME and the MS. The $K_{eNB}$, on the other hand, is derived in order to secure the communication between the eNB and the MS.

The value of $K_{eNB}$ depends on the eNB identity [53]. This key is therefore unique and cannot be reused by another eNB. This approach increases the security of the LTE network because the compromise of one eNB does not affect the security of others. However, this has an important implication in the proactive handover preparation mechanism which we propose and describe in Chapter 4.

2.3.7 Location Update

In the LTE technology, the serving MME of the MS is recorded in a network database, specifically the HSS. Tracking the serving MME enables the LTE network to exchange control messages with the MS, because the MME is the entity in the CN which terminates the signaling interface with the MS.
When an MME receives an Attach Request message from the MS, it first authenticates the MS and then updates the MS’s location information in the HSS. The MME sends an Update Location Request message to the HSS to inform it that the MS is now connected to this MME\cite{55}. The HSS then records the serving MME’s information and acknowledges the location update request message.

2.3.8 Default Bearer

Following a location update, the MME establishes a default bearer for the MS\cite{53}. The default bearer is used for signaling purposes between the MS and the LTE network. The default bearer is also used for data flows which are not allocated to a dedicated EPS bearer.

Figure 2.18 depicts the procedure for creating the default bearer. The procedure is performed during initial LTE attachment and is initiated after the location update process. To initiate the default bearer establishment, the MME first selects a suitable S-GW and a P-GW, as described by 3GPP in\cite{55}. The MME uses the MS’s subscription information to determine the QoS parameters of the default bearer. The MME then sends a Create Session Request message to the S-GW. The S-GW creates a new entry in its EPS Bearers table and forwards the message to the P-GW.

Similar to other EPS bearers, the establishment of the default bearer requires interaction with the dynamic Policy and Charging Control (PCC) framework of the LTE network, if it is deployed. In this framework, the Policy and Charging Rule Function (PCRF) makes policy decisions on the establishment of network bearers and the allocation of resources. The P-GW indicates the request for the establishment of an EPS bearer, in this case the default bearer. Subsequently, the PCRF downloads the user’s subscription information from the Subscription Profile Repository (SPR) and decides on establishment of the default bearer with certain QoS parameters. The P-GW then adds a new entry in its EPS Bearers table which allows routing of data packets to and from the MS. In cases where the dynamic PCC framework is not implemented, the P-GW uses its locally stored policies to decide on the creation of the default bearer.

The P-GW then sends a Create Session Response message to the S-GW. If the IP allocation is combined with the default bearer creation, this message also includes the assigned IP address. The S-GW forwards this message to the MME. The MME stores the required information and requests the eNB to allocate radio resources and establish radio bearer to satisfy the QoS parameters of the default bearer.

The creation of radio bearers is performed in next steps. First the eNB sends the RRC Connection Reconfiguration message which includes Attach Accept, the identity of the radio bearer, its QoS parameters, and the new IP address (if allocated).
Then the MS stores this information and indicates the completion of radio bearer establishment. The eNB sends the Initial Context Setup Response message to the MME.

The MS also sends the Direct Transfer message to the eNB which includes the Attach Complete message which is then forwarded to the MME. Upon reception of both Attach Complete messages, the MME informs the S-GW that the MS is ready for reception of downlink packets, if they have been buffered. The S-GW replies with an acknowledgment and sends any buffer data packets towards the MS.

2.3.9 QoS Support and Dedicated Bearer Establishment

In LTE, QoS is supported by the creation of EPS bearers which carry data packets across the network. An EPS bearer is equivalent to the legacy Packet Data Protocol (PDP) context which was used in 3G networks. An EPS bearer may contain many Service Data Flows (SDF) which are identified by a 5-tuple: the source and destination IP addresses, the source and destination ports, and the protocol ID [56]. However the same QoS policy is applied to all of SDFs within an EPS bearer. The operator can choose which SDF is mapped to a specific dedicated bearer. If an SDF does not have a dedicated bearer, it uses the default bearer.

The QoS parameters of an EPS bearer include the Allocation Retention Priority, the Guaranteed Bit Rate, the Maximum Bit Rate, and the QoS Class Identifier (QCI). While the EPS bearer determines the service flow QoS characteristic in the entire LTE network, the QCI specifies the over-the-air treatment of data packets. The purpose of the QCI and its parameters is to provide a QoS experience which is consistent in both the CN and the RAN part of the LTE network.

QCI parameters include Bearer Type, L2 Packet Delay Budget, and L2 Packet Loss Rate [53]. The bearer type specifies if the data rate should be guaranteed. The delay budget has a strict meaning only for real-time applications, and the loss rate determines if techniques such as Automatic Repeat Request (ARQ) and Hybrid ARQ (HARQ) should be used.

The dedicated bearer creation process is based on the dynamic Policy and Charging Control (PCC) framework. The policy control is a set of rules and their application to determine access to resources [53]. In early versions of UMTS, the policy control was MS-based, that is, the MS determined its required QoS parameters and asked the network for its provision. However, as the IMS role as a standard for multimedia applications became more important, the 3GPP developed a more flexible framework which is referred to as the Policy and Charging Control (PCC) framework [57]. Aside from filtering packets and collecting charging information, this framework is used to ensure a consistent QoS in the LTE network. In the
Figure 2.18: Default bearer creation procedure
PCC framework, the PCRF node makes policy decisions based on information it receives from applications servers, network nodes such as the P-GW, and subscriber information stored in the SPR.

The procedure for network-initiated EPS bearer creation is depicted in Figure 2.19. The procedure starts when an Application Server (AS) contacts the PCRF to create an EPS bearer to carry data packets of an application. The PCRF makes policy decisions, determines QoS parameters, and informs the P-GW to enforce the policy. This message is forwarded to the S-GW and then the MME. Subsequently, the eNB is asked to set up radio bearers with the desired QCI parameters by sending RRC Connection Reconfiguration message to the MS. After the establishment of the radio bearer, the P-GW is informed by a Create Dedicated Bearer Response message.

To reduce the signaling delay, the creation of default and dedicated bearers may be combined [55]. In this situation, the PCRF places the responses to the creation of the default and dedicated bearers in one message and forwards it to the P-GW.
2.3.10 Mobility Support in LTE

LTE uses a network-controlled MS-assisted approach to manage the mobility of a terminal which is in the active mode [53]. The serving eNB uses its measurements and MS reports to determine a target eNB. However, a proactive make-before-break mechanism is used to reduce handover delay and packet forwarding is implemented to eliminate packet loss.

If the serving and target eNBs can communicate directly (over the X2 interface) the serving eNB sends a *HANDOVER REQUEST* message to the target eNB and asks for resource allocation. This message contains the $K_{eNB}$, which is calculated at the serving eNB. It also contains the Radio Resource Control (RRC) context and the S1 related context.

The target eNB performs admission control and responds with a *HANDOVER REQUEST ACKNOWLEDGE* message which contains the target eNB security algorithm identifiers and possibly a dedicated random access preamble. The serving eNB then asks the MS to perform the handover. The MME is then informed and it orders the S-GW to send packets to the new eNB. The signaling is explained in more detail in 3GPP TS 36.300 [54]. In this process, the S-GW functions as an anchor which remains in the media path to hide the MSs movement in the E-UTRAN from communicating nodes.

Mobility within the LTE network is out of scope of this research and is not covered here.

2.4 IP Multimedia Subsystem (IMS)

The IP Multimedia Subsystem (IMS) is the defacto standard platform for the delivery of multimedia services in modern cellular networks [18]. IMS is the result of efforts by standardisation entities such as the European Telecommunications Standards Institute (ETSI), the 3GPP, the Internet Engineering Task Force (IETF), and Open Mobile Alliance (OMA) [58].

The design of the IMS was inspired by the success of the Internet and the widespread use of multimedia applications over general IP systems. However, IMS is not a competitor to the Internet; rather it tries to improve the interaction of multimedia applications with Internet architecture to provide better QoS. Improved service security and an open interface for service development are other important IMS features.

IMS does not respect the end-to-end Internet philosophy that suggests intelligence should be located in end devices and not in network nodes [59]. This means some of the service intelligence can be located in the Application Servers (AS) which
are located in the core network as part of the IMS. This approach has some advantage; end devices can be simpler and applications servers will have better cooperation with other network entities to ensure security and quality of service.

The IMS specifications were first published in 2002 by the 3GPP in Release 5 of UMTS. The 3GPP proposed IMS to enable operators to provide packet-based multimedia services over mobile networks. However, IMS was also used in fixed wired IP-based networks. In 2004, the ITU-T created the Next Generation Network (NGN) Focus Group to study fixed line access to IMS [60]. The focus group proposed the service stratum of of NGN based on IMS. IMS therefore is able to realise fixed and mobile convergence.

The IMS has the following features:

- Multimedia session management using the Session Initiation Protocol (SIP) [61]
- Client authentication and authorisation based on 3GPP and IETF protocols.
- Encryption of signaling between the client and the IMS network.
- Compression of signaling packets for reduction of transmission delay over wireless links.
- Open application programming interface for service development.

2.4.1 IMS Architecture

The logical IMS architecture can be seen in Figure 2.20. In this architecture, the IMS client is a SIP endpoint which implements 3GPP specific SIP headers and other IMS-related mechanisms such as IMS registration and signaling compression. The IMS client is identified by a SIP Uniform Resource Identifier (URI) which has a format of \textit{SIP:username@domain}.

The IMS core network is composed of three Call Session Control Functions (CSCF): the Proxy CSCF (P-CSCF), the Serving CSCF (S-CSCF), and the Interrogating CSCF (I-CSCF). Note that IMS nodes in Figure 2.20 are logical functions; in a physical implementation one or more components may be deployed in one physical network node.

The P-CSCF is the first point of attachment to the IMS system. All SIP messages from and towards the client should pass through the P-CSCF. The IMS client finds the IP address of this node during network attachment. The standard DHCP mechanism, the Media Independent Handover (MIH) protocol, or other network configuration mechanisms can be used to obtain the address of the serving P-CSCF.
The P-CSCF is usually located in an IP network which provides the connectivity service. For example, in the LTE network the P-CSCF can be located in the core network and close to the P-GW. The P-CSCF inspects all SIP messages to ensure they are properly constructed and include any required headers. The P-CSCF may also extract media parameters from the SIP messages to determine the QoS parameters required by the multimedia application. It can then contact the PCRF to initiate the creation of network and radio bearers.

The P-CSCF establishes a Security Association (SA) with the IMS client and encrypts all IMS signaling messages. The parameters for the SA are negotiated between the P-CSCF and the client during the IMS registration process. The SA is dependent on the IP address of the MS which causes some problems when the MS moves between different subnetworks.

The S-CSCF is the central node of the IMS. This node functions as a SIP server which controls SIP sessions. It also acts as a SIP Registrar, meaning it maintains a binding between the clients IP addresses and their SIP accounts [59]. All IMS communications of a client goes through its designated S-CSCF. The S-CSCF routes all SIP messages to appropriate Application Servers (AS) or to other destinations. The decision to route messages to ASs are based on Initial Filter Criteria which are configured in the S-CSCF.

The S-CSCF is also in charge of the user’s authentication. This process is performed when the client registers with the IMS network. The authentication process includes interaction between the S-CSCF and the Home Subscriber Server (HSS). The HSS is located in every IMS network and contains user information and their
subscribed services. We describe the IMS registration process later in this section.

The I-CSCF is located at the edge of the IMS and can be reached from external networks. The IP address of the I-CSCF is listed in the Domain Name System (DNS) record of the IMS domain and therefore can be resolved from the identity of the IMS client. The I-CSCF routes SIP messages to the appropriate S-CSCF.

The IMS architecture also includes the AS which runs one or more IMS services. The AS usually sits between SIP communication between the client and its corresponding host and possibly alters SIP messages to provide a specific service. The AS receives SIP messages from the S-CSCF. It also has a connection to the I-CSCF to exchange signaling with IMS nodes and end points in cases where a S-CSCF is not assigned to the MS. The AS may also access subscriber information by interfacing with the HSS.

### 2.4.2 IMS Registration Process

The IMS registration process is performed to allow the MS to use services provided by the IMS. Through this process, the MS binds its current IP address to its public identity (its URI). The MS can then be reached in the current serving network. In the IMS registration process, the S-CSCF and the MS also mutually authenticate each other. The authentication process is similar to the LTE AKA method and the USIM card can be used for subscription information. During the registration message exchange, the MS and the P-CSCF agree on the security mechanism, establish a Security Association (SA) and determine the message compression method.

Figure 2.21 shows the IMS-level user registrations process [60]. To initiate the process, the MS sends a SIP REGISTER message to the P-CSCF in the serving network. The P-CSCF uses the MS URI to resolve the I-CSCF IP address and forwards the message to this node. The I-CSCF queries the HSS to determine whether the MS is authorised to access IMS services from the current serving network. Based on the response received from the HSS, the I-CSCF forwards the register message to a selected S-CSCF.

The S-CSCF then contacts the HSS to download the user's Authentication Vector (similar to the LTE AKA procedure) and update the HSS location information to indicate that it is serving the MS. The HSS replies with the authentication vector. The S-CSCF replies back to the MS with a SIP 401 (unauthorised) message which contains a challenge for the MS. The MS returns another SIP REGISTER message which includes the response to the challenge.

The S-CSCF receives the second SIP REGISTER message and verifies the identity of the MS. It then informs the HSS that the MS is now registered to the IMS and downloads the user's profile. To complete the IMS registration process, the
S-CSCF sends a SIP OK message to the MS.

Upon successful completion of IMS registration process, the MS is able to securely and efficiently communicate with the IMS network nodes and application servers. SIP messages between the MS and the P-CSCF may be compressed by the SigComp algorithm, RFC 3320 [62]. This compression reduces signaling delay, particularly over low-rate radio links. SIP messages are also encrypted using the SA established between the MS and the P-CSCF.

In the current version of IMS, the SA between the MS and the P-CSCF depends on the MS’s IP address [10]. This means if the MS changes its IP address, for example during a vertical handover, it should terminate all ongoing multimedia applications and deregister from IMS. The MS should then use its new IP address to perform a new IMS registration. The IMS relies on the underlying transport network to hide MS mobility and the change of IP addresses from the IMS nodes. Hiding MS mobility can be achieved by using the Mobile IP [16] protocol. However, this protocol is not widely implemented in mobile networks.

In the next section, we overview the problem of mobility management in NGMN and discuss how the IMS can be modified to allow session continuity even when the underlying network uses mobility management mechanisms such as the Mobile IP protocol.
2.5 Mobility Management in NGMN

Efficient mobility management is one of the fundamental requirements for NGMN [33, 2]. Mobile operators are responding to the ever-growing subscriber demand for rich multimedia communication services by integrating various wireless technologies such as WiFi, WiMAX, and LTE into their network infrastructure, and providing seamless mobility between them [63]. A mobility management mechanism is required to allow roaming among different radio technologies. In this section we provide an overview of different mobility types, network scenarios, and describe the mobility management procedure in NGMN.

2.5.1 Mobility Types, Scenarios, and Definitions

A Mobility management solution is a framework and its mechanisms utilised to ensure the *reachability* of a Mobile Station (MS) and its user, and the *migration* of any ongoing data session from one network connection to another. Mobility management is required for the following mobility types:

- **Terminal Mobility.**
  This is the most familiar and common type of mobility which is caused by the movement of the MS [33]. Because of this movement, and other varying factors such as radio conditions and network load, the terminal often exits the coverage of its current Point of Attachment (PoA) or a more suitable PoA becomes available. The mobility management mechanism is required to update the MS’s location information in the network and transfer any ongoing data session to the new PoA.

- **Network mobility.**
  Network Mobility (or NEMO) is a scenario where an entire IP subnet network is itself moving and changing its point of attachment to the network [64]. This type of mobility is common in inside-vehicle wireless networks.

- **Personal mobility.**
  This type of mobility refers to the movement of a single user among multiple terminals. To support personal mobility, the mobility management mechanism should be able to identify the user and deliver telecommunication services regardless of the user’s location and current terminal.

- **Service mobility.**
  Service mobility is a situation where a data session, or parts of it, are transferred from one device to another. This transfer is often required to improve
the user’s experience of a telecommunication service, for example transferring the video component of a multimedia call from a phone to a television.

Mobility management solutions may support one or all mobility types. For example, the Mobile IP (MIP) [16] only supports terminal mobility while the Session Initiation Protocol (SIP) [61] supports all these types of mobility. Network operators may therefore choose to combine different mobility solutions to support different mobility types. In this research, we focus on terminal mobility and we use the term mobility to refer to this type of mobility.

Because of terminal mobility, the MS may be forced to change its Point of Attachment (PoA). A PoA is a network node which provides a Link layer (L2) radio connection to the MS. The MS may leave a PoA, called the old or current PoA, and join another, called the target or new PoA, before initiating a data session (pre-session mobility) or during the session (mid-session mobility). Managing pre-session mobility is relatively less complicated and only involves updating the MS’s location information in the network. On the other hand, ensuring service continuity for mid-session mobility is more complicated and requires additional functionality to facilitate inter-technology handover control.

Based on session continuity requirements, mobility management scenarios are categorised as follows [33]:

- Nomadic.
  In nomadic mobility only pre-session mobility is required. In this scenario, when the MS changes its PoA, it should be able to access the same set of services and be reachable by its peers. In this type of mobility, service continuity is not needed and the MS does not communicate with others during the movement.

- Handover.
  In the handover scenario, the MS should be able to continue a data session while it changes the serving PoA. In this scenario, mid-session mobility should not force the MS to drop its data session with the Corresponding Host (CH). However, during the handover, the MS may experience significant service interruption.

- Seamless handover.
  Seamless handover refers to a scenario where the MS does not experience service degradation while performing a handover in mid-session mobility.

Seamless handover is relatively easy to achieve in a homogeneous Radio Access Network (RAN) which uses only one wireless technology. Roaming among PoAs of this RAN results in intra-technology handovers, also referred to as horizontal handovers.
Usually intra-RAN mobility involves L2 handovers and is transparent to the IP and higher layers.

On the other hand, heterogeneous wireless networks, employ different radio technologies which makes mobility management more challenging. Mobility in heterogeneous networks often results in inter-technology handovers, also referred to as vertical handovers. The terms “horizontal” and “vertical” handovers were first used in overlay network scenarios where the MS is served by small-range, mid-range and wide-range wireless networks (such as WiFi, cellular, and satellite networks) and performs upward (and downward) vertical handovers to larger (and smaller) networks [1].

In inter-technology handovers, the MS is forced to perform full link-layer scanning and synchronisation with the target PoA, resulting in high handover delays. In some cases the current and target PoAs belong to RANs connecting to different Core Networks (CN) possibly managed by different operators. This mobility scenario is referred to as inter-CN and inter-administrative handovers [65]. The handover process is more time-consuming and includes authentication and new IP-address allocation procedures. While the authentication process increases handover delay, the change of IP address may result in connection teardown.

The business model of NGMN also contributes to the complexity of the mobility management. As specified by ITU-T in [39], in the general scenario for NGMN business model, the service stratum (commonly based on IMS) is independent from the underlying transport stratum and requires a separate authentication procedure. In fact, the user has subscriptions with both the IMS and access network, although in some the user receives a unified subscription bill [66]. This double authentication process results in more complexity and a higher handover delay. The transport layer itself may also have a complicated business model with RANs and CNs being owned by different operators [13]. Mobility among RANs requires direct roaming agreements between these operators. This flexible business model in NGMN results in complicated exchange network access control messages between the MS and the Authentication Authorization and Accounting (AAA) servers.

Despite the challenges, heterogeneity has many benefits including improved radio coverage, increased network capacity, and reduced service costs. Figure 2.22 shows three common network scenarios which a heterogeneous network is implemented for improving access service delivery. In the complementary coverage scenario, RANs are used to extend their radio coverage in different geographical locations. The RANs may have overlapping areas to allow seamless mobility management. In the overlay networks scenario, two or more RANs cover one region and provide the MS with the freedom to select the most suitable RAN for a particular service [37]. In the hotspot scenario, a RAN with limited coverage is used to improve radio coverage
or offload traffic from the main network. An example of the hotspot scenario is the deployment of WiFi cells to complement cellular networks [67]. Another example is the implementation of WiMAX femtocells to increase the capacity and the coverage of WiMAX networks [68].

The mobility management mechanism should be optimised to address the requirements of each of the above scenarios. For example, the hotspot requirement is to enable the MS to join the RAN with limited coverage quickly to enjoy better connectivity service. The complementary coverage requirement, on the other hand, is to reduce packet loss as much as possible. In next section, we overview the mobility management procedure and its application in each scenario.

2.5.2 Mobility Management Procedure Overview

A complete mobility management solution should include two general processes: location management and handover control [33]. The aim of the location management process is to identify the current MS’s location information which in IP-based net-
work is inferred from the MS's current IP address. Using this information, network servers and other end points can initiate a data session with the MS. The location management by itself can ensure a successful pre-session mobility management.

The handover control process, on the other hand, is related to mid-session mobility management and is used to migrate a data connection from the serving RAN to the target RAN. The main aim of the handover control process is to reduce service interruption during a change of radio interface [33].

The handover procedure generally includes tasks such as detecting a need for handover (trigger detection), collecting mobility information on surrounding RANs (network discovery), selecting the best available RAN (network selection), associating with the target RAN, performing AAA procedures, receiving IP configuration, establishing network bearers, migrating the data session to the new RAN, and finally releasing network resources in the old RAN. These tasks are categorised, as specified by ITU-T in [69] and used by authors in [70, 37, 52], into three handover stages:

- **Network discovery.** In this stage, also referred to as handover initiation, the MS or the network determines that a handover is required and selects the optimum target network (or networks) by collecting mobility information.

- **Handover preparation.** In this stage, the new radio interface is prepared for data transmission. To do so, the target network initial attachment procedure is performed and the MS obtains a new IP address. The location management process is also often performed at this stage.

- **Handover execution.** In this stage, the data session is transferred to the newly established radio link and the old link becomes obsolete.

Although this categorisation of handover tasks is the common notation, other variations are also found in the literature. For example Lampropoulos et al. in [71] consider an extra *handover completion* stage. In the following, we describe the common three stages of the handover process in more detail.

### 2.5.3 Network Discovery

Network discovery is the process of finding the optimum target RAN and the PoA within this network. The network discovery process is executed once when the MS powers on and is then performed continuously to keep the MS Always Best Connected (ABC). The concept of ABC was introduced by Gustafsson [13] to characterise the ultimate goal of the NGMN mobility management which significantly differs from the goal in traditional mobile networks. In NGMN, the MS is not only always connected, but is also connected using the *best* radio link.
The definition of “best” is ambiguous and is different for each user [72]. A network discovery and selection mechanism should consider many parameters to determine which radio connection best satisfies the user’s requirements. The following parameters have been used in the literature to assess suitability of wireless links:

- **Physical and link layer parameters.** These are the most commonly used parameters. Examples of such parameters include Received Signal Strength (RSS), Signal-to-Noise Ratio (SNR), Bit Error Rate (BER), Frame Error Rate (FER), and Packet Error Rate (PER). The network selection algorithms use various PHY and Link layer parameters as an indication of the experienced reliability of the radio side of the network. These parameters are often measured (or estimated) from the radio interfaces of the MS or PoAs.

- **Users location and mobility pattern.** A user’s location information such as the speed of movement is another network selection factor which is more important in heterogeneous networks than in homogeneous networks [65]. This is because heterogeneous networks consist of cells and RANs with different coverage areas. A fast-moving MS should be discouraged from joining a RAN with small coverage [73].

- **QoS parameters.** These parameters include class of service (for example, real-time or best effort), reserved (or guaranteed) bit rates, maximum bit rate, delay budget, and jitter. Network capability in ensuing certain QoS parameters can be estimated only for the radio link or for the end-to-end connection. The QoS parameters demanded by an application can be collected directly from an application. If the application requirements and the network capabilities match, the RAN is considered as suitable. A handover trigger arises when there is an imbalance. Note that QoS parameters are dynamic in nature and handover triggers because of QoS variation are expected.

- **Power consumption and battery status.** In many cases the lifetime of a mobile terminal is an important factor for users. In such cases, handover to a less power-consuming wireless technology is desirable [74]. In addition to network discovery, battery information is useful in handover preparation and execution stages. For example, simultaneous usage of two radio interfaces during the handover period can be avoided if battery power is an important issue.

- **Operator policies.** Based on network conditions and other factors, an operator may discourage or encourage a MS to join a wireless network. A change in the operator’s policy is another source of handover triggers.
• **User preference.** In ABC concept, the term “best” is somewhat ambiguous. Users have different preferences and requirements which makes one connectivity option more desirable than another. For example, one user may prefer a cheaper service while another may require a high-quality connection. Therefore, the user’s preferences should be taken into account. However, we cannot expect the user to have the technical knowledge to make a decision when considering the parameters listed above [69]. Instead, the user should be provided with a simple application interface to select the priority of the high-level parameters. A change of user’s preference, although infrequent, results in a handover trigger.

• **Monetary costs.** Cost is often a major concern for users and should be considered in network selection.

• **Network condition.** In some cases, handovers are performed for the purpose of load balancing between alternative RANs when network conditions require the relief of congestion [75] and load balancing [76].

• **Handover frequency.** To avoid a ping-pong effect the MS should take into account the handover frequency [74]. Frequent handovers result in an increase in handover signaling and service interruption.

• **PoA, RAN, and operator identity.** Network identifiers and the operator administering a network may affect the outcome of the network selection [69], for example in cases where the user is not willing to connect to less-known operators.

• **Security mechanisms.** Security features at the link layer of a RAN are certainly important when an MS is selecting it. If there is a incompatibility between the candidate RAN security policies and the MS capabilities, a new connection cannot be established and the candidate network should not be selected. The network selection algorithm should filter out all incompatible candidate networks [77].

• **Provided services.** Connection service provision cannot be inferred by scanning link-layer information. The MS might be able to establish an L2 connection successfully but this does not mean that a specific service will be accessible. For example, an MS which is currently connected to an LTE network might detect a WiFi signal and connect to this network only to find that Internet connectivity is not provided in the new network. This information should be obtained from higher layer mechanisms.
The MS may scan radio channels to obtain some network related parameters. However, in an NGMN heterogeneous environment with many RANs and radio technologies, scanning suffers from high power consumption and delay and cannot be used for collecting all parameters [37, 72]. For a seamless handover, more efficient network discovery solutions are required. These are discussed in next chapter.

The network selection algorithm uses these mobility-related parameters and selects the best network. Common network selection mechanisms range from simple algorithms which consider only one-parameter to more complex multiple-parameter algorithms which rely on fuzzy logic and game theory decision-making techniques [78]. The different network selection algorithms are out of the scope of this research and are not discussed in this thesis.

2.5.4 Handover Preparation

Handover preparation is the second stage in the handover process and includes tasks required to prepare the new link to carry the multimedia application. The target network preparation tasks include:

- PHY and L2 synchronisation and power adjustment. The MS performs time and frequency synchronisation by listening to broadcast information and pilot signals. This process often involves the exchange of several ranging messages between the MS and the PoA.

- Initial PoA association and basic capability negotiation. The MS and the PoA may exchange their basic capabilities.

- Authentication and authorisation. The MS and the target network mutually authenticate each other and security keys are generated and distributed.

- Initial context or bearer setup. The target network creates some context which allows the MS to send and receive occasional data packets.

- Allocation of an IP address.

- Dedicated bearer setup. Dedicated bearers with specific QoS parameters are created in the RAN and the CN to satisfy the requirements of the multimedia application.

- Other processes such as location update, receiving time of the day, etc.

In NGMN, in addition to target network preparation, the MS is also required to register with the IMS to be able to use a multimedia service. The rationale behind this is the NGMN design principle of independence of the service stratum from the
underlying transport stratum. Since the IMS is generally considered a separate network, the MS should be authenticated by the IMS as well as the underlying network. The target network preparation and IMS registration processes can result in a significantly higher handover delay beyond the tolerance of multimedia applications. For a seamless handover, the handover preparation delay should be minimised. Possible solutions are discussed in the next chapter.

2.5.5 Handover Execution

In the handover execution stage, the multimedia session is transferred to the newly prepared link. For this purpose, multimedia data packets should be redirected to the MS’s new IP address. In the literature, there are different handover execution schemes which can be categorised based on their impact on the network infrastructure and the protocol layer in which they operate:

- **Network-based approaches.** The network manages the change of IP address and redirects the packets to the MS. The Mobile IP (MIP) [16] and its variants such as the Proxy Mobile IP (PMIP) [79] fall into this category. Network-based approaches are usually implemented in the IP layer.

- **End-to-end approaches:** The MS and the CH, as end points of the multimedia session, update the IP address information and packets are sent to the current IP address. The standard SIP mobility management [17], the mobile Stream Control Transmission Protocol (mSCTP) [80, 81], and TCP Migrant [82] are examples of end-to-end mobility solutions. These solutions are implemented in the transport and application layers.

- **Network assisted approaches.** In some cases, the performance of the end-to-end mobility management solutions is improved by assistance from network entities. For example, Salsano et al. in [83] offer a SIP-based handover execution with network assistance. IMS mobility solution, called the Service Continuity and Centralization (SCC) [84], also falls under this category.

Handover execution schemes exhibit different advantages and drawbacks. A comparison of the more common protocols was presented by Eddy in [15], and by Mohanty et al. in [85]. Table 2.2 summarises different characteristics of these schemes. It is apparent that none of these protocols outperforms others under all conditions.

Transport layer mobility protocols such as mSCTP and TCP-Migrant, have lower signaling delay and do not require modification in the transport layer. The main disadvantage is that they are not currently deployed in common operating systems. They also do not include integrated location management mechanisms and they
Table 2.2: Comparison of handover control solutions

<table>
<thead>
<tr>
<th>Feature</th>
<th>MIP</th>
<th>mSCTP</th>
<th>TCP-Migrant</th>
<th>SIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network support required?</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>OS modification required?</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Location management included?</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Handover signaling delay</td>
<td>high</td>
<td>high</td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td>Wide-spread adoption?</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Mobility transparency for applications</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Security</td>
<td>good</td>
<td>good</td>
<td>good</td>
<td>good</td>
</tr>
</tbody>
</table>

rely on mechanisms such as the dynamic Domain Name System (DNS) to map the identity of the MS to its current IP address.

The Mobile IP and SIP protocols are considered the leading protocols for mobility management [18]. MIP is implemented in the transport stratum of the NGMN and makes handover transparent to all data services including real-time multimedia applications and non-real time data applications. However, MIP requires modifications to the network infrastructures and the operating systems of mobile nodes. This limitation hinders the widespread implementation of the protocol. There are also many performance issues such as a high delay for registration and handover, triangular routing, the creation of a single point of failure in the network (called the Home Agent), excessive load on the HA and inconsistency with network firewalls.

In designing the IMS, the 3GPP considered mobility management a task for the underlying transport stratum using MIP. For this, the original IMS specification (TS 23.228 [10]) requires the change of IP address to be transparent to the IMS. In IMS specifications, the Security Association (SA) between the MS and the IMS edge node (the P-CSCF) depends on the registered IP address. If the MS’s IP address changes, the SA is dropped and the ongoing multimedia session is broken.

Since MIP was not widely implemented, the 3GPP recently proposed mobility management in the service stratum. The proposed Service Centralization and Continuity (SCC) mechanism [84] enhances the IMS with a SIP-based handover execution capability. Since the SIP protocol is implemented in the application layer, it does not impose complicated changes to the MS’s operating system or the transport stratum. SIP is widely implemented because it is also used as the IMS session management protocol. Apart from terminal mobility, SIP can be used for session and personal mobility [86]. Using this advantage, the proposed SCC solution provides a viable solution for mid-session connection transfer for NGMN. To introduce the SCC mobility solution, we first overview the SIP protocol.

The SIP protocol defines several message types, referred to as *SIP methods* which are used to control a multimedia session. Each message contains several *SIP headers* and the message body. For example, the INVITE method is used to create a new
multimedia session and contains headers including From, To, and CALL-ID which identify the sender, receiver, and the multimedia session. To ensure interoperability of SIP devices, the RFC 3261 [61] lists the mandatory SIP methods and headers. SIP uses the Session Description Protocol (SDP) [87] to describe the parameters of the audio and video components of the session, the IP addresses of the source and destination, and the UDP ports to be used in data transmission.

SIP presents a complete mobility management solution by supporting both the location management and the handover control processes. For location management, the SIP REGISTER method is used to bind the SIP user identities to their current IP addresses. For handover control, the three standard REFER, UPDATE, and re-INVITE SIP methods can be used. The re-INVITE method is widely used in literature for handover execution [88, 89, 90]. This method is shown in Figure 2.23. When the MS obtains a new IP address, it sends an INVITE message with the CALL-ID referring to the existing session and SDP parameters which include the new IP address. This message is often called the re-INVITE message. The CH updates its location information of the MS and replies with a SIP OK message. Finally, the MS acknowledges the completion of SIP handover execution process with a SIP ACK message.

The SCC solution uses a similar approach with some assistance from the IMS network. In this solution, a SCC Application Server (SCC AS) anchors SIP signaling between the MS and the CH. This means when a multimedia session is established, the SCC AS creates two session legs. The leg between the MS and the SCC AS is referred to as the access leg while the signaling connection between the SCC AS and the CH is called the remote leg.
When the MS roams among different RANs and renews its IP address, it is required to perform an IMS registration with its new IP address. This registration process creates an SA between the MS and the new P-CSCF. It then sends a SIP re-INVITE message to the SCC AS to request the transfer of a specific multimedia session to the newly established connection and update the remote leg. This procedure is depicted in Figure 2.24.

Although session continuity can be obtained with this scheme, the signaling delay for a new IMS registration and SCC handover execution significantly increases the overall handover delay. Since SIP messages are text-based with relatively large sizes (a few hundred bytes to a few kilobytes) they can cause a high signaling delay. The interception of SIP messages by network nodes including the IMS nodes and the SCC AS also contributes to the high handover delay. As a result, achieving seamless handover (with no service interruption) by using the SCC mechanism is challenging.

In the next chapter, we focus on the problem of seamless mobility management in heterogeneous wireless networks which use the IMS for service provision. Through our literature review, we describe the possible approaches to achieve seamless mobility and we point out the shortcomings of previously adopted schemes.
Chapter 3

Seamless Handover Control

3.1 Introduction

Seamless handover, in terminal mobility, is defined as the connection migration from one wireless network to another without causing any service interruption. In other words, in seamless handover the mobility management protocol provides network transparency to the user of the MS [88].

In this chapter, we overview prior research works on seamless mobility management in IP-based heterogeneous networks. While the problem of mobility management has been mostly addressed in single-radio technology wireless networks (e.g. LTE and WiMAX), seamless inter-technology handover with low delays and low packet loss is still a research challenge. In the past decade, extensive research efforts has been made to minimise the handover delay and packet loss. We describe some of these research works and identify some remaining research gaps.

We start by describing the challenges of efficient seamless mobility which have been identified in the literature. These challenges include the lack of inter-layer cooperation in the traditional Internet protocol stack, inefficiencies in measuring and collecting mobility-related information, high handover delay, and significant packet loss and connection unreliability during the handover period.

After identifying problems, we overview solutions proposed previously for different aspects of seamless inter-technology handovers. In Section 3.4, we discuss handover delay optimisation solutions using cross-layer techniques, more efficient collection of mobility information, and proactive handover preparation. Although solutions proposed previously can significantly reduce the handover delay and allow the Mobile Station (MS) to remain always best connected, we identify some shortcomings such as the unsuitability for the Next Generation Mobile Networks (NGMN) or the lack of optimal performance in reducing the handover delay.

In Section 3.5, we overview related works on using Soft Handover (SHO) schemes
which simultaneously utilise multiple radio interfaces both to reduce the handover delay and to minimise the packet error rate during the handover period. We describe various SHO schemes which enable the MS to join the target network and leave the old network smoothly. Despite their advantages, it is revealed that SHO methods cannot be utilised in all vertical handover scenarios, and that they generally suffer from the problem of excessive resource consumption.

After identifying research gaps, in the next chapter we present our seamless mobility management framework which takes us one step towards the seamless integration of wireless technologies by improving existing delay optimisation methods and enhancing the performance of SHO schemes in resource consumption.

3.2 The Challenge of Seamless Mobility

Seamless mobility management in IP-based heterogeneous networks is challenging mainly because of the inefficiencies in the traditional Internet protocol stack and heterogeneity in wireless networks. The Internet protocol stack was designed for fixed hosts and the use of mobile terminals was not envisaged [14, 91]. The IP layer and transport layer protocols such as TCP and UDP were devised for situations where the IP addresses do not change. The task of mobility management was not assigned to any specific protocol layer [15]. As a result, a mid-session handover which involves a change of IP address often results in connection teardown with subsequent service degradation reachability problems.

In this regard, the following limitations have been identified for the traditional Internet protocols in managing terminal mobility:

- **Packet routing based on IP address.**
  In the traditional Internet suite, IP packets are routed based on the destination IP address where the IP address uniquely identifies a node and its serving network. The IP address functions as both the host identifier and the locator [14]. With this approach, when the MS changes its location and joins another network, it is forced to obtain a new IP address identifying the new network; this causes reachability problems. Therefore, a mobility management solution is required to allow the MS to retain a permanent identity while changing its location.

- **Limitation of applications and higher-layer protocols.**
  Some Internet applications and higher-layer protocols are aware of node IP addresses and a change of IP address disrupts their operation. For example, SIP uses the initial INVITE message to create a multimedia session. This message contains various parameters including the MS’s IP address and the
UDP ports used for data transmission. To establish a multimedia session, the SIP application in the MS and the CH interact with their operating systems to open the required sockets which are identified by local IP addresses and port numbers. A change of IP address disrupts the multimedia session as the operating system is unable to deliver data packets to multimedia applications. Another example of the dependency of traditional Internet applications on IP addresses is the Dynamic Name System (DNS) protocol where the host name is statically bound to the IP address and mobility causes reachability problems.

- **Lack of mobility awareness and cross-layer cooperation.**

  The modular design of the Internet protocol suite simplifies protocol specifications and provides flexibility in utilising suitable protocols for different applications. For example, UDP can be used for real-time multimedia applications while TCP is used for reliable communication. This classification of networking tasks, however, has an adverse effect on the efficiency of the mobility management mechanism [14]. A well-known problem of this approach is the operation of TCP during a network handover. Since the TCP layer is unaware of mobility, it does not adjust its operation to the newly established network link [91]. Another example is the operation of multimedia applications which do not change their codecs to match the QoS capabilities of the new link.

In addition to these limitations of the traditional Internet protocol stack, the recent trend in designing radio technologies and implementing mobile network infrastructure also makes mobility management more challenging. Providing ubiquitous access to data services requires the coexistence of different wireless technologies and heterogeneity in the access networks. As such, the mobile terminal market is moving towards multi-radio platforms [92] and we have seen the proliferation of multi-modal mobile devices [18, 93]. However, wireless technologies are usually independently designed, implemented, and administered [94]. In order to join a target network and perform a handover, the MS is required to perform the full initial network access procedure.

The advantage of this approach is that each network and radio technology stands alone and can function independently. The disadvantage is that the network entry procedure is not optimised to utilise existing network connectivity via other radio interfaces. As such, the MS must perform the full initial network attachment process which, depending on the radio technology, include tasks such as L2 synchronisation and initial ranging, basic capabilities information exchange, authentication process, location update, default bearer setup, additional information exchange, assignment
of an IP address, and dedicated bearer setup to ensure consistent QoS. Among these
tasks, the authentication process and bearer setup have a high impact on network
attachment delay, mainly because they include the exchange of several signaling
messages between the MS and network, and require retrieving user subscription
information from network data bases. For example, Hur et al. report that a typical
mutual authentication process takes around one second [50]. However, according
to ITU-T recommendation G.114 [19], a high quality real-time multimedia requires
packet latency of less than 400 milliseconds.

In addition to high handover preparation delay, inter-technology handovers also
suffer from high packet loss. During the handover period, the old and the new radio
interfaces often experience poor radio conditions and shadow fading has a dominant
effect. While the MS waits for the new radio link to be prepared and to become
stable, it suffers from high packet loss on the old interface. Excessive handover
preparation and execution delays also have another important implication: since
inter-technology handovers are time-consuming and involve much signaling, frequent
handovers should be avoided. This means on some occasions the MS is forced to
tolerate some degree of packet loss before changing its radio interface.

Figure 3.1 shows a typical handover scenario in which the MS moves away from
PoA1 and towards PoA2. The observed average SNRs of Interface 1 (IF1) and
Interface 2 (IF2) are depicted on the left-hand side. In this scenario, if the MS
selects the stronger radio interface at each given time, it performs handovers back
and forth between PoA1 and PoA2. This results in a notorious ping-pong effect
which can increase packet loss and signaling cost [21].

Similar to traditional handover control schemes in homogeneous mobile networks,
mechanisms such as SNR hysteresis and dwelling-time [95] can be used to avoid the
ping-pong effect in heterogeneous networks. In the hysteresis method (Figure 3.1)
frequent handovers are avoided by introducing a Hysteresis Margin (H) [96]. The MS
performs a handover if the SNR of current interface is less than the Lower Threshold
($T_L$) and the SNR of the new interface is stronger than the Higher Threshold ($T_H$)
where $T_H = T_L + H$. In Figure 3.1, if the hysteresis method is used, a handover
occurs only at point 2. The disadvantage is that the MS may suffer from a service
outage between points 1 and 2, even when the new interface has better reception.

Similarly, in the dwelling-timer method, the MS may may suffer from high packet
loss before performing a handover. In this method, the MS is required to stay con-
ected with the current PoA for a minimum predefined time. For better performance
the dwell time can be dynamically changed by considering parameters such as the
bandwidth [95]. In either case, to avoid frequent inter-technology handovers which
are time consuming and costly, the MS should tolerate some degree of packet loss
and this results in service interruption.
To summarise, efficient seamless inter-technology handover techniques should satisfy the following requirements [85, 33]:

- Minimum handover delay. Vertical handover latency is often beyond the tolerance of real-time multimedia applications. The handover delay should be significantly reduced.

- Minimum packet loss. Inter-technology handovers often include high packet loss resulting in a significant degradation in the quality of multimedia services. A smooth transfer of the ongoing session with minimum packet loss is required for a seamless handover.

The above requirements should be satisfied while considering other challenges such as strong security, consistent QoS across various radio technologies, efficient power management, and dynamic spectrum allocation [97]. Ensuring strong security is particularly challenging because wireless technologies use various authentication and encryption mechanisms. Although a multi-modal MS is required to implement the security mechanisms of the supported radio technologies, sharing security information (for example, encryption keys) between networks is complex. Likewise, matching QoS parameters between wireless networks while maintaining the QoS is challenging. The MS should determine the required QoS parameters and create network bearers before migrating a connection.

NGMN compatibility imposes additional challenges on mobility management frameworks. The service stratum should be considered independent of the transport stratum. As a result, the handover process includes extra signaling and delay to enable the MS to access these strata.

In the last decade, much research has focused on endeavors to devise NGMN mobility management frameworks and mechanisms which address these challenges. In the following sections, we discuss mobility frameworks and architectures proposed previously.
3.3 Seamless Mobility Frameworks and Architectures

Various mobility management frameworks and architectures have been proposed to enable inter-technology handovers. Earlier works can be found on interconnection architectures for WLAN and cellular technologies [98, 99, 100, 101]. Salkintzis et al. [100, 101] overview early standardisation efforts which define six levels of integration between WLAN and cellular networks. In the first level, only billing information is shared between two networks while in the sixth level the MS is able to perform mid-session seamless handovers. They also propose two interconnection architectures, namely, tight coupling and loose coupling.

In tight coupling architecture, as depicted in Figure 3.2, two Radio Access Networks (RAN) are directly connected to a single CN. The secondary RAN (the WiFi network) connects to the main CN using an access gateway which implements the AAA and QoS mechanisms of the main network. Seamless mobility is easier to achieve in the tight coupling architecture because the same mechanisms are deployed across both networks, and data packets are routed to the main CN. However, this architecture is suitable only for situations where one operator owns both networks [100].

The loose coupling architecture provides more flexibility and has broader application [28]. In this architecture, the wireless networks function independently, and route data packets directly to the Internet. Standard IETF protocols are used for the purpose of user authentication and mobility management. This interconnection architecture is more suitable for NGMN network environment where different operators may own and manage RANs by using different security and network control mechanisms. Nevertheless, since the two networks are not directly connected and use different security mechanisms, high handover delay occurs.

A number of mobility management frameworks and protocols have been proposed to improve different aspects of vertical handover performance including billing management, network selection, handover latency, QoS, energy consumption, signaling traffic, security, scalability, and manageability. Regarding network selection, Gustafsson and Johnson in [13] introduced the concept of Always Best Connected (ABC) and note that various parameters should be considered for access network selection. They also present various scenarios for billing agreements between the MS and network operators including a scenario where an ABC service provider (a virtual access network provider) hides multi-operator network environment from end users.

Casetti et al. in [102] proposed the AISLE framework which considers the available bandwidth of the target link for network selection. Their method can only
Figure 3.2: Scenarios for system architecture in interworking of wireless networks
be applied to a multi-homing transport protocol such as Stream Control Transmission Protocol (SCTP) [103]. In the AISLE framework, the MS creates several data paths to the remote destination, also called the Corresponding Host (CH). The MS selects one of the connections as the primary connection and uses it for data transmission. The MS constantly estimates the bandwidth of both data connections by transmitting back-to-back SCTP control packets over all paths and measuring the dispersion of packet reception time at the CH side. Based on this measurement, the MS may change its primary connection and perform a handover to achieve higher throughput.

Similarly, Andersson et al. in [104] propose the Multimedia Mobility Manager ($M^4$) framework which uses the Round Trip Time (RTT) of control messages to estimate the bandwidth of several radio links and then select the one with the higher throughput. In particular, they use the Binding Update messages of MIP to estimate the data rate of each radio interface. If a higher available rate determines a handover should occur, this is executed using the standard MIP mechanism.

Using MIP or SCTP control messages has the advantage of being radio technology independent. However, the AISLE and $M^4$ frameworks suffer from several shortcomings. First, a limited number of network selection criteria are considered. Second, the measurement of radio resources by measuring parameters such as current bandwidth is often inaccurate and prone to change. Modern wireless technologies such as LTE and WiMAX contain centralised radio resource management frameworks and in this environment, the MS can query network elements for available resources.

Ciubotaru and Muntean in [105] propose the Multimedia Mobility Management System (M3S) framework which considers more network selection criteria including parameters such as QoS capability, power efficiency, and service cost. In the M3S framework, an MS receives a multimedia streaming service from an application server. Based on the multimedia packets received, the MS calculates the Quality of Multimedia Streaming (QMS) metric. The QMS is the weighted sum of various interface selection parameters. The MS reports calculated QMS values to the multimedia application server which then sends data packets to the interface of choice. Although M3S uses more parameters for better selection of radio interfaces, it still suffers from the other shortcomings of the AISLE and $M^4$ frameworks. Moreover, M3S does not use standard mechanisms for the exchange of network selection parameters.

Given the shortcomings of the above solutions, a standard framework for the fast exchange of comprehensive mobility information between the MS and network nodes is required. The IEEE 802.21 [106], also referred to as the Media Independent Handover (MIH) protocol, provides such a framework. An overview of the MIH
The MIH protocol can be used to provide the MS with extensive mobility-related information. The static information such as the coverage area of a RAN, the operator and existing roaming agreements, the supported services, and service costs can be stored in MIH Information Servers and delivered to the MS long before performing a handover. More dynamic information, such as the current network load and available radio resources, can be obtained when the MS approaches the target PoA. The MS uses the MIH Command Services to communicate with the target PoA which implements an MIH Function (MIHF). The MS may use the existing radio connection to transmit the MIH commands and the target PoA can receive them over its backbone link.

The MIH protocol has been used in many research works on seamless mobility [110, 111, 107, 112, 113, 114, 77]. However, it should be noted that the MIH protocol does not provide a complete mobility management solution. A mobility management protocol such as MIP or SIP is still required for handover execution. The MIH protocol is often used to collect the MAC or PHY layer information. The higher-layer mobility management protocol is then triggered to perform the actual handover process. This results in cross-layer mobility management frameworks which have proved to be effective solutions in providing seamless mobility. This concept was used by Kalle et al. in [115] for proposing the SWIFT mobility architecture. They use the MIH event services to predict an imminent handover and improve the performance of the Mobile IP Version 6 (MIPv6) protocol. A similar approach was also used by Mussabbir et al. [113]. SIP mobility management performance enhancement using the MIH protocol has also been reported by Almosawi et al. [116]. These solutions reduce the handover delay by predicting imminent handovers; however, they do not optimise the handover process itself. Another shortcoming is that they do not consider the QoS, an important factor for real-time multimedia applications.

However, The MIH protocol can also be used to create mobility management frameworks with integrated QoS mechanisms. The MIH protocol provides command services which enable the MS or the current RAN to query the target RAN for available resources. The target network can also be requested to reserve the required resources in advance. Baek et al. in [112] use these capabilities, with some modifications, to reserve radio resources in the target RAN before committing a handover. However, the authors do not consider the interaction with the IMS and the Policy and Charging Control (PCC) framework.

IMS is becoming the de facto standard framework for providing multimedia services in IP-based mobile networks [117]. It contains functionalities for authenticating users, establishing and controlling multimedia sessions, and ensuring consistent QoS by interacting the PCC framework. By using the IMS as the basis for the mobil-
ity management framework, a complete solution with integrated security and QoS capabilities can be created. Considering these advantages, and motivated by the slow adoption of MIP, the 3GPP enhanced the IMS with the Service Continuity and Centralization (SCC) framework [84] which realises session continuity as the MS changes its serving network. However, due to the extensive signaling required to perform a handover, achieving seamless mobility is challenging.

Much existing research focuses on using the IMS for mobility management [58, 118, 119, 86, 120, 121]. A preliminary IMS-based mobility management framework, called the IN-SMA architecture, was proposed by Kalmanek et al. in [118]. They implement a Mobility Server and provide mobility management as an advanced IMS service. Although the authors claim the framework can be used for seamless mobility, they do not discuss the required mechanisms for handover delay optimisation.

Udugama et al. in [121] propose the NetCAPE framework which attempts to achieve seamless mobility by improving network discovery and the selection process. They do not use SIP-based mobility management. Instead they use the MIP which hides MS mobility from the IMS. The NetCAPE framework improves the delay in network discovery by monitoring radio interfaces and detecting inter-technology handovers. In the NetCAPE framework, the authors propose a policy engine which consists of different functional blocks including an IMS interaction module for obtaining the requirements of multimedia applications, a GUI for collecting the user’s preference, and a LINE interface module which interacts with and monitors radio interfaces. In this solution, when a radio link is disconnected, instead of waiting for MIP Router Advertisement messages, the MS can immediately initiate a handover process by selecting a target network. Although, the handover detection time is decreased, the NetCAPE framework is not overly effective in minimising the overall handover delay. As discussed in the previous section, handover preparation and execution impose significant delays on the handover process. Therefore, it is insufficient to reduce handover detection time – a more proactive mobility management framework is required for seamless handover. Moreover, the NetCAPE approach is based on tight coupling architecture which limits its application in NGMN.

The MA-PCSCF framework proposed by Arnaud and Negro [122] attempts to achieve seamless mobility through Soft Handover (SHO) and packet duplication in the transport stratum. However, the solution requires modifying IMS nodes particularly the P-CSCF.

Bella vista et al. in [58] propose the IHMAS framework which is fully IMS-compliant. The authors use a SIP-based handover control mechanism and the MIP is not required. To reduce significant modifications on the IMS infrastructure and protocols, the IHMAS framework uses an IMS Application Server (AS) to provide mobility service. The proposed IHMAS framework is capable of providing proactive
Hard Handovers (HHO) and Soft Handovers (SHO). In the proactive HHO, when the MS intends to perform a handover, media packets are forwarded to the target network and buffered in a Media Proxy (MP). When the MS finally joins the target network, it immediately receives buffered packets. The authors argue that during the proactive HHO procedure, the MS may consume its local buffer to support session continuity. Therefore, it seems that this mechanism is more suitable for multimedia streaming applications. For real-time applications, the SHO mechanism can be used in which the MP in the current RAN duplicates and forwards packets to the target RAN. The MS then receives duplicated multimedia streams over its current and new radio interfaces.

The IHMAS framework proposed by Bellavista et al. has several advantages. First, because of being IMS-compliant it is well suited to the NGMN network environment in which the IMS is commonly used as the basis for the service stratum. Second, the IHMAS framework eliminates the need for MIP for mobility management, instead using SIP which is commonly implemented for session management. Third, it supports reactive and proactive HHO and SHO mechanisms to satisfy the requirements of different applications. Despite these advantages, the IHMAS is not a complete solution for seamless mobility. The authors do not consider QoS management and the delay associated with the creation of network bearers in order to guarantee consistent QoS across both the old and new RANs. To verify IHMAS performance in achieving seamless handovers, the authors used Bluetooth and WiFi technologies which do not have specific procedures for the creation of radio bearers. Moreover, the IHMAS framework does not include optimised network registration mechanisms to minimise handover delay.

Based on the shortcomings of frameworks proposed previously, we conclude that for efficient mobility management in NGMN, a framework is required which the following characteristics:

- Support for different interworking architecture. An efficient mobility management framework should be flexible in supporting a variety of interconnection architectures including both tight and loose coupling options.

- Integration with the QoS control framework. The mobility management framework should interact with existing network access control systems and QoS control frameworks. Through this interaction, the target network can be prepared to accommodate the multimedia session.

- Minimum handover delay. The security and QoS management mechanisms employed in the mobility management framework should be optimised to impose minimum delay on the overall handover procedure.
In the next section, we discuss some of the available solutions for network registration and QoS management in mobility management frameworks.

3.4 Handover Delay Optimisation: Cross-layer Approaches

The handover procedure incorporates three stages: network discovery, handover preparation and handover execution. Previous research attempted to achieve seamless handover by minimising the delay associated with each stage. Delay optimisation solutions usually require interaction between mechanisms operating at different protocol layers resulting in cross-layer handover optimisation techniques. A complete seamless handover control solution often combines various cross-layer techniques. In this section we describe some of the promising handover delay optimisation schemes and discuss their advantages and potential shortcomings.

3.4.1 Network Discovery Optimisation

The optimisation of network discovery stages can be achieved by the early detection of handover triggers, fast and proactive collection of mobility information, and the optimum selection of the target RAN and PoA. Typical handover triggers include the degradation of radio conditions and decaying Received Signal Strength (RSS). Much prior research work uses PHY layer information to detect the necessity of inter-technology handovers [21, 25, 85, 123].

Chang et al. in [21] use RSS prediction to initiate vertical handovers. The authors use polynomial regression-based curve fitting to predict future samples of the RSS of different radio interfaces. Guo et al. in [123] also detect decaying RSS by applying a technique based on the Fast Fourier Transform (FFT). Similarly, Lee et al. in [25] propose predicting RSS values using an Auto-Regressive Integrated Moving Average model (ARIMA). Lee et al. compare the predicted values of RSS of the old and the new PoAs to detect a handover a short time before the actual handover occurs. As a result, the MIP handover procedure can be initiated around 100 milliseconds faster than a reactive scheme.

Although RSS prediction is helpful in fast handover detection, this short-term prediction cannot significantly improve the handover performance, sufficiently enough to achieve a seamless handover. Long-term RSS prediction using this technique is not possible because radio channel conditions under both fast and slow fading effects become uncorrelated after a few milliseconds and a few hundred milliseconds respectively. However, longer-term prediction of imminent handover can be achieved using
higher-layer information. The user’s location information and mobility pattern can significantly contribute to achieving a seamless handover [124].

The idea is to use the MS’s location information in combination with the coverage area of different wireless technologies to predict imminent handovers to specific candidate PoAs.

Based on this concept, Hsieh et al. proposed Seamless Mobile IP (S-MIP) [26]. The S-MIP system architecture contains a Decision Engine (DE), a network node which makes handover decisions and coordinates the handover process. The DE determines the MS’s movement pattern. If the MS is moving linearly, the DE then determines the target Access Router (AR) and instructs it to prepare for an imminent handover. If the MS is stationary at the boundary area of two PoAs with overlapping radio coverage, the MS obtains two IP addresses from each network and the DE requests both ARs to maintain binding with the MS. If the MS moves stochastically, all neighboring ARs are requested to anticipate hosting the MS. Since in most cases, terminal mobility is restricted by man-made constructions (for example, roads and buildings), predictable movement occurs more often. Therefore, by using the current location of a MS, its direction of movement, and a civic map, possible future locations can be predicted.

Akyildiz and Wang in [125] use the user mobility profile to predict future visited zones if the exact cell cannot be determined. To further improve the prediction accuracy, Celebi et al. in [126] suggest that statistical learning such as neural networks and Markov models can be used to predict the MS trajectory based on previous movement information. These solutions can be used to predict future locations of the MS and estimate radio conditions. Imminent handovers, therefore, can be predicted. However, we note that network selection in NGMN is not only based on radio conditions.

In addition to radio conditions, optimum PoA selection in NGMN involves parameters including network traffic load, QoS capabilities, supported services, and their costs. Service interruption may occur if careful network selection is not performed. For example, the MS may connect to a PoA and discover that the new RAN does not provide IMS connectivity or that radio resources cannot be granted. For a seamless handover, it should be possible to collect mobility-related information with minimum delay and, if possible, prior to actually performing the handover.

As mentioned before, the IEEE 802.21 (the MIH protocol) provides a framework for the fast collection of mobility information [106]. Using the MIH protocol, the MS can obtain network level information such as service costs and the roaming agreements of available RANs. The MS can use the predicted locations to limit the number of RANs and the scope of provided information. To expedite the network selection procedure, the MS may use the received information to filter out PoAs with
insufficient capabilities. For example, Izquierdo et al. in [77] propose using the MIH Information Services to obtain security policies of candidate RANs and filter out incompatible PoAs. As a result, the MS does not attempt to connect to unsuitable PoAs and thus increase network discovery and handover preparation latencies.

3.4.2 Proactive Preparation Optimisation

After a fast and proactive network discovery, the MS attempts to prepare a radio link to the newly selected RAN. The handover preparation stage typically includes association with the target PoA, registration to the target network (including authentication and authorisation processes), creation of required network bearers to satisfy QoS requirements of the multimedia session, and registration to the IMS through the target network. These tasks significantly contribute to the overall handover delay. For example, a full authentication procedure using an Extensible Authentication Protocol (EAP) method may take 1000 milliseconds which by itself is beyond the tolerance of multimedia applications [50]. For this reason, there has been significant amount of research on optimising different handover preparation tasks at different protocol layers.

Some of these research proposals include fast authentication schemes [127, 128, 129], various context transfer methods [120, 130], combined network/IMS registration proposals [22], IMS pre-registration schemes [24], decoupling of mobility management processes from radio connectivity [131], and finally pre-registration schemes [23] including both pre-authentication and pre-configuration. These methods are indeed effective in significantly reducing the handover delay. However, some of them suffer from a number of shortcomings while the performance of others can yet be improved.

Two well-known fast authentication approaches include the re-authentication schemes and pre-authentication methods [127]. In the re-authentication method, as proposed by Narayanan et al. in [132], the target network uses security obtained from the old network to authenticate the MS and derive the rest of security keys. This method requires some modification to authentication servers to allow the transfer of the security keys from the old network to the target network. When the MS joins the new network it does not need to perform the full authentication procedure. Instead, the MS uses the less time-consuming re-authentication process. Using this concept, Rodriguez et al. in [110] propose security context transfer using the MIH protocol. In their solution, when a handover is required, the MS uses an MIH Command Service and requests the target PoA to retrieve the security context from the old PoA. Since the standard MIH protocol does not include such a feature, Rodriguez et al. propose new MIH messages including AAA Context Request and
The re-authentication schemes suffer from a number of disadvantages. First, the old network must be willing to share its generated master key, and the target network willing to accept the validity of this key. This arrangement requires a mutual trust between operators of different networks [110]. Second, the target network should use cryptographic methods which match the ones of the old network to the extent that reusing security keys becomes possible. Third, the re-authentication approach may be susceptible to the domino affect [128], a situation where a compromise of security in one network jeopardises the security of other networks.

The pre-authentication method, as described by Ohba in [129], does not require mutual trust between operators or compatible cryptographic schemes in the current and target networks. In this method, the MS uses its old radio interface to perform the authentication procedure with the AAA server in the target network. In the pre-authentication scheme, the old network functions as a relay between the MS and new AAA. In some cases, the standard authentication method of the target network should be modified to allow pre-authentication. For example, in a handover to LTE, the EAP-AKA authentication method should be used instead of the standard AKA procedure. However, the procedure remains almost intact, and the security of neither the target or current networks is compromised. The pre-authentication approach is recognised as a viable delay reduction solution by the IEEE in WiMAX (IEEE 802.16e) standard [133] and by the 3GPP2 in [134]. Despite its advantages, the pre-authentication scheme can only be used in situations where the MS and both the current and target networks support it.

The re-authentication and pre-authentication concepts can also be used for other network and IMS registration tasks. Similar to the re-authentication approach, Kempf in [135] proposes network context transfer from the old Access Router (AR) to the new AR. The network context includes AAA, header compression, and QoS information which enables the new AR to replicate the state information in the MS’s new network path resulting in less signaling overhead and delay. Context transfer for QoS purposes is applicable in situations where the old and new networks use similar QoS mechanisms. However, as described in sections 2.3 and 2.2, LTE and WiMAX networks usually use technology-specific mechanisms to create network and radio bearers to ensure QoS and therefore, little information can be shared between these technologies. Context transfer, however, can be useful for IMS registration process.

Context transfer for IMS registration has been proposed by Renier et al. in [120] and Farahbakhsh et al. in [130]. In this method, during the handover preparation process the IMS context including session states and the Security Association (SA) parameters are transferred from the current P-CSCF to the new P-CSCF. With
this approach the number of messages required for IMS registration is reduced by half. In the full IMS registration procedure, as described in Section 2.4, the MS sends and receives four SIP messages to complete the registration procedure. In order to authenticate the user, the S-CSCF needs to query the HSS and obtain the Authentication Vector. The S-CSCF is also required to inform the HSS of user registration if the authentication process is successful. This results in excessive registration delay. In the context transfer approach, as proposed by Renier et al., the MS only sends and receives two SIP messages. It first sends a Re-Register message to the new P-CSCF which initiates the procedure. The new P-CSCF queries the old P-CSCF for IMS context and then sends a Route Update to the S-CSCF. Finally the S-CSCF updates the HSS record of the user registration and replies with a SIP OK message.

The IMS Re-Register method suffers from a number of shortcomings. First, the P-CSCFs must be modified to support context request/response. Second, the MS and the new P-CSCF cannot negotiate the SA parameters. The full IMS registration process is used for both user authentication and the negotiation of SA parameters between the MS and the new P-CSCF. But in the context transfer approach, the new P-CSCF should use same security parameters. Third, in the current version of IMS, the SA depends on the MS’s IP address; removing this dependence is a significant modification to the IMS specifications and may cause security issues.

Another solution to minimise IMS registration delay is the one-pass registration scheme proposed by Yi-Bing et al. in [22]. In this scheme, the IMS registration procedure is combined the network authentication process. The MS first performs the regular 3GPP authentication process to join the network. The MS then sends a SIP Register message to the P-CSCF. This message is intercepted and amended by the network node which has already authenticated the MS. This allows the S-CSCF in the IMS network to authenticate the MS faster and return the SIP OK message to the MS. Although this method reduces the number of required SIP messages by half, there are some disadvantages. The network node must be modified to support the SIP protocol which will further enable it to modify SIP messages. Moreover, the IMS network must now rely on the transport network’s ability to authenticate the user. This approach, however, conflicts with the fundamental design requirement of NGMN where the service stratum is independent from the underlying network.

To overcome these problems, the IMS pre-registration approach, based on the pre-authentication concept, can be used. In this method, as proposed by Ito et al in [24], when the MS selects the target PoA, it proactively obtains a new IP addresses from the target network. Then, the MS uses its current radio interface to establish a data tunnel to the target PoA. This tunnel is used for standard IMS registration through the target network. Since IMS allows for the registration of
several IP addresses, the MS can maintain its new IMS registration and use it when it eventually joins the target PoA.

The approach used in IMS pre-registration can also be used for other network registration tasks such as IP address configuration, QoS setup, and location update. In this regard, Dutta et al. in [23] propose a pre-registration mobility framework which includes both pre-authentication and pre-configuration. In their framework, two nodes called the Authentication Agent (AA) and the Configuration Agent (CA) are implemented in each wireless network. As the name suggests, the AA is responsible for proactively authenticating clients using the pre-authentication approach. The CA, on the other hand, communicates securely with the MS to deliver the configurations required, including the newly assigned IP address. Although the authors do not explicitly address the QoS management mechanisms, it seems that the secure link between the MS and the CA enables the MS to proactively prepare the target network to accommodate the ongoing multimedia session to be transferred.

The idea of using one radio interface to prepare another interface, as used in the pre-registration mechanism, seems a viable approach to reduce the handover preparation delay. Traditionally, wireless technologies were designed to be standalone, meaning the MS does not need any other network connection to join a wireless network. All configurations can be performed over the air. However, this approach results in excessive link preparation delay. With ubiquitous network access, the MS often has multiple network connectivity. The pre-registration solution exploits this network environment and optimises the network registration process. However, previous pre-authentication and pre-registration schemes can still be improved to further reduce handover delay.

In the pre-registration methods proposed by Dutta et al in [23], there is no separation between the radio-dependent tasks and non-radio procedures. As a result, the pre-registration process can be initiated after the selection of the target RAN and the PoA within the network. To elaborate on this issue, we consider some steps of the handover preparation process that can be expedited by decoupling radio-independent tasks from radio-dependent processes. In step one of the handover preparation process, the MS is authenticated in the target RAN. In this step, the identity of the target PoA is often needed to generate the required security keys which allows traffic encryption and integrity checking of data packet transmission over the new radio link. As described in Section 2.2 and 2.3 for WiMAX and LTE technologies, these security keys depend on the identity of the target PoA. Therefore the MS should first select the target PoA and then initiate the authentication process. Similarly, in creating network bearers to ensure satisfying QoS parameters, the common approach in wireless networks is to verify availability of radio resources at the target cell and again, the target PoA should be selected first. To overcome
these shortcomings, we have proposed a two-step pre-registration process called the PRIME mechanism which is detailed in the next chapter.

3.4.3 Proactive Handover Execution

The proactive approaches used in the network discovery and handover preparation stages can also be used in the handover execution stage. In this stage, routing information is updated and data packets are delivered to the target network. In proactive handover execution schemes, some of the required tasks are performed while the MS continues to receive data packets over the current interface.

Fast Mobile IPv6 (FMIPv6) [131], developed by IETF MIPSHOP Working Group, uses the concept of proactive handover execution to reduce the handover delay of the MIPv6. In this protocol, the MS first detects the presence of a new PoA. Proactive network discovery solutions may be used for this purpose. The MS then initiates the proactive handover preparation process by querying its current Access Router (AR) to determine the AR associated with the new PoA. Using the information received, the MS creates a new Care of Address (CoA) which is an IP address that can be used in the target network. In the next steps, the MS performs proactive handover execution by executing the L3 handover even if the L2 handover is not completed yet. For this purpose, the MS sends a Fast Binding Update (FBU) message over its current interface to register the new CoA. The current AR acknowledges receipt of the binding update message and immediately tunnels data packets to the new RAN. The new AR receives and buffers data packets and waits for the MS to complete the link-layer handover.

When the MS joins the new AR, it sends a specific message to the new AR to announce its presence in the target network. Subsequently, the new AR can forward buffered packets. Therefore, in the FMIPv6 protocol, the L3 handover execution latency is removed from the overall handover delay. This delay reduction due to proactive handover execution is not sufficient for a seamless handover and the FMIPv6 should be combined with optimisation techniques for network discovery and handover preparation. For that reason, Dutta et al. in [23] integrate proactive tunnel establishment into their pre-registration scheme for seamless handovers. Their approach for proactive handover execution is similar to FMIPv6 and based on packet forwarding and buffering at the new AR in anticipation of a handover.

Despite its advantages, proactive handover execution with packet forwarding suffers from a number of shortcomings. Firstly, the FMIPv6 requires the exchange of several signaling messages between the MS and the old AR. Since the radio condition of the wireless interface may change rapidly, it is possible that the MS may lose its radio connectivity before completing the FMIPv6 procedure [113]. As
a result, the tunnel between the old AR and the new AR is not created. In these cases, the FMIPv6 specifies that the MS must resort to reactive handover execution and service interruption may occur.

Secondly, it is often difficult to determine precisely when proactive handover execution and packet forwarding should be initiated. If packet forwarding starts too early, the MS may not have an active or stable radio connection with the new PoA. Therefore it cannot receive data packets over the new interface or it will experience high Packet Error Rate (PER). If packet forwarding is initiated too late, the current interface becomes unavailable and service interruption occurs. This problem is highlighted when shadow fading has a dominant effect, usually at the edges of cells where handovers are more common.

Thirdly, similar to other HHO schemes, FMIPv6 is susceptible to the ping-pong effect in which the MS performs frequent handovers between the old and new PoAs. If proactive tunnel establishment is used, data packets are forwarded back and forth between the old AR and the new AR. As a result, data packets may be delivered with unacceptable latency or eventually dropped.

To overcome these problems, the Soft Handover (SHO) approach can be used. SHO has been shown to significantly reduce the PER which the MS observes during the handover period. In the next section, we overview some of the related research which use the SHO approach to minimise the PER.

3.5 Packet Loss Reduction: SHO Approaches

A Soft Handover (SHO) is defined as a handover process in which the MS is associated with more than one PoA concurrently [136]. The Hard Handover (HHO), on the other hand, is a scenario where the MS is associated with only one PoA. The SHO approach can be used in mobile terminals with multiple radio interfaces which can be used simultaneously. Currently many mobile devices from different manufacturers contain multiple wireless interfaces and it is expected that in the near future all Internet-connected portable devices will support several wireless technologies [63]. The prevalence of multi-mode mobile terminals presents a great opportunity to reduce the handover delay to near zero and greatly improve packet delivery performance [137]. In the follow subsections, we describe some of SHO applications in inter-technology handovers.

3.5.1 SHO for Network Discovery And Preparation

In performing inter-technology handovers, the SHO approach can be used for the all or part of the handover process. For example, the MS may use SHO for network
discovery and handover preparation, and HHO for handover execution. This SHO approach is also called the *make-before-break* approach [75, 138].

To start the make-before-break procedure, a multi-mode MS uses one of its radio interfaces to discover the target RAN while simultaneously receiving data packet over another interface. The MS may use the standard scanning procedure of the target target radio technology to obtain information on the target RAN. However, as Yiping et al in [72] explain, information collected through scanning is not suitable for ABC purposes in heterogeneous networks. The scanning process is time- and power-consuming and the information collected is usually not sufficient for optimum network selection. As such, the location-based access discovery architecture proposed by Yiping et al. or other proactive schemes such as the MIH protocol are still required in the SHO network discovery process.

To prepare the radio interface in a SHO, the MS may perform the standard network entry procedure of the target RAN. Seamless handover can be achieved if the MS can maintain its current radio connectivity until the new link is ready. However in some cases, signal degradation over the current interface occurs too quickly and the MS loses connectivity before preparing the target interface. Abrupt signal degradation is common for fast moving mobile terminals or RANs with a limited coverage such as WLANs. Azhari et al. in [139] propose a solution for call continuity in WLAN-to-cellular handovers. In this solution, the author assumes the MS is already attached to a cellular network and only the session establishment delay should be overcome. They propose a Vertical Handover Support Node (VHSN) to retrieve the ongoing voice call and redirect it to the local cellular base station which sends voice packets to the MS. Meanwhile, the MS establishes a new voice session over the cellular network and completes the handover process.

This solution is limited to a specific interworking scenario and voice call services. More general solutions include faster handover preparation mechanisms such as the regional registration schemes for SIP [140] and MIP [141] which can be used to reduce the handover preparation delay. However, to the best of our knowledge, regional registration for IMS has not been proposed in the literature. The one-step authentication method, proposed by Yi-Bing et al. in [22], can be used for delay reduction. But as mentioned before, it is not suitable for NGMN environments. Moreover, optimising the IMS registration process is not sufficient to ensure a seamless handover. Other network registration tasks such as authentication and QoS setup impose delays beyond the tolerance of real-time multimedia applications.

In summary, SHO network discovery and handover preparation is not capable of achieving seamless handover in all cases. In some cases, the signal of the current RAN degrades rapidly and the new link preparation cannot be completed on time. As a result, packet loss occurs which causes significant service interruptions. We
should also stress that SHO schemes do not improve the optimality of the MS’s radio connection in ABC scenarios: the MS is forced to use its old interface while preparing the new link, even if the new RAN provides a better connectivity service. This shortcoming is highlighted in a situation where the new RAN has limited coverage and is temporarily available. In this case, the handover preparation latency is significant compared with the duration of the opportunity of transmission over a faster RAN. One solution, as proposed by Lee et al. in [73], is to avoid the selection of such RANs by fast moving MSs. Alternatively, the network selection and preparation processes should be optimised.

3.5.2 SHO for Handover Execution

In SHO handover execution schemes, for a period of time the MS has several active radio interfaces for data transmission and reception. Receiving multiple data streams enables the MS to maintain the best possible radio connectivity with minimum delay and the lowest Packet Error Rate (PER). There has been an impressive amount of research work on the approach to SHO handover execution. These research works include proposed packet duplication at different protocol layers, performance analysis of SHO schemes, and the incorporation of data coding into packet duplication mechanisms.

SHO handover execution in physical and link layers was proposed in the early 1990s in CDMA-based networks [142]. With the UMTS technology, the MS communicates with up to three base stations to realise a SHO [143]. In this type of the handover, the MS should be able to simultaneously decode up to three physical layers and merge the information received to extract data packets. This approach requires all base stations to be synchronised and transmit the same data frame at each given time [144]. SHO at the MAC layer is also challenging, even in cases where interfaces use the same radio transmission technology. This is because data streams are always encrypted at MAC modules and cannot be directly combined. Therefore, SHO schemes which include packet duplication and merging above the MAC layer (for example, the IP layer) are more suitable for NGMN heterogeneous access networks.

Liu et al. in [27] and [145] attempt to achieve SHO by proposing an Interworking (IW) sublayer which is implemented on top of the MAC layer in the MS and network nodes. The IW peers are responsible for packet duplication and merging, and controlling retransmission over wireless links. Similarly, SHO through packet duplication in the network layer has been proposed by Hamdaoui and Ramanathan in [144]. The solution, called Soft Handover over IP (SHIP), include a network node (SHIP FA) which duplicates and forwards data packets to the current and target
RANs.

The disadvantage of the above solutions is that they require the introduction of new protocols and significant modifications to the MS’s operating system and network infrastructure. This shortcoming can be removed by using standard protocols such as MIP and SIP. Mobile IP Version 4 (MIPv4) allows simultaneous registration of multiple Care of Addresses (CoAs) [16]. With this feature, the MS can register its acquired IP addresses with the Home Agent (HA). In the MIP Registration messages, the MS can set the bit “s” to indicate that the HA should consider all CoAs as contemporary. As a result, when data packets arrive at the HA, it duplicates and forwards them towards all CoAs. With this n-cast approach, packet loss due to handover delay and ping-pong effect is minimised. To control network load, the HA may reject the MS’s simultaneous registration requests if a limit has reached. A similar mechanism for Fast MIPv6 has been proposed by El Malki and Soliman in [30].

SHO using SIP has been proposed by many researchers [17, 88, 29, 89, 83]. SIP as a mobility management protocol was originally proposed by Schulzrinne and Wedlund in [17]. In the same paper, the authors suggest packet duplication to minimise PER during the handover procedure. They recommend using a network node, called the RTP translator, which can create a duplicated data stream towards the MS. The RTP translator can rewrite SIP messages in order to hide MS mobility from the CH.

Banerjee et al. [88] propose SHO for SIP-based multimedia sessions. The proposed approach requires the modification of base stations by adding a Back-to-back User Agent (B2BUA) which intercepts SIP messages from and towards the MS. If a handover is needed, the serving base station duplicates RTP packets and forwards them to the target RAN. Similarly, Salsano et al. in [83] propose modifying the Session Boarder Controller (SBC) to anchor both SIP signaling and data streams. While hiding MS mobility from the CH, the SBC is able to duplicate data packets and send them to MS’s new and old interfaces.

The above solutions require the introduction (or modification) of a network node for packet duplication. This approach has two important benefits. First, packet duplication is not required at the CH which may be a mobile device with limited bandwidth. In fact, the CH is not required to support soft handover and no modification at the end node is needed. Second, since the MS is no longer required to exchange handover signaling with the CH, the excessive signaling delay associated with an end-to-end protocol such as SIP is reduced. The MS is only required to inform the packet duplication agent to begin receiving multimedia packets over its newly established radio interface.

The disadvantage is that SHO support from the underlying transport network
is required. To overcome this problem, some researchers have proposed network-independent SHO schemes. Koh et al. in [89] propose an end-to-end packet duplication seamless mobility solution based on SIP. The solution, called Mobile SIP (mSIP), introduces a new handover header to the SIP protocol which allows the MS to inform the CH that it has acquired a new IP address. Consequently, the CH replicates data packets and forwards them to both IP addresses.

Packet duplication at end points is also used in transport layer SHO schemes. Koh et al. in [81] propose Mobile SCTP (mSCTP) which extends the SCTP protocol to allow SHO with packet duplication at the MS and the CH. Taking a similar approach, Aydin et al. in [146] propose the Cellular SCTP (cSCTP) mobility solution. These end-to-end SHO schemes do not require mobility support from wireless networks. SHO support is only required at the MS and the CH. Both network-based and end-to-end packet duplication schemes significantly reduce the experienced PER in the handover execution period.

The PER performance of SHO schemes has been the subject of much research [147, 32, 31, 144]. Roy and Chandra in [147] provide a closed-form formula for the probability of radio coverage outage defined as the probability of an event in which that the MS is not able to communicate with any PoA. They show mathematically that using a SHO scheme involving multiple PoAs is effective in mitigating the adverse effect of shadow fading on outage probability. Roy and Chandra, however, do not consider practical issues such as signaling delay. Belghoul et al. in [32] compare the PER of Fast MIPv6 with and without bicasting by considering signaling delay. As a reference, they also present the PER of an ideal SHO in which the MS is always connected to both RANs. Their simulation results show that simple packet duplication reduces PER of the Fast Mobile IP by nearly seventy per cent.

Huang et al. in [31] show that the overall radio connectivity can be significantly improved by more complex packet replication. Instead of simply replicating data packets, Huang et al. consider a SHO mechanism which applies a (2K, K) Reed-Solomon [148] code to generate K check bits for a K-bit data packet. The coded bits and the check bits are separately inserted into two data packets and forwarded to the old and new RANs. The MS receives both packets and combines them at the IP layer to extract the original k-bit data packet. For SHO performance evaluation, the authors define the Packet Delivery Gain (PDG) which is the ratio of the average PER of the HHO scheme over the PER of SHO approach:

$$PDG = 10 \log \left( \frac{\text{PER}_{\text{HHO}}(L_1, \ldots, L_n)}{\text{PER}_{\text{SHO}}(L_1, \ldots, L_n)} \right)$$

79
the PDG graph based on varying PER over radio interfaces. As expected, it is shown that high packet delivery gain can be achieved in certain circumstances, for example when one of the links has low PER. Similar results were reported by Matsuoka and Yoshimora in [149]. They show that SHO with Forward Error Correction (FEC) using Reed-Solomon coding significantly improves the average peak signal-to-noise ratio of video frames in a multimedia session.

Hamdaoui and Ramanathan in [144] also present a performance evaluation for simple packet replication and Reed-Solomon coding in homogeneous networks. They first use simulation results to show the Reed-Solomon SHO method outperforms replication SHO and HHO schemes. Then, through mathematical analysis, they present an interesting argument. Since SHO results in much lower PERs, under certain conditions, it is possible to reduce the total transmission power of PoAs and still obtain a PER equivalent to the one of HHO. To draw this conclusion, the authors assume that 1) users are uniformly distributed, 2) radio cells are adjacent to each other with hexagonal structure, 3) there is no shadow fading, and 4) PoAs can communicate to adjust their transmission power. However, these assumptions are not applicable to NGMN with heterogeneous access networks. In NGMN, RANs have different cell sizes with overlapping radio coverage. Each RAN independently tries to provide the lowest possible PER for each connected MS. Since the MS holds multiple independent radio connections, the power consumption significantly increases and radio resources are wasted. Excessive resource consumption is the major shortcoming of SHO schemes.

Radio resources are scarce in wireless networks. The radio cells are often congested and service providers strive to cope with the increasing demands of mobile traffic. In these circumstances, employing SHO schemes using packet replication places further stress on wireless networks. Radio resources are particularly scarce at the edges of cells where handovers are more likely to occur. Irmer et al. in [150] present their LTE field trial results which indicate that the cell border throughput on a 10 MHz radio channel is around 300kbps, significantly less than the average sector throughput. This limited radio capacity may prevent the application of a SHO handover execution. MS battery life is another important factor which is adversely affected by SHO.

To minimise resource consumption, some SHO schemes stop packet duplication as soon as the first data packet arrives on the new radio interface. Park et al. [151] utilise the SIP-based SHO mobility solution which disconnects the old interface when the new interface receives data packets. Another example is the Cellular IP scheme proposed by Campbell et al. in [152]. The handover execution mechanism which is called the *semi-soft handover* involves packet bicasting in the network so that when the MS joins the new RAN, it can immediately receive buffered data.
packets. These approaches are effective in reducing the handover execution delay. However, the disadvantage is that the ping-pong effect may still cause high PER. If the target radio interface is unstable and the MS is forced to return to the old RAN, the MS needs to exchange signaling messages to perform a new handover process or to resume packet duplication.

Ciubotaru and Monteau in [105] propose an innovative handover scheme for multimedia streaming which can be categorised as a semi-soft scheme. Their scheme, called SASHA, was proposed as part of their Multimedia Mobility Management System (M3S) framework which uses multiple radio connection to exploit all available radio resources. In SASHA, the MS maintains radio connection with several PoAs, each sending some of the data packets depending on the Quality of Multimedia Streaming (QMS) factor. The QMS is the weighted sum of radio interface parameters such as QoS, power consumption, and cost. As the MS moves deeper into the coverage area of a PoA, the QMS of the associated radio interface improves which enables the PoA to deliver more data packets. Eventually, all media packets are received from the new PoA and the old connection becomes obsolete. The advantage of SASHA is that the multimedia session is smoothly transferred and the overall PER is reduced. Since only one interface is used for the transmission of each packet, the PER of the SASHA mechanism is higher than SHO schemes such as mSIP [89]. However, SASHA creates less traffic load and consumes less power.

However, the SASHA scheme has several shortcomings. In SASHA the handover is performed by end-to-end communication between the MS and the multimedia server. This approach suffers from high handover delay which is intrinsic in end-to-end mobility solutions. Moreover, the QMS is calculated at the server-side and based on feedback information received from the MS. As a result, this scheme is too slow to react to fast changing radio channel conditions due to shadow fading effects.

Another semi-soft handover is Fast Base Station Switching (FBSS) which is used in the WiMAX technology based on the IEEE 802.16e standard [133]. In FBSS, the MS is able to send and receive data frames from multiple base stations within a diversity set. However, at each given time interval, the MS only communicates with one base station, referred to as the anchor BS. The MS can switch to any BS within the diversity set with little delay. However, adding a new base station to the diversity set incurs considerable delay for target base station synchronisation, registration, and security association [153]. Moreover, FBSS requires strict cooperation between the base stations of the diversity set. In particular, the base stations should be synchronised, share all network entry information, and share connection information such as connection ID and service flow ID [6].

In summary, we note that the SHO approach is able to significantly reduce service interruptions of vertical handovers. This approach has been extensively proposed.
and studied in literature. Despite its advantages, SHO schemes make heavy use of battery and radio resources. Since these resources are scarce in wireless networks, the SHO approach may have limited application in wireless networks, particularly at the edges of radio cells where inter-technology handovers are more common. Semi-soft handover schemes have been proposed: these are either too slow or require tight integration of base stations. In our work, we develop a semi-soft handover execution mechanism which addresses these shortcomings. Our solution, which is called the Conservative Soft Handover (cSHO), is used in our UPTIME mobility management framework to reduce PER during handover period. We describe the UPTIME framework and its cSHO mechanism in the next two chapters.
Chapter 4

UPTIME mobility framework

4.1 Introduction

In this chapter, we propose a novel mobility management framework specifically designed for Next Generation Mobile Networks (NGMN). The aim is to achieve seamless mobility during inter-technology handovers in which noticeable service interruptions are avoided. To achieve this goal, handover delay and packet loss should be reduced to less than the QoS requirements of multimedia applications. Other requirements include strong security and efficient radio and battery resource management.

In the previous chapter, we discussed the shortcomings of seamless mobility frameworks which currently can be found in the literature. The shortcomings include the lack of support for NGMN architectural principles (for example, the separation of signaling and media), utilising non-standard signaling protocols and framework, the lack of integration with centralised Policy and Charging (PCC) platforms to ensure Quality of Service, incorporating sub-optimum handover delay minimisation mechanisms, and imposing significant changes in wireless networks.

In this chapter, we propose the Uninterrupted Proactive connection Transfer for IMS Mobility Enhancement (UPTIME) frameworks which addresses some of the above shortcomings. The UPTIME framework addresses weaknesses of previous proposals in the following ways:

- The UPTIME framework separates signaling from media functions and utilises open standards such as the Session Initiation Protocol (SIP) and Media Independent Handover (MIH) protocol.

- The proposed solution is IP Multimedia Subsystem (IMS) compliant in that the standard IMS infrastructure and mechanisms are reused for efficient mobility management. As a result, modifications to the IMS platform are minimised.
The UPTIME framework includes interaction with PCC frameworks to ensure consistent QoS across radio networks.

Compared with previously adopted schemes, the UPTIME framework further reduces handover preparation delay and packet loss by incorporating novel mechanisms.

In the UPTIME framework we combine proactive handover preparation approach with a soft handover execution technique. Our handover preparation scheme, called the Pre-Registration for IMS Mobility Enhancement (PRIME), demonstrates how proactive cross-layer handover preparation can be integrated into the modern mobile networks. Our modified soft handover mechanism for packet loss reduction is called the Conservative Soft Handover (cSHO) scheme. By incorporating the PRIME and cSHO mechanisms into the UPTIME framework, we present a mobility management framework which is capable of performing seamless inter-technology handovers.

The rest of this chapter is organised as follows. In Section 4.2 we describe our major design considerations and rationale behind the UPTIME architecture and operation. In Sections 4.3 and 4.4, we described the UPTIME system architecture and overview mobility management procedures. In Sections 4.5 we explain how the PRIME mechanism is used for fast handover preparation. Finally, Section 4.6 presents two signaling flow examples to demonstrate the operation of the UPTIME framework and the PRIME mechanisms. In these examples handovers to the LTE and WiMAX technologies are considered. The cSHO mechanism for handover execution will be described in the next chapter.

4.2 Design Considerations

The UPTIME framework is specifically designed for real-time multimedia applications with strict QoS requirements. In NGMN, these applications are provided through the IMS infrastructure and use the Session Initiation Protocol (SIP). Considering only real-time multimedia applications enables us to optimise the SIP and IMS signaling to achieve a seamless handover. More specifically, we design the UPTIME framework with the following requirements in mind:

Optimisation of support for multimedia applications. The UPTIME framework is designed specifically to address the challenge of a seamless handover for real-time multimedia applications. Focusing on these applications has significant implications when selecting the mobility management protocol. In particular, it imposes strict handover delay and packet loss limits. The UPTIME framework utilises the SIP protocol for handover execution. SIP exhibits some benefits compared with
other protocols such as Mobile IP (MIP), especially for multimedia applications [83]. However, SIP mobility management can be applied only on multimedia applications which use this protocol for session management. Support for non-SIP applications is not considered in this research. We note that these non-SIP applications generally do not have strict delay and loss requirements and can be supported using other mobility management protocols. Our approach fits well with the mobility toolbox scenario, proposed by Perera et al. in [154], in which different mobility solutions are used for different applications. Note that the proposed handover preparation scheme (PRIME) can be used in combination of different handover execution methods including the MIP. However, the signaling details would be slightly modified.

NGMN compatibility. The proposed UPTIME framework follows the NGMN requirements for mobility management solutions which were identified by ITU-T in [33]. ITU-T recommends using one of the existing standard IP-based mobility solutions which is independent of wireless technologies and is able to perform fast handovers. ITU-T guidelines also dictate the separation of the transport stratum from the service stratum. As such, in UPTIME we separate media handling nodes (packet duplication, buffering, and forwarding) from handover control functions. This means the UPTIME server which controls the handover procedure is located in the service stratum of the NGMN and interacts with packet processing nodes using standard and open signaling interfaces.

IMS compliance. IMS is emerging as the standard solution for multimedia applications over heterogeneous networks [155]. As such, the UPTIME framework is designed to be IMS-compliant. We reuse IMS signaling for mobility management and avoid significant changes to the IMS infrastructure. For this purpose, the mobility management feature of UPTIME is deployed as an advanced IMS service. This means the server which controls the handover procedure is deployed as an IMS application server and provides mobility service using the standard IMS interfaces and procedures.

Utilisation of open standards. In line with ITU-T guidelines [33], open standards are used in UPTIME for the purpose of exchanging handover messages between different entities. Open standards are those which are available to the general public and the ratification process is consensus-driven. We consider protocols that were proposed and accepted by standardisation organisations such as IEEE, IETF, and 3GPP to be open standards. Using these protocols ensures the interoperability of the devised solution with the rest of the NGMN infrastructure and makes it more viable for implementation. As such, we use the SIP and the Media Independent
Handover (MIH) [106] protocols. We take advantage of the fact that both protocols are extensible meaning new features can be added as required. For example, 3GPP extended SIP protocol by proposing some headers in RFC 3455 [156].

**Heterogeneity in access network.** NGMN comprises different RANs which utilise different radio technologies. To integrate a variety of radio technologies into the access network, we design the UPTIME framework to be independent of the underlying radio transport technologies. This means the use of UPTIME mechanisms is not limited to any particular wireless technology. Throughout this research we use the LTE and WiMAX technologies to illustrate the operation of UPTIME. At the time of writing, these are the most common and suitable technologies to support real-time multimedia applications in high mobility scenarios. However, the same mobility management concepts we describe can be applied to other wireless technologies such as IEEE 802.11.

We should also mention that the performance of UPTIME mechanisms may be affected by the choice of the wireless technology and its ability to transfer handover messages with reasonably low transmission delay. The UPTIME performance may significantly suffer when used over technologies that do not support priority transmissions. However, these technologies are not suitable for real-time multimedia applications and so are unlikely to be used in such an environment.

**Cross layer approach towards handover optimisation.** No single-layer mobility solution appears to be completely adequate to meet all requirements of seamless mobility [15]. In the UPTIME, we take a cross-layer approach to mobility management. Link layer and higher layer information is collected using the MIH protocol, handover control is performed using the SIP protocol, while IP-layer packet duplication and buffering is used for packet delivery.

**Support of multi-interface MS.** The current demand for pervasive wireless access is driving the mobile terminal market towards a multi-radio platform [92]. In particular, the market has seen a proliferation of multi-modal mobile devices [18]. In UPTIME framework we take advantage of this trend without completely excluding single-interface terminals. The MS can benefit from a *make-before-break* strategy and SHO scheme if it has several radio interfaces. Otherwise, though our cSHO method will not be available, the PRIME mechanism can still be used for delay reduction.

**Administration independence.** We consider the general scenario where RANs are under the administration of different operators. This scenario has additional requirement on the mobility management solution compared with the simpler case
in which one operator controls all RANs. We also consider the IMS network as being under the administration of its own operator. As a result, in optimising UPTIME procedures, we do not assume that the IMS and RANs necessarily share one subscription database.

**MS-controlled network-assisted handover.** UPTIME is an MS-controlled and network-assisted handover mechanism: the MS initiates and controls the handover procedure while the network facilitates the handover preparation and execution processes. Although network-controlled approaches provide operators with a tighter grasp on network operation and better load balancing capability, the MS-controlled schemes are more suitable for NGMN in which the MS has the freedom of connecting to different RANs controlled by different operators [157]. In the UPTIME framework, we enable operators to apply their policies at the IMS level or through the Policy and Charging Control (PCC) framework.

**Support for various radio coverage scenarios.** The UPTIME framework can be implemented to support inter-technology handovers in various radio coverage scenarios including complementary coverage, hotspot scenario, and overlay wireless networks. UPTIME is used in complementary coverage scenario to improve radio connection reliability. Meanwhile, it can be implemented in hotspot and overlay scenarios to assist them MS in connecting to the optimum network more quickly.

**Backward compatibility.** It is highly desirable that mobility management protocols for NGMN should be backward compatible with the existing networks [33]. The UPTIME framework is designed for backward compatibility with the service stratum (the IMS), the transport stratum (RANs and CNs), and mobile terminals. This means:

- An UPTIME-enabled IMS network can connect to both supporting and non-supporting RANs.
- Both UPTIME-enabled MSs and regular MSs can be simultaneously served by the network. PRIME and cSHO mechanisms are only available to supporting MSs.
- If the network does not support UPTIME, the MS requests are simply ignored without causing interoperability issues.
- The UPTIME-enabled MS can communicate with peers that do not necessarily support soft handovers.
In the next section, we describe the UPTIME system architecture which is designed based on these requirements.

4.3 UPTIME System Architecture

Before presenting technical details, we overview the operation of the UPTIME framework in a typical scenario. A user initiates a video call while traveling. Using utilities provided by the UPTIME frameworks, the user configures their handset to select the RAN with the highest QoS capabilities regardless of associated service costs. Based on these user preferences, the MS initially selects an LTE network and establishes a radio connection to carry the multimedia session data packets. However, the LTE coverage along the travel route is intermittent. Fortunately, where LTE coverage is unsatisfactory, there is complementary WiMAX coverage.

As the user roams within the LTE network the received radio signal quality varies. The UPTIME-enabled MS analyses the user’s movement trajectory and historic location data to determine locations in which service interruption is likely. The MS queries the UPTIME mobility management server, located in the IMS infrastructure, to identify potential candidate RANs which provide radio coverage in those locations. The MS selects the WiMAX network and initiates the PRIME mechanism to register with the WiMAX Core Network (CN). When the MS approaches the target WiMAX Base Station (BS), it completes the preparation of the WiMAX interface.

When the MS’s WiMAX interface is ready for data transmission, the cSHO mechanism is used to transfer the multimedia session smoothly to the WiMAX network. For this purpose, the MS requests packet duplication in the network. Both WiMAX and LTE networks receive duplicated data packets and buffer them at the PoAs. At each given period of time, the MS receives data packets over the interface with the best radio condition. The MS can easily change its active interface by sending one control message to associated PoAs. Finally, when the WiMAX interface is stable, the packet duplication is stopped and UPTIME procedure is completed.

To realise the above procedure, we develop the UPTIME system architecture containing new entities in the MS and wireless networks. The aim is to follow NGMN architectural requirements, use open standards, ensure minimum handover delay and packet loss, and avoid unnecessary modifications to the network infrastructure. Figure 4.1 depicts the overall UPTIME architecture which is based on the NGMN architectural paradigm. System architecture is divided into service control and transport strata. The components in the service stratum are in charge of controlling users’ mobility while the transport stratum entities perform packet duplication, buffering, and transmission.
In the service stratum, we introduce an IMS Application Server (AS), called the Handover Server (HOS), which provides the UPTIME mobility management as an advanced IMS service. Since the HOS utilises the standard IMS interfaces and mechanisms, it imposes minimum changes to the IMS [58]. This server may even be integrated in an existing AS such as a multimedia call manager.

In the MS protocol stack, we develop the Handover Adaptation Layer (HAL) which initiates and controls the handover process through cooperation with the HOS. This MS-controlled network-assisted approach used in UPTIME is better suited to multi-operator heterogeneous network environments [157]. The HAL also manages radio interfaces to make mobility transparent to multimedia applications running on top of this adaptation layer.

In the transport stratum, the two new components are the Media Duplication Function (MDF) and the Point of Attachment Handover Agent (PHA). The MDF is deployed in the CN or the RAN and is responsible for packet duplication. The MDF is a purely media-handling function operating under the control of the HOS. This approach is in line with the NGMN requirement of signaling/media separation and is more scalable compared with approaches which collocate signaling and media in one node, for example the solution proposed by Salsano et al. in [83].

The PHA is implemented in UPTIME-enabled PoAs and is in charge of data packet buffering and transmission. The PHA packet buffering functionality is initialised by the HOS acting on behalf of the MS. The PHA packet transmission op-
eration is controlled directly by the MS. The UPTIME approach in packet buffering and transmission at the edge of the transport network (that is, in PoAs) significantly reduces handover signaling delay and enables the MS to react to sudden changes in radio channels.

In line with NGMN requirements, the communication links between UPTIME components are based on open standards, that is, the SIP and the MIH protocols. In particular, the SIP protocol is used between the MS and the HOS. For other links the MIH protocol is utilised. To realise various UPTIME procedures, both SIP and MIH protocols should be modified with new features and functionalities. Fortunately, both standards are extensible and new extensions can be added to core specifications. We continue this section by describing the UPTIME components more in detail.

4.3.1 Handover Server

In the UPTIME framework, we propose the HOS which is implemented as a SIP Application Server (SIP AS). This server contains and executes the UPTIME mobility service. The rationale for the HOS implementation is to include a network node which communicates with the MS securely to provide mobility information and coordinate handover execution. To minimise modifications to underlying RANs, the HOS is implemented within the IMS platform which is accessible from all RANs. The HOS connects to the IMS service control platform using a standard interface often used for service creation. Utilising standard interfaces is important because it removes necessary modifications to the IMS infrastructure [58].

The HOS also plays the following important roles:

- The HOS coordinates handover preparation and execution mechanisms of the UPTIME framework.

- The HOS supports the MIH protocol to query candidate RANs and collect mobility-related information. This protocol is also required for coordinating packet duplication in the transport network and packet buffering in PoAs.

- The HOS functions as an information server to provide the MS with mobility-related information. The HOS may use various techniques to accumulate this information and the exact mechanism is not specified in the MIH protocol [106].

- To facilitate proactive handover preparation, the HOS has signaling interfaces with authentication servers in RANs and relays control messages between the MS and these servers.
• The HOS may function as a proxy DHCP server to enable proactive IP configuration.

• Functioning as a standard IMS application server, in line with 3GPP and WiMAX Forum specifications for the Policy and Charging Control (PCC) [57] and [51], the HOS communicates with the PCC frameworks to create necessary network bearers and ensure consistent QoS over different RANs.

The technical description of the HOS interaction with the rest of the IMS platform is as follows. The HOS connects to the IMS nodes, and more specifically to the Serving Call Session Control Function (S-CSCF), using the standard IMS Service Control (ISC) interface [10]. This standard interface allows the HOS to send and receive SIP messages from IMS clients (such as the MS) and communicate with other session management nodes. Whenever the S-CSCF receives a SIP message from the MS, it uses the Service Identifiers combined with initial Filter Criteria to determine the relevant application server, in this case the HOS [10]. The S-CSCF then routes the messages to the HOS. This results in a secured SIP-based connection between the MS and the HOS. In 3GPP terminology used in [84] this connection is called the access leg. Similarly a SIP-based connection is created between the HOS and the CH. This connection is called the remote leg.

When the MS decides to perform a handover, it performs our PRIME mechanism to proactively create a new access leg over the target RAN. The MS then signals the HOS to transfer the multimedia session to the new access leg. To coordinate the handover procedure, the HOS always remains in the signaling path between the MS and the CH. In other words, the HOS anchors signaling between the MS and the HOS and functions as a SIP Back to Back User Agent (B2BUA) [61] between the MS and the CH. Staying in the signaling path also enables the HOS to have accurate knowledge of the multimedia session’s QoS parameters. As we describe later in this chapter, this knowledge is important in preparing the new RAN to satisfy the QoS requirements of the ongoing multimedia session.

As the final remark on the HOS, we note that the signaling delay between the MS and the HOS can be minimised by having a direct interface between the HOS and the P-CSCF. In this way, SIP messages do not travel through the S-CSCF and possibly the I-CSCF. The UPTIME framework can be deployed in both the standard and optimised IMS architecture.

4.3.2 Media Duplication Function

The Soft Handover (SHO) handover execution of the UPTIME framework is based on packet duplication in the network rather than end-to-end packet replication as
proposed by others in [81, 89, 146]. The advantage is reduced resource consumption in the Corresponding Host (CH) side. We propose the MDF which is implemented in each UPTIME-capable wireless network and is used for network-based packet duplication. The MDF can be implemented as a separate node or be integrated in network nodes which are in the media path towards the MS.

In NGMN architecture there are several entities, both in the CN and RAN sections, which can contain the MDF functionality. In the IMS architecture, either of the Media Resource Function Processor (MRFP) and the Transition Gateway (TrGW) can be used for packet duplication. These nodes are often used for media transcoding and therefore are able to handle simple packet duplication. In the transport network, RAN gateways (also called Access Gateways) or network routers can be enhanced to support packet duplication functionality. For example, in the LTE technology the MDF can be integrated in the S-GW which functions as the access gateway. In any case, in the UPTIME framework we assume the MDF is in the media path and can be invoked for packet duplication.

As a media handling entity in the transport stratum of NGMN, the MDF is controlled by the service control stratum, and more specifically by the HOS. This approach is in line with the architectural guidelines of NGMN for efficient mobility management which requires the separation of control from transport functions to improve architectural scalability and flexibility [33]. To this end, we avoid solutions, such as the one proposed by Salsano et al. in [83], in which one node (i.e. the MDF) anchors both the signaling and media flows.

The ITU guidelines for NGMN mobility management frameworks [33] also require utilising a standard interface between the control function (the HOS) and the transport function (the MDF). Both the SIP and MIH protocols can be used for this purpose. The standard SIP signaling, as described by the Camarillo et al. in RFC 4117 [158], allows the replication of a media stream using the transcoding service. However, this solution causes signaling overhead and increases handover delay.

In the UPTIME framework we avoid extended signaling delay of SIP-based solutions and instead we use the MIH protocol to request packet duplication at the MDF. Since the current version of the standard does not support packet duplication, we propose a new extension which we explain later in this chapter.

The internal architecture of the MDF is depicted in Figure 4.2. The MDF contains an MIH Function (MIHF) which is used to receive duplication requests from the HOS. The MDF contains a Duplication Table which contains information on mobile terminals which require packet duplication. The list is populated with a new entry every time the HOS requests packet duplication for a particular station. Each table entry includes the current and target IP addresses of an UPTIME MS.
The Duplication Agent replicates all IP packets with a destination address which matches one of the entries in the duplication table.

We also note that in the communication link between the MDF and the HOS, the MIH messages are carried over higher-layer protocols such as the TCP and UDP. This is because the HOS and the MDF are not necessarily located within the same subnet. For the purpose of this communication, the IETF MIPSHOP work group published the RFC 5677 [159] which recommends the transport of MIH messages over the TCP and UDP protocols.

### 4.3.3 PoA Handover Agent

In the UPTIME framework, we use a modified SHO mechanism which overcomes the problem of excessive resource consumption of traditional SHO schemes. The solution which we call the Conservative Soft Handover (cSHO) is based on packet buffering in PoAs which serve the MS. The proposed PHA module is implemented in cSHO-enabled PoAs to enable them to buffer and transmit multimedia packets in a controlled fashion. To support backward compatibility, the UPTIME framework can also include PoAs which do not contain a PHA module. However, these PoAs can only be used in traditional SHO or Hard Handover (HHO) schemes and not the cSHO mechanism.

The rationale behind including the PHA in the UPTIME framework is to enable the MS to react to sudden signal fluctuations due to shadow fading, and activate or deactivate its radio interfaces accordingly. The UPTIME framework allows packet buffering at the edge of the transport stratum of NGMN, that is, in PoAs. To activate an already prepared radio interface and receive multimedia packets, the MS only needs to send one control message to associated PoAs which can be accomplished in few milliseconds. This approach remedies the excessive signaling delay of end-to-end mobility management methods such as SIP and enables the MS to react
quickly to changing radio conditions, mostly caused by shadow fading.

The operation of the PHA is initialised by the HOS. The HOS informs the PHAs of the current and target PoAs that a particular mobile terminal is using the cSHO mechanisms. Subsequently, PHA modules buffer the MS’s multimedia packets if the radio link between the MS and the PoA is deactivated. The PHA forwards data packets if the MS activates the radio interface.

The technical description of the PHA module includes its protocols used for communication UPTIME components, buffer sizes and dequeuing method, and PHA interaction with MAC scheduler to improve packet delivery performance. The PHA communicates only with the HOS and the MS. The protocol used for this communication should have the following properties:

- Following ITU-T guidelines for NGMN mobility management frameworks [33], the utilised protocol should be based on open standards.
- ITU-T requires the protocol to be independent of the underlying radio technologies. As a result, the mobility management protocol can be used in heterogeneous network.

The MIH protocol, also called the IEEE 802.21 standard [106], satisfies the above requirements. However, the current version of the MIH protocol does not include packet buffering. Fortunately the standard is extensible meaning new extensions can be added to the core specification of the protocol. Later in this chapter we discuss our proposed extensions to the standard.

To support the MIH protocol, the PHA may contain its own MIH Function (MIHF). Alternatively, the PHA can utilise the MIH services provided by the MIHF which is implemented within the PoA or an MIH Point of Service (PoS) which is located deeper into the RAN. Since communication with an external PoS increases signaling delay, implementing the MIHF in the PHA (or in the PoA) is more desirable.

Regarding packet buffering, as depicted in Figure 4.3, the PHA uses separate buffers to store multimedia packets destined for different mobile terminals. The PHA dequeues these buffers when associated radio interfaces become activated. The buffer size should be large enough to accommodate recent data packets required by multimedia applications. The number of required packets depends on the packet latency requirements and frame rates of multimedia applications.

According to ITU-T recommendation G.114 [19], high quality real-time voice applications require one way packet transmission delays of less than 150 milliseconds. For network planning however, the upper limit of 400 milliseconds should be considered. As such, the size of the PHA buffer for each MS should be large enough to contain data packets equivalent to 400 milliseconds of the multimedia session.
For example, in the case of the G.723.1 codec [160] with a rate of 33.3 frame/sec, the buffer capacity should be bigger than 14 packets.

When the radio interface of the MS becomes activated, the PHA dequeues data packets using a simple First In First Out (FIFO) method and delivers them to the MAC module of the PoA. The MAC module then forwards the packet to the MS. For the maximum performance of the UPTIME handover execution, the PHA may interact with MAC scheduler to indicate the availability of the MS. The MAC scheduler can then avoid scheduling packet transmission for a mobile terminal which has deactivated its radio connection by communicating with the PHA.

4.3.4 Handover Adaptation Layer and MS protocol stack

In the UPTIME framework, we propose the Handover Adaptation Layer (HAL) which is a mobility management agent inside the MS. The HAL has the following functionalities:

- Controlling handover process by communicating with the HOS and PHA
- Collecting mobility information and making handover decisions (network discovery)
- Making handover transparent to multimedia applications by managing network connections and activating or deactivating radio interfaces

To perform the above functionalities, the HAL is implemented as a virtual layer (also called middleware) in the MS protocol stack. A similar approach has been used in [105, 121, 161, 162, 123, 163]. However, in contrast to some solutions such as [163, 161, 123], the HAL is implemented on top of the transport layer and in
the user level. As a result, modifications to the kernel of the operating system are avoided.

Another advantage of the HAL is its ability to hide mobility from multimedia applications. In traditional SIP-based mobility management solutions, the multimedia application monitors radio interfaces and utilises SIP signaling to perform mid-session handovers. In the UPTIME framework, however, the HAL acts on behalf of the applications and adds or deletes radio links based on RANs’ availabilities.

The HAL and its components (blocks) are shown in Figure 4.4. The depicted blocks only demonstrate the logical architecture of the HAL and the categorisation of handover tasks. Although we use same architecture in developing the HAL simulation model, the real life implementation may differ. The HAL operation can be summarised as follows: To initiate the UPTIME procedure, the HAL collects mobility information and stores it in the Information Manager. The mobility information is obtained through the MIH Interface or from peripheral modules such as the GPS module. The processed information is fed to the Decision Maker which selects the best target networks. Subsequently, the Mobility Manager initiates handover preparation by communicating with the HOS through SIP messages which are constructed using the IMS Stack. Finally, to perform the handover execution, the Link Manager switches between radio interfaces while hiding it from multimedia applications. The technical description of each block is provided in the following.

**Information Manager**  The Information Manager collects and stores information on important handover parameters, such as those described in Section 2.5.3. This
block collects mobility information by communicating with radio interfaces through the MIH protocol. This information includes physical and link layer parameters such as SNR, Bit Error Rate (BER), and Packet Error Rate (PER). Higher level information such as network load, RAN service capabilities, and associated costs are obtained from MIH information servers and other network nodes. Additional information can be collected by querying the peripheral components in the MS. For instance, the location information and direction of movement can be obtained from the GPS module.

The user preferences may be obtained directly from the user, for example by using a Graphical User Interface (GUI). The operator policy information, as suggested by the 3GPP in [3GPP_IMS_stage2_R8], can be collected by utilising network management protocols like the Device Management protocol [164]. The accumulated information is processed and passed to the Decision Maker.

**Decision Maker**  The Decision Maker runs the handover decision algorithm. In the handover initiation stage it selects the optimum target networks. It then determines when packet duplication should be initiated and terminated, and which radio interface should be used to achieve the minimum PER. The outcome of each decision is sent to the Mobility Manager which takes an action.

**Mobility Manager**  The Mobility Manager is the central control unit of the HAL. This block handles different handover signaling and performs the actions determined by the Decision Maker block. When the Decision Maker requests an action, the mobility manager executes all necessary signaling for the lowest delay. For example, during the handover stage when packet duplication is needed, the Mobility Manager interacts with the MS Stack to generate a suitable SIP INVITE message. The Mobility Manager is connected to the MIH Interface and the IMS Stack blocks to create the required MIH and SIP control messages.

**Link Manager**  The aim of this block is to make mobility transparent to multimedia applications running on top of the HAL. These applications continue receiving data packets from the HAL as the Link Manager switches between radio interfaces and handles the change of IP addresses.

One of the drawbacks of SIP-based mobility management is that multimedia applications must be “mobility-aware” whereas in solutions such as the Mobile IP they can remain “mobility-unaware” [165]. In SIP protocol, when the IP address changes, the application interacts with the operating system to open new sockets to obtain and deliver multimedia packets. The procedure is as follows. To initiate a multimedia session, the application generates a SIP INVITE message and sends
it to the Corresponding Host (CH). This message contains the Session Description Protocol (SDP) [87] parameters which describe session parameters such as utilised codecs, IP addresses, and UDP port numbers. The application interacts with the operating system to open the TCP and UDP sockets specified in the SIP INVITE message. These sockets are bound to the MS’s local IP address and port number. The operating system is then able to deliver data packets to the application as they arrive.

When the MS’s local IP address changes, the multimedia session is interrupted. The multimedia application must be aware of terminal mobility. When the MS acquires a new IP address, the SIP application opens a new UDP socket, bound to the new IP address, and sends a new SIP INVITE message, called the re-INVITE message, to the CH.

The more desirable solution is to make mobility transparent to multimedia applications [14]. The Link Manager block provide mobility transparency by providing a virtual socket feature. When the multimedia session is initiated, the Link Manager opens a UDP socket on behalf of the application. Every time a new radio link is ready, the link manager opens a new socket which is bound to the newly obtained IP address. The application, however, is not informed about this change. When a change of active interface is required, the link manager maps the initial UDP socket to the new socket and redirects packets to the new interface. The Policy routing feature which exists in most operating systems can be utilised for the purpose of redirecting packet to a specific interface [123].

We note that mobility should not be completely hidden from multimedia applications. It is common to have different different QoS parameters over different radio interfaces. As such, for better service experience, multimedia applications should be informed to adjust their operation (for example, to adjust the multimedia coding rate) based on the QoS capabilities of the new radio access network. The applications which are capable of adapting to the conditions of radio interfaces are referred to as Access QoS Aware applications [56]. As such, we envisage the inclusion of an Application Programming Interface (API) which allows the HAL to communicate with multimedia applications and inform them of new link’s QoS parameters.

The Link Manager also deals with removing duplicated packets when they are occasionally received. The UPTIME framework uses a modified SHO scheme based on packet duplication and buffering. Receiving duplicated packets may occur as out-of-order data packets arrive at the current and target PoAs. To eliminate packet duplication, the Link Manager records the sequence numbers and drops packets already received. To avoid additional delays, the link manager does not contain any buffer for reordering packets. This task is performed at the application level.
**IMS Stack**  An IMS Stack block is an implementation of the standard SIP protocol, 3GPP-specific SIP extensions [156], and IMS procedures [10]. The IMS stack typically provides IMS/SIP functionalities such as IMS registration, construction of SIP messages, tracking SIP dialogs, and receiving and parsing SIP messages. Similar to SIP stacks, an IMS stack facilitates the development of IMS applications by providing services which can be used by other applications. Applications therefore are developed without the full implementation of the SIP protocol and IMS specifications. An example of an IMS stack is the Java Specification Request 281, called the *IMS Services API*, which provides a high-level API for application developers [59].

In the HAL architecture, the mobility manager utilises the services of the IMS Stack to generate SIP messages and communicate with the HOS. Since we extend the SIP protocol with a new handover header, the IMS stack should also be modified accordingly.

**MIH Interface**  Lack of cross-layer awareness and information exchange is one of the main limitations of the traditional Internet protocol suite [14]. The MIH protocol can be used to overcome this limitation and allow higher layers to collect information from radio interfaces and control interface operations. In the HAL, we use services provided by the MIH protocol for network discovery and handover execution. The MIH Interface block is responsible for utilising MIH services by communicating with the MIH Function (MIHF) within the MS protocol stack.

As specified by IEEE 802.21 standard [106], the interaction between an MIH user (that is, the HAL) and the MIHF occurs through a set of service primitives called the MIH Service Access Point (*MIH_SAP*). The MIH Interface uses various primitives to interact with the MIHF to obtain mobility information or execute specific tasks.

In the following, we provide some examples and list important MIH primitives which are used by the MIH Interface. When the HAL is initialising, the MIH Interface uses the *MIH_Capability_Discover* primitive to discover the capabilities of the MIHF. This allows the MIH Interface to determine the supported MIH events, commands, and information services and avoids requesting commands which are not recognised by the MIHF or the network node providing MIH services (the PoS). Subsequently, the MIH interface subscribes to events which are important in the UPTIME framework by issuing an *MIH_Event_Subscribe.request* primitive and listing the required events. More information is obtained using the *MIH_Get Information.request* primitive. Table 4.1 lists other important MIH_SAP primitives which are used in different UPTIME procedures.

The MIH Interface block is also used by the HAL to construct MIH messages required for handover execution. For example, during handover execution, the HAL
Table 4.1: MIH_SAP primitives used in UPTIME

<table>
<thead>
<tr>
<th>Primitive</th>
<th>Used by</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIH_Capability_Discover</td>
<td>HAL</td>
<td>To discover MIH capabilities of MIHF or PoS</td>
</tr>
<tr>
<td>MIH_Register.request</td>
<td>HAL</td>
<td>To ask the MIHF to register with a PoS</td>
</tr>
<tr>
<td>MIH_Event_Subscribe.request</td>
<td>HAL</td>
<td>To ask the MIHF or PoS to be informed if an MIH event happens.</td>
</tr>
<tr>
<td>MIH_Get_Information.request</td>
<td>HAL</td>
<td>To ask for information from an MIHS server</td>
</tr>
<tr>
<td>MIH_Link_Detected.indication</td>
<td>MIHF</td>
<td>To indicate that the first PoA of a RAN has been detected.</td>
</tr>
<tr>
<td>MIH_Link_Up.indication</td>
<td>MIHF</td>
<td>To inform the HAL that an L2 link has been established.</td>
</tr>
<tr>
<td>MIH_Link_Down.indication</td>
<td>MIHF</td>
<td>To indicate an L2 connectivity was lost.</td>
</tr>
<tr>
<td>MIH_Link_Get_Parameters.request</td>
<td>HAL</td>
<td>To ask for the value of some link parameters such as SNR.</td>
</tr>
<tr>
<td>MIH_Link_Actions.request</td>
<td>HAL</td>
<td>To order the execution of a handover action such as packet forwarding.</td>
</tr>
<tr>
<td>MIH_MN_HO_Candidate_Query.request</td>
<td>HAL</td>
<td>To ask the serving PoS to query candidate networks for QoS resources and IP configuration.</td>
</tr>
</tbody>
</table>

communications with associated PoAs to order packet transmission and buffering. This communication is based on the MIH protocols and the MIH Interface block is responsible for generating required messages.

4.4 UPTIME Procedures Overview

As previously mentioned in Section 3.2, seamless handover has two important requirements. Firstly, the handover delay should be minimised. Secondly, packet loss should be avoided [85]. The above requirements should be satisfied while observing other issues such as security, QoS, and power management [97]. In this section we give an overview of UPTIME mechanisms which ensure the above requirements.

We follow ITU-T guidelines for NGMN mobility management, presented in [69], and categorise the UPTIME procedures into three stages: network discovery, handover preparation, and handover execution. We then propose mechanisms to optimise each of these stages.

To develop UPTIME procedures for seamless mobility we consider the following principles:

- The MS functions proactively and discovers potentials target RANs and prepares new radio links before a handover becomes necessary.

- To further reduce handover preparation delay, we subdivide network discovery and network registration mechanisms into radio-dependent and radio-independent tasks. The radio-independent tasks are performed in early stages.
of handover trigger detection.

- To avoid packet loss during handover execution stage while minimising energy consumption, a modified soft handover scheme is used.

Based on the above principles, the UPTIME mobility management procedure can be regarded as a combination of proactive cross-layer handover preparation schemes, similar to solutions presented in [21, 125, 24, 131, 23], and soft handover execution approaches similar to those described in [27, 88, 83]. However, we further improve the performance of previously adopted processes. In particular, we further reduce handover preparation delay by proposing the Pre-Registration for IMS Mobile Enhancement (PRIME) procedure. We also propose the Conservative Soft Handover (cSHO) to reduce energy consumption of traditional SHO methods.

As a more general mobility management solution, the UPTIME framework can contain both or either of the PRIME and cSHO procedures. For example, the PRIME procedure can be combined with a traditional SHO scheme for handover preparation and execution. This scenario may occur in cases where the target and current PoAs do not support the cSHO procedure.

Figure 4.5 depicts the overall UPTIME process for seamless mobility. The first two stages of the UPTIME process, the network discovery and handover preparation, are performed using the PRIME procedure. In traditional network discovery schemes, the MS detects a handover trigger (for example, degrading SNR) and discovers surrounding RANs. The goals of network discovery stage include early handover trigger detection and optimum network selection. For network selection, the MS determines the optimum RAN and the best PoA within this network.

However, in the PRIME mechanism, as described in the next section, we separate RAN detection from PoA discovery. This two-phase network discovery process results in reduced handover preparation delay. In the first phase, the MS speculatively selects one or more RANs. This phase is performed in the early stages of a potential handover and the MS is not required to be located in the coverage area of target RANs. Later on, when the MS approaches the vicinity of candidate PoAs, it performs the PoA discovery procedure. As we discuss later on, the MIH protocol provides necessary tools for both RAN and PoA discovery processes. The advantage of this approach is that some tasks of the handover preparation mechanism are performed well before the handover becomes imminent.

The handover preparation stage of the PRIME mechanism is also performed in two phases. The first phase includes non radio-dependent tasks such as the registration to the IMS and the Core Network (CN) of the target network. This phase is initiated after RAN discovery and includes authentication, IP configuration, and network bearer setup for QoS support. The second phase is radio-dependent
and includes frequency and time synchronisation, the establishment of the PoA security association, and radio-bearer setup. This phase can be started after the selection of the target PoA. The two-phase handover preparation approach results in a significant reduction in the handover preparation delay when compared with the conventional pre-registration method as presented in [23].

In the handover execution, the third stage of the UPTIME procedure, the goal is to transfer the connection from the old PoA to the new one smoothly and avoid packet loss as much as possible. As discussed in Section 3.5, this goal is best achieved by using a soft handover approach. In cSHO mechanism which we use in the UPTIME framework, as depicted in Figure 4.5, the MS initiates packet duplication in the network and buffering in the current and target PoAs. Then, based on the conditions of radio interfaces, the MS determines which interface has lower Packet Error Rate (PER) and should be activated. The MS alternatively switches between interfaces until the new radio connection becomes stable. The handover process is then completed. The MS also has the option of halting the handover process which means the resources in the inactive RAN remain reserved.

As can be observed in Figure 4.5, the UPTIME procedure has a four-stage decision making process. At stage 1, the MS selects one or more candidate RANs. At stage 2, it decides if packet duplication should be commenced on the target RAN.
At stage 3, the MS continuously selects the active interface. Finally, at stage 4, the MS determines if the handover process should be terminated. Compared with two-stage methods in which the MS only decides on target RAN and handover completion time, the UPTIME mechanism has two advantages. First, the MS is not required to wait to arrive in the coverage area of the target network. It can initiate the pre-registration process well in advance of the handover execution time. Second, the MS can utilise the target radio link as soon as it becomes occasionally available. The MS does not need to wait for a stable data connection over the target link.

In the next section, we describe the UPTIME procedure. We then provide two comprehensive signaling flow examples for LTE and WiMAX technologies. The cSHO mechanism is covered in Chapter 5.

### 4.5 Handover Preparation with PRIME Mechanism

As the name suggest, the Pre-Registration for IMS Mobility Enhancement (PRIME) mechanism is based on the pre-registration approach as used by Dutta et al. in [23] and Ito et al. in [24]. The novelty of the PRIME procedure is that it takes advantage of the NGMN architecture and decouples RAN selection and preparation from PoA preparation. The result is a significant reduction in handover delay.

As we described in Section 2.1, the NGMN transport stratum is divided into the Core Network (CN) and Radio Access Network (RAN) sections. The RAN section contains radio transmission functionalities while the CN provides IP connectivity, user authentication, and mobility and QoS support. In line with this architecture paradigm, in PRIME we divide registration tasks into the radio-independent and radio-dependent tasks. The first group of tasks can be performed without the involvement of the target PoA. These tasks include the AAA process, IP configuration, QoS and network bearer setup at the CN side. The registration to the IMS (through the target network) can also be performed after these tasks. The radio-dependent tasks are the ones that require communication with the target PoA. As such, these tasks can only be performed after the selection of the target PoA.

As the PRIME mechanism significantly reduces the handover delay, it improves optimum network connectivity in both complementary and overlay (or hotspot) coverage scenarios. In complementary coverage where the MS leaves the coverage area of one RAN, the MS is able to connect to the target PoA before losing its current radio link. In overlay scenario, the MS can benefit from faster or cheaper radio connectivity as soon as the target PoA becomes available.

We note that both the conventional pre-registration schemes and the PRIME
methods require slight modifications to both the new and old wireless network. The old network should be able to relay pre-registration messages to the new network while the new network should support proactive authentication and IP configuration. However, handover delay reduction is significant and pre-registration schemes are usually required for seamless handovers. If the old and new networks do not support the PRIME mechanism or other pre-registration methods, the MS can still use the standard registration procedure to prepare its radio interface.

The PRIME mechanism includes the following tasks:

- RAN discovery and selection. the MS discovers surrounding RANs and selects one or more of them as target networks.
- Core Network Pre-Registration (CN-PR). the MS registers with the CN of target networks and obtains a new IP address.
- IMS Pre-Registration (I-PR). the MS uses the new IP address to register with the IMS through a tunnel to the target network.
- PoA discovery and selection. the MS arrives to the vicinity of the candidate PoAs and selects one of them as the target PoA.
- PoA Pre-Registration (P-PR). the MS connects to the target PoA and completes the radio bearer creation process.

RAN and PoA discovery are performed using the standard MIH mechanisms. However, decoupling radio-dependent and radio-independent handover preparation tasks requires slight modifications to the network registration procedures. We continue this section by discussing each PRIME tasks and their technical details.

4.5.1 RAN Discovery Procedure

The target RAN discovery process is generally time-consuming. To achieve seamless mobility, this process should be performed proactively, that is, while the MS is still connected to the current RAN. Proactive and efficient RAN selection has two requirements. First, a long-term handover prediction is required. As discussed in Section 3.4, the long term handover prediction is commonly based on the estimation of the MS's future locations. Second, a protocol is required to enable the flow of mobility information across RANs. The IEEE 802.21 standard (the MIH protocol) provides a framework and necessary mechanisms which allow the MS to discover target RANs through the current RAN.

The long-term prediction of handovers, as discussed in the literature review, can be achieved using a geographical map and the user's direction of movement,
or by utilising statistical learning methods and estimating the user's movement trajectory. In cases where the exact location cannot be estimated, the mobility profile of the user can be used to determine probable zones that the MS will visit. Since a RAN with potentially many PoAs covers a large area, it is likely that the MS can use future location estimation to determine potential target RANs. However, the MS may not be able to pin-point specific target PoAs in the early stages of network discovery.

To determine candidate RANs, the MS uses the MIH protocol. In particular, the mobility information provided by the Media Independent Information Service (MIIS) is used for this purpose. As specified by IEEE 802.21 specifications, MIIS information is mostly static in nature and includes RAN parameters such as the radio technology used (for example, WiFi, WiMAX), roaming partners services and their costs, and QoS capabilities. This information is represented in Information Elements (IE). Table 4.2 lists some of the RAN-specific IEs that can be used in the PRIME procedure.

In UPTIME framework, the HOS functions as an MIH information server and provide the mobility information. In this framework, the IEs are carried in MIH messages and sent over the secure SIP-based connection which is already established between the MS and HOS. Encapsulating MIH messages in SIP messages has been previously proposed by Rodrigues in [111]. The advantage is reduced signaling delay to establish a secure connection with the information server. Although the encapsulation of MIH messages inside SIP messages is not a standard solution, its implementation only affects the HAL and the HOS and no modification to the IMS architecture is required.

Using received IEs, the HAL selects a suitable target RAN which has roaming agreements with the MS's home network, supports access to the IMS, and has reasonable costs. The MS also considers the QoS capabilities of the RAN such as supported Class of Service and minimum packet transmission delay. At this stage, however, the MS cannot determine if the QoS requirements of the multimedia application can be satisfied. The RAN's capability to accommodate the MS depends on the radio conditions and the network load at the target PoA which is not known at this stage.

During RAN discovery process, the MS can also obtain the IP address of the Proxy Call Session Control Function (P-CSCF) which provides access to the IMS. This information may affect the outcome of the RAN selection algorithm as RANs which are covered by the MS's current P-CSCF are more attractive. Another important selection parameter is the protocols supported by the RAN. In the current version of the MIH protocol, the support for the Mobile IPv4, Mobile IPv6, and Internet Key Exchange version 2 (IKEv2) can be indicated with some IE values.
Table 4.2: Examples of the MIH Information Elements used in UPTIME

<table>
<thead>
<tr>
<th>Name of IE</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IE_ROAMING_PARTNERS</td>
<td>Operators with roaming agreements with the target RAN</td>
</tr>
<tr>
<td>IE_NETWORK_TYPE</td>
<td>Type of the RAN including WiFi, WiMAX, LTE, etc.</td>
</tr>
<tr>
<td>IE_COST</td>
<td>Service cost</td>
</tr>
<tr>
<td>IE_NETWORK_QOS, IE_NETWORK_DATA_RATE, IE_NET_FREQUENCY_BANDS</td>
<td>QoS, data rate, and bandwidth of the target RAN</td>
</tr>
<tr>
<td>IE_NET_CAPABILITIES</td>
<td>capabilities such as security, and emergency calls, and MIH support</td>
</tr>
<tr>
<td>IE_NET_IMS_PROXY_CSCF</td>
<td>Address of the P-CSCF which provides IMS access</td>
</tr>
<tr>
<td>IE_NET_MOB_MGMT PROT</td>
<td>support for mobility management protocols such as Mobile IP, or UPTIME (extended MIH protocol)</td>
</tr>
</tbody>
</table>

reserved for future protocols such as our PRIME and cSHO mechanisms. The MS therefore can select a RAN which better supports seamless mobility.

4.5.2 CN Pre-Registration

The CN Pre-Registration (C-PR) is initiated after RAN selection. To perform the CN-PR process, the MS is not required to be within the radio coverage of the target RAN. In other words, the MS is able to perform CN-PR tasks which are non-PoA-dependent, without a direct radio link to the target RAN.

With the pre-registration to the target CN the MS achieves three goals. Firstly, the MS is authenticated in the target CN. Secondly, the MS receives the IP addressing configuration. Thirdly, network bearers are set up to support the required QoS parameters. The description of each task follows.

**Authentication** To perform proactive authentication, also called pre-authentication [127], the MS should be able to communicate with the Authentication Authorization and Accounting (AAA) server of the target network. A suitable protocol is needed for the exchange of authentication messages over IP networks. To be applicable for different radio technologies, this protocol should support different authentication methods. The standard Extensible Authentication Protocol (EAP) [47] provides these features.

The EAP protocol is not an authentication method; it is merely used to carry authentication messages between the MS and AAA servers. The protocol specifies using AAA proxy servers (located in visited networks) to relay messages between
the MS and AAA server in the home network [5]. The authentication signaling also passes through the Network Authenticator Servers (NAS) located in RANs which controls access to network resources.

The EAP protocol supports various authentication methods including Subscriber Identity Module (SIM) based schemes which are commonly used in mobile networks. Both the 3GPP and WiMAX Forum recommend using this protocol for user authentication in LTE [40] and WiMAX technologies [6]. For authentication method, the SIM-based LTE Authentication and Key Allocation (AKA) [55] scheme can also be used in the WiMAX technology. Authentication using the EAP-AKA scheme is an approach that we also consider within the PRIME mechanism. However, the PRIME solution is extensible to other authentication methods.

In the PRIME mechanism, the HOS functions as an AAA proxy and relays authentication messages to the AAA server located in the target network. To initiate the procedure, the MS sends its identity to the HOS using the EAP response/identity message. The identity is reconstructed using the International Mobile Subscriber Identity (IMSI) of the MS and commonly has a format of IMSI@MyOperator.com [132]. The HOS uses the MS’s identity to route the message to the AAA server located in the operator’s domain (for instance, MyOperator.com). Consequently, the AAA server and the MS run the AKA procedure to mutually authenticate each other. Upon successful authentication, both the MS and the server generate a master security key. The other security keys are derived from this master key. For more details, refer to Sections 2.2 and 2.3 for discussion of LTE and WiMAX security procedures.

It is important to note that at this point, some security keys cannot be derived because they rely on the identity of the target PoA. For example, in the LTE technology, the Access Security Management Entity (ASME) key (K_{ASME}) is constructed in the MS and the Mobility Management Entity (MME) in the CN. However, the eNB key (K_{eNB}) cannot be derived because it depends on the identity of the target eNB. As such, all security keys which are required for traffic encryption and are derived from the K_{eNB} cannot be constructed [53].

Similarly in the WiMAX technology, a successful EAP authentication generates a master session key (MSK) in the MS and the AAA server. However, the BS identity is needed to generate the Authorization Key (AK) which is transferred to the target BS [50].

In the PRIME mechanism, we propose a two-step authentication process. First, the MS is authenticated and the master key is derived. Second, when the MS seeks connection to the target PoA, security keys generated and delivered. This approach reduces authentication delay.

Since the MS is authenticated in the CN of the target network, a new IP address
can be assigned to the MS and other configurations can be delivered. We explain these processes in the following paragraphs.

**IP address and other configurations** After the authentication process, the MS can receive a new IP address from the target network. The exact signaling to receive the IP configuration depends on network mechanisms and operator policies. In the LTE technology, the IP address can be obtained as part of the default bearer setup (see Section 2.3). Alternatively, the standard DHCP mechanism may be used. In the WiMAX technology, the DHCP mechanism is commonly used for IP address assignment.

If DHCP is used, since the MS is not connected to the target network, the HOS functions on behalf of the user and receives the IP address. The address of the DHCP server in the target network can be acquired using the MIH protocol. The HOS then functions as a DHCP client and obtains a new IP address and forwards it to the MS. To reduce signaling delay, the HOS may include the newly obtained IP address in the acknowledgment message which is sent at the end of pre-registration scheme.

In the CN-PR process, the MS may also perform other tasks of initial network attachment procedures which do not involve the target PoA. For example, in the LTE technology, the *location update* procedure is performed to update the information stored in the Home Subscriber Station (HSS) database located in the LTE CN. A *default bearer* is also created for occasional data and signaling transmission.

**QoS setup** To ensure a consistent QoS in the target network, the CN-PR process also includes the creation of network bearers. By creating network bearers, the CN nodes become aware of QoS characteristics required by the multimedia application. In this process, the HOS negotiates the QoS parameters with the Policy and Charging Control (PCC) framework of the target network. As such, the operator is able to enforce its policies and modify the data session parameters.

It is important to note that at this stage the handover is “speculative” as the MS may not be in the coverage area of the target RAN and the target PoA is not selected. Therefore, network resources are not reserved at this stage. The creation of network bearers only indicate that the MS, based on its subscription information and network policies, is authorised to consume certain resources. However, the actual radio resources are allocated when the MS determines the target PoA. This approach is similar to the MS's *IDLE* mode in LTE technology in which the network bearers are preserved with their QoS parameters [53]. When the MS becomes *ACTIVE* again, there is no need to create network bearers.

In PRIME mechanism, after a successful authentication and receiving an *EAP-
Success message, the HOS initiates the process of creating network bearers. The process is performed by interacting with the PCC framework. In this procedure, the HOS acts as an Application Function (AF) [57] and requests the Policy and Charging Rules Function (PCRF) node to create one or more network bearers. This is in line with the IMS framework in which application servers (in this case the HOS) and the P-CSCF can act as the AF. Since the HOS anchors the SIP signaling between the MS and the Corresponding Host (CH), it is able to determine the QoS requirements of the multimedia application. In particular, the HOS extracts the QoS information from the Session Description Protocol (SDP) parameters included in the SIP INVITE message.

To authorise the creation of a service flow, the PCRF downloads the MS subscriber information from a network repository, for example, the Home Subscriber Server (HSS). The PCRF then makes a policy decision and either authorises or denies the establishment of the network bearer. If the bearer creation is authorised, the PCRF sends a request to the Policy and Charging Enforcement Function (PCEF). The PCEF is the network element which initiates control signaling to create the network bearers. In the LTE technology, the MME is the network node which functions as the PCEF. It receives the request from the PCRF and orders the Serving Gateway (S-GW) to create a dedicated bearer. In the WiMAX technology, the Access Service Network Gateway (ASN-GW), or more specifically the Service Flow Authorization (SFA) module inside the ASN-GW, is responsible for creating the service flow.

As mentioned before, at this stage of the PRIME mechanism, the target PoA has not yet been selected and therefore radio resources cannot be verified. As such, radio bearers cannot be established. The PCEF node therefore stores the information received from the PCRF and waits for further instructions. When the target PoA is discovered, the MS requests the HOS to contact the PCEF and complete the resource reservation process.

4.5.3 IMS Pre-Registration

Before transferring a multimedia session to the target RAN, the MS must be registered to the IMS through this RAN. This registration process creates a Security Association (SA) between the MS and the P-CSCF in the new RAN. The MS is then able to securely use IMS services. The IMS registration also allows the IMS operator to enforce its policies. For example, the MS can be barred from accessing certain radio technologies.

Since the IMS registration process is time consuming, the PRIME mechanism uses the IMS Pre-Registration (I-PR) scheme, as proposed by [24], for delay reduc-
tion (see Section 3.4 for more details). Since the IMS registration usually does not depend on the target PoA, the MS can perform this registration well before the detection of the target PoA.

In a typical scenario, the target RAN connects to the IMS through a P-CSCF. As mentioned previously in Section 4.5.1, the MS can use the MIH protocol to discover the IP address of the P-CSCF in the target RAN. When the MS performs obtained in new IP address, it is ready to perform IMS registration through the target P-CSCF.

The I-PR process is similar to the standard IMS registration procedure. The only difference is that the HOS functions as a proxy and relays signaling messages between the MS and the new P-CSCF. To start the I-PR procedure, the MS sends a SIP Re-Register message to the HOS. This message is forwarded to the new P-CSCF. Consequently a regular IMS registration procedure is performed.

4.5.4 PoA Discovery and Selection

The PRIME mechanism uses a two-phase network discovery mechanism. In the first phase, described in Section 4.5.1, the MS selects one or more target RANs. In the second phase, the target PoA is selected. This phase is usually executed when the MS approaches or arrives in the radio coverage area of the target PoA.

For PoA discovery, the MS uses different MIH services [106]:

- The MIH Information Service (MIIS) is used to obtain static information on candidate PoAs. This information is presented in Information Elements (IE) received from the HOS and includes parameters such as the IP and MAC address of the PoA, its physical location, and radio channel range.

- The Command Service (the MICS) is used to query the PoA and obtain more dynamic information. For this purpose, the local MIHF directly communicates with the MIH Point of Service (PoS) in the target PoA or RAN. For example, the MS may query available radio resources at the target PoA.

- The MIH Event Service (the MIES) is used to inform the HAL of specific events such as the detection of the first PoA of target RAN.

We now describe this in more detail. The MS can send an MIH message \textit{MIH\_MN\_HO\_Candidate\_Query request} to the target network, probably through the current RAN, to check the status of radio resources at the candidate PoAs. In this message, the MS indicates the preferred PoAs and the minimum required QoS resources. The MIH Point of Service (PoS) in the target network replies with a list of PoAs which satisfy the indicated QoS requirements and should be considered for the handover.
The MS selects one or more PoAs and waits for their availability. Through the use of an MIES called the *MIH_Link_Detected*, the MS detects the availability of the first PoA of the target RAN. Subsequently, our proposed HAL in the MS protocol stack requests the local MIH Function (MIHF) to report the link layer information of the detected radio channels. The MIHF uses the *MIH_Link_Parameters_Report* event service to inform the HAL when link layer parameters, such as the SNR or PER, surpass a predefined threshold. Alternatively, the HAL may also use the MIH command service, the *Link_Get_Parameters*, to actively monitor the signal strength of candidate PoAs.

Based on observations of the radio signal strength, the MS is able to select the target PoA. To improve the performance of the PoA selection process the MS can also apply a short-term prediction of the radio signal, as proposed by Chang and Chen in [21].

A typical handover decision algorithm combines different observed and predicted parameters to select the optimum PoA. A cost function which combines different handover metrics is vital for multi-technology and multi-operator environment [75]. The UPTIME framework and the PRIME mechanism can include different type of network selection algorithms. However determining the exact network decision algorithm is behind the scope of this research.

### 4.5.5 PoA Pre-Registration

The PoA Pre-Registration (P-PR) is the final stage of the PRIME handover preparation process. In the PRIME mechanism, the MS registers with the CN and the IMS before detecting the target PoA. When the target PoA becomes available, the MS can quickly create a new radio connection by performing the following tasks:

- The MS requests the delivery of security keys to the target PoA.
- The creation of radio and network bearers is initiated.
- The MS performs time and frequency synchronisation with the target PoA.
- The MS establishes a Security Association (SA) with the target PoA.
- Radio bearers are created.

Since most of the handover preparation tasks (for example, authentication, IP address assignment, IMS registration) have been already performed, radio link preparation is much faster than standard network entry processes described in Section 2.2 and 2.3.

The P-PR procedure is depicted in Figure 4.6. This procedure starts when the MS enters the coverage area of the target RAN and detects the target PoA.
To initiate the process, the MS sends a SIP UPDATE message to the HOS. This message contains information on the target PoS including the PoA Identifier (PoA ID) which is obtained in the PoA discovery phase using the MIH protocol. The PoA ID can be presented in the format of the LINK_ADDR data type of the MIH protocol, as specified in the IEEE 802.21 standard [106], and is a choice between the PoA’s MAC address and Cell ID.

Subsequently, the HOS requests the AAA server to distribute security keys to the target PoA. Since the MS is already authenticated in the target CN, no further signaling with the MS is required. The AAA server generates the required security keys by using the PoA ID and transfers the security keys to the target PoA. Counterpart security keys are generated in the MS without communication with the AAA server. Now that both the MS and the new PoA possess required keys, a simple handshake is needed to establish a Security Association (SA) at both ends of the radio link. This handshake is performed after radio synchronisation.

To complete the QoS setup, the HOS sends a request to the PCRF. As the MS is already authorised in the CN, this request can be sent before or after the key distribution request. At this stage of the PRIME procedure, allocating radio resources and establishing radio bearers are the only remaining tasks in the QoS setup procedure. For this purpose, the HOS’s request is forwarded to the RAN Gateway (the ASN-GW in the WiMAX and the S-GW in the LTE). The gateway then requests the target PoA to allocate the requested radio resources. The admission control module of the PoA considers local resources and either grants or rejects the request. Since the MS queries candidate PoAs before making a handover decision, it is expected
that the new PoA will grant the requested resources. Upon successful admission, the MS may need to perform a handshake with the new PoA and complete the creation of radio resources. To communicate with the target PoA, the MS is required to perform radio synchronisation.

Since the air interface of modern wireless technologies such as LTE and WiMAX is based on the Orthogonal Frequency Division Multiplexing (OFDM), both time and frequency synchronisations are required. The MS achieves time and frequency synchronisation by listening to pre-defined preambles broadcast over radio frames. The next step is collecting target cell system information which can be expedited by using the MIH protocol and obtaining this information from the current PoA. For example, in an LTE technology the old PoA may provide the MS with important parameters such as a new Cell Radio Network Temporary Identity and the System Information Block of the target eNB [54]. A similar approach has been used in the WiMAX technology [133].

After radio synchronisation and information collection, the MS may send a ranging request message to the new PoA to adjust its timing and transmission power level. To reduce the signaling delay for this message, the current and target PoAs may cooperate and provide the MS with a dedicated transmission opportunity for this message. Both the LTE and WiMAX technologies include this delay reduction technique [6, 53] which can be extended to inter-technology handovers by utilising the MIH protocol.

Integrating these delay reduction techniques combined with CN-PR and I-PR processes of the PRIME mechanism can result in a handover preparation delay which satisfies the requirements of a seamless handover. In the next section, we provide the signaling flow for LTE and WiMAX handover preparation using the PRIME mechanism. It is shown that the number of signaling messages required to establish a radio connection is significantly reduced.

4.6 PRIME Signaling Flow for LTE and WiMAX

To demonstrate the PRIME signaling flow in typical handover preparation scenarios, we consider two examples. In the first example, the MS is being served by a WiMAX network and performs a handover to an LTE network. In the second example, an LTE to WiMAX handover is presented. The presented PRIME signaling flow can be compared with the standard LTE and WiMAX network entry processes described in Section 2.2 and 2.3. This comparison forms the basis of our performance analysis for the UPTIME mechanism.
4.6.1 LTE Handover Preparation

The overall UPTIME process for LTE link preparation can be seen in Figure 4.7. The MS first performs the target RAN discovery and selection process. It then executes the CN-PR procedure to register with the LTE Enhanced Packet Core (EPC). Consequently, it pre-registers with the IMS through the P-CSCF located in the LTE network. This node is usually collocated (or closely located) with the Packet Data Network Gateway (P-GW). Finally, when the target eNB becomes available, the MS completes the LTE link preparation by performing the P-PR process.

In the PRIME mechanism we slightly modify the standard LTE initial registration process. This modifications allows registration with the EPC even when the target eNB is not determined. The Core Network Pre-Registration (CN-PR) is depicted in Figure 4.8. The signaling flow is based on procedures specified in the 3GPP Technical Specification [57, 55, 167, 168]. For more details on the standard LTE network procedures, refer to Section 2.3. This signaling flow is for basic configurations with a minimal number of messages. Extra messages may be required to implement additional features such as more security procedures or online charging.

The CN-PR procedure is performed in the following 23 steps:

1. The PRIME reregistration to an LTE network is initiated when the MS sends a SIP UPDATE message to the HOS. In this message, the MS requests the initiation of the pre-registration process by including our proposed SIP handover header. The MS also includes its identity using a Network Access Identifier (NAI). The NAI has a format of name@home and determines the identity of the MS and the home network.

2. The HOS uses the NAI to determine the target network. It then forwards the MS identity to the authentication server, in this case the Mobile Management Entity (MME).

3. and (4) The MME requests the Authentication Vector (AV) from the HSS and receives this information which is required for MS authentication.

5. and (6) The MME also retrieves the MS’s subscription profile to determine the MS’s eligibility to establish an Enhanced Packet System (EPS) bearer.

7. to (10) In these steps the MS and the MME mutually authenticate each other.

11. If the mutual authentication is successful, the MME responds with an EAP-Success message. At this stage, the $K_{\text{ASME}}$ can be constructed in the MME and the MS. From this key, the $K_{\text{NAS_enc}}$ and $K_{\text{NAS_int}}$ are obtained; these are security keys for the encryption and integrity check of control messages between the MS and the MME. To secure the communication between the MS and the target eNB, another security key called the eNB Key ($K_{\text{eNB}}$) should be generated. However, the $K_{\text{eNB}}$
Figure 4.7: UPTIME process overview for LTE handover preparation
Figure 4.8: CN-PR process for LTE handover preparation
depends on the identity of the target eNB and cannot be constructed at this stage.

(12) and (13) The MME updates the location information stored in the HSS and receives a reply back. The HSS may reject the update location request if it determines the MS should not be granted access to certain services at some locations [55]. If location update is rejected, the MME rejects the MS registration request. Otherwise, the MME constructs and stores a context for the MS.

(14) and (15) The HOS uses its knowledge of the multimedia session and determines the required QoS parameters. Then it informs the PCRF node in the LTE network by sending an Application/service info message, as specified by the 3GPP Policy and Charging Control (PCC) framework [57]. The HOS can send the service information any time after the successful authentication procedure.

(16) to (20) After the reception of the update location acknowledgment in step (14), the MME initiates the creation of the default bearer. The default bearer provides the MS with an “always on” connectivity. This bearer is mostly used for signaling and infrequent data transmission. To create the default gateway, the MME sends a request to the Serving Gateway (S-GW) which subsequently informs the Packet Data Network Gateway (P-GW). The P-GW then contacts the PCRF (assuming the PCC framework is implemented) to verify the QoS parameters of this bearer against the user’s subscription information and the operator’s policies. If the PCC framework is not implemented, the MME makes policy decisions on the bearer creation. The PCRF downloads the user’s profile from the Subscription Profile Repository (SPR), makes policy decisions and then either authorises or rejects the session creation request.

To reduce signaling delay, the PCRF may combine the create session request of the initial registration process with the dedicated bearer activation procedure, as specified by the 3GPP in [57]. This means that while a default bearer is being established, one or more dedicated bearers are also created to carry the data packets of the multimedia session. The PCRF uses the service information obtained from the HOS in step (14) to authorise the creation of dedicated bearers with the required QoS parameters.

(21) If IP address assignment is performed as part of default bearer activation, the PDN GW provides the MS with an IP address. The assignment of an IP address at this stage reduces the signaling delay because it is combined with the bearer creation procedure. However, alternative mechanisms such as the DHCP protocol can be used for the purpose of IP address allocation. In this case, the S-GW stores bearer information and forwards the create session response to the MME.

(22) At this stage, the MME is aware that the MS is authorised to use the requested services. Dedicated bearers with desirable QoS parameters are then established in the core network. However, since the target PoA has not yet been
selected, radio resources cannot be verified. The MME therefore should wait for the
initiation of the next phase of the PRIME mechanism. To indicate the success of
the CN-PR procedure, the MME sends a confirmation message to the HOS. In this
message, the MME includes the new IP address for the MS.

(23) The HOS sends a SIP OK message to the MS to inform it of a successful
CN-PR procedure. The HOS also includes the assigned IP address. If the procedure
is not successful, the HOS include the failure reasons. For example, the HOS may
indicate that the target network has rejected requested QoS parameters. The MS
may then select another RAN.

After a successful CN-PR procedure, the MS performs the IMS Pre-Registration
(I-PR) process. In the I-PR process, the MS uses its new IP address to register
with the IMS network though the P-CSCF in the LTE network. As mentioned
previously, the HOS functions as a relay between the MS and the new P-CSCF. To
reduce signaling delay, a direct link between the HOS and the new P-CSCF can be
established. Alternatively, the standard IMS architecture can be used in which the
HOS communicates with the new P-CSCF through the S-CSCF. See Section 2.4 for
the signaling flow for IMS registration. The underlying radio technology (in this
case LTE) does not affect the I-PR signaling flow.

After I-PR process and when the target eNB becomes available, the MS performs
the PoA Pre-Registration (P-PR) mechanism, depicted in Figure 4.9, to complete
the preparation of the handover preparation process. The 14 steps of this procedure
are:

(1) When the MS detects the target PoA, it initiates the P-PR process by sending
a SIP UPDATE message to the HOS.

(2) The HOS requests the MME to generate security keys and distribute them.
For this reason, the HOS includes the identity of the target eNB in the request.

(3) The MME and the MS then generate the $K_{eNB}$ key. The MME initiates the
process of distributing security keys, completing the initial attachment process,
and establishing radio bearers by contacting the target eNB. For this purpose, the MME
sends the Initial context setup request message to the target eNB. In this message
the MME includes the security key $K_{eNB}$. Using this key, the target eNB and the
MS are able to generate other keys for the encryption and integrity check of data
packets transmitted over the radio link.

(4) The MME also sends the Bearer setup request message to the eNB which
includes the QoS parameters of the EPS bearer along with other parameters.

(5) The eNB maps EPS bearer QoS parameters to the Radio Bearer QoS pa-
rameters. Then if the radio resources can be allocated, it sends a RRC Connection
Reconfiguration message to the MS. We note that in selection of the target eNB,
the MS first queries the availability of the radio resources. As such, the target eNB
Figure 4.9: P-PR process for LTE handover preparation
should be able to allocate the radio resources when requested by the MME.

6. The MS acknowledges the activation of the radio bearer.

7. and 8. The eNB acknowledges the activation of the default bearer and the dedicated EPS bearer.

9. The MS sends the Direct Transfer message to the eNB which contains both the Attach Complete and Session Management Response messages as responses to the initial attach procedure and dedicated bearer setup procedure respectively.

10. The eNB sends the attach complete / session management response message to the MME.

11. and 12. The MME informs the S-GW that the radio bearers have been established. The MME also includes the eNB address so that data packets can be forwarded to the target eNB.

13. The MME acknowledges the completion of P-PR process to the HOS.

14. The HOS sends a SIP OK message to the MS to indicate that the P-PR process is now complete.

4.6.2 Handover to WiMAX

The overall UPTIME procedure for a WiMAX handover is similar to the process depicted in Figure 4.7 for an LTE handover, with slight variations. For example, in the WiMAX technology there is no explicit location update procedure. The detailed signaling flow for WIMAX CN-PR is depicted in Figure 4.10. We use WiMAX procedures as specified by the WiMAX Forum in [5] and [51]. We also assume that the WiMAX network implements the Policy and Charging Control (PCC) framework including the PCRF, the Service Flow Authorization (SFA), and the Service Flow Management (SFM) functionalities. As mentioned in Section 2.2, in WiMAX Release 1.5, the SFA is implemented in the ASN Gateway (ASN-GW) and the SFM is located in BSs.

The 19 steps of the WiMAX CN-PR procedure include:

1. to 10. Similar to the handover to the LTE technology, the MS and the AAA server in the WiMAX CSN mutually authenticate each other. We assume the authentication is based on the EAP-AKA method which is also used in the LTE technology. As a result of successful authentication, the MS and the AAA server generate a Master Session Key (MSK). In WiMAX technology, the MSK is used to generate all other security keys. From this key, the MS and the ASN-GW (when it receives the MSK from the AAA server) are able to generate the Pairwise Master Key (PMK). However, at this stage the Authentication Key (AK) cannot be generated because it depends on the target BS Identity (BSID) which is not yet known. Without the AK to derive other security keys, the MS and the ASN-GW
Figure 4.10: CN-PR process for WiMAX handover preparation
should wait for the detection of the target BS.

(11) and (12) The HOS functions as a DHCP client and receives an IP address on behalf of the MS.

(13) to (17) To initiate the establishment of the network bearers, the HOS sends the service information (the QoS parameters) to the PCRF in the target WiMAX network. The PCRF acknowledges receipt of the message and downloads the user’s profile from the SPR database. The PCRF then makes a policy decision on the requested QoS parameters and sends the results to the ASN-GW. At this stage, since the target BS is not detected, radio resources cannot be verified and radio bearers cannot be established. The ASN-GW stores the bearer context and informs the HOS that the QoS policies have been received.

(18) and (19) The HOS receives a response from the ASN-GW which indicates the CN-PR has been successfully completed. Consequently, the HOS sends a SIP OK message to the MS.

After the CN-PR process, the MS registers with the IMS through the WiMAX network. In this process the HOS functions as a proxy between the MS and the P-CSCF in the target network. When the MS detects the target BS, it initiates the P-PR process, depicted in Figure 4.11. The concept is similar to the LTE P-PR process; the MS associates with the BS and performs handshakes to establish a SA and required radio bearers.

The 21 steps of the WiMAX P-PR procedure include:

1) The MS informs the HOS to initiate the P-PR process. In this message the MS includes the BSID of the selected BS.

2) The HOS sends a request to the AAA server to generate the PMK security key and distribute it to the Authenticator (the ASN-GW).

3) and (4) The ASN-GW receives the PMK and generates the AK and sends it to the BS.

5) The ASN-GW also initiates the establishment of radio bearers as described by the WiMAX Forum in [5].

6) and (7) The ASN-GW acknowledges the distribution of the AK to the BS. Consequently, the HOS informs the MS to proceed with the P-PR process.

8) and (9) The MS performs time and frequency synchronisation with the target BS by listening to frame preambles and broadcast information. Then the MS performs the WiMAX ranging process. The MS sends a Ranging Request (RNG-REQ) to the BS and receives a reply back in the Ranging Response (RNG-RSP) message.

10) and (11) In these steps, the MS and the BS exchange information about their capabilities, for example, supported modulation and coding schemes.

12) to (14) Since both the MS and the target BS possess the authentication key, only a three-way handshake is required to establish a Security Association
Figure 4.11: WiMAX P-PR signaling flow
(SA) between these entities. As Hur et al. describe in [50], the SA-TEK-Response message includes the SA-TEK-Update field which contains the Traffic Encryption Key (TEK) for a secure communication between the MS and the BS. After this step, the messages between the MS and the BS can be encrypted.

(15) and (16) The MS then sends a Registration Request (REG-REQ) message to the target BS, as required in the WiMAX technology. To further reduce the P-PR registration process in the WiMAX technology, the MS can send the REG-REQ and SBC-REQ to the target network during the CN-PR process. However, further changes in the WiMAX network are required.

(17) to (19) To establish radio bearers, the target BS and the MS perform a three-way handshake, as specified by IEEE in the WiMAX standard [133]. In this process, the BS first performs the admission control and allocates required radio resources. The BS assigns a Service Flow ID (SFID) and includes it in the Dynamic Service Allocation Request (DSA-REQ) message. This message also contains a Connection ID (CID) and the admitted QoS set. The MS records the CID and accepts the QoS parameters by sending the DSA-RSP message. To complete the process, the BS sends a DSA-ACK message to the MS.

(20) and (21) The BS sends a Path_Reg_RSP message to the ASN-GW and receives an acknowledgment. At this stage, the radio and network bearers are created and ready for data transmission.

When the MS prepares the radio connection using the PRIME mechanism, the radio interface is ready for data transmission and the MS is registered to the IMS. The next stage of the handover process is handover execution. In the UPTIME framework, we use the Conservative Soft Handover (cSHO) for this purpose.

In the next chapter we take a closer look at the cSHO mechanisms. This mechanism complements the PRIME scheme by executing a smooth transfer of the multimedia session to a newly created radio connection. The aim of the cSHO handover execution process is to reduce packet loss as much as possible while conserving radio resources.
Chapter 5

Conservative Soft handover

5.1 Introduction

In this chapter, we describe our Conservative Soft Handover (cSHO) mechanism for handover execution. As discussed in Chapter 3, seamless handover requires minimum handover preparation delay and packet loss [33, 85]. In the previous chapter we proposed the PRIME mechanism for handover delay reduction. In this chapter we focus on packet loss minimisation. We describe our cSHO mechanism which reduces packet loss while conserving radio resources and battery power.

Inter-technology handovers often suffer from high packet loss. In particular, packet loss occurs in cases where the radio signal from the current and target Radio Access Networks (RANs) fluctuate significantly. Signal fluctuation is particularly common at the edges of radio cells where vertical handovers are more likely to occur. Due to signal fluctuations, the MS may occasionally experience packet loss over the current or target radio interfaces.

As discussed in Section 3.5, both Hard Handover (HHO) and SHO schemes fail to properly mitigate packet loss in inter-technology handovers. In HHO schemes, because of the high handover delay and signaling costs in inter-technology handovers, the MS is not able to move frequently between RANs and avoid packet loss. On the other hand, in SHO, the simultaneous use of two interfaces minimises packet loss. However, resources are used excessively for the transmission of duplicated packets.

A mobility management solution suitable for the Next Generation Mobile Network (NGMN) should include minimum packet loss and efficient resource management [33]. Furthermore, the mobility solution should use utilise standard protocols for handover control. In this chapter, we describe the cSHO mechanism which satisfies these requirements with the following approach:

- Multimedia packets are duplicated in the network and delivered to both the current and target Point of Attachment (PoA).
Based on radio conditions at any short period of time, one of the interfaces is used for data transmission.

The MS quickly switches between radio interfaces to experience lowest Packet Error Rate (PER)

The Session Initiation Protocol (SIP) and the Media Independent Handover (MIH) protocols are extended and used for exchanging control messages.

This chapter is organised as follows. In Section 5.2, we explain the rationale behind cSHO mechanism and our design considerations. In Section 5.3, we describe the cSHO procedure. Section 5.4 provides technical details on interface selection for minimum packet loss. Finally, in Section 5.5, necessary modifications to the standard SIP and MIH protocols are discussed.

5.2 Motivations, Rationale, and Design Principles

Because of radio signal fluctuations due to shadow fading and other effects, the Hard Handover (HHO), in some cases, is not able to achieve seamless handover. Figure 5.1 demonstrates a scenario where two Radio Access Networks (RANs) are used to provide coverage. The MS includes two radio interfaces, Interface 1 (IF1) and Interface 2 (IF2), which can be used independently and simultaneously. In locations where RAN1 does not have good coverage, the Mobile Station (MS) seamlessly connects to RAN2. In an HHO, when the MS detects that the signal from the current RAN (that is, RAN1) is degrading, it joins the target RAN (RAN2) which provides better signal quality. In this scenario, the motivation for handover is maintaining suitable radio connectivity.

In the handover process, the MS attempts to determine the optimum time for handover execution. In an ideal (but imaginary) scenario, as the MS moves away from the current PoA in a straight line, the received Signal-to-Noise Ratio (SNR) of the current PoA decreases. Meanwhile the SNR of the target PoA increases constantly and eventually surpasses the SNR from the current PoA. In this scenario, determining the optimum handover execution time is as easy as comparing SNR, or the Packet Error Rate (PER), of two RANs.

In a more realistic scenario however, radio signals fluctuate rather than decreasing or increasing constantly. The main reason for this is shadow fading. The shadow fading effect, which is often modeled by a random process with a standard deviation of 8 to 12 dB, can have a dominant effect at radio cell edges where handover is most likely to occur. In the presence of shadow fading, it is difficult to determine the optimum handover execution time. If the handover is performed too early, the
Figure 5.1: A vertical handover scenario to improve radio coverage

signal from the target interface is not yet stable and packet loss occurs. If the handover is executed too late, service interruption may be experienced over the current interface. In Figure 5.1, service interruption occurs over IF1 or IF2 if the handover process is executed too late (after point 1) or too soon (before point 2).

Furthermore, in HHO schemes frequent handovers, also known as the ping-pong effect, should be avoided. Because of the high signaling delay, frequent handovers may increase packet loss. They also incur a significant signaling cost and increase network load. Therefore, mechanisms such as SNR hysteresis and dwelling-time are often used to avoid the ping-pong effect [95].

As discussed in the previous chapter, the above drawbacks are overcome using a Soft Handover (SHO) approach in which both interfaces are used for the transmission of a duplicated media stream. The result is a significant reduction of packet loss [147, 32, 31, 144]. However, the downside is the consumption of scarce radio resources and battery power. Because of the scarcity of radio resources, particularly in cell edges where handovers are more likely to occur, SHO schemes which require duplicating IP packets may cause an unacceptable network traffic load.

Moreover, SHO schemes are not suitable for offloading data traffic from the main network to the target network which covers a hotspot. In the hotspot scenario, depicted in Figure 5.2, the high traffic demand is met with overlay wireless networks. The RAN with smaller coverage (RAN2) is used to offload data traffic economically from the main network (RAN1). The smaller RAN often provides higher data rates, shorter delays and lower service costs. These advantages motivate the MS to perform a handover to RAN2 as soon as it becomes available. The SHO mechanism can still be used to improve radio connectivity and minimize packet loss. However, since SHO requires using two radio interfaces, the benefits of the hotspot scenarios are lost: data traffic is not offloaded from RAN1, the service cost is not reduced, and the MS cannot use higher-rate multimedia codecs. As such, it is more reasonable for the MS to wait for a stable radio signal from RAN2 before performing a handover.

To overcome the above limitations of SHO schemes, we propose the cSHO mech-
anism which has better packet delivery performance than HHO but does not require the transmission of duplicated packets. In particular, we consider the following design requirements:

- **Compatibility with the UP TIME framework.** The cSHO mechanism is designed for the UPTIME framework and follows NGMN general requirements such as the separation of signaling and media, IMS compliance, and utilising open standards.

- **Radio resource consumption.** The cSHO scheme is designed to significantly reduce the resource consumption of SHO schemes.

- **Battery power usage.** Using two radio interfaces in the SHO method results in battery power drainage, due to doubled radio transmission power and always-on interfaces. However, prolonging battery life is one of critical issues of multi-radio mobile terminals [97]. We designed the cSHO mechanism to save on the battery power by activating and deactivating radio interfaces.

- **Packet delivery compromise.** SHO schemes result in minimum packet loss, but have high resource consumption. HHO mechanisms, on the other hand, cause higher packet loss, but conserve radio and battery resources. There is a trade-off between resource consumption and Packet Error Rate (PER) performance. We designed the cSHO method to have a good balance between PER performance and resource consumption. Under the same conditions, the cSHO scheme should have a considerable PER advantage over the HHO scheme. However, when compared with SHO, the performance may not be as good.

The rationale behind the cSHO mechanism is to enable the MS to switch alternately between the current and target PoAs to maintain the best possible radio connectivity. In the cSHO mechanism, media packets destined for the MS are duplicated in the network and buffered at the current and target PoAs. At each given period of
time, the MS monitors the SNR of radio links and selects the radio interface with the lowest PER for packet transmission. In cases where a change of active interface is required, the MS sends a simple signaling message to the current and target PoAs and immediately receives data packets over the newly activated link. In comparison with the SHO scheme, the cSHO mechanism consumes far fewer resources because only one interface is active at each given period of time. In comparison with the HHO method, the cSHO mechanism benefits from fast interface switching. In the HHO approach, activating a new link involves a considerable signaling cost and high handover delay. As such, frequent handovers (the ping-pong effect) must be avoided. In cSHO however, we make switching interfaces and reacting to sudden channel variations practical by satisfying the following requirements.

- Signaling delay for interface switching should be significantly reduced.
- Signaling load should be minimised.

As we identified in Section 2.5.2, SIP-based handover execution solutions suffer from high handover delays due to using relatively large text-based messages and end-to-end signaling exchange. The handover execution delay is a few hundred milliseconds which makes it impractical to follow shadow fading variations. In cSHO, we significantly reduce the interface switching delay by some of the mobility management functionalities at the edge of the transport network, that is, in the PoAs. In this solution, the MS only needs to communicate with the directly connected PoA rather than exchanging information with the Corresponding Host (CH). Moreover, instead of using multiple large SIP messages, interface switching in cSHO is performed with the transmission of only one MIH message. The result is a reduction of the signaling exchange delay to the latency of the wireless link.

In the cSHO mechanism, before the actual handover occurs, data packets are duplicated in the transport network and sent to the current and target PoAs. Using the MIH protocol, the MS instructs the currently active PoA to transmit data packets and the inactive PoA to buffer them. When a change of radio interface is required, the MS instructs both PoAs to change their roles; the newly activated PoA starts sending data packets and the other PoA buffers them. The number of signaling messages is reduced from several end-to-end messages to only one over-the-air message.

Another aspect of the cSHO mechanism is the fact that it is a cross-layer handover execution method. While SIP (as an application layer protocol) is used to control the overall handover procedure, the IP layer packet buffering and transmission along with MIH-based interface switching is deployed to improve packet delivery performance.
The cSHO mechanism also benefits from the *gradual* transfer of the data session to the new radio link. As the MS moves deeper into the radio coverage area of the new PoA, the average SNR of the new interface increases and the average SNR of the current interface decreases. During the handover period of the cSHO scheme, as the radio condition of one interface improves, the interface is utilised more often. Eventually, the MS completely joins the new PoA and the old radio link becomes obsolete. This smooth transition results in a lower packet error rate when compared with an HHO.

The drawback of the cSHO method is the involvement of PoAs in the handover execution process. In other words, the cSHO mechanism should be supported by both the current and target PoAs. In comparison, traditional SIP-based HHO and SHO schemes do not require network support. However, this additional complexity at PoAs is required to significantly improve session quality and reduce resource consumption.

We also note that the performance of the cSHO method can be affected by underlying wireless technologies. The cSHO mechanism requires minimum radio transmission delay for MIH messages to enable the MS to react to variations in the quality of the radio channel. Large transmission delays over radio interfaces mean that the MS is not able to quickly switch its radio interfaces and use the interface with the lowest PER. Fortunately, modern wireless technologies such as WiMAX and LTE have relatively low delays. For example, the LTE technology requires RAN packet latency, defined as delay of data packet transmission between the MS and the RAN gateway, to be less than 5 milliseconds [53].

In the next section, we provide an overview of cSHO handover execution scheme and describe its signaling flow.

### 5.3 cSHO Procedure and Signaling Flow

The handover process in a complete mobility control solution, such as the UPTIME framework, includes three stages: 1) network discovery, 2) handover preparation, 3) handover execution. In the UPTIME framework, the network discovery and handover preparation stages are performed using the PRIME mechanism. The outcome of the PRIME mechanism includes:

- The target RAN and PoA are selected.
- The MS is registered in the target network and the new radio link is ready for data transmission.
- The MS is registered to the IMS network.
For the third stage of the handover process, the cSHO mechanism is performed. To initiate the process, the MS determines when packet duplication should be started. The MS decision is based on two factors. First, packet duplication should be initiated before the disconnection of the current interface. Otherwise, service interruption may occur. Second, the radio interface of the target PoA should be strong enough for data transmission. However, intermittent degradation of signal quality is acceptable. If the duplication process starts too early, network resources on the backbone link of the target PoA are consumed which is not desirable. The MS may use an average SNR combined with location information to estimate the cSHO initiation time.

To start packet duplication, as depicted in Figure 5.3, the Handover Adaptation Layer (HAL) in the MS sends a SIP re-INVITE message to the Handover Server (HOS). This is similar to the standard IMS Service Continuity and Centralization (SCC) mechanisms specified by the 3GPP in [84]. However, the re-INVITE message in the cSHO mechanism includes a new SIP header called the handover header.

This header is our proposed extension to the SIP protocol and will be described in Section 5.5. The handover header contains necessary information for requesting packet duplication and buffering. More specifically, the handover header includes the following information:

- **Handover type.** The MS indicates which handover type should be used. As mentioned in the previous chapter, backward compatibility is one of the important design considerations for the UPTIME framework. In cases where the cSHO mechanism is not supported by current or target PoAs, the MS may use the traditional SIP-based SHO which does not require network support.

- **Required action.** The handover header contains information on the action requested by the HAL. In this case, the HAL is requesting the initiation of the packet duplication process. The HAL may also request handover termination or suspension.

- **Current and target PoA identity.** These fields allow the HOS to identify the associated PoAs and communicate with them to order data packet buffering.

- **MS current and new IP addresses.** This information is used for packet duplication and handover execution.

The SIP re-INVITE message is forwarded to the HOS through IMS nodes which include the Proxy Call Session Control Function (P-CSCF) and Serving Call Session Control Function (S-CSCF). For the sake of presentation clarity, we have not included these nodes in the signaling flow depicted in Figure 5.3.

The HOS receives the SIP re-INVITE message and collects the necessary information. It then generates a new SIP re-INVITE message and forwards it to the
Figure 5.3: cSHO procedure

CH. As a result, the CH updates the multimedia session information and replies with a SIP OK message. Finally the HOS sends a SIP ACK message to the CH and completes the update of session information at the CH side. The process is identical to the three-way handshake for updating the remote leg in the standard IMS SCC handover process described in Section 2.5.2.

Through the remote leg update process, the CH becomes aware of the MS change of IP address. The multimedia application in the CH opens new UDP sockets which enables it to continue to receive multimedia packets from the underlying operating system.

In the cSHO procedure, the CH does not duplicate media packets. We avoid the end-to-end packet duplication approaches such as solutions proposed in [146, 81, 89]. Instead we use packet duplication in the network. This is because end-to-end packet duplication consumes resources at the CH side as well as the MS side. In addition, the RAN serving the CH may not be able to accept the allocation of more resources for packet duplication.

To order packet duplication in the network, the HOS sends a request to the Media Duplication Function (MDF). In the UPTIME framework, the MDF is a network element located in the transport stratum of NGMN and is controlled by the HOS located in the service stratum. In line with ITU-T recommendations presented in [33], the communication link between the HOS and the MDF is based on the MIH protocol which is an open standard. As explained in Section 4.3, we prefer the MIH
protocol over SIP because it causes relatively less delay to initiate packet duplication.

As seen in Figure 5.3 in step 4, the HOS sends an MIH message to the MDF and requests packet duplication. Unfortunately, the current version of the MIH standard does not include features for requesting packet duplication in wireless networks. As such, we extend the MIH protocol with a set of new features which allows packet duplication, buffering, and transmission. These are further explained in the next section.

In the next step, the MDF starts packet duplication. As discussed in Section 4.3, the MDF includes a *Duplication List* and a *Duplication Agent*. With every request from the HOS, a new entry is created in the Duplication List. The list entry contains the current and new IP addresses of the MS and the IP address of the CH. The existence of an entry for the MS indicates that the MS is in the cSHO handover execution status. The Duplication Agent of the MDF replicates any IP packets which encapsulate UDP packets of multimedia sessions destined for the MS. Data packets of other applications, such as web browsing and file transfer, are not duplicated. As these applications do not have strict QoS requirements, the MS can use other mobility management solutions such as Mobile IP for them.

In step 6 of the cSHO mechanism, the HOS informs both the current and target PoAs that the MS is in the cSHO handover execution stage. The MIH protocol is used for this purpose. As with packet duplication, the current version of the MIH protocol does not provide features for requesting packet buffering. We introduce packet buffering as a new MIH feature.

In Figure 5.3, we assume the PoA and the PoS are co-located. As such, the MIH messages from the HOS are sent directly to the PoA. In scenarios where the PoS is implemented deeper in the network and as a separate node, the HOS request the PoS and then the PoS communicates with the PoA to request packet buffering. The latter implementation option slightly increases signaling delay and is not desirable.

Finally, in step 7 the HOS informs the MS that the packet buffering and duplication process has been completed. This is achieved by exchanging SIP OK and ACK messages between the MS and the HOS. In the SIP OK message, the HOS also include the handover header and indicates that the requested handover action has been successfully performed. We note that the HOS may change the type handover which the MS requested. For example, in some scenarios based on network conditions, user subscription, and operator’s policies the HOS may determine a cSHO handover is not acceptable. In this case the HOS may request the MS to perform a HHO or SHO scheme. The MS acknowledges receipt of the SIP OK message by a SIP ACK message.

At this stage, both the current and target PoA receive duplicated data packets from the MDF. However, only the current PoA transmits packets to the MS. When
the MS determines that interface switching is required, it sends two MIH messages to the current and target and PoAs. Interface selection at this stage is based on SNR samples of radio interfaces and is discussed in Section 5.4.

The MIH message sent to each PoA indicates the MS request for either data buffering or data transmission. When the MS requests data transmission, it also includes the sequence number of the expected packet which is the last received sequence number incremented by one. The PoA then dequeues data packets in a simple FIFO fashion. The PoA drops data packets with sequence numbers less than the expected one. It then sends the expected packet and all subsequent packets. The Real Time Protocol (RTP) sequence numbers of multimedia packets can be used for tracking received packets. Using RTP sequence numbers to reduce service interruptions during the handover process is an approach suggested by Qi et al. in [169]. Alternatively, as suggested by Vakil et al. in [29], the MS may use the RTP time stamp to indicate the expected data packet.

With the above approach, the activation of radio interfaces is performed quickly and the MS immediately begins receiving data packets from the new radio link. This fast interface activation allows the MS to react to channel variation due to shadow fading. At this stage of the cSHO process, the MS makes interface switching decisions at specific time intervals. The duration of these time intervals is selected so that the radio channel can be considered almost constant between two consecutive decision making points. As such, the MS can estimate which interface is likely to have a lower PER in the following time interval. In the following section, we discuss the interface selection mechanism and the duration of handover decision intervals.

To avoid the loss of radio connectivity in cases where MIH messages are lost over radio channels, the cSHO mechanism relies on the explicit deactivation of radio interfaces. The PoA with the active interface continues data transmission until it receives an MIH message indicating the deactivation of that radio interface. Occasionally, when the MIH message is lost over the radio link, the PoA is not informed of interface deactivation and continues data transmission. As a result, for a short period of time the MS may receive duplicated packets over both radio interfaces. However, in the next interface switching interval the PoA will most likely receive an MIH message and stops packet forwarding. The HAL manages the occasional reception of duplicated packets and so multimedia applications do not notice this issue.

When the MS decides that the new radio link is stable, it can stop or suspend the cSHO process. To do so, as shown in Figure 5.4, the MS communicates with the HOS which coordinates the cSHO handover process. The MS initiate the process by sending a SIP UPDATE message to the HOS. This message contains the handover header which indicates the required action type. The action type can indicate han-
dover termination or suspension. Consequently, the HOS performs the requested action.

To stop the cSHO process, the HOS requests the MDF to remove the MS from the Duplication List. It then requests the current and target PoAs to complete the cSHO handover process, that is, to stop packet buffering. As mentioned before, the communication of the HOS with PoAs and the MDF is based on the MIH protocol. In particular, the HOS uses the standard \textit{MIH\_N2N\_HO\_Complete request} message to announce the completion of cSHO process.

To suspend the cSHO process, in addition to stopping packet duplication and buffering, the HOS also requests the old RAN, the RAN which the MS is leaving, to retain radio resources and other required information. This allows the MS to return to the previous RAN if the new RAN becomes unsuitable. The MIH protocol provides necessary services for this purpose. In particular, the HOS uses the \textit{MIH\_N2N\_HO\_Complete} request message with the ResourceRetention-Status flag to TRUE \cite{106}.

In the next section, we describe the radio interface selection method of the cSHO mechanism.

### 5.4 Interface Selection in cSHO

The cSHO mechanism aims to reduce the overall Packet Error Rate (PER) which the MS experiences during the handover process. As such, the interface selection is based on PER estimation using the received Signal-to-Noise Ratio (SNR) of the current and target radio interfaces. The Handover Adaptation Layer (HAL) in the MS samples the SNR of both interfaces and estimates which interface will have the lowest PER in the next \textit{handover decision} time interval.
Figure 5.5: Sampling SNR of radio interfaces

As depicted in Figure 5.5, handover decision intervals have a duration of $T_d$ (decision time). In each interval, the received SNR is sampled every $T_s$ second. $T_s$ is much less than $T_d$ meaning the received SNR is sampled several times in each handover decision interval. Using these SNR samples, interface switching decisions are made every $T_d$ seconds. The aim is to predict the PER of each radio interface in the next time interval. The prediction is based on averaging SNR samples over a time interval.

For accurate PER prediction, the HAL should be able to collect enough SNR samples to enable it to estimate the SNR average correctly. With a correct average SNR calculation, the HAL can determine the average PER experienced by the MS on each radio link. Moreover, the duration of handover decision intervals should be short enough to enable the MS to react to fast variations in radio channels. Determining suitable values for the SNR sampling rate and the interface selection frequency depends on the characteristics of radio channels.

The SNR received by a moving MS, as depicted in Figure 5.6, changes due to three main effects [6]. The pathloss effect represents attenuation in signal strength due to the distance between the MS and the PoA. Shadow fading results from the reflection and blockage of signal by surrounding objects such as buildings and trees. Fast fading is due to multipath reception of the radio signal and the Doppler spread.

The speed of MS movement affects the speed of radio channel variations due to Doppler spread and shadowing effect. The channel coherence time parameter, or its spatial counterpart the coherence distance, indicates the speed of channel fluctuations. The Doppler spread has a much smaller coherence time than shadow fading. For example, for a speed of 45 km/h in the 5.8 GHz frequency band, the Doppler coherence time is around 4 ms [104]. In other words, every 4 ms the radio channel significantly changes. Higher speeds result in shorter channel coherence time and faster channel variation. This fast channel fluctuation imposes limitations on supporting high mobility speeds. The WiMAX and LTE technologies are expected to
support high performance for speeds up to 125 km/h and 120 km/h [52].

The coherence distance of the shadow fading effect in vehicular environments, as recommended by the ITU-T in [170], can be considered to be 20 m. As such, the shadow fading coherence time for the speed of 45 km/h is 1.6 seconds. For faster movements, the coherence time will be smaller. However for common speeds of up to 120 km/h, the coherence time will be few hundreds of milliseconds.

Because of its short coherence time, the MS cannot perform inter-technology handovers according to fast channel variations. However, the MS may be able to react to the shadow fading effect provided that fast signaling for interface switching is implemented. As such, in the cSHO mechanism, the effect of fast fading is averaged out from SNR samples. A simple low path filter can be applied to the received signal to remove fast fading component while following shadow fading variations. The time constant of such a filter should be significantly larger than the channel coherence time of due to Doppler spread and much smaller than the coherence time of shadow fading [171]. For example, a simple exponential weighted moving average can be used as the low pass filter:

$$\tilde{\gamma}[n] = \alpha \gamma[n] + (1 - \alpha)\tilde{\gamma}[n - 1],$$

where $\tilde{\gamma}$ is the average SNR at time $t = nT_s$, $\gamma[n]$ is the nth SNR sample, and $\alpha$ is the smoothing factor which is a constant between 0 and 1. $\alpha$ can be calculated from the required time constant of the filter (denoted by $\tau$):

$$\alpha = exp(-T/\tau),$$

In the cSHO mechanism, the duration handover decision interval ($T_d$) equals the filter time constant ($\tau$). This enables the HAL to react to the shadow fading effect while creating minimum signaling load. If the $T_d$ is set to lower values than $\tau$, the

---

Figure 5.6: Three components of the received signal [6]
MS sends MIH messages more frequently than required because the fast fading has been removed from the output of exponential moving average filter.

SNR samples are taken every $T_s$ seconds by measuring the pilot signal sent in radio frames [172]. Both WiMAX and LTE technologies use pilot tones for the estimation of average received power and radio synchronisation. In comparison with blind channel estimation methods, pilot-based schemes are faster and more suitable for high speed wireless technologies [52]. Pilot tones are scattered in time and frequency to allow accurate channel measurement. The maximum time distance between subsequent pilots is selected to be less than the channel coherence time [52]. Therefore, SNR sampling can be performed with adequate rates to allow fast fading to be averaged out.

After measuring the average SNRs, the next step is to estimate the PER of each radio interface. We note that the SNR of different wireless technologies cannot be directly compared [37] because each technology utilises a different Modulation and Coding Scheme (MSC). Other techniques such as Hybrid Automatic Repeat Request (HARQ) and Forward Error Correction (FEC) also affect the PER performance. Therefore one value for SNR may result in different PERs over different radio interfaces. As such, in cSHO mechanism, we estimate and compare the PER of radio interfaces and select the interface with the lower PER.

The PER can be estimated using the PER-SNR curves. An example of such graph is depicted in Figure 5.7. The figure shows two PER-versus-SNR curves for WiMAX and LTE technologies. The PER-SNR curve of the WiMAX technology has been reported by Zaggoulas et al. in [7] and belongs to a scenario in which the radio interface uses the 16QAM modulation scheme with a coding rate of $3/4$. The LTE PER performance, as reported by Nagaraj et al. in [8], is for a QPSK modulation scheme with a coding rate of $8/9$ combined with the HARQ technique.

PER-SNR curves for various modulation techniques are stored in the HAL. After collecting SNR samples and running them through the averaging filter, the HAL uses these curves to estimate the average PER experienced over each radio interface. Since the duration of the handover decision interval is much less than the coherence distance of shadow fading, it is expected that the average SNR, and therefore the average PER, remains almost unchanged in the next interval. As such, the HAL selects the interface with the lowest PER as the active interface.

As mentioned in the previous section, if a change of active radio interface is required, the HAL sends MIH messages to the current and target PoAs and demands data transmission and buffering. In the next section, we describe these messages and the required extensions to the IEEE 802.21 standard to allow allow packet duplication, buffering, and transmission.
5.5 Extensions to SIP and MIH Protocols

Using standard protocols for signaling exchange is one of the critical requirements of an efficient mobility management framework for Next Generation Mobile Network (NGMN) [33]. Moreover, the principle of separation of service control and media functionalities in NGMN requires the use of standard protocols for communication between the service and transport strata [20]. As such, our UPTIME mobility management framework and its handover preparation and execution mechanisms are developed based on standard protocols such as the Session Initiation Protocol (SIP) [61] and Media Independent Handover (MIH) [106] standards.

Both protocols are extensible which means new extensions can be added to them without modifications to the core specifications. In this section, we present the technical details of the proposed extensions to the SIP and MIH protocols.

5.5.1 Proposed SIP Handover Header

The SIP protocol comprises basic capabilities and extended features. Basic capabilities of SIP are specified in the IETF RFC 3261 [61] which ensures the interoperability of SIP-based products and solutions. The protocol operation is based on dialogs between SIP peers, that is, between the Mobile Station (MS) and its Corresponding Host (CH). Each dialog consists of SIP Requests and SIP Responses. SIP requests are also called methods.

The RFC 3261 defines six basic methods [173]. These methods (that is, SIP requests) include REGISTER, INVITE, ACK, CANCEL, BYE, OPTIONS. This RFC also specifies the structure of SIP messages. Each SIP method consists of several headers and one message body. SIP headers provide the necessary information
for the SIP method. For example, the SIP INVITE method includes the TO and FROM headers. By following the basic specifications of the SIP protocol, SIP peers can communicate with each other and SIP servers can route messages to the correct destinations.

However, not all multimedia services can be implemented using basic SIP methods and headers. SIP implementations often support other methods such as the UPDATE method as specified in IETF RFC 3311 [174]. The 3GPP also uses the flexibility of the SIP protocol and in RFC 3455 [156] new methods and headers for IMS clients are proposed.

We take a similar approach and realise the UPTIME framework by extending the SIP protocol with a new handover header. This header is included in the SIP re-INVITE, UPDATE, and OK methods. The handover header enables the MS and the Handover Server (HOS) to communicate and coordinate the handover process. For example, as shown in Figure 5.3, the MS includes this header in the SIP re-INVITE message to request the HOS to initiate packet duplication in the network.

The introduction of this new header is not a significant modification to the SIP implementations in the IMS network. Since this header is only exchanged between the MS and the HOS, only these two entities should support this header. Other network nodes such as IMS Call Session Control Functions (CSCF) and the CH are not required to support the handover header.

The handover header is depicted in Figure 5.8. The header contains the following fields, some or all of which may be present in different SIP messages:

- **Type**. This field specifies the type of handover execution mechanism requested by the MS or granted by the HOS. The available options for this text field include cSHO, HHO, and SHO. Other choices can be added if more mechanisms are proposed for the UPTIME framework.

- **Action**. The MS includes this field in the SIP Re-INVITE message to request the HOS to take an action. The “start” and “stop” options are used to initiate and terminate the cSHO procedure. If the MS wishes to complete the cSHO procedure while retaining network resources in the old RAN, the MS selects the “hold” option instead of “stop”. To rejoin a RAN which has retained resources, the MS uses the “continue” option of the Action field in the handover header.

- **New_poa_id and old_poa_id**. This field is an identifier which uniquely determines the current and target PoA. In the 3GPP specifications a combination of cell ID and network ID can be used to uniquely determine the PoA. In WiMAX technology, the MAC address of the PoA can be used for this purpose. Using this field the MS can inform the HOS of the identity of PoAs.
The HOS can then locate and communicate with the current and target PoA to request packet buffering.

- **New_ip_address and old_ip_address.** These fields specify the MS's current and new IP addresses. This information is required by the HOS to perform the handover process by updating the multimedia session information at the CH.

Figure 5.9 shows an example of handover header usage in a standard SIP re-INVITE message. To focus on the handover header, we only show part of the SIP message. In this figure, we use a sample SIP re-INVITE message presented by the 3GPP in [175]. The re-INVITE message starts with the name of the SIP method, in this case, INVITE. Then the standard “Via” header is included which determines the interim nodes that the message passes through. This header is updated by network nodes as the SIP message is forwarded to the next hop. The message structure continues with other standard headers such as From, To, and call-ID. Finally, as highlighted in Figure 5.9, the Handover header is included.

It should be noted that the SIP protocol does not mandate the order of SIP headers within a message. Therefore the position of SIP headers including the handover header may be different than Figure 5.9.
5.5.2 Proposed MIH Link Actions

In the UPTIME framework, the MIH protocol is used for MS-PoA communication and for links between the HOS and the MDF and PoAs. In particular, in the cSHO mechanism the MIH protocol is used by the HOS to request packet duplication in the MDF. The HOS also uses the MIH protocol to communicate with the PoAs and order packet duplication. Finally, to activate or deactivate radio interfaces, the MS uses MIH messages to directly instruct the current and target PoAs.

The current version of the MIH protocol provides services for mobility information collection and enables the flow of such information between various wireless technologies. The protocol also specifies messages for requesting resource reservation and handover execution. However, the MIH protocol has limited features for handling media packets. For example, in a Hard Handover (HHO) process, the MS may use an MIH message to request the old PoA to forward buffered data packets to the target PoA [106]. Nevertheless, the standard does not provide features for requesting packet duplication and buffering. Our aim in this section is to extend one of the standard MIH messages to support these features. The standard allows protocol extensions by allowing Vendor Specific Information Elements (IE) and reserving some of available bits in each data field.

The MIH protocol allows a link layer action to be requested at a remote node. For example, an MIH Point of Service (PoS) may request a mobile terminal to lower its transmission power or disconnect a radio link. This MIH command service seems suitable for requesting other actions such as data stream duplication and packet buffering. The technical description of this MIH service is as follows.

To request a link action at a network node, the MIH user sends a MIH_Link_Actions request message. The IEEE 802.21 standard [106] specifies the format of this message, as depicted in Figure 5.10. Like other standard MIH messages, the MIH user includes its own identifier (Source Identifier) and the network node identifier (Destination Identifier) in this message. The MIH user then lists the required actions by including the LinkActionsList field. This lists contains one or more actions each of them presented using a specific data type called the LINK_AC_TYPE. Currently the MIH standard includes only four link actions. Available options are LINK_DISCONNECT, LINK_LOW_POWER, LINK_POWER_DOWN, and LINK_POWER_UP.

MIH link actions may also be combined with a valid attribute. For example, an MIH user (such as a mobile terminal) may request a remote node (such as a base station) to disconnect a radio link and forward buffered data packets by including LINK_DISCONNECT combined with DATA_FWD_REQ attributes. In this example, the MIH user may also request the retention of radio resources by using the
attribute LINK_RES_RETAIN.

As mentioned previously, the current standard options for the combination of link actions and their attributes are not enough to support UPTIME procedures such as the cSHO and PRIME mechanisms. Therefore we propose new link actions to be added to the standard. Fortunately, the standard includes enough reserved bits for this purpose. As specified in the IEEE 802.21 standard, the LINK_AC_TYPE is an 8-bit integer number which currently has only the first five values (0 to 4) are assigned to defined link actions. The proposed MIH link actions are listed in Table 5.1. In this table we have also included the options which are currently available in the MIH protocol.

We conclude this chapter by summarising the cSHO operation and its main characteristics. The proposed cSHO mechanism is capable of reducing the packet loss of handover process by enabling the MS to quickly change its active radio interface. In cSHO mechanism, data packets are duplicated in the network and forwarded to both current and target PoAs. At each given period of time, the MS selects the radio interface with the lower PER as its active interface. Since both PoAs receive the multimedia data stream, the activation of a radio interface requires the transmission of a single signaling message from the MS to the associated PoAs. With this minimised interface activation delay, the MS is able to react to sudden changes in the radio channel quality due to shadow fading. The proposed cSHO mechanism has the following characteristics:

- Compared with an SHO scheme, the cSHO method consumes much less radio resources and battery power. This is because at each given period of time only one radio interface is utilised for packet transmission.
• Compared with an HHO scheme, the cSHO mechanism has a lower PER. Because of the large handover delay and signaling cost, frequent handovers are not possible in the HHO scheme. Therefore, the MS may experience some degree of packet loss before it is allowed to change its radio interface. On the other hand, in the cSHO mechanisms the interface activation delay and signaling cost is minimised which enables the MS to use the interface with the lower PER.

• The cSHO mechanism is implemented using open standards such as SIP and MIH protocols. Some extensions to these protocols have been proposed.

The cSHO mechanism can be combined with the PRIME scheme proposed in the previous chapter to present a comprehensive seamless mobility solution. In this solution, handover preparation latency is reduced by the PRIME mechanism while packet loss is avoided using the cSHO method. This combination allows mobile operators to mitigate service interruptions in intra-technology handovers.

In the next section, we analyse the performance of cSHO and PRIME mechanisms in reducing handover delay and packet loss.
6.1 Introduction

In this chapter, we analyse the performance of our Uninterrupted and Proactive connection Transfer for IMS Mobility Enhancement (UPTIME) framework. The UPTIME framework aims to reduce both handover preparation delay and handover execution packet loss. These two parameters determine the effectiveness of a mobility management framework in achieving seamless mobility. In addition to handover delay and packet loss minimisation, an efficient mobility management solution must also consider other important factors such as radio resource consumption and battery power usage. In this chapter, we also study the performance of the UPTIME framework in terms of handover delay, packet loss, and the consumption of bandwidth and power.

The UPTIME framework, as a general mobility management framework, introduces two novel mechanisms to reduce handover delay and packet loss. These mechanisms include the Pre-Registration for IMS Mobility Enhancement (PRIME) scheme and the Conservative Soft Handover (cSHO) method. The PRIME mechanism, described in Chapter 4, deals with minimising handover preparation. PRIME is based on proactive handover preparation in which the target network is prepared while the Mobile Station (MS) is still connected to the current Radio Access Network (RAN). Compared with previously adopted pre-registration methods such as the one proposed by Dutta et al. in [23], the PRIME mechanism further reduces handover preparation delay by being more proactive and performing target network and IMS registration processes in the early stages of handover, long before registering with the target Point of Attachment (PoA) registration.

To evaluate the performance of the PRIME method in terms of handover preparation delay, we considered inter-technology handovers to WiMAX and LTE technologies and calculated the number of signaling messages required to establish a new
radio link to the target PoA. We demonstrated that the PRIME mechanism, under certain conditions, requires fewer signaling messages than previously suggested solutions. The result is a significant reduction of handover preparation delay, which enables the MS to transfer a multimedia session to the target RAN more quickly.

The cSHO mechanism, described in Chapter 5, reduces packet loss during the handover execution process while also minimizing resource consumption. In cSHO, a reliable radio connectivity is maintained by fast activation and deactivation of network interfaces. Since the MS continuously switches to the interface with better radio conditions, the cSHO mechanisms outperforms conventional Hard Handover (HHO) schemes in terms of the rate of packet loss. Moreover, when compared with simple Soft Handover (SHO) schemes, the cSHO method consumes fewer radio and battery power resources because it uses only one interface at any given time.

To evaluate the performance of the cSHO mechanism, we first developed an analytical model for calculating the Packet Error Rate (PER) of the cSHO, HHO, and SHO mechanisms. The analytical results confirmed that the cSHO scheme considerably outperforms the HHO method in minimizing the PER.

We then simulated inter-technology handovers to WiMAX and LTE technologies for two purposes. First, through simulation results we confirmed the accuracy of the analytical model for PER calculations. Second, we measured bandwidth and power consumption of the cSHO and compared it with that for SHO and HHO schemes. Our simulation results showed that the cSHO mechanism reduces resource consumption by nearly half of that for SHO. We used the OPNET Modeler [34] for simulation purposes.

This chapter is organised as follows. In Section 6.2 we analyse the handover delay of the PRIME mechanism and compare it with previously adopted solutions. In Section 6.3, we describe our analytical model for analysing the PER performance of cSHO, HHO, and SHO handover execution schemes. In Section 6.4 we describe the simulation environment including the UPTIME functional entities which we developed using OPNET Modeler software. Finally, Section 6.5 contains our simulation results.

6.2 Handover Delay Analysis

In this section, we focused on delay analysis for our UPTIME framework proposed in Chapter 4. The aim was to demonstrate the effectiveness of the proposed mechanisms in reducing handover delay. In particular, we showed that the proposed Pre-Registration for IMS Mobility Enhancement (PRIME) mechanism can enable the Mobile Station (MS) to achieve two goals:

1. Less service interruption by reducing the delay of new radio link preparation,
2. Faster connection to more suitable Radio Access Networks (RANs).

As explained in this section, Soft Handover (SHO) schemes may reduce service interruption for certain scenarios. However, it is not able to assist the MS to join a more suitable RAN more quickly. The PRIME mechanism, however, can achieve both goals.

6.2.1 Introduction and definitions

A complete handover process includes network discovery, handover preparation, and handover execution stages. As such, the handover delay \( D_{ho} \) can be written as:

\[
D_{ho} = D_{disc} + D_{prep} + D_{exe}
\]  

(6.1)

Where \( D_{disc} \), \( D_{pr} \), and \( D_{exe} \) are respectively network discovery, handover preparation, and handover execution delays. \( D_{disc} \) greatly varies with the number of available RANs, network selection criteria, and information collection mechanisms. In this research, we focused on improving the handover preparation delay and minimising packet loss due to handover execution. We therefore do not include \( D_{disc} \) in our delay analysis.

Figure 6.1 depicts the operation of traditional Hard Handover (HHO), Soft Handover (SHO), Pre-Registration, and PRIME mechanisms in preparing a radio link and executing a handover. In this figure, the MS uses Interface 1 (IF1); at \( t = T_{dec} \) the handover decision is made and Interface 2 (IF2) is prepared for connection to the target Point of Attachment (PoA). In the HHO and SHO scheme, the MS performs the standard initial network entry procedure of the target RAN and registers with the IP Multimedia Subsystem (IMS). The MS then executes the handover by using the standard Session Initiation Protocol (SIP) or the IMS Service Continuity and Centralization (SCC) [84] process. Since the MS does not have any records in the target network, it is required to perform the full handover procedure with a delay of \( D_{full} \).

The SHO mechanism, depicted in Figure 6.1 (b), may eliminate service interruption by simultaneously utilising the target interface for handover signaling while receiving data packets over current interface. The service interruption is removed if the current interface (IF1) remains active until the multimedia session is transferred to the target interface (IF2). In other words, the condition \( T_{dec} + D_{full} < T_{out} \) should be met where \( T_{out} \) is the time when radio outage is experienced over the current interface.

In the Pre-Registration (PR) scheme, as proposed by Dutta et al. in [23], the MS performs some of the handover preparation tasks over the current interface. The MS may then disconnect from IF1 and directly communicate with the target
RAN over IF2. The overall handover delay for the PR mechanism is $D_{PR}$. However, for a noticeably shorter period of time the MS is unable to receive data packets. Alternatively, to reduce service interruption the MS may combine the PR method with the SHO scheme and continue receiving data packets over the current interface. This is the approach that we use in the PRIME mechanism.

In the PRIME mechanism, the MS proactively registers with the Core Network (CN) and IMS in the early stages of the handover process when the target PoA is not yet selected. When the target PoA becomes available, the MS completes the handover preparation procedure and transfers the multimedia session. During this process, the MS continues receiving data packets over the current interface. The advantage of the PRIME mechanism over the PR method is that the delay associated with the Core Network Pre-Registration (CN-PR) and IMS Pre-Registration (I-PR) are removed from the overall handover preparation process which is initiated after detection of the target PoA. The PRIME mechanism also outperforms the traditional SHO method for the following reasons:

- Since the PRIME preparation delay is less than that of the traditional SHO scheme, link outage before handover completion is less likely.

- Under the PRIME scheme, the MS is more quickly connected to the more suitable RAN. This means in hotspot scenarios, the MS is able to benefit from higher data rates of the target PoA as soon as it becomes available.

The latter point is important in the Always Best Connected (ABC) philosophy of modern heterogeneous wireless networks [13]. In traditional mobile networks,
uninterrupted radio connectivity is important and handover delay is defined as the
time interval between the reception of the last data packet over the current interface
and the arrival of the first data packet over the target interface [67, 113, 176].
In ABC scenarios, the optimality of the radio connection is also important and
handover delay is defined as the time duration in which the MS discovers a better
PoA until it connects to this PoA [110]. The PRIME mechanism improves both
network connectivity and service optimality.

In the following subsections, we consider typical inter-technology WiMAX and
LTE handovers and calculate the handover delay by counting the number of signaling
messages required for each handover scheme.

### 6.2.2 Handover Delay of the HHO Scheme

The overall handover delay of an unoptimised HHO scheme includes the standard
target network entry, IMS registration, and handover execution by using IMS SCC
mechanisms

\[
D_{\text{HHO}} = D_{\text{TR}} + D_{\text{IMS}} + D_{\text{SCC}}
\]

(6.2)

where \(D_{\text{TR}}\), \(D_{\text{IMS}}\), and \(D_{\text{SCC}}\) are respectively the target RAN registration, IMS
registration, and SCC handover execution delays. These delay components can
be expressed in terms of the number of exchanged signaling messages and their
associated delay. The following parameters affect the signaling delay:

- Number of messages transmitted by the MS \((N_{tx})\), received by the MS \((N_{rx})\),
  and exchanged between network elements \((N_{ne})\). These parameters can be
determined by examining the standard procedures of different wireless tech-
nologies. In some cases, standards do not mandate the exact signaling flow to
provide network implementation flexibility. In these cases, we consider typical
implementations.

- Radio transmission delay \((T_{tx})\) and reception delay \((T_{rx})\). These factors depend
  on transmission rate, message size, and channel access delay. \(T_{tx}\) and \(T_{rx}\)
  may be evaluated using detailed simulation of wireless networks or through
  mathematical analysis with simplifying assumptions.

- Packet latency between network entities \((T_{ne})\). We consider \(T_{ne}\) to emphasise
  that packet transmission delay over backbone may include lower delays when
  compared with a radio link with limited radio capacity and varying data rate.
  A common approach to model \(T_{ne}\) is to assume a fixed intra-domain latency
  and a larger value for inter-domain communication [177].
Message processing delay ($T_p$). Depending on the processing power and queuing delay at each network node, transmission of signaling packets may be further delayed. For modeling purposes, $T_p$ can be modeled by a fixed value or by simulating buffers and CPU power.

Based on these parameters, the handover delay of the HHO scheme can be written as

$$D_{HHO} = \sum_{i=1}^{N_{tx}} T_{tx}^i + \sum_{i=1}^{N_{rx}} T_{rx}^i + \sum_{i=1}^{N_{ne}} (T_{ne}^i + T_p^i)$$  \hspace{1cm} (6.3)

where the index $i$ refers to the $i$th handover message. Equation 6.3 can be applied to standard signaling flows for each wireless technology. Without loss of generality, and for clarification purposes, we provide two example for HHO in the WiMAX and LTE technology. We apply Equation 6.3 to the standard WiMAX and LTE network entry procedures in a typical scenario. Tables 6.1 and 6.2 contain our results. Refer to Sections 2.2 and 2.3 respectively for WiMAX and LTE procedures.

As can be seen in Tables 6.1 and 6.2 for handovers to WiMAX and LTE technologies, the required total number of messages ($N_{msg}$) is 82 and 69 respectively. In these examples, our assumption is that each network functional entity is implemented in a single network node. The actual number of messages may increase in scenarios where: 1) several nodes are used to implement one entity, 2) multiple instances of one functional entity intercept signaling messages, or 3) proxy nodes are implemented. For example, in the network authentication process, we assume that there is no proxy server between the MS and the authentication server in the home network. We present the simulation results for handover delays of these scenarios in Section 6.5.

We note that the HHO scheme usually involves the exchange of many control messages. As such, the handover delay for radio connectivity may be beyond the tolerance of real-time multimedia applications. The HHO scheme also has disadvantages regarding radio connection optimality; when a more suitable RAN becomes available, a large number of messages need to be sent before the MS is able to receive a better connection.

### 6.2.3 Handover Delay of Pre-Registration scheme

The Pre-Registration (PR) method, as described by Dutta et al. in [23], reduces service interruption by exchanging some of the handover signaling over the current interface. In the PR method, the current RAN functions as a proxy to the target network and forwards the handover messages to the Authentication and Configuration Agent (ACA), located in the target CN. Since the messages travel through the
Table 6.1: Number of control messages for WiMAX handover

<table>
<thead>
<tr>
<th>Process</th>
<th>$N_{tx}$</th>
<th>$N_{rx}$</th>
<th>$N_{ne}$</th>
<th>$N_{msg}$</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handover preparation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ranging/capability exchange</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Authentication</td>
<td>2</td>
<td>3</td>
<td>11</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>SA handshake</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>BS register</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>DHCP</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>assuming no exchange of DISCOVER and OFFER messages</td>
</tr>
<tr>
<td>Bearer setup</td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>11</td>
<td>assuming the existence of dynamic PCC framework</td>
</tr>
<tr>
<td>IMS registration</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>14</td>
<td>considering message exchange with I-CSCF</td>
</tr>
<tr>
<td>Overall handover preparation</td>
<td><strong>12</strong></td>
<td><strong>13</strong></td>
<td><strong>33</strong></td>
<td><strong>58</strong></td>
<td></td>
</tr>
<tr>
<td>SCC handover execution</td>
<td>3</td>
<td>3</td>
<td>18</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td><strong>15</strong></td>
<td><strong>16</strong></td>
<td><strong>51</strong></td>
<td><strong>82</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.2: Number of control message for LTE handover

<table>
<thead>
<tr>
<th>Process</th>
<th>$N_{tx}$</th>
<th>$N_{rx}$</th>
<th>$N_{ne}$</th>
<th>$N_{msg}$</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handover preparation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial random access</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Attach request</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Authentication</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>10</td>
<td>assuming no identity check</td>
</tr>
<tr>
<td>Location update</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>assuming no old location information in network</td>
</tr>
<tr>
<td>Bearer setup, IP address allocation</td>
<td>2</td>
<td>1</td>
<td>13</td>
<td>16</td>
<td>assuming combined initial attachment, IP address allocation, and bearer setup procedure</td>
</tr>
<tr>
<td>IMS registration</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>14</td>
<td>considering message exchange with I-CSCF</td>
</tr>
<tr>
<td>Overall handover preparation</td>
<td><strong>8</strong></td>
<td><strong>6</strong></td>
<td><strong>31</strong></td>
<td><strong>45</strong></td>
<td></td>
</tr>
<tr>
<td>SCC handover execution</td>
<td>3</td>
<td>3</td>
<td>18</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td><strong>11</strong></td>
<td><strong>9</strong></td>
<td><strong>49</strong></td>
<td><strong>69</strong></td>
<td></td>
</tr>
</tbody>
</table>
current RAN instead of being sent directly to the target network, an additional delay is imposed on the overall procedure. However, the MS continues to receive data packets until the completion of the PR procedure. Let us assume that $D_{PR}$ is the time required to perform the pre-registration tasks. If these tasks can be performed before an outage occurs over the current radio interface (that is, $D_{PR} < D_{out}$), the handover delay is optimised to $D_{opt}$ where $D_{opt} < D_{full}$.

We can apply the formula 6.3 to the PR scheme. Table 6.3 and 6.4 summarise the number of exchanged messages for the PR scheme in WiMAX and LTE handover scenarios. As can be seen from the table, the overall number of messages ($N_{msg}$) increases. This is because messages are relayed over the current RAN to the target RAN. However, during the transmission of the majority of these messages, the MS is able to receive data packets provided that the current interface remains active. In the example of handover to WiMAX, 46 messages are sent over the current interface. Only 40 messages are sent over the target link when the MS is disconnected from the current RAN, compared with 82 messages sent in standard HHO.

The above delay optimisation is achieved only if the current interface remains active until the pre-registration process is completed. Otherwise, the MS is required to perform any incomplete handover preparation tasks when it joins the target network. This results in a less than optimum handover preparation delay (denoted by $D'_{opt}$). As such, the handover delay of the PR scheme can be written as
Table 6.4: PR mechanism and the number of control messages for LTE handover

<table>
<thead>
<tr>
<th>Process</th>
<th>(N_{tx})</th>
<th>(N_{rx})</th>
<th>(N_{ne})</th>
<th>(N_{msg})</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-authentication</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Location update</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>assuming no old location information in the network</td>
</tr>
<tr>
<td>Pre-configuration REQ/RES</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Pre-configuration: bearer setup</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>IMS pre-registration</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>18</td>
<td>assuming the exchange of DISCOVER and OFFER messages</td>
</tr>
<tr>
<td>PR overall</td>
<td>5</td>
<td>5</td>
<td>40</td>
<td>50</td>
<td>number of messages exchanged over old interface</td>
</tr>
<tr>
<td>Optimised HO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial random access</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Completion of bearer setup</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>SCC handover execution</td>
<td>3</td>
<td>3</td>
<td>18</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Optimised HO overall</td>
<td>7</td>
<td>6</td>
<td>23</td>
<td>36</td>
<td>number of messages exchanged after disconnection of old interface</td>
</tr>
<tr>
<td>Overall</td>
<td>12</td>
<td>11</td>
<td>63</td>
<td>86</td>
<td></td>
</tr>
</tbody>
</table>

\[
D_{PR} = \begin{cases} 
D_{\text{opt}} & D_{PR} < T_{\text{out}} \\
D'_{\text{opt}} & D_{PR} > T_{\text{out}} 
\end{cases} 
\]  

(6.4)

To summarise, one limitation of the PR scheme is that the MS may not have enough time to complete the pre-registration process before the current interface becomes disconnected. Another disadvantage is that the PR scheme may increase the handover delay from a connection optimality point of view; when a better RAN becomes available, more messages are required to prepare the new radio link. This is because control messages are relayed to the target RAN over the current RAN which results in additional delay.

6.2.4 Handover Delay of the SHO Mechanism

For the SHO scheme, the MS simultaneously uses two radio interfaces and performs the handover procedure while remaining connected to the current RAN. As per the HHO method, the MS performs the standard network entry procedure. As such, SHO tasks for handover preparation are identical to those of the HHO scheme. The SHO handover execution process, on the other hand, requires additional signaling for initiating packet duplication. Since packet duplication can be initiated by a simple request/response exchange of messages between the handover server and packet duplication function, this additional signaling is minimal.

The main advantage of the SHO scheme is that the MS continues packet reception while performing the handover procedure, provided the current interface
remains active until the handover process is complete. The handover delay for radio connectivity may therefore be reduced to zero if the condition \( D_{\text{full}} < T_{\text{out}} \) is met where \( D_{\text{full}} \) is the delay of the full SHO process including packet duplication signaling. Otherwise, handover delay is non-zero. That is to say

\[
D_{\text{SHO}} = \begin{cases} 
D_{\text{full}} - T_{\text{out}} & D_{\text{full}} > T_{\text{out}} \\
0 & D_{\text{full}} < T_{\text{out}} 
\end{cases} \quad (6.5)
\]

\[
D_{\text{full}} = D_{\text{TR}} + D_{\text{IMS}} + D_{\text{SCC}} + D_{\text{dup}} \quad (6.6)
\]

Since in some cases \( D_{\text{full}} \) is large, the SHO scheme may experience service interruption. Either way, the SHO scheme does not improve the handover delay from the point of view of connection optimality; when a more suitable RAN becomes available, the MS is still required to perform a full initial attachment procedure with the target network and register with the IMS network. Because of this, the user is not able to receive the better services offered by the target RAN as soon as possible.

### 6.2.5 UPTIME Handover Delay Analysis

Handover preparation in the proposed UPTIME framework is performed using the PRIME mechanism which includes Core Network Pre-Registration (CN-PR), IMS Pre-Registration (I-PR), and PoA Pre-Registration (P-PR) stages. CN-PR and I-PR are radio-independent tasks performed in the early stages of the handover when the target PoA has not yet been selected. As such, handover preparation after PoA detection at \( t = T_{\text{dec}} \) only includes the P-PR process

\[
D_{\text{prep}}(\text{PRIME}) = D_{\text{P-PR}} \quad (6.7)
\]

Handover execution in the UPTIME framework is performed using the cSHO mechanism. The cSHO mechanism includes a SIP-based handover execution procedure, signaling for packet duplication/buffering, and over-the-air activation of radio interfaces. The handover execution method is similar to the standard IMS SCC scheme and includes the same number of SIP messages. Requesting packet duplication and buffering involves the simultaneous transmission of request messages to the duplication functions in both the current and target PoAs. Interface activation requires the simultaneous transmission of a Media Independent Handover (MIH) message to the current and target networks. As such, the cSHO handover execution delay can be written as

\[
D_{\text{exe}}(\text{cSHO}) = D_{\text{SCC}} + D_{\text{dup}} + D_{\text{MIH}} \quad (6.8)
\]
Table 6.5: The number of control messages in an UPTIME handover to the WiMAX technology

<table>
<thead>
<tr>
<th>Process</th>
<th>$N_{tx}$</th>
<th>$N_{rx}$</th>
<th>$N_{ne}$</th>
<th>$N_{msg}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN-PR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentication</td>
<td>2</td>
<td>1</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>IP address allocation</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Bearer setup</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>CN-PR acknowledgment (SIP OK)</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>PR overall</td>
<td>2</td>
<td>2</td>
<td>31</td>
<td>35</td>
</tr>
<tr>
<td>I-PR</td>
<td>2</td>
<td>2</td>
<td>38</td>
<td>42</td>
</tr>
<tr>
<td>P-PR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initiation of P-PR</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>Ranging/capability exchange</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>SA handshake</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>BS register</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Bearer setup completion</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Optimised HO overall</td>
<td>8</td>
<td>8</td>
<td>13</td>
<td>29</td>
</tr>
<tr>
<td>SCC handover execution</td>
<td>3</td>
<td>3</td>
<td>18</td>
<td>24</td>
</tr>
<tr>
<td>Packet duplication</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>MIH transmission</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Overall</td>
<td>16</td>
<td>15</td>
<td>102</td>
<td>133</td>
</tr>
</tbody>
</table>

As with the SHO scheme, in the UPTIME framework the handover process (including PRIME and cSHO mechanisms) is performed while the MS is connected to the current RAN. Handover delay can be reduced to zero provided that the current interface remains active until the handover process is completed. That is to say

$$D_{\text{UPTIME}} = \begin{cases} 
D_{\text{prep (PRIME)}} + D_{\text{exe (cSHO)}} - T_{\text{out}} & D_{\text{prep (PRIME)}} + D_{\text{exe (cSHO)}} > T_{\text{out}} \\
0 & D_{\text{prep (PRIME)}} + D_{\text{exe (cSHO)}} < T_{\text{out}} 
\end{cases}$$

Tables 6.5 and 6.6 show the number of exchanged messages for a typical UPTIME handover to WiMAX and LTE technologies. As can be seen for the WiMAX case, the PRIME mechanism only requires the exchange of 29 messages to prepare a link for the target PoA. This is lower than the standard network entry process which is used in the traditional SHO scheme. For this reason, it is more likely that the PRIME mechanism can prepare a radio link before the current interface experiences radio outage.

Another advantage of the UPTIME mobility framework over the traditional SHO schemes is the reduction of handover delay from in terms of connection optimality. When a better RAN becomes available, the preparation of the radio link involves a delay of $D_{\text{P-PR}}$ which is lower than the standard handover preparation delay of $D_{\text{TR}} + D_{\text{IMS}}$.

In the next sections, we provide a simulation environment and present numerical
Table 6.6: The number of control messages in an UPTIME handover to the WiMAX technology

<table>
<thead>
<tr>
<th>Process</th>
<th>$N_{tx}$</th>
<th>$N_{rx}$</th>
<th>$N_{ne}$</th>
<th>$N_{msg}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN-PR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentication</td>
<td>2</td>
<td>1</td>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td>Location update</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Bearer setup</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>CN-PR acknowledgment (SIP OK)</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>PR overall</td>
<td>2</td>
<td>2</td>
<td>35</td>
<td>39</td>
</tr>
<tr>
<td>I-PR</td>
<td>2</td>
<td>2</td>
<td>38</td>
<td>42</td>
</tr>
<tr>
<td>P-PR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initiation of P-PR</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>Initial random access</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Completion of bearer setup</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Optimised HO overall</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>SCC handover execution</td>
<td>3</td>
<td>3</td>
<td>18</td>
<td>24</td>
</tr>
<tr>
<td>Packet duplication</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>MIH transmission</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Overall</td>
<td>12</td>
<td>10</td>
<td>102</td>
<td>124</td>
</tr>
</tbody>
</table>

results for handover delay of various schemes.

6.3 Packet Delivery Analysis

SHO schemes achieve minimum packet loss by the simultaneous use of multiple radio interfaces and the transmission of duplicated media streams. However, SHO schemes suffer from the problem of excessive resource consumption. HHO schemes, on the other hand, suffer from high packet loss during the handover process. The proposed cSHO mechanism has better packet delivery performance than HHO and lower resource consumption than SHO. In this section, we analyse and compare the Packet Error Rate (PER) of different handover schemes.

We start by defining a PER gain factor, similar to Huang et al. in [31]. We define the gain as the inverse ratio of PER of one scheme to that of the HHO scheme. The HHO scheme is selected as the base handover scheme because it exhibits a higher packet error rate than either the cSHO and SHO methods. As such, the gain of the cSHO scheme, denoted by $G_{cSHO}$, can be presented as

$$G_{cSHO} = \frac{\text{PER of HHO}}{\text{PER of cSHO}}$$ (6.10)

Similarly, the PER gain for the SHO mechanism, denoted by $G_{SHO}$, is defined as

$$G_{SHO} = \frac{\text{PER of HHO}}{\text{PER of SHO}}$$ (6.11)

We expect the SHO scheme to have the best packet delivery performance (that
is, the lowest PER) followed by the cSHO method and then the HHO scheme. Therefore, intuitively we expect

\[ G_{\text{SHO}} > G_{c\text{SHO}} > 1 \]

To calculate the exact values for PER gains, we need to first determine the PER of each handover execution scheme. The overall PER of each scheme depends on the PER of the MS's current and target radio interfaces denoted by Interface 1 (IF1) and Interface 2 (IF2) receptively. The PER of each interface in turn is a function of the signal-to-noise ratio (SNR) received over that interface. We denote the SNR of the IF1 and IF2 by \( \gamma_1 \) and \( \gamma_2 \).

We take the following steps to determine the PER gain of the cSHO and SHO schemes:

1. A mathematical model for \( \gamma_1 \) and \( \gamma_2 \) is presented. This model provides a mathematically friendly probability distribution function for \( \gamma_1 \) and \( \gamma_2 \) stochastic processes.

2. To simplify the PER calculations, we represent \( \gamma_1 \) and \( \gamma_2 \) with a two-dimensional Finite State Markov Chain Model (FSMC) model. This model enables us to determine the probability of using the current and target radio interfaces during the handover process.

3. The overall average PER of each handover execution scheme is calculated from the probability of using each interface and its PER.

In the following subsections, we describe the above steps.

### 6.3.1 Radio Channel Modeling

By definition the SNR of a radio interface at time \( t \) is:

\[ \text{SNR}(t) = P_r(t) - P_n(t) \]  \hspace{1cm} (6.12)

Where the \( P_r(t) \) and \( P_n(t) \) are received signal power and the noise power, both expressed in a logarithmic scale. As discussed in Section 5.4, the received power is affected by fast fading due to Doppler spread and multipath reception, shadow fading and pathloss phenomena. At each given short period of time, the latter two factors determine the average signal power while multipath fading causes fluctuations around this average [104]. For cSHO, fast fading is smoothed out from the SNR samples by applying an averaging filter. As such, the received signal power can be expressed as
\[ P_r(t) = P_0 + 10 \alpha \log_{10}(\frac{d_0}{d}) + \chi \] 

(6.13)

Where \( P_0 + \chi \) is the measured received power at the distance \( d_0 \) from the transmitter, \( \chi \) is the shadow fading random process, \( \alpha \) is the pathloss exponent, and \( d \) is the current distance from the transmitter [7]. As recommended by the ITU-T in [170], the shadow fading process can be modeled by a log-normal random process. That is to say, the \( \chi \) random process, presented in logarithmic scale, has a Gaussian or normal distribution with the mean of 0 and standard deviation of \( \sigma_s \):

\[ \chi \sim N(0, \sigma_s^2) \] 

(6.14)

The spatial auto-correlation function of the random process \( \chi \) is modeled by

\[ R(\Delta d) = \exp(-\frac{\Delta d}{d_c} \ln(2)) \] 

(6.15)

where \( \Delta d \) is the distance between two locations where auto-correlation is calculated and \( d_c \) is the decorrelation length which depends on the environment. A typical value of 20 metres is considered for \( d_c \) in the vehicular environments [170].

From the formulae 6.12 and 6.13, one can conclude that \( \gamma_1 \) (the SNR of the IF1) is a Gaussian random process with a mean of \( P_0 + 10 \alpha \log_{10}(\frac{d_0}{d_1}) - P_{n1} \) and the standard deviation of \( \sigma_s \)

\[ \gamma_1 \sim N(P_0 + 10 \alpha_1 \log_{10}(\frac{d_0}{d_1}) - P_{n1}, \sigma_s) \] 

(6.16)

where \( d_1 \) is the distance between the MS and the current PoA, \( P_{n1} \) is the noise and interference power of IF1, \( \alpha_1 \) is the pathloss exponent for IF1, and \( P_0 \) is the received power at the reference distance.

Likewise, \( \gamma_2 \) is written as

\[ \gamma_2 \sim N(P_0 + 10 \alpha_2 \log_{10}(\frac{d_0}{d_2}) - P_{n2}, \sigma_s) \] 

(6.17)

where \( d_2 \) is the distance between the MS and the target PoA and \( P_{n2} \) is the noise and interference power of IF2. Since the target RAN may use a different radio frequency and antenna height, the pathloss exponent of IF1 (\( \alpha_2 \)) and the reference received power may be different than that of IF1. However, the important point to note is that the received SNR over radio interfaces can be modeled by a Gaussian random process with a mean determined by the pathloss and a variance determined by shadow fading. The spatial correlation functions of \( \gamma_1 \) and \( \gamma_2 \) are calculated using the formula 6.15. Later in this section, we use this information to derive the probability of using an interface during the handover process.
6.3.2 Finite State Markov Chain Model

The probability of using a particular radio interface depends on the probability of the received SNRs of IF1 and IF2 having certain values. For example, in the Hard Handover (HHO) scheme the MS performs a handover if the SNR of the current interface is less than a lower threshold and the SNR of the target interface is above a higher threshold. To determine the probability of using each radio interface, we use a Finite State Markov Chain (FSMC) to model the received SNRs (and $\gamma_2$). Modeling shadow fading with a FSMC is a mathematically simplifying approach also used by the authors in [178, 179, 147, 180].

A FSMC is defined by $\{S, \pi, T\}$ where $S$ is a state space, $\pi$ is a vector consisting of steady state occupancy probabilities, and the matrix $T$ contains inter-state transition probabilities [181]. To use the FSMC for modeling the received SNR, the SNR is quantised into finite number of levels. Each level indicates a specific state of the radio channel [182]. The channel is said to be in a particular state if the SNR has a value that belongs to the corresponding level. As the SNR fluctuates, the radio channel transits into different states of the FSMC model. The probability of being in each state (the steady state probability, $\pi$) and the probability of SNR changing to a particular value (the transition probability) can be determined from the Probability Density Function (PDF) of the SNR.

We first model the SNR of IF1 ($\gamma_1$). To create the FSMC model, we need to determine three components: 1) the state space, 2) the steady-state occupancy, and 3) the state transition probability matrix. We denote a state space with $\{a_i\}$, which corresponds to a set of quantisation levels denoted by $\{\alpha_i\}$. IF1 is said to be in the state $a_i$ of the FSMC model if $\gamma_1 \in (\alpha_{i-1}, \alpha_i)$. Since we are only interested in PER calculations for handover schemes, as shown in Figure 6.2, we consider four quantisation levels. The thresholds for quantisation levels are the outage SNR of IF1 ($\gamma_{\text{out1}}$), the Lower Threshold for HHO scheme ($T_{L1}$), and the Higher Threshold ($T_{H1}$). $\gamma_{\text{out1}}$ is an SNR value at which all data packets are lost over radio interface and the communication link between the MS and the PoA is disconnected. With these thresholds, the state space $\{a_i\}$ is defined as:

$$
\begin{align*}
\begin{cases}
  a = a_1 & \gamma_1 < \gamma_{\text{out1}} \\
  a = a_2 & \gamma_{\text{out1}} < \gamma_1 < T_{L1} \\
  a = a_3 & T_{L1} < \gamma_1 < T_{H1} \\
  a = a_4 & T_{H1} < \gamma_1
\end{cases}
\end{align*}
$$

The steady-state occupancy probability of states $\{a_i\}$, denoted by $\pi_{a_i}$, indicates the probability of IF1 being in a particular state. By definition:
where \( f_{\gamma_1}(\gamma) \) is the probability density function of the random process \( \gamma_1 \) and the set \([\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4] = [-\infty, \gamma_{out1}, T_{L1}, T_{H1}, +\infty] \).

The last step of creating the FSMC model is determining the state transition probabilities. We note that the SNR of IF1 is sampled at regular intervals and compared with the SNR of IF2 for making a handover decision. We denote two adjacent samples of \( \gamma_1 \) taken at \( t = k \) and \( t = k + 1 \) by \( \gamma_k^1 \) and \( \gamma_{k+1}^1 \). The problem of determining the probability of transition between state \( a_i \) to \( a_j \) is equivalent of calculating the probability of \( \gamma_{k+1}^1 \) falling into state \( a_j \) provided that \( \gamma_k^1 \) belongs to \( a_i \). That is to say:

\[
T_{a_ia_j} = \frac{Pr\{\gamma_{k+1}^1 \in a_j \mid \gamma_k^1 \in a_i\}}{Pr\{\gamma_k^1 \in a_i\}} = \frac{Pr\{\gamma_{k+1}^1 \in a_j, \gamma_k^1 \in a_i\}}{Pr\{\gamma_k^1 \in a_i\}} = \frac{1}{\pi_{a_i}} \left( \int_{\alpha_i}^{\alpha_j} \int_{\alpha_i}^{\alpha_j} f(\gamma_k^1, \gamma_{k+1}^1) d\gamma_{k+1}^1 d\gamma_k^1 \right)
\] (6.20)

In this equation, \( f(\gamma_k^1, \gamma_{k+1}^1) \) is the joint probability density function of two jointly normal random variables, \( \gamma_k^1 \) and \( \gamma_{k+1}^1 \). This function required the correlation of random variables \( \gamma_k^1 \) and \( \gamma_{k+1}^1 \) which can be determined by using the formula 6.15 and calculating the spatial distance between the consecutive SNR samples. Assuming that the MS is moving with a constant speed of \( \nu \) m/s, the spatial distance of samples
\[ \gamma_1^k \text{ and } \gamma_1^{k+1} \]
\[ \Delta d = \nu T_d \quad (6.21) \]

where \( T_d \) is the handover decision interval. By inserting \( 6.21 \) into \( 6.15 \), the correlation of \( \gamma_1^k \) and \( \gamma_1^{k+1} \) is obtained. Then \( f(\gamma_1^k, \gamma_1^{k+1}) \) and consequently the transition probabilities, are calculated.

Equations \( 6.18 \) to \( 6.20 \) complete the creation of a FSMC model for IF1. With a similar process, IF2 is modeled by a FSMC identified by states \( \{b_i\} \), steady-state occupancy probabilities of \( \{\pi_{b_j}\} \), and transition probabilities of \( \{T_{b_ib_j}\} \).

The average PER for handover schemes depends on the SNRs of both interfaces. Therefore, based on the four-state FSMC of each interface, we construct a two-dimensional FSMC presented in Figure 6.3. In this figure, the state \( S_{ij} \) represents a situation where IF1 is in state \( a_i \) and IF2 is in state \( b_j \):

\[ (\text{IF1, IF2}) \in S_{ij} \Leftrightarrow (\text{IF1} \in a_i) \land (\text{IF2} \in b_j) \quad (6.22) \]

The occupancy probability of being in \( S_{ij} \), denoted by \( \pi_{ij} \), is calculated from the independence of IF1 and IF2 states:

\[ \pi_{ij} = \pi_{ai} \pi_{bj} \quad (6.23) \]

Similarly, the probability of transition from \( S_{ij} \) to \( S_{kl} \), denoted by \( T_{S_{ij}S_{kl}} \), can be found

\[ T_{S_{ij}S_{kl}} = T_{a_i a_k} T_{b_j b_l} \quad (6.24) \]

Equations \( 6.22 \) to \( 6.24 \) determine our two-dimensional FSMC model for joint IF1/IF2 channel states. Using this joint FSMC model, in the subsequent sections, we calculate the average PER for each handover scheme.
6.3.3 Packet Error Rate (PER) Of HHO

The overall Packet Error Rate (PER) of the HHO scheme depends on the probability of being in a specific state and using each radio interface. We define $\psi_1^{ij}$ and $\psi_2^{ij}$ as the probability of using IF1 and IF2 when the two-dimensional FSMC model is in state $S_{ij}$:

$$\psi_1^{ij} = Pr\{\text{using IF1}|S = S_{ij}\}$$  \hspace{1cm} (6.25)

$$\psi_2^{ij} = Pr\{\text{using IF2}|S = S_{ij}\}$$  \hspace{1cm} (6.26)

The average PER of the HHO scheme in state $S_{ij}$, denoted by $PER_{HHO}^{ij}$, can be found as:

$$PER_{HHO}^{ij} = \psi_1^{ij} \bar{p}_1(\gamma_1|S = S_{ij}) + \psi_2^{ij} \bar{p}_2(\gamma_2|S = S_{ij})$$  \hspace{1cm} (6.27)

Where $\bar{p}_1(\gamma_1|S = S_{ij})$ and $\bar{p}_2(\gamma_2|S = S_{ij})$ refer to the average PERs of IF1 and IF2 respectively.

The probability of using IF1 and IF2 (i.e., $\psi_1^{ij}$ and $\psi_2^{ij}$) can be found by noting that at $t = k + 1$ the MS uses specific interface only if, 1) at $t = k$ the same interface was being used and no handover occurred, 2) at $t = k + 1$ another interface was used and a handover was performed. In the HHO scheme, a change of interfaces occurs only in states $S_{14}$, $S_{24}$, $S_{41}$, and $S_{42}$ where one of the interfaces has a strong signal and the other one receives a poor SNR. As such, we define handover states set, denoted by $H$, as a subset of Markov states where a handover is possible:

$$H = \{S_{14}, S_{24}, S_{41}, S_{42}\}$$  \hspace{1cm} (6.28)

The Non-handover states set, denoted by $N$ is defined as:

$$N = \{S_{ij} \notin H\} \hspace{1cm} 1 \leq i \leq 4 \hspace{1cm} 1 \leq j \leq 4$$  \hspace{1cm} (6.29)

The probability of using IF1 and IF2, $\psi_1^{ij}$ and $\psi_2^{ij}$, are calculated using the following formulae (see Section 6.3.6 for the proof):
\[ \pi_{ij} \psi_{ij}^{1} = \sum_{S_{mn} \in N} \pi_{mn} \psi_{ij}^{mn} T_{S_{mn}S_{ij}} + \frac{\pi_{41}}{\pi_{ij}} T_{S_{41}S_{ij}} + \frac{\pi_{42}}{\pi_{ij}} T_{S_{42}S_{ij}} \] (6.30)

\[ \pi_{ij} \psi_{ij}^{2} = \sum_{S_{mn} \in N} \pi_{mn} \psi_{ij}^{mn} T_{S_{mn}S_{ij}} + \frac{\pi_{14}}{\pi_{ij}} T_{S_{14}S_{ij}} + \frac{\pi_{24}}{\pi_{ij}} T_{S_{24}S_{ij}} \] (6.31)

(6.30) and (6.31) give us a set of equations which are sufficient to find \( \psi_{ij}^{1} \) and \( \psi_{ij}^{2} \) for all states. Consequently, using formula 6.27, we are able to calculate \( \text{PER}_{\text{HHO}}^{ij} \) for all states. Finally, the average PER for the HHO scheme, \( \text{PER}_{\text{HHO}} \), is calculated by averaging the PER over all \( S_{ij} \) states:

\[ \text{PER}_{\text{HHO}} = \sum_{S_{ij}} \text{PER}_{\text{HHO}}^{ij} \times \pi_{ij} \] (6.32)

### 6.3.4 Packet Error Rate (PER) of cSHO

In this section we calculate the PER of the Conservative Soft Handover (cSHO) scheme. In the cSHO method, the MS does not consider the Lower Threshold \( (T_L) \) and the Higher Threshold \( (T_H) \) which are used in the Hard Handover (HHO) schemes. Packet duplication and buffering mechanisms of the cSHO method enables the MS to quickly changed its radio interfaces with creating minimum signaling traffic load. As such, the ping-pong effect is not a problematic issue and the SNR hysteresis approach is not needed. Instead, in the cSHO method, the interface with the stronger signal is always used which results in lower PER.

In the FSMC model, in state \( S_{ij} \), IF1 is used if \( i > j \) and IF2 is used if \( j > i \). In cases where \( i = j \) (i.e. \( S_{11}, S_{22}, S_{33} \)) any of the radio interfaces may be in use. In these states, we cannot determine the probability of using a specific interface with the approach presented in the previous section. Instead, we use a different method.

We note that the cSHO mechanism selects the interface with the lowest PER. Therefore, in an ideal scenario where the MS immediately starts using a radio interface as soon as it becomes the better interface, the cSHO PER at any given time during the handover process is the minimum of the PERs of each interface. Therefore in this ideal scenario, the instantaneous cSHO PER can be written as:

\[ p(\gamma_1, \gamma_2) = \min(p_1(\gamma_1), p_2(\gamma_2)) \] (6.33)
Where $p_1(\gamma_1)$ and $p_2(\gamma_2)$ are the instantaneous PERs of IF1 and IF2 and $p(\gamma_1, \gamma_2)$ is the instantaneous PER for the cSHO mechanisms. This equation is valid for all values of $\gamma_1$ and $\gamma_2$ and therefore can be applied to all states of the FSMC model. The average cSHO PER for the ideal scenario can be calculated by averaging the random process $p(\gamma_1, \gamma_2)$.

The realistic scenario for the cSHO mechanism differs from the ideal scenario in two ways. First, the interface with the lowest PER is selected at discreet $T_d$ intervals. However, radio conditions of network interfaces may change before the next handover decision interval. Second, activation of the selected network interface involves some signaling delay. Nevertheless, since $T_d$ is much smaller than the channel coherence time due to shadow fading, and signaling delay for interface activation is minimal, Equation 6.33 provides a good approximation for instantaneous cSHO PER which can be written as:

$$p(\gamma_1, \gamma_2) \geq \min(p_1(\gamma_1), p_2(\gamma_2))$$

(6.34)

Where we approach the lower bound as the duration handover decision intervals ($T_d$) and signaling delay decrease. We continue this section with finding the lower bound for the PER of the cSHO mechanism. In Section 6.5 where we present the simulation results, we examine the accuracy of this estimation.

The average PER of the cSHO scheme, denoted by $PE_{R_{cSHO}}$ can be found from the Complementary Cumulative Distribution Function (CCDF$^a$) of the stationary random process $p(\gamma_1, \gamma_2)$, denoted by $F_c(p)$:

$$E\{p_ij(\gamma_1, \gamma_2)\} = \int_0^1 F_c(p)dp$$

(6.35)

$F_c(p)$ can be found as follows:

$$F_c(p) = Pr\{\min(p_1, p_2) > p\}$$

$$= Pr\{p_1 > p\} \times Pr\{p_2 > p\}$$

$$= F_{\gamma_1}(\gamma_1) \times F_{\gamma_2}(\gamma_2)$$

(6.36)

Where $F_{\gamma_1}(\gamma)$ and $F_{\gamma_2}(\gamma)$ are the Cumulative Distribution Function (CDF) of SNR samples of IF1 and IF2 respectively. The numbers $\gamma_1$ and $\gamma_2$ satisfy the condition of $p_1(\gamma_1) = p_2(\gamma_2) = p$ and can be found from the PER-SNR curves of Modulation and Coding Schemes (MCS) used by each interface.
6.3.5 Packet Error Rate (PER) of SHO

In the SHO mechanism, the MS receives duplicated independent streams of data packets over radio interfaces. An erroneous data packet is received only if both interfaces receive the packet with error. Since the SNR of radio interfaces are independent, their PERs are also independent. Therefore, the average PER for a SHO scheme can be simply found by:

\[
\text{PER}_{\text{SHO}}^{ij} = p_1(\gamma_1)p_2(\gamma_2)
\]  

(6.37)

6.3.6 Proof for Given Formulae

In this appendix we prove formulae (6.30) and 6.31. We first derive Equation (6.30) which is used for determining the probability of using IF1 at each given Finite State Markov Chain (FSMC) model. With a similar approach, Equation 6.31 for IF2 can be derived.

The probability of being in state \(S_{ij}\) and using IF1 at time \(k+1\), denoted by \(IF_1[k+1]\), can be found as follows:

\[
\Pr\{IF_1[K+1], S[k+1] = S_{ij}\} = \\
\sum_{S_{mn} \in N} \Pr\{IF_1[K], S[k] = S_{mn}\} T_{S_{mn}S_{ij}} \\
+ \Pr\{S[k] = S_{41}\} T_{S_{41}S_{ij}} \\
+ \Pr\{S[k] = S_{42}\} T_{S_{42}S_{ij}}
\]  

(6.38)

In other words, the probability of using IF1 being in state \(S_{ij}\) can be found by noting that: 1) The MS continues using IF1 if no handover happens in the previous state, and 2) A handover to IF1 only happens in states \(S_{41}\) and \(S_{42}\).

(6.38) can be rewritten as

\[
\Pr\{IF_1[K+1]|S[k+1] = S_{ij}\}\Pr\{S[k] = S_{ij}\} = \\
\sum_{S_{mn} \in N} [\Pr\{IF_1[K]|S[k] = S_{mn}\} \times \Pr\{S[k] = S_{mn}\} T_{S_{mn}S_{ij}}] \\
+ \Pr\{S[k] = S_{41}\} T_{S_{41}S_{ij}} + \Pr\{S[k] = S_{42}\} T_{S_{42}S_{ij}}
\]  

(6.39)

In the steady state condition, the probability of using IF1 and being in a specific state does not depend on time. Therefore, we can eliminate the time index and
write Equation 6.39 as

\[
\Pr\{\text{IF1} \mid S = S_{ij}\} \Pr\{S = S_{ij}\} = \\
\sum_{S_{mn}\in N} \Pr\{\text{IF1} \mid S = S_{mn}\} \Pr\{S = S_{mn}\} T_{S_{mn}S_{ij}} \\
+ \Pr\{S = S_{41}\} T_{S_{41}S_{ij}} + \Pr\{S = S_{42}\} T_{S_{42}S_{ij}} \quad (6.40)
\]

Using the definition of \(\psi_{i}^{ij}\), we have

\[
\pi_{ij} \psi_{i}^{ij} = \sum_{S_{mn}\in N} \pi_{mn} \psi_{1}^{mn} T_{S_{mn}S_{ij}} \\
+ \frac{\pi_{41}}{\pi_{ij}} T_{S_{41}S_{ij}} + \frac{\pi_{42}}{\pi_{ij}} T_{S_{42}S_{ij}} \quad (6.41)
\]

Using same approach we can derive (6.31) for IF2.

### 6.4 Simulation Environment

We used OPNET MODELER® [34] to confirm the accuracy of the Packet Error Rate (PER) analysis presented in Section 6.3. We also provided numerical results for the signaling delay of the UPTIME framework and its alternative including the Pre-Registration (PR) scheme and traditional Hard Handover (HHO) and Soft Handover (SHO) methods.

In this section, we discuss the simulation environment including the OPNET Modeler interface, developed models for UPTIME functional entities, the radio channel model, the mobility pattern and network traffic.

In the simulation process, we used the ten-step systematic approach to network performance evaluation presented by Jain in [183] and used by Ransbottom in [184]. In particular, the scope of the simulation was restricted to inter-technology handovers; mobility within a Radio Access Network (RAN) was not simulated. The performance metrics selected were PER, signaling delay, power consumption and bandwidth usage. Other parameters such as processing power and buffer usage were not studied. We designed simulation scenarios and network traffic load to exclude the effect of other factors such as link failure and network topology change on simulation results. Finally, we verified and interpreted simulation results by comparing them with analytical results and expected behaviors.
6.4.1 OPNET Modeler Overview

OPNET Modeler takes a top-down approach to network simulation. In the top level, the Network Editor is used to build the network topology and insert network nodes and links at the right locations. Network-wide parameters such as traffic loads and application usage profiles are set at this level. Figure 6.4 shows a screen shot of the network editor and our simulation scenario.

In the next level and in the Node Editor the internal modules of each node are defined. The modules of a node are commonly connected to each other by packet streams which model inter-module information exchange. The last level, called the process level, is where processes inside modules are defined. A process model contains the codes and variables of each module which ultimately determine the behavior of the encompassing node.

In OPNET, each process is a state machine, driven by simulation events, also called interrupts. The process is initialised in its first state and then waits for simulation events. Each time an interrupt occurs, the process is invoked, executes the C codes of the current state, transits to the next state, and returns control to the main simulation process. This event-driven and combined graphical-textual presentation of process models provides an easy interface for developing new network nodes.

In order to model radio transmission and wireless channels, OPNET Modeler uses Radio Transceiver Pipelines. Pipelines are text-based files containing C codes. For each packet transmission between the radio transmitter and receiver, pipeline codes are invoked and executed one after other to simulate phenomena such as transmitter antenna gain, propagation delay, receiver antenna gain, received power, background and interference noises, bit error rate, error allocation, and error correction. This sequential modeling of different wireless effects has proved to be accurate for our purpose.

6.4.2 Network Level Simulation Architecture

Figure 6.5 shows the network architecture used in our simulation scenarios. We deployed an LTE, a WiMAX, and an IMS network. For each of the WiMAX and LTE networks only one radio cell was deployed because intra-RAN mobility is out of the scope of the simulation scenarios. The WiMAX Base Station (BS) served 10 fixed WiMAX-only stations which were used to create network traffic that might affect packet transmission latency. Network elements of the WiMAX Core Network (CN) such as the Authentication Accounting and Authorization (AAA) server, DHCP server, the Policy and Charging Rule Function (PCRF) were modeled by a single server which supported custom applications to simulate WiMAX signaling flow. The
Figure 6.4: OPNET top-down approach in network simulation
effect of packet latency within the WiMAX CN was modeled in defining custom applications.

Similarly, the LTE eNB served 10 LTE-only fixed stations and connected to the Evolved Packet Core (EPC) node which is part of the OPNET LTE model. We implemented the LTE CN server node to simulate signaling between the MS and network entities within the core network such as the Mobility Management Entity (MME), PCRF, and the Signaling Gateway (S-GW).

The IMS network contained two server nodes, the Serving Call Session Control Function (S-CSCF) and the Handover Server (HOS) which was a regular IMS application server. Other nodes were not explicitly implemented, but the delay of communicating through them was considered in defining custom applications.

All stations in the network ran the same real-time multimedia application which included both voice and video components. They communicated with the Corresponding Host (CH) which was modeled by a 10 Mbps Ethernet station connected to an Internet Service Provider (ISP) network. We did not explicitly model the ISP network; rather we considered it as a network cloud imposing additional delay on the communication with the CH.
Table 6.7: Network level parameters of simulation scenarios

<table>
<thead>
<tr>
<th>Parameter</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of fixed LTE stations</td>
<td>10</td>
</tr>
<tr>
<td>number of fixed WiMAX stations</td>
<td>10</td>
</tr>
<tr>
<td>number of dual-mode WiMAX/LTE mobile stations</td>
<td>10</td>
</tr>
<tr>
<td>intra-network packet latency</td>
<td>10 msec</td>
</tr>
<tr>
<td>inter-network packet latency</td>
<td>30 msec</td>
</tr>
<tr>
<td>distance between PoAs</td>
<td>3 km</td>
</tr>
<tr>
<td>approximate radius of radio cells</td>
<td>1700 km</td>
</tr>
</tbody>
</table>

IP clouds were added to the simulation network environments to model the packet latency between various nodes. Depending on the location of network entities and system architecture, the real-life packet latency may greatly vary which impacts the performance of handover preparation and execution schemes. We approximated this latency variation by using two values for intra-network and inter-network communications. All entities within a network (such as LTE or WiMAX) could communicate with each other with a fixed delay of $T_{\text{net}}$ while entities located in different networks had a packet latency of $T_{\text{inter}}$. This approach was also used by Melnyk et al. in [177]. We set the packet latency attribute of IP clouds to 10 milliseconds which resulted in 10 and 30 milliseconds for $T_{\text{net}}$ and $T_{\text{inter}}$ parameters respectively. In this model, we assumed that the IMS network, the LTE and WiMAX networks, and the ISP network were all independent. We also assumed that the P-CSCF was located in core networks.

Table 6.7 summarises the network-level parameters used in simulation scenarios.

### 6.4.3 UPTIME Node Models

To simulate the UPTIME framework, we created new OPNET models including:

1. a dual-mode LTE/WiMAX mobile terminal model which contained our proposed Handover Adaptation Layer (HAL),
2. a PoA Handover Agent (PHA) model for the WiMAX BS and LTE eNB, and
3. a Media Duplication Function (MDF) model.

The Handover Server (HOS) was implemented using a general OPNET server model. In the following, we describe the models developed in more detail.

The dual-mode MS model was created by modifying an LTE MS and adding a WiMAX interface. The MS model developed, Figure 6.6, contained a HAL module which was implemented above the IP layer for easier implementation. All received data packets were delivered to the HAL module. This module either buffered packets or sent them to the IP module for transmission. The HAL also indicated to which radio interfaces packets should be directed; it also simulated the *Virtual Socket* feature.
Figure 6.6: Dual-mode LTE/WiMAX MS model with the HAL module
The HAL received remote interrupts from the radio receivers and collected SNR samples and records power consumption statistics. This process was used for simulating monitoring radio interfaces using MIH primitives.

The operation of the HAL module was configured in the OPNET Network Editor through a node called the Handover Configuration. The HAL could operate in three modes:

1. HHO mode. In this mode the HAL performed a hard handover based on the collected SNR samples and a hysteresis.

2. SHO. The HAL duplicated data packets received from higher layers and forwarded them to both interfaces.

3. cSHO. The HAL ran our proposed cSHO algorithms and exchanged MIH messages with the old and target PoAs.

The parameters of each mode were provided through the attributes of the Handover Configuration node, as seen in Figure 6.7. For example, in the cSHO mode, the duration of the handover decision period and coefficients of the exponential weighted moving average filter could be set.

The process inside the HAL module, depicted in Figure 6.8, simulated the block diagram of the HAL described in Section 4.3.4. The first two states of this process initialised the HAL by defining variables and reading configuration parameters. The
HAL process then transitioned to the *Idle* state and waited for network events such as interrupts from radio interfaces reporting SNR values or the arrival of an IP packet. The HAL process then proceeded to a specific state and performed any required actions. For example, the HAL periodically transitioned to the *Decision* state and decided if packet duplication should be initiated or the best interface selected. If the transmission of an MIH message was required, the HAL process proceeded to the *MIH* state which simulated the MIH Interface block and constructed the MIH message.

In addition to the HAL, we developed the PHA module and process. The PHA module was created using an OPNET queue process with number of subqueues matching the maximum number of dual-mode MSs. In the WiMAX BS node, this module was located above the IP module and in the LTE eNB was located above the GPRS Tunneling Protocol (GTP) module.

The PHA process, depicted in Figure 6.9, included a list of mobile terminals using the cSHO mechanism. Based on this list, which was frequently updated by the HOS, and the status of radio link between the MS and the encompassing PoA, the PHA process decided whether received IP packets should be buffered or transmitted to their mobile destination.

The MIH message exchange between the PHA and the MS was explicitly modeled. For simulation efficiency and for verification purposes, it was possible to use remote interrupts instead of MIH messages.
Finally, we developed the MDF model by modifying an Ethernet switch model. We added a new module to the switch to duplicate the data packets towards MSs. This module held a list of MSs which were in cSHO or SHO handover states and duplicated packet streams accordingly. The MDF was instructed by remote interrupts from the HOS to add or remove an MS from the list. The packet latency between the HOS and the MDF nodes was modeled by incorporating fixed delays in remote intervals.

6.4.4 MS mobility pattern

In simulation scenarios, the mobility pattern of MSs was designed to achieve two goals:

1. Maximise handover instances. To emphasise the effect of handovers on experienced PER, we designed a mobility pattern which resulted in a relatively high number of handover instances. This goal was achieved when MSs moved straight out of the coverage of the current cell towards the target PoA. We also noted that handover was more frequent at the edge of radio cells. Therefore, in simulation scenarios, the MS trajectory was mostly limited to the area in middle of the line between the BS and the eNB.
2. Generate randomness. We designed the mobility pattern to avoid unnecessary synchronisation of movements among different MSs. Two parameters injected randomness into movements: first a random destination point in the target cell, followed by a random waiting time at the destination.

Figure 6.10 shows the mobility pattern of a sample MS. To start its movement, the MS picked a random destination point in the mobility box in the target cell. The MS moved along a straight line (simulating typical movement in a street) towards the destination. At the destination, the MS waited for a random period of time before it started to move back to the previous cell. The same procedure was repeated for the entire duration of the simulation.

The dimensions and position of mobility boxes were selected by taking into account the coverage area of the PoAs. The aim was to initially have the MSs within the coverage area of the associated PoA. The speed of movement was selected to simulate the mobility of a vehicular environment. Table 6.8 lists the parameters which were used in the simulation scenarios.

### 6.4.5 Shadow fading model

Shadow fading has an important effect on the PER performance of different handover schemes, especially on the cSHO mechanism. To model shadow fading accurately, we used the ITU recommendations presented in [170]. Shadow fading was modeled...
with a random process with the PDF and auto-correlation function presented in
formulae 6.14 and 6.15.

Shadow fading can be simulated both in two-dimensional (2D) and one-dimensional
(1-D) fashions. The 2-D model, for example as presented by Cai et al. in [185], is the
more accurate model for a general mobility pattern and system-wide performance
evaluation. The 1-D model, on the other hand, is easier to implement and involves
fewer computations [186]. The 1-D model is accurate in cases where the MS moves
along a trajectory with a straight lines. As such, we used the 1-D model in our
simulation scenarios in which the MS traveled between the LTE and WiMAX cells
in straight lines.

Two important parameters of the shadow fading model are the standard deviation
($\sigma_s$) and the correlation distance ($d_c$). $\sigma_s$ has been reported to be in the range of 5 to
12 dB with a typical value of 8 dB for macrocellular applications [186]. We used this
value in our simulations. $d_c$ which roughly indicates the maximum distance between
correlated locations and in vehicular environments is around 20 metres [170], a value
we used in the simulation scenarios.

To generate a stochastic process with the desired characteristics, we first gener-
ated a vector of independent, identically distributed (i.i.d.) random variables using
the normal distribution.

6.4.6 LTE and WiMAX Model Assumptions and Parameters

6.4.6.1 WiMAX

The OPNET model for WiMAX has the following features which affected the accu-

racy of our simulation results:

- QoS modeling through service class definition and mapping higher layer traffic
to classes.
- Scheduling services. All five WiMAX scheduling services including UGS,
eRTPS, RTPS, NRTPS, Best Effort are supported.
- Bandwidth request. Both contention-based and piggyback bandwidth request
are supported.
- Automatic Retransmission Request (ARQ) and Hybrid ARQ (HARQ) are sup-
ported.
- Adaptive Modulation and Coding (AMC). Modulation and coding schemes
from QPSK with coding rate of 1/2 to 64QAM with coding rate of 3/4 are
supported. The change of modulating and coding can be adaptive (based on
received SNR) or fixed.
- **Power Savings Mode.** The MS can be in normal operation, active/listening when the MS is not transmitting, and sleep mode.

- **Packet loss modeling.** Packet loss can be modeled using pipelines or simply by assigning a packet loss ratio.

- **PHY layer overhead are modeled and impairments such as interference, pathloss, and fast fading are simulated.**

- **Open loop power control for uplink channels is supported to reduce the power consumption of MSs and radio interference.**

The WiMAX model in the OPNET Modeler v16 has some limitations. For example, only Time Division Duplex (TDD) operation mode is supported and Frequency Division Duplex (FDD) is not included in the model. Also, network assisted ranging and periodic ranging features are not supported.

### 6.4.6.2 LTE

The LTE model in OPNET has the following features:

- **EPS bearer definition and service data flow classification is modeled.** The MS can have up to eight EPS bearers. The default bearer creation is combined with an initial attach process.

- **LTE MAC features such as random access, scheduling request, buffer status reporting, and channel dependent scheduling are supported.**

- **PHY channels with bandwidth from 1.4 MHz to 20 MHz are supported.**

- **PHY frame structure is modeled with a frame length of 10 milliseconds and a subframe length of 1 millisecond.**

- **PHY measurement of SNR values is supported to allow the adaptation in modulation and coding.**

- **Transmission power is modeled according to number of OFDM carriers.** The measurement of power consumption is also supported.

- **PHY impairments such as pathloss, multipath fading, and interference are modeled.**

The LTE model has some limitations: The size of control messages is estimated rather than calculated exactly. The effect of LTE header compression is simulated by random reduction of message sizes. Features such as TDD operation and handovers within LTE network are not supported.
Table 6.9: Attributes of the voice stream in the multimedia application

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>VoIP codec</td>
<td>the speech coder</td>
<td>G.723.1</td>
</tr>
<tr>
<td>codec rate</td>
<td>output bit rate of speech coder</td>
<td>5.3 kbps</td>
</tr>
<tr>
<td>frame size</td>
<td>duration of voice frames in codec</td>
<td>30 msec</td>
</tr>
<tr>
<td>frames per packet</td>
<td>number of VoIP frames in data packets</td>
<td>1</td>
</tr>
<tr>
<td>silence suppression</td>
<td>silence suppression for saving bandwidth</td>
<td>none</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>frame interarrival time</td>
<td>the approximate radius of radio cells</td>
<td>50 msec</td>
</tr>
<tr>
<td>frame size</td>
<td>distance between two radio</td>
<td>900 bytes</td>
</tr>
<tr>
<td>codec rate</td>
<td></td>
<td>144 kbps</td>
</tr>
</tbody>
</table>

Table 6.10: Attributes of the video component of the multimedia application

6.4.7 Network Traffic Model

OPNET Modeler provides tools to model network traffic accurately. The software again uses a top-down approach in implementing data applications. In the top level, a Profile Configuration node is used to define the behavior of users in using different applications. When a Traffic Profile is assigned to an MS, it determines what applications the user utilises, when applications start and finish, and if applications are run simultaneously or one after each other (serially).

In the next level, using the Application Configuration node, applications are defined in term of tasks that are accomplished. For example, a web application may contain tasks such as signing in, downloading some data, and signing out. In the application level we also determine what transport layer, TCP or UDP, is used by the application. In the last level, the Task Configuration node is used to define the exact messages exchanged to complete each task. In this stage, parameters such as the number of messages, message size, and processing time are determined.

We used pre-defined OPNET applications to model the multimedia application and we created custom applications to simulate the message flow of different handover schemes in different wireless technologies. We simulated a real-time multimedia session, such as a video call, between MSs and a CH. The realistic application includes two-way voice and video components. Table 6.9 and 6.10 describe the different parameters of voice and video streams.

6.4.7.1 Network and Handover Control Signaling

To simulate signaling messages used in network entry, IMS registration, and handover control, we defined custom applications. The attributes of these custom applications include the number of messages and generated packets, the size of messages,
Table 6.11: Attributes of the video component of the multimedia application

<table>
<thead>
<tr>
<th>SIP message</th>
<th>Size (bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>INVITE</td>
<td>810</td>
</tr>
<tr>
<td>REGISTER</td>
<td>225</td>
</tr>
<tr>
<td>UPDATE</td>
<td>260</td>
</tr>
<tr>
<td>Unauthorised</td>
<td>100</td>
</tr>
<tr>
<td>OK</td>
<td>100</td>
</tr>
<tr>
<td>ACK</td>
<td>60</td>
</tr>
</tbody>
</table>

and processing delays. In defining a custom application, we could determine the source and destination of each transaction and all intermediate nodes resulting in accurate modeling of the signaling flow.

The size of exchanged messages in different procedures varies according to the wireless technology and implementation options. Therefore, in defining our custom application, this parameter could only be approximated. For SIP messages which are text-based, we used values reported by Munir et al. in [119] and listed in table 6.11. The values presented are for SIP message compressed by the SigComp algorithm, RFC 3320 [62], mandatory for MSs. The assumption is that SigComp reduces the size of the initial SIP message (the first Register and Invite messages) by 55% and subsequent messages by 80%.

The size of the MIH message exchange of the MS and PoAs to activate or deactivate radio links was assumed to be 20 bytes, large enough to account for the MIH header and the limited information in the body of the message. For other control messages, we assumed the message size to be 100 bytes.

6.5 Simulation Results

In this section, we present the numerical results for the PRIME mechanism delay analysis, the accuracy of the analytical model for Packet Error Rate (PER) evaluation of the cSHO mechanism, and the packet delivery performance of various handover schemes. At the end, we studied the bandwidth usage and power consumption of the cSHO mechanism and compared it to that of the HHO and SHO schemes.

6.5.1 PRIME Handover Preparation Delay

Table 6.12 presents numerical results for the handover preparation delay for different mechanisms. To obtain these results, we considered typical handovers to WiMAX and LTE technologies. Four handover preparation schemes including traditional SHO and HHO methods, the Pre-Registration (PR) scheme, and cSHO mechanisms
were studied.

In the traditional SHO and HHO schemes, the MS performed a normal network entry and IMS registration. In the WiMAX technology the MS required 881 milliseconds to associate with the PoA, perform the authentication process, receive an IP address and establish radio and network bearers. This delay was 609 milliseconds in the LTE technology mainly because IP allocation and dedicated bearer establishment could be combined with the initial network attachment process. The normal IMS registration process took 327 and 345 milliseconds over the LTE and WiMAX interfaces respectively. This indicated that in our simulation the LTE link was slightly faster than the WiMAX link. The overall handover preparation delay ($D_{\text{prep}}$) was 936 and 1226 milliseconds for LTE and WiMAX technologies respectively. When the HHO scheme was used, the MS was not able to send or receive data packets during this period of time and service interruption occurred. In the SHO scheme, service interruption was avoided if the current interface remained active for around 1 to 1.2 seconds.

For the Pre-Registration (PR) method, the MS used its old interface to perform some of handover preparation tasks. For pre-registration to the target network we used the signaling flow presented by Dutta et al. in [23] with the added bearer establishment process. The pre-registration to IMS process was derived from the 3GPP specification TS 23.203 [168]. The MS required 1.2 seconds to prepare the LTE and WiMAX interfaces using the PR scheme.

For the UPTIME mechanism, however, the MS only required 279 milliseconds to prepare a radio link to an LTE eNB. This is because the core network and IMS registration processes, which take 476 and 381 milliseconds respectively, can be performed well in advance of the PoA detection. This results in 80% and 70% less handover preparation delay when compared with the PR and direct registration in the SHO scheme.

For WiMAX handovers, $D_{\text{P-PR}}$ was 472 milliseconds, 61% less than the direct registration and PR schemes. As can be seen, the performance improvement was less than that of LTE because the WiMAX technology required an extensive exchange of control messages between the BS and the MS to perform the initial ranging, capability exchange and three-way handshake for security association, security key requests, BS registration, and finally the three-way handshake for bearer establishment.

The handover execution phase using the IMS SCC procedure took 291 milliseconds for a handover to LTE which was slightly higher than the 281 milliseconds for a handover to WiMAX. This was because the SCC signaling was carried over the current radio link, which in our simulation was slightly slower than the LTE link. An additional 40 milliseconds was required for packet duplication which increased the SHO handover execution delay. The same process was used in the UPTIME.
Table 6.12: Handover preparation delay for UPTIME and pre-registration method

<table>
<thead>
<tr>
<th>Delays (seconds)</th>
<th>LTE</th>
<th>WiMAX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SHO/HHO</td>
<td>PR</td>
</tr>
<tr>
<td>Preparation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_{TR}$</td>
<td>609</td>
<td>-</td>
</tr>
<tr>
<td>$D_{MS}$</td>
<td>327</td>
<td>-</td>
</tr>
<tr>
<td>$D_{P-PR}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$D_{CN-PR}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$D_{I-PR}$</td>
<td>-</td>
<td>381</td>
</tr>
<tr>
<td>$D_{T-PR}$</td>
<td>-</td>
<td>805</td>
</tr>
<tr>
<td>$D_{prep}$</td>
<td>936</td>
<td>1166</td>
</tr>
<tr>
<td>Execution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_{SCC}$</td>
<td>291</td>
<td>291</td>
</tr>
<tr>
<td>$D_{dup}$</td>
<td>40</td>
<td>-</td>
</tr>
<tr>
<td>$D_{MIH}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$D_{exe}$</td>
<td>331</td>
<td>291</td>
</tr>
</tbody>
</table>

However, in this case packet duplication was initiated before the actual change of radio interface. The average delay of MIH packet transmission was 9 and 12 milliseconds for LTE and WiMAX technologies respectively. This was the delay required to activate one interface and deactivate the other.

6.5.2 Accuracy of Developed Analytical Model

To evaluate the accuracy of our PER analysis presented in Section 6.3, we simulated the random movement of 10 dual-mode MSs roaming freely between an LTE cell and a WiMAX cell for the duration of 3000 seconds. The long simulation duration was chosen to reduce calculation errors in our simulation scenarios which included a considerable number of random variables.

In the simulated scenario, the WiMAX interface (IF1) used 16QAM modulation with a coding rate of $3/4$ while the LTE interface (IF2) used QPSK modulation with coding rate of $8/9$ and benefited from the Hybrid ARQ (HARQ) technique. The Modulation and Coding Scheme (MCS) of WiMAX and LTE technologies were selected to be different to emphasis that the SNR of different radio interfaces cannot be directly compared. These specific MCS values were selected to enable us to use the PER-SNR curves (see Figure 6.11) reported by the authors in [7] and [8] for WiMAX and LTE respectively.

We first fixed the long-term average SNR (which includes only the pathloss effect) of IF1 and IF2 to 18 dB and 14 dB respectively. This allowed the accurate measurement of the steady-state occupancy probability of the Finite State Markov Chain (FSMC) model and determined the transition probabilities. These values should match the result obtained from the analytical model.

For the HHO scheme, a 4 dB hysteresis was considered. A smaller hysteresis would result in a ping-pong effect. The state occupancy and transition matrices of the FSMC model were calculated using (6.19) through (6.24), while (6.30) and
Figure 6.11: PER vs SNR curves for WiMAX and LTE [7] and [8]

Table 6.13: Simulation and analytical results comparison for PER

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Analytical</th>
<th>Simulation</th>
<th>Difference (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_{ij}$</td>
<td>-</td>
<td>-</td>
<td>1.7</td>
</tr>
<tr>
<td>$T_{S_{i},S_{mn}}$</td>
<td>-</td>
<td>-</td>
<td>9</td>
</tr>
<tr>
<td>IF1 usage</td>
<td>41%</td>
<td>40.5%</td>
<td>1.3</td>
</tr>
<tr>
<td>IF2 usage</td>
<td>59%</td>
<td>59.5%</td>
<td>1</td>
</tr>
<tr>
<td>HHO PER</td>
<td>4.4%</td>
<td>4.62%</td>
<td>5</td>
</tr>
<tr>
<td>cSHO PER</td>
<td>2.5%</td>
<td>2.8%</td>
<td>12</td>
</tr>
<tr>
<td>SHO PER</td>
<td>1.8%</td>
<td>1.9%</td>
<td>5.5</td>
</tr>
<tr>
<td>$G_{cSHO}$</td>
<td>1.76</td>
<td>1.65</td>
<td>6.2</td>
</tr>
<tr>
<td>$G_{SHO}$</td>
<td>2.44</td>
<td>2.43</td>
<td>0.5</td>
</tr>
</tbody>
</table>

(6.31) were used to calculate the average probability of using of IF1 and IF2. We then determined the average PER for each handover scheme using (6.27), (6.33), and (6.37). We compare the analytical model with simulation results in Table (6.13).

The average PER for cSHO is 12% higher than the lower bound presented in (6.34). This is due to the discrete sampling of interface quality. The handover decision making interval in the simulated results was set to 100 msec which resulted in this acceptable inaccuracy.

The results presented in table 6.13, are for a long-term average when the SNR of both interfaces is fixed. These results demonstrate the accuracy of the analytical model at that exact values for average SNRs. We expected that the cSHO PER would exhibit the same behavior for other values of the SNR. To evaluate this hypothesis, we repeated the calculations and simulation scenarios for different values of the average SNR on IF1 and plotted the results collected in Figure 6.12 and Figure
6.13. Similar results were observed when the average SNR on IF2 was varied. These results confirmed that the PER performance of the cSHO mechanism was higher than that of the HHO method and close to the PER of the SHO scheme. They also validated the accuracy of the analytical model in estimating the average PER of the different schemes.

In the next section, we evaluated the PER performance of various handover schemes in more realistic scenarios where the mobile terminals were free to move and communicate with the CH.

6.5.3 Packet Delivery Evaluation

To present a realistic simulation scenario, we considered a scenario where the MSs moved freely between LTE and WiMAX base stations located 3000 metres apart; the SNR of the interfaces varied according to the distance to the PoAs. We set the maximum transmission power of the base stations to obtain similar coverage for LTE and WiMAX with some overlap in the middle.

Figure 6.14 shows the number of received VoIP packets of an MS located 1100 metres from the LTE base station which moves towards the WiMAX BS at t=120 seconds. The MS stops at t=160 seconds. It can be seen that receiving duplicated packets in the SHO scheme significantly improved the service quality as only 19 packets were lost. These occasional disruptions occurred in areas where there was minimal LTE or WiMAX coverage. When using cSHO, 39 packets were lost but the service interruption period was short enough for a seamless handover. For HHO, 109 packets were lost and interruptions of up to 1.5 seconds were observed.
Figure 6.13: The gain of SHO and HHO methods

Figure 6.14: Received VoIP packets for different handover schemes
Depending on the location of the MS and its movement, the MS may have observed a different packet loss than that presented in Figure 6.14. To have a better assessment of the cSHO performance, we simulated a more extensive scenario where 10 dual-interface MSs randomly moved between the LTE and WiMAX cells for over 10 minutes. Since MSs are moving at the edge of WiMAX and LTE cells, there was always a possibility of handover. We also added 10 WiMAX-only and 10 LTE-only stations to inject extra background traffic in both networks.

Figure 6.15 plots the PER for different handover schemes for the duration of the simulation. We only considered packet loss after $t=200$ seconds to allow all MSs to start their movement to random locations. As observed, the average PER for HHO was approximately 8.7 percent while the comparative values for cSHO and SHO were 4.1 percent and 2.8 percent respectively. This resulted in decreases in the observed PER for HHO of 53% for cSHO and 68% for SHO.

The length of packet loss bursts is also an important performance criterion that is an indication of the length of service interruption during handover instances. Figure 6.16 shows the Cumulative Distribution Function (CDF) for the length of packet loss bursts for VoIP and video streams. In 90% of cases, the HHO had bursty losses with a length of less than 30 packets by comparison with 17 packets for cSHO and 12 packets for SHO. These were equivalent of burst losses of the length of 560 milliseconds, 310 milliseconds and 230 milliseconds for HHO, cSHO and SHO respectively.

From these figures, we can say that cSHO performs better than HHO, but not as well as SHO. However, the advantage of cSHO over SHO is its lower power usage...
Figure 6.16: CDF for length of packet loss bursts

and bandwidth consumption. Figure 6.17 shows the energy that dual-interface MSs use for radio transmission on their LTE and WiMAX interfaces. As expected, the energy used during cSHO is marginally more than that used during HHO because for cSHO, only one stream of media packets is sent over the air interface at any time. By contrast, the energy usage of 140.7J for SHO is considerably higher than the energy usage during HHO (79.1J) and cSHO (85.4J). This results in energy savings of 39% for radio transmission using cSHO when compared with SHO.

The overall bandwidth usage of different handover schemes is plotted in Figure 6.18 as a percentage of the combined LTE and WiMAX capacity. We observe that bandwidth usage of cSHO is marginally greater than HHO at 4.4% and 4.1% respectively, while the requirements of 8.5% for SHO indicates that cSHO requires 48% less radio resources when compared with SHO. This indicates that the results observed regarding energy savings are due primarily to saving of radio resources.
Figure 6.17: The radio transmission energy usage of all MSs

Figure 6.18: The bandwidth usage of different schemes
Chapter 7

Conclusion

7.1 Research Summary

Seamless mobility is one of the important requirements of Next Generation Mobile Networks (NGMN). It enables operators to integrate new radio technologies into their existing infrastructure and improve network capacity and capabilities. The integration of different wireless technologies results in a heterogeneous network environment.

A heterogeneous wireless access network consists of several Radio Access Networks (RANs) which may use different radio technologies and exhibit different characteristics and capabilities. The Mobile Station (MS) searches among surrounding RANs, which may have overlapping or overlapping radio coverage, and selects the most suitable RAN taking into consideration parameters such as service cost, security, Quality of Service (QoS) capabilities, and reliability. As the user moves, network selection parameters vary and better RANs may become available. The MS constantly monitors candidate RANs and performs inter-technology handovers to keep the user Always Best Connected (ABC) [13].

The ABC scenario is viable if users do not perceive service interruption during inter-RAN handovers. However, inter-technology handovers in IP-based wireless networks are very time-consuming. To perform an inter-technology handover, the MS must first prepare the new radio link. The required tasks to achieve this include target RAN discovery, time and frequency synchronisation with the target Point of Attachment (PoA), mutual authentication between the MS and the target Core Network (CN), the receipt of IP address configurations, and the creation of the network bearers to ensure consistent QoS across the current and target RANs. These tasks contribute considerably to the overall handover delay. For example, a typical mutual authentication process takes about one second [50]. As a result, the handover delay is well beyond the tolerance of real-time multimedia applications.
The handover process is even more time-consuming in NGMN where the service control stratum is independent of the underlying transport stratum. The service stratum of NGMN, which is usually based on the IP Multimedia Subsystem (IMS) framework, authenticates subscribers separately and establishes secure signaling connections with them. This results in additional handover delay.

The extensive signaling required for handover preparation and execution combined with rapidly varying radio channels often causes intolerable packet loss. As the current radio channel degrades rapidly, there is sometimes insufficient time to prepare the new radio link and migrate the connection to the target RAN. Furthermore, when the MS joins the target RAN, the new radio connection may fluctuate because radio links are often not reliable at cell borders where handovers are commonly executed. However, to avoid frequent handovers (the ping-pong effect) which create signaling traffic and may increase packet loss, the MS is not allowed to re-join the old RAN immediately. For that reason, high Packet Error Rate (PER) is common in inter-technology handovers.

In this research, we focused on the provisioning of seamless real-time multimedia services over heterogeneous wireless networks. We identified some shortcomings of solutions for seamless mobility management previously adopted. In particular, we noted that some of the previously proposed mobility management frameworks are not well suited to NGMN requirements including independence of the service stratum from the underlying transport stratum, separation of signaling and media functions, strong security, and consistent QoS management.

Moreover, previously adopted solutions for minimising handover preparation delay, such as the pre-registration schemes [23], are not optimal in reducing handover delay and mitigating service interruption. The pre-registration schemes were proposed to allow the MS to prepare a radio connection to the target Point of Attachment (PoA) while being connected to the current PoA. However, in the conventional pre-registration schemes, all link preparation tasks are performed after the selection of the target PoA. The delay of the pre-registration method can be reduced with a better NGMN-compliant scheme which performs the radio-independent handover preparation tasks prior to detection of the target PoA.

Finally, the previously proposed Soft Handover (SHO) mechanisms for the reduction of packet loss make excessive use of radio resources. This limits the application of SHO schemes in heterogeneous wireless networks where cell capacity is limited, especially at cell borders. SHO methods also drain battery power as they require the transmission of duplicated media streams to and from the MS. A more resource-conservative inter-technology handover execution scheme is required to allow the smooth transfer of real-time multimedia applications between the RANs of a NGMN heterogeneous networks.
We noted that the provision of seamless real-time multimedia requires the accomplishment of two major goals. First, the handover delay should be minimised as much as possible to avoid the MS joining the target RAN before the current RAN becomes unavailable. Second, packet loss during the handover process should be avoided. These goals should be achieved while observing other requirements such as minimum resource consumption, uncompromised security, and minimum modifications to network infrastructure and protocols.

7.2 Contributions

Seamless integration of different radio technologies is one of the important requirements of Next Generation Mobile Networks (NGMN). This allows mobile operators to employ a variety of radio technologies to better meet users' demand for pervasive high quality multimedia services. We contributed to the research towards seamless provision of real-time multimedia applications over heterogeneous networks by proposing an efficient mobility management framework for NGMN.

Our proposed Uninterrupted Proactive connection Transfer for IMS Mobility Enhancement (UPTIME) framework was designed to satisfy NGMN requirements for a mobility management framework. In this framework, we also included two novel mechanisms for handover delay minimisation and packet loss mitigation. The first mechanism, called the Pre-Registration for IMS Mobility Enhancement (PRIME) method, enhances the performance of traditional pre-registration schemes and further reduces handover preparation delay. The second mechanism, called the Conservative Soft Handover (cSHO) scheme, improves the resource consumption performance of conventional SHO schemes.

In the following sections, we summarise our contributions and the benefits of the proposed mobility management framework and mechanisms.

7.2.1 UPTIME Mobility Framework

The Uninterrupted Proactive connection Transfer for IMS Mobility Enhancement (UPTIME) framework was designed specifically for NGMN environments. The UPTIME system architecture is in line with NGMN guidelines for network architecture. Most notably, the principle of separation of signaling and media is observed. Moreover, open standards such as the Session Initiation Protocol (SIP) and Media Independent Handover (MIH) protocol have been used to implement the UPTIME framework. To summarise, the UPTIME framework has the following characteristics and advantages:

- NGMN compliance. The UPTIME framework follows the NGMN design prin-
ciples and NGMN requirements for an efficient mobility management protocols as specified by ITU-T in [20] and [33]. This principles include the separation of signaling from media functions and the independence of the service stratum from the underlying transport stratum.

- IP Multimedia Subsystem (IMS) compatibility. The UPTIME framework utilises the IMS framework and mechanisms to execute inter-technology handovers. This approach minimises modifications on network infrastructure. More specifically, a Handover Server (HOS) is implemented using the standard IMS interfaces to facilitate inter-technology handovers. It is envisaged that mobile operators would offer the UPTIME mobility management service as an advanced IMS service without significantly modifying the standard IMS platform.

- Radio technology independence. The proposed UPTIME framework can be applied to different wireless technologies. We used LTE and WiMAX technologies to demonstrate the operation of the UPTIME framework. However, the proposed architecture is also suitable for alternative current and future technologies such as WiFi, LTE-Advanced, and WiMAX2.

- Using open standards. Open standards including SIP and MIH protocols have been used for signaling between various UPTIME nodes. This approach ensures that the proposed mechanisms can be implemented with fewer modifications to existing network signaling methods. We noted that both protocols are extensible meaning new features can be added without modifying the core specifications of the standards. We used this characteristic to propose new extensions to SIP and MIH protocols.

### 7.2.2 PRIME Handover Preparation Mechanism

The proposed PRIME mechanism was developed to reduce the handover preparation delay in a heterogeneous network environment. The PRIME mechanism is based on the pre-registration scheme for handover preparation. However, it improves the performance of previous pre-registration solutions, such as the one proposed by Dutta et al. in [23], by taking a novel approach.

In developing the PRIME method, we made use of a common architectural paradigm in modern wireless technologies such as LTE and WiMAX. In these technologies, the network structure is divided into the Core Network (CN) and the Radio Access Network (RAN) parts. The radio-independent tasks such as authentication and authorisation, IP address configuration, and location management are handled
Table 7.1: Number of control messages required to prepare a radio connection

<table>
<thead>
<tr>
<th></th>
<th>WiMAX</th>
<th>LTE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NO. message</td>
<td>delay (ms)</td>
</tr>
<tr>
<td>Pre-Registration</td>
<td>62</td>
<td>1240</td>
</tr>
<tr>
<td>P-PR phase of PRIME</td>
<td>29</td>
<td>472</td>
</tr>
</tbody>
</table>

by the CN section while the radio-dependent tasks are controlled by the RAN section. We used this architectural paradigm to significantly reduce the handover delay. In the PRIME method, radio-independent tasks are performed long before detection of the target Point of Attachment (PoA). When the target PoA is detected and a handover is imminent, the MS performs the remaining handover preparation tasks.

The PRIME mechanism includes the following three phases of handover preparation:

1. Core Network Pre-Registration (CN-PR). The MS selects one or more target networks and registers with them using the currently active radio interface. In this phase, the MS is authenticated in the target network and receives a new IP address.

2. IMS Pre-Registration (I-PR). The new IP address is used for an IMS registration through the target network.

3. PoA Pre-Registration (P-PR). When the target PoA becomes available and the MS decides to perform a handover, radio synchronisation to this PoA is executed. Since the MS is already authenticated in the target network and IMS, the new radio link can be established with minimum delay.

We described the PRIME operation in LTE and WiMAX technologies as two examples of candidate radio technologies for NGMN. We also presented typical signaling flows for inter-technology handovers to LTE and WiMAX. These signaling flows highlight the advantage of the PRIME mechanism over the previously adopted pre-registration schemes. As summarised in the table 7.1, the PRIME mechanism results in much less link preparation delay mainly because most of the tasks are performed prior to the selection of the target PoA.

7.2.3 cSHO Handover Execution Mechanism

The cSHO mechanism was proposed to mitigate the main drawback of conventional Soft Handover (SHO) schemes. SHO methods require the transmission of duplicated media streams from and towards mobile nodes. It was shown that such approach significantly reduces Packet Error Rate (PER) in handover periods. However, in conventional SHO schemes, excessive radio resources are consumed and the mobile
node’s battery is drained. Given the scarcity of these resources in mobile networks, we developed the cSHO mechanism to mitigate this drawback.

The cSHO mechanism enables the MS to switch quickly between its two radio interfaces in order to experience a lower PER. The solution requires duplication of data packets in the wireless network and buffering them in the current and target PoAs. During the handover period, the MS remains associated with both PoAs and at each given period of time uses the interface with the best radio conditions to send and receive data packets. When the MS decides to change its active interface, it only needs to send a direct handover message to both old and new PoAs. Since the signaling delay is minimal, the MS is able to react to rapid fluctuations in radio signal strength caused by the shadow fading effect.

It is true that when compared with the SHO scheme in which both interfaces are constantly in use, the cSHO mechanism has a higher PER. However, compared with a traditional end-to-end Hard Handover (HHO) method, cSHO has much lower PER.

The advantage of cSHO over SHO becomes evident when investigating the resource consumption of both schemes. The cSHO scheme significantly reduces the required radio resources and battery power consumption of the conventional SHO methods. Table 7.2 uses simulation results presented in Chapter 6 to summarise the cSHO performance and compare it with that of HHO and SHO schemes. It is observed that the cSHO energy and radio resource consumption is close to that of the HHO scheme while its PER is still acceptable when compared with the SHO method.

Table 7.2: Comparison of HHO, SHO, and cSHO performance and their resource consumptions

<table>
<thead>
<tr>
<th></th>
<th>PER (%)</th>
<th>cell capacity usage (%)</th>
<th>energy consumption (J)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHO</td>
<td>8.7</td>
<td>4.1</td>
<td>79.1</td>
</tr>
<tr>
<td>cSHO</td>
<td>4.1</td>
<td>4.4</td>
<td>85.4</td>
</tr>
<tr>
<td>SHO</td>
<td>2.8</td>
<td>8.5</td>
<td>140.7</td>
</tr>
</tbody>
</table>

7.3 Limitations and Future Directions

Seamless provision of multimedia services over heterogeneous networks requires an efficient mobility management framework which enables mobile operators to integrate a variety of radio technologies into their existing network infrastructure. In this thesis, we have proposed such a framework.

The UPTIME framework provides a seamless handover solution for NGMN. Service interruptions during inter-technology handovers are minimised by reducing the
handover preparation delay and minimising packet loss. With this approach, terminal mobility and inter-technology handovers become transparent to users. As such, a Mobile Stations (MS) can constantly discover new RANs and seamlessly join them to keep its user Always Best Connected (ABC).

In this research, we only focused on the mobility management for real-time multimedia applications. These applications have strict QoS requirements and may suffer from inter-technology handovers. Considering only real-time multimedia applications allowed us to design a mobility management framework which is optimised for such applications and meets their specific demands. We note that other data applications such as file transfer also require a mobility management solution. Without such a solution, the change of the serving RAN may adversely affect an application’s performance. Other mobility management mechanisms can be specifically developed for such applications. This approach results in a scenario where the MS is equipped with a mobility toolbox which contains several mobility management solutions each suitable for a particular application [154].

The disadvantage of a mobility toolbox approach is that it creates additional complexity in mobile terminals and network infrastructure. As mobile networks become more capable and end devices become smarter, such advanced mobility management scenarios becomes more viable. The work presented in this thesis can be considered as one of the major items of such a mobility toolbox. The integration of other mobility management protocols in the UPTIME framework can be regarded as a future work.

In the proposed UPTIME framework, we developed two mechanisms: the PRIME mechanism was developed for minimising the latency of the handover preparation stage and the cSHO scheme was proposed for reducing the packet loss of the handover execution stage. However, the exact processes of the network discovery stage including the exact handover decision algorithm and network selection criteria were not discussed. The performance of the UPTIME framework may be affected by the effectiveness of the network discovery algorithm and its ability to determine the optimum target network and a suitable PoA within this network. As discussed in Chapter 3.2, several predictive network selection algorithms can be found in the literature [21, 123, 26, 125, 126]. Evaluating the effect of the inclusion of such algorithms in the UPTIME framework can be considered as a future work.

The UPTIME framework was designed by assuming a common network deployment scenario in which RANs are under the administration of different network operators. These RANs often function independently and provide the MS with radio connectivity without tight cooperation with other RANs. A more advanced scenario is the joint radio resource management scheme in which various RANs of a heterogeneous network cooperate to maintain optimum radio connectivity with
mobile terminals [187]. Based on parameters such as network traffic load and instantaneous radio conditions, the MS is provided with a data transmission opportunity over one of the serving RANs. This scenario results in much more network complexity but it can provide higher network capacity and better radio connection reliability. The incorporation of an advanced radio resource management scheme and its interaction with the mobility management mechanisms is a possible future research work.

In the PRIME mechanism, we propose a fast pre-registration method which enables the MS to be authenticated in the target RAN and the IMS network before an inter-technology handover occurs. In this method, the MS uses the current RAN as a proxy and performs the full authentication procedure of the target radio technology. With this approach the delay associated with the authentication procedure can be removed from the overall handover delay. However, in some cases where the target RAN cannot be detected in the early stages, the MS might not have enough time to perform the authentication process. In this case, a more complex scenario where user authentication is performed by an ABC service provider can be considered [13]. The ABC service provider is a virtual access network operator which has a subscription agreement with users. This operator can authenticate users and ensures that the user is authorised to access the services of surrounding RANs. Compared with the PRIME mechanism in which RANs authenticate users independently, this scenario presents additional complexity and requires tighter cooperation between network operators. However, it may result in less handover preparation delay. The delay reduction of this scenario can be investigated in future work.

In analysing the performance of the cSHO mechanism, we have only considered the involvement of the current and target PoAs in serving the MS. However, the cSHO mechanism can be applied to more than two PoAs. In a more general case, data packets can be replicated in the network and sent to several PoAs which belong to different RANs. In this case, since the MS maintains several independent radio links, it may experience less service outage. The disadvantage is that more signaling is required to prepare several radio links and activate or deactivate them. Furthermore, more backbone capacity is required for the transmission of replicated packets between associated PoAs. However, as the capacity of backbone links increases, this multi-PoA cSHO mechanism may become more viable and its performance gain should be studied.

Devising and deploying the above techniques will result in a highly efficient mobility management framework which will enable mobile operators to achieve NGMN great promises including the seamless integration of various technologies and more reliable radio connections. This thesis has made a contribution to that work.
Bibliography


List of publications

