A Framework for Secure Mobility in Wireless Overlay Networks

by

Hejun Chen
A Framework for Secure Mobility in Wireless Overlay Networks

by

Hejun Chen

Dissertation

submitted in fulfillment
of the requirements
for the degree

Magister Technologiae

in

Information Technology

in the

Faculty of Engineering, the Built Environment and Information Technology

of the

Nelson Mandela Metropolitan University

Promoter: Prof. Reinhardt A. Botha

January 2006
Declaration

I, Hejun Chen, hereby declare that:

• The work in this dissertation is my own work.

• All sources used or referred to have been documented and recognized.

• This dissertation has not previously been submitted in full or partial fulfillment of the requirements for an equivalent or higher qualification at any other recognized educational institute.

__________________________

Hejun Chen
DECLARATION
Abstract

Various wireless networks are widely deployed worldwide. Current technologies employed in these networks vary widely in terms of bandwidths, latencies, frequencies, and media access methods. Most existing wireless network technologies can be divided into two categories: those that provide a low-bandwidth service over a wide geographic area, for example UMTS, and those that provide a high bandwidth service over a narrow geographic area, for example 802.11. Although it would be desirable to provide a high-bandwidth service over a wide coverage region to mobile users all the time, no single wireless network technology simultaneously satisfies these requirements. Wireless Overlay Networks, a hierarchical structure of wireless personal area, local area, and wide area data networks, is considered as an efficient and scalable way to solve this problem.

Due to the wide deployment of UMTS and 802.11 WLAN, this study attempts to combine them to implement the concept of Wireless Overlay Networks. Furthermore, the information transmitted over this Wireless Overlay Networks is protected in terms of authentication, integrity and confidentiality. To achieve this goal, this study aims to combine GPRS, Mobile IP and IPSec to propose a framework for secure mobility in Wireless Overlay Networks.

The framework is developed in three steps:

- Firstly, this study addresses the problem of combining GPRS and Mobile IP, so that GPRS users are provided with Mobile IP service. This results in presenting a uniform Mobile IP interface to peers regardless of whether mobile users use UMTS or 802.11 WLAN.

- Secondly, this study discovers the existing problem when combining Mobile IP and IPSec, and proposes a Dual Home Agent Architecture to achieve secure mobility.
Finally, based on the output of the previous two steps, a complete framework is proposed, which achieves secure mobility in Wireless Overlay Networks, specifically, in UMTS and 802.11 WLAN.

The framework also implements seamless handover when mobile users switch between UMTS and 802.11. This results in UMTS and 802.11 WLAN looking like a single network when participating in this framework, and presents seamless and secure mobility.
Acknowledgements

This work would not have been possible alone. I would like to thank the following people for their contribution to this dissertation:

- My parents, for their love, support and motivation.
- My promoter, Professor Reinhardt Botha, for his guidance, patience, support and supervision.
- Mrs Adéle Botha, for her support and help proofreading the academic paper and dissertation, especially considering English is my second language.
- My colleagues and friends at the Nelson Mandela Metropolitan University, for their assistance.
- My future wife, who always motivates me with a happy future, even though I do not know who and where she is at this point in time.
- The Nelson Mandela Metropolitan University, for support and financial assistance.
- All others who helped me in any way and who are, due to limited space, not mentioned here.
Contents

Declaration ......................................................... i

Abstract ................................................................ iii

Acknowledgements .............................................. v

I Background ......................................................... 1

1 Introduction ....................................................... 3
   1.1 Problem Statement ......................................... 3
   1.2 Background ................................................ 4
   1.3 Scenario ..................................................... 9
      1.3.1 Handover .............................................. 11
      1.3.2 Routing ............................................... 12
      1.3.3 Security .............................................. 13
      1.3.4 Mobility .............................................. 14
   1.4 Goal and Scope ............................................. 15
   1.5 Objectives .................................................. 17
   1.6 Methodology ............................................... 17
   1.7 Related Work .............................................. 18
   1.8 Overview of This Study .................................. 20

2 GPRS — A Key Milestone towards 3G .................. 23
   2.1 Introduction to 3G ......................................... 24
      2.1.1 Overview of 3G Standards ......................... 24
      2.1.2 UMTS Network Architecture ..................... 26
      2.1.3 Interfaces and Protocols .......................... 31
      2.1.4 Addresses and Identifiers ........................ 33

vii
2.1.5 Handover ................................................. 37  
2.2 GPRS Overview ............................................. 40  
  2.2.1 Conceptual View of GPRS ................................. 40  
  2.2.2 GPRS System Architecture ............................... 42  
  2.2.3 Interfaces and Protocols ............................... 44  
2.3 GPRS Data Routing and Address Conversion ................. 45  
  2.3.1 Scenario ................................................ 46  
  2.3.2 Identifiers ............................................. 46  
  2.3.3 Data Transfer from MS ................................. 48  
  2.3.4 Data Transfer to MS .................................... 48  
  2.3.5 Summary ............................................... 49  
2.4 GPRS Procedures ............................................ 49  
  2.4.1 Attach Procedure ...................................... 49  
  2.4.2 Detach Procedure ..................................... 53  
  2.4.3 Activate PDP Context .................................. 53  
  2.4.4 Deactivate PDP Context ............................... 55  
2.5 GPRS Handover ............................................. 56  
2.6 Conclusion ............................................... 58  

3 Wireless Overlay Networks 61  
  3.1 Handover in Wireless Overlay Network .................... 62  
    3.1.1 Horizontal Handover .................................. 63  
    3.1.2 Vertical Handover .................................... 63  
    3.1.3 Comparison between Horizontal and Vertical Handover 64  
    3.1.4 Seamless Handover ................................... 65  
  3.2 Models of Wireless Overlay Network ....................... 66  
    3.2.1 Tunnelled Networks .................................. 67  
    3.2.2 Hybrid Networks ..................................... 68  
    3.2.3 Heterogeneous Networks ............................... 68  
  3.3 Integration of WLAN and 3G Cellular Network ............... 69  
    3.3.1 Tight Coupling Interworking ......................... 69  
    3.3.2 Loose Coupling Interworking ......................... 71  
  3.4 Conclusion ............................................... 72  

4 IP ......................................................... 73  
  4.1 Architectural Model ...................................... 74
## CONTENTS

### 4.2 IP Design Principles
- 4.2.1 Connectivity ........................................ 76
- 4.2.2 End-to-End Principle ............................... 77
- 4.2.3 Layering and Modularity ........................... 77
- 4.2.4 Summary ............................................ 79

### 4.3 IP Addressing and Routing ....................................... 81

### 4.4 Mapping IP Address to Physical Addresses ....................... 83
- 4.4.1 Proxy ARP ........................................ 84
- 4.4.2 Gratuitous ARP .................................... 84

### 4.5 IP Drawbacks .................................................. 84
- 4.5.1 Billing and Management ............................. 84
- 4.5.2 Quality of Service .................................. 85
- 4.5.3 Address Exhaustion .................................. 86
- 4.5.4 IP Mobility .......................................... 86
- 4.5.5 Session Management ................................. 87
- 4.5.6 Security Issues ..................................... 87
- 4.5.7 Summary ............................................ 88

### 4.6 IP Network Security ............................................. 89
- 4.6.1 Information Security ................................. 90
- 4.6.2 Attacks ............................................ 91
- 4.6.3 Firewall ............................................ 94
- 4.6.4 Virtual Private Network ......................... 98

### 4.7 Conclusion ..................................................... 99

### 5 Mobile IP .......................................................... 101
- 5.1 Overview of Mobile IP .................................. 101
  - 5.1.1 Agent Discovery .................................. 104
  - 5.1.2 Registration ...................................... 106
  - 5.1.3 Routing ........................................... 111
- 5.2 Reverse Tunnelling ....................................... 113
- 5.3 Optimization ............................................. 116
  - 5.3.1 Route Optimization ............................... 116
  - 5.3.2 Smooth Handover ................................ 118
  - 5.3.3 Hierarchical Foreign Agent Scheme ............. 120
- 5.4 A Hierarchical Approach to IP mobility ................. 120
CONTENTS

5.5 Conclusion .............................................. 123

6 IPSec .................................................. 125
6.1 Overview of IPSec ..................................... 126
  6.1.1 Security Services .................................. 126
  6.1.2 Security Association .............................. 127
6.2 IPSec Scenarios ....................................... 133
  6.2.1 Host-to-Host ..................................... 133
  6.2.2 Network-to-Network ............................. 134
  6.2.3 Host-to-Network ................................. 134
6.3 IPSec Standards ..................................... 135
6.4 AH Protocol .......................................... 136
  6.4.1 AH Protocol Format .............................. 136
  6.4.2 Connectionless Integrity and Data Origin Authentication 138
  6.4.3 Replay Protection ............................... 141
6.5 ESP Protocol .......................................... 141
  6.5.1 ESP Protocol Format .............................. 142
  6.5.2 Confidentiality ................................. 143
  6.5.3 Other Security Services ......................... 143
6.6 Key Management ..................................... 144
6.7 Evaluation with IP Design Principles ............... 145
6.8 Conclusion .......................................... 148

II The Framework ....................................... 149

7 Integration of GPRS and Mobile IP .................. 151
  7.1 Introduction of Mobile IP into GPRS ................. 152
    7.1.1 Step 1 — Offering Mobile IP Service ............ 153
    7.1.2 Step 2 — Intermediate GPRS-MIP System ........ 153
    7.1.3 Step 3 — Using Mobile IP for Intra System Mobility . 155
  7.2 Overview of the Model ............................. 156
    7.2.1 Hierarchical Handover ........................ 157
    7.2.2 Protocol Architecture ......................... 158
    7.2.3 Procedures .................................... 159
  7.3 Initial Stage ....................................... 159
CONTENTS

7.4 Routing ................................................. 162
  7.4.1 CN → MN .......................................... 162
  7.4.2 MN → CN .......................................... 163
7.5 Handover ............................................. 164
7.6 Conclusion ........................................... 165

8 Integration of IPSec and Mobile IP 167
  8.1 Background .......................................... 167
  8.2 Two Modes of Combining Mobile IP and IPSec Tunnels .... 171
    8.2.1 Mobile IP Tunnel inside IPSec Tunnel ............... 171
    8.2.2 IPSec Tunnel inside Mobile IP Tunnel ............... 173
  8.3 Five Scenarios of Deploying HA and IPSec Gateway ..... 175
    8.3.1 Home Agent Inside the Intranet behind a VPN Gateway 175
    8.3.2 Home Agent and IPSec gateway on the VPN Domain
        Border ................................................. 177
    8.3.3 Combined VPN Gateway and Home Agent ............ 178
    8.3.4 Home Agent outside the VPN Domain ............... 179
    8.3.5 Combined VPN Gateway and Home Agent on the Lo-
        cal Link .............................................. 180
    8.3.6 Evaluation of Scenarios ............................ 181
  8.4 Dual Home Agent Architecture .......................... 181
    8.4.1 Network Environment .............................. 182
    8.4.2 Registration ...................................... 183
    8.4.3 Handover .......................................... 185
    8.4.4 Routing ........................................... 186
  8.5 Conclusion ........................................... 186

9 A Complete Framework 189
  9.1 Overview of the Framework ............................ 189
    9.1.1 Address Scheme .................................... 191
    9.1.2 Protocol Stack .................................... 192
    9.1.3 Security Services .................................. 193
    9.1.4 Hierarchical Mobility ............................. 194
    9.1.5 Authentication Agreement .......................... 195
    9.1.6 Handover .......................................... 195
  9.2 Network Layer Handover ............................... 197
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.2.1</td>
<td>Handover from Inside to Outside</td>
<td>197</td>
</tr>
<tr>
<td>9.2.2</td>
<td>Handover from Outside to Inside</td>
<td>199</td>
</tr>
<tr>
<td>9.2.3</td>
<td>Smooth Handover</td>
<td>200</td>
</tr>
<tr>
<td>9.3</td>
<td>Internal Mobility</td>
<td>200</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Routing between i-CN and i-MN</td>
<td>200</td>
</tr>
<tr>
<td>9.3.2</td>
<td>Routing between i-CN and x-MN</td>
<td>202</td>
</tr>
<tr>
<td>9.4</td>
<td>External Mobility</td>
<td>205</td>
</tr>
<tr>
<td>9.4.1</td>
<td>Routing between x-CN and i-MN</td>
<td>205</td>
</tr>
<tr>
<td>9.4.2</td>
<td>Routing between x-CN and x-MN</td>
<td>207</td>
</tr>
<tr>
<td>9.5</td>
<td>Conclusion</td>
<td>208</td>
</tr>
</tbody>
</table>

### III Epilogue

#### 10 Conclusion

10.1 Revisiting the Objectives | 214
10.1.1 Effective IP Mobility | 214
10.1.2 Seamless Handover | 214
10.1.3 Offering Mobile IP Service in GPRS | 214
10.1.4 Incorporation of Mobile IP and IPSec | 215
10.2 Future Work | 215
10.3 Final Word | 216

### IV Appendices

A Academic Paper | 219

References | 235
List of Figures

1.1 Classic 3G Layer Diagram (Wisely, Eardley, & Burness, 2002, p. 27) ................................. 5
1.2 Nokia D211 Multiple-mode PCMCIA Card ................................. 7
1.3 3G Scenario ................................................................. 10
1.4 Mobility inside and outside the Intranet ........................................... 14
1.5 Internal and External Mobility ................................................. 15
1.6 Mobility Classification ...................................................... 15
1.7 Layout of Dissertation ...................................................... 21

2.1 High Level UMTS Architecture (Korhonen, 2001, p. 206) ...... 26
2.2 UMTS Network Architecture (Kaaranen, Ahtiainen, Laitinen, Naghian, & Niemi, 2001, p. 10) ................................. 27
2.3 UMTS System Hierarchy (Eberspacher, Vogel, & Bettstetter, 2001, p. 29) ................................. 28
2.4 UMTS Interfaces (Korhonen, 2001, p. 207) ................................. 32
2.5 UMTS Handover ................................................................. 38
2.6 High Level Conceptual Model of the GPRS Network (Salkintzis, 2004, p. 3-3) ................................. 41
2.7 GPRS System Architecture (Bates, 2002, p. 82) ......................... 42
2.8 GPRS Protocol Stack in UMTS (Kaaranen et al., 2001, p. 344) 44
2.9 GPRS Protocol Stack in GSM (Eberspacher et al., 2001, p. 253) 45
2.10 Protocol Stack with the Key Identifiers in the Scenario ............ 47
2.11 GPRS Attach Procedure (Salkintzis, 2004, p. 3-8) .................... 50
2.12 Message Sequence Diagram in Establishing a PDP Context (Salkintzis, 2004, p. 3-11) ..................... 54
2.13 GPRS Handover ................................................................. 56
2.14 Inter-SGSN Handover (Eberspacher et al., 2001, p. 251) ........... 57
LIST OF FIGURES

3.1 Wireless Overlay Network Structure (Stemm & Katz, 1998) . . 62
3.2 Horizontal vs Vertical Handover (Stemm & Katz, 1998) . . . 63
3.3 Seamless Handover in Wireless Overlay Network . . . . . . . 66
3.4 Architecture Models for Wireless Overlay Network (Wu & Havinga, 2001) .................................................. 67
3.5 Tight Coupling and Loose Coupling . . . . . . . . . . . . . . 69

4.1 The Internet Architecture Model (Comer, 1991, p. 57) . . . 74
4.2 The Hourglass Architecture of the Internet Protocol Stack
(Deering, 1998) ................................................................. 75
4.3 TCP/IP Internet Layering Model (Comer, 1991, p. 146) . . 78
4.4 A Comparison of the OSI and TCP/IP Models (Stallings,
2004, p. 40) ................................................................. 79
4.5 TCP/IP Concept (Stallings, 2004, p. 41) ......................... 80
4.6 Formats of Internet Protocol (IP) Datagram and Address . . 81
4.7 Concept of the ARP ..................................................... 83
4.8 Fattened Waist (Deering, 1998) ................................... 89
4.9 Packet Filtering Routers (Whitman & Mattord, 2003, p. 277) 96
4.10 Screened Host Firewall (Whitman & Mattord, 2003, p. 280) . 97
4.11 Dual-homed Host Firewall (Whitman & Mattord, 2003, p. 281) 97
4.12 Screened Subnet (DMZ) (Whitman & Mattord, 2003, P. 282) . 98

5.1 A Typical Mobile IP Network Configuration (Dimopoulou & Venieris, 2004, p. 5-3) .............................................. 103
5.2 Mobility Agent Advertisement Format (Dimopoulou & Ve-
nieris, 2004, p. 5-5) ....................................................... 105
5.3 Registration Procedure in Mobile IP .............................. 108
5.4 Registration Request Message Format (Perkins, 1996b) ...... 108
5.5 Routing in Mobile IP ................................................... 113
5.6 Reverse Tunnelling .................................................... 114
5.7 Route Optimization ................................................... 118
5.8 Smooth Handover ..................................................... 119
5.9 MIPv4 Regional Registration Architecture (Dimopoulou & Ve-
nieris, 2004, p. 5-39) .................................................. 121
5.10 Hierarchical IP Mobility ............................................. 122
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>IPSec Architecture (Thayer, Doraswamy, &amp; Glenn, 1998)</td>
<td>126</td>
</tr>
<tr>
<td>6.2</td>
<td>IPSec Topologies (Kent &amp; Atkinson, 1998c)</td>
<td>128</td>
</tr>
<tr>
<td>6.3</td>
<td>Security Association Mode</td>
<td>130</td>
</tr>
<tr>
<td>6.4</td>
<td>Transport Adjacency</td>
<td>133</td>
</tr>
<tr>
<td>6.5</td>
<td>End-to-End Security in Network-to-Network Topology</td>
<td>134</td>
</tr>
<tr>
<td>6.6</td>
<td>Road Warrior Scenario</td>
<td>135</td>
</tr>
<tr>
<td>6.7</td>
<td>Overview of IPSec Documents (Schafer, 2003, p. 199)</td>
<td>136</td>
</tr>
<tr>
<td>6.8</td>
<td>Format of AH Protocol (Kent &amp; Atkinson, 1998b)</td>
<td>137</td>
</tr>
<tr>
<td>6.9</td>
<td>IP Packet Format with AH (Kent &amp; Atkinson, 1998b)</td>
<td>140</td>
</tr>
<tr>
<td>6.10</td>
<td>Format of ESP Protocol (Kent &amp; Atkinson, 1998a)</td>
<td>142</td>
</tr>
<tr>
<td>6.11</td>
<td>IP Packet Format with ESP (Kent &amp; Atkinson, 1998a)</td>
<td>144</td>
</tr>
<tr>
<td>6.12</td>
<td>Node-to-Node and End-to-End Encryption (Canavan, 2001, p. 45)</td>
<td>147</td>
</tr>
<tr>
<td>7.1</td>
<td>First Step</td>
<td>154</td>
</tr>
<tr>
<td>7.2</td>
<td>Second Step</td>
<td>154</td>
</tr>
<tr>
<td>7.3</td>
<td>Third Step</td>
<td>155</td>
</tr>
<tr>
<td>7.4</td>
<td>Model of Integrating GPRS and Mobile IP</td>
<td>157</td>
</tr>
<tr>
<td>7.5</td>
<td>Hierarchical Handover</td>
<td>158</td>
</tr>
<tr>
<td>7.6</td>
<td>Protocol Architecture</td>
<td>158</td>
</tr>
<tr>
<td>7.7</td>
<td>Procedures</td>
<td>159</td>
</tr>
<tr>
<td>7.8</td>
<td>Initial Stage</td>
<td>161</td>
</tr>
<tr>
<td>7.9</td>
<td>Routing between CN and MN</td>
<td>163</td>
</tr>
<tr>
<td>7.10</td>
<td>Handover Sequence Diagram</td>
<td>164</td>
</tr>
<tr>
<td>8.1</td>
<td>Mobile IP Tunnel inside IPSec Tunnel</td>
<td>172</td>
</tr>
<tr>
<td>8.2</td>
<td>IPSec Tunnel inside Mobile IP Tunnel</td>
<td>173</td>
</tr>
<tr>
<td>8.3</td>
<td>Network Environment</td>
<td>175</td>
</tr>
<tr>
<td>8.4</td>
<td>Home Agent behind the VPN Gateway (Adrangi &amp; Levkowetz, 2005)</td>
<td>176</td>
</tr>
<tr>
<td>8.5</td>
<td>Home Agent and the VPN Gateway on the Border (Adrangi &amp; Levkowetz, 2005)</td>
<td>177</td>
</tr>
<tr>
<td>8.6</td>
<td>Combined VPN Gateway and Home Agent (Adrangi &amp; Levkowetz, 2005)</td>
<td>179</td>
</tr>
<tr>
<td>8.7</td>
<td>Home Agent outside the VPN Domain (Adrangi &amp; Levkowetz, 2005)</td>
<td>179</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>8.8</td>
<td>Combined VPN Gateway and Home Agent on the Local Link</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>(Adrangi &amp; Levkowetz, 2005)</td>
<td></td>
</tr>
<tr>
<td>8.9</td>
<td>Dual Home Agent Architecture (Dutta et al., 2004)</td>
<td>182</td>
</tr>
<tr>
<td>8.10</td>
<td>Mobile IP Registration in the Internal Domain</td>
<td>184</td>
</tr>
<tr>
<td>8.11</td>
<td>Interaction of Protocols for Dual Home Agent Architecture</td>
<td>185</td>
</tr>
<tr>
<td>8.12</td>
<td>Curved IPSec Tunnel When x-MN Moves to Another Foreign</td>
<td>186</td>
</tr>
<tr>
<td></td>
<td>Network</td>
<td></td>
</tr>
<tr>
<td>9.1</td>
<td>Framework</td>
<td>190</td>
</tr>
<tr>
<td>9.2</td>
<td>Address Scheme of an implementation</td>
<td>192</td>
</tr>
<tr>
<td>9.3</td>
<td>Protocol Stack</td>
<td>193</td>
</tr>
<tr>
<td>9.4</td>
<td>Mobility</td>
<td>195</td>
</tr>
<tr>
<td>9.5</td>
<td>Handover</td>
<td>196</td>
</tr>
<tr>
<td>9.6</td>
<td>Handover Scenario</td>
<td>197</td>
</tr>
<tr>
<td>9.7</td>
<td>Sequence Diagram in Handover from Inside to Outside</td>
<td>198</td>
</tr>
<tr>
<td>9.8</td>
<td>Routing between i-CN and i-MN</td>
<td>201</td>
</tr>
<tr>
<td>9.9</td>
<td>Routing between i-CN and x-MN</td>
<td>203</td>
</tr>
<tr>
<td>9.10</td>
<td>Routing between x-CN and i-MN</td>
<td>206</td>
</tr>
<tr>
<td>9.11</td>
<td>Routing between x-CN and x-MN</td>
<td>208</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Air Interface Standards of IMT2000 Family .............. 25

4.1 Security Mechanisms at Different Layers of the TCP/IP Model
(Wisely et al., 2002, p. 108) ............................... 88

6.1 Security Service Matrix ................................. 127

6.2 Use of Transport and Tunnel Mode ........................ 131
Part I

Background
Chapter 1

Introduction

No single wireless network technology simultaneously provides a low latency, high bandwidth, and wide area data service to a large number of mobile users. Wireless Overlay Networks, a hierarchical structure of wireless personal area, local area, and wide area data networks, address the problem of providing network connectivity to a large number of mobile users in an efficient and scalable way. Furthermore, the network connectivity needs to be offered with the guarantee of security not being violated, regardless of which network mobile users attach to. This, in high-level terms, defines the principle problem of this study: “How is secure mobility achieved in wireless overlay networks?”.

1.1 Problem Statement

To address the problem of secure mobility in wireless overlay networks, there are several sub-problems that need to be addressed:

- How can mobility be achieved in an effective manner regardless of whether movement is in a wide area or a narrow area?
- How can seamless handover be realized between different wireless networks?
- How can the security characteristics of information transferred over wireless networks be protected, specifically when multiple wireless networks are involved?
As these issues are rather broad, the next sections will further explore current developments in the field, as well as describe a specific scenario which will then be used to set more specific objectives in Section 1.5.

1.2 Background

It is widely recognized that the mobile cellular telecommunication systems have evolved along a path: 1G → 2G → 3G. The first generation, 1G, is based on analogue or semi-analogue technologies. Examples include the Nordic Mobile Telephone (NMT) system and the American Mobile Phone System (AMPS) (Schiller, 2003, pp. 9 – 15). The emphasis of services offered by 1G was on speech and speech-related services. 1G networks were developed with national scope only. Often the main technical requirements were agreed upon by the governmental telecommunication and domestic industry without wider publication of the specifications. 1G networks, therefore, were incompatible with each other.

With the common intention of a more global mobile communication system, various international specification bodies specified individual communication standards. Despite diversity between these standards, a common characteristic is that they are all digitized. Thus a generic name, 2G, umbrellas these standards. Two mainstream 2G standards are Code Division Multiple Access (CDMA) IS-95 and the Global System for Mobile Communication (GSM) (Eberspacher et al., 2001). The emphasis of 2G is on compatibility and international transparency; the system is regional (like European-wide) or semi-global. Besides the traditional speech service these networks are able to provide some data services and more sophisticated supplementary services.

It was at this time that the need of data services, especially the need of access to the Internet, increased. This simulated the advent of various enabling data service technologies. A generic name, 2.5G, covers all these technologies, including high-speed circuit-switched data (HSCSD), general packet radio services (GRPS), and enhanced data rates for global evolution (EDGE) (Korhonen, 2001, p. 5 – 8). Among them, GPRS is a key milestone for GSM data services, as it introduced a brand-new, packet-switched domain (Bates, 2002, pp. 52 – 53). The GPRS backbone is completely built on IP-based networks, which is more suited to data services.
However, due to the regional nature of the standardization, the concept of globalization did not succeed completely. Inter-operability between the various networks became a big problem.

The third generation, 3G, is expected to complete the globalization process of mobile communication. The main idea behind 3G is to prepare a universal infrastructure, able to carry various existing and also future services. The infrastructure should be designed to allow technology changes and evolution to the network without rendering existing services unstable. The emphasis thus is more on data services (Korhonen, 2001, p. 9).

The 3G landscape can be captured by the classic picture of a layered architecture of radio cells in Figure 1.1. There are mega-cells for satellites, macro-cells for wide-area coverage (rural areas), micro-cells for urban coverage, and pico-cells for indoor use. Typically, a mixture of public and private use occurs (Wisely et al., 2002, p. 27).

As envisioned above, 3G intends to provide a world-wide universal mobile communication system. It will offer true broadband data: video on demand, videophones, and high bandwidth games. 3G systems differ from 2G voice and text messaging services in terms of both the bandwidth and data capabilities.

With the goal of a world-wide universal mobile communication system, 3G was originally conceived as being a single world-wide standard. However,
by the time 3G was born, there were two proposed 3G systems and five air interface standards (Korhonen, 2001, p. 13). Besides the various Wireless Wide Area Network (WWAN) access technologies offered by 3G, there also exists various Wireless Local Area Network (WLAN) access technologies, which provide high-bandwidth over a narrow geographical area, relative to WWAN’s low-bandwidth over a wide geographical area (Stemm & Katz, 1998). The question needs to be asked why a single network cannot provide all services with high-bandwidth over a wide geographical area at all times, as desired in 3G’s ‘Martini’ version — ‘anytime, anyplace, anywhere’ (Wisely et al., 2002, p. 29). The following two aspects provide reasons why.

One aspect is economical. The existing 2G operators and manufacturers desire to reuse as much existing equipment, development effort, and services as possible (Wisely et al., 2002, p. 29), so that they can maximize the revenue of their investment. This is their primary concern.

The other aspect is technological. As stated by Stemm and Katz (1998), no single wireless network technology simultaneously provides a low-latency, high-bandwidth, wide-area data services to a large number of mobile users. Although technologies progress faster and faster, this statement is still true at this point in time.

Thus, currently end users cannot enjoy services anytime, anyplace, anywhere without the complication of carrying a multitude of devices in their pockets. However, if this goal is reviewed with a prudent look, it appears as if it does not necessarily require a single network. What is required is a ubiquitous service regardless of using a single network or multiple networks.

Two aspects play a role to attain an ubiquitous service. Firstly, technology. Integration technology is required to integrate all existing heterogeneous network resources in such a way that each service is delivered via the network that is most efficient for that service (Wu & Havinga, 2001). Secondly, business. A number of roaming agreements between operators are required to enable seamless roaming between heterogeneous networks in terms of authentication, authorization and accounting.

As far as technology is concerned, there are two key factors that significantly impact on the integration of heterogeneous networks. One is on the physical layer and user end while the other is on the network layer.

The user terminal should be equipped with a multiple wireless air interface
module so that it can switch to different networks easily and transparently without user interaction. The Nokia D211, shown in Figure 1.2, is this kind of multiple-mode equipment. It is a PCMCIA card which can connect to the GPRS or WLAN network according to the real time condition. A laptop equipped with this card will be able to access both GPRS and WLAN networks.

An innovative technology, Software-Defined Radio (SDR), has already emerged to enable more flexible multiple-mode equipment. SDR is defined as a radio with enough programmability, with “enough” being a purposefully loose term (Berezdivin, Breinig, & Topp, 2002). It was reported that Sandbridge Technologies in the USA has successfully implemented a complete “3G” multimedia handset design utilizing its SB3010(TM) flexible baseband processor to perform all baseband and multimedia functions in software (Glossner et al., 2005). This will dramatically stimulate the application of SDR technology in the multiple-mode terminal.

Besides the physical layer and user end factor, another key factor for the integration of heterogeneous networks is on the network layer. A protocol should exist to glue all the heterogeneous networks together. The Internet Protocol (IP), originally designed to interwork between different networks, seems to be such a protocol. The explosive growth of the Internet also shows
IP’s strong vitality. It is predicted that the key role of IP in future mobile data systems will probably be the provision of efficient and cost effective interworking between heterogeneous overlay networks (Guardini, D’Urso, & Fasano, 2000).

However, traditionally, IP-based networks have been designed with the assumption that they are fixed. Although the concept works well in a fixed environment, when it comes to mobile environments, a major problem arises because an IP address has two contradictory semantics (Perkins, 1998). On the one hand, the IP address serves as an identifier of the mobile terminal for the upper protocols — this should be the same regardless of the sub-network it attaches to. On the other hand, the IP address also serves as a locator of the mobile terminal for routing the datagram to its destination — this requires change to reflect the sub-network the mobile terminal has moved to. A solution, Mobile IP (Perkins, 1996b), has been proposed by the Internet Engineering Task Force (IETF) to address this problem. The essential idea is that each Mobile Node is allocated two IP addresses:

- Home Address (HoA), which is permanently assigned to a Mobile Node, and acts as its identifier, and
- Care-of Address (CoA), which is temporarily assigned to a Mobile Node, and acts as its locator reflecting its actual location.

This leads to the introduction of two new entities into the Internet infrastructure:

- Home Agent (HA), which resides on a Mobile Node’s home network. The Home Agent tunnels packets destined for the Mobile Node’s home address to its Care-of Address when it is away from the home network, and maintains current location information for the Mobile Node.

- Foreign Agent (FA), which resides on a Mobile Node’s visited network, that is, the foreign network. The Foreign Agent provides routing services to the Mobile Node while registered. The Foreign Agent detunnels and delivers packets to the Mobile Node that were tunnelled by the Mobile Node’s Home Agent.
1.3. SCENARIO

From an information security perspective, referring to the NSTISSC Security Model (Whitman & Mattord, 2003, pp. 10 – 15), the aforementioned solution only addresses the problem partially, that is, it ensures availability only. This calls for a security mechanism in terms of authentication, confidentiality, and integrity characteristics of the user data. The need is emphasized when considering that roaming data will be delivered between different administrative domains which lack sufficient mutual trust.

For the purpose of protecting the authenticity, confidentiality, and integrity of the data, the IETF has proposed a security architecture for the Internet, IPSec (Kent & Atkinson, 1998c). Like IP, it was originally designed for fixed networks, and therefore lacks mobility support. At a glance, a combination of Mobile IP and IPSec seems to be a possible solution to mobility and security support in the integration of heterogeneous networks in the 3G era (Rantala, 2004). However, the research community has indicated that there is a big challenge in combining Mobile IP and IPSec (Adrangi & Levkowetz, 2005).

1.3 Scenario

Starting with the classic 3G layer diagram in Figure 1.1, and delving further into the network details, a more detailed scenario diagram can be derived, as illustrated in Figure 1.3. All further discussions in this study will be based on this scenario.

Figure 1.3 adopts some notations from Vaarala and Klovning (2005), and extends its usage based on the same representation.

From a horizontal perspective, the scenario is divided into two separate administrative domains, an untrusted external domain (it is often the Internet) and a trusted internal domain (it is often an intranet). In general, the internal domain is a corporate intranet. The corporate intranet can deploy wireless networks inside the corporate administrative domain. Each Mobile Node (MN) has a Security Association with the internal domain and belongs to a home network. When a Mobile Node roams inside the internal domain, it is designated as i-MN in this study. The internal subnetwork, which it visits, is designated as i-Foreign Network. The Correspondent Node (CN) inside the internal domain is designated as i-CN. When the Mobile Node roams outside
the internal domain, it is designated as x-MN. Any external foreign network is designated as the x-Foreign network. Similarly, the Correspondent Node outside the internal domain is designated as x-CN.

Note that i-MN and x-MN represent different roles of the same entity rather than different entities. A Mobile Node always belongs to the internal domain, regardless of whether it is designated as i-MN or x-MN. This relationship is represented as a Security Association between the Mobile Node and the internal domain. In contrast, x-CN always represents an external domain node, while i-CN always represents an internal domain node.

From a vertical perspective, according to the coverage range, the scenario is divided into four layers (Stemm & Katz, 1998). Higher layers in the hierarchy provide a lower bandwidth per unit area connection over a larger geographic area. The top layer is satellite network. It provides the widest regional-area coverage, which forms the mega-cells. The second layer is Wireless Wide Area Networks (WWANs). It provides a metropolitan-area coverage, which forms the macro-cells. The third layer is Wireless Local Area Networks (WLANs). It generally provides a campus-area coverage, which forms the micro-cells. The bottom layer is Wireless Personal Area Networks (WPANs). It provides a room-size coverage, which forms the pico-cells. Of the various layers the satellite network and WWANs, generally, are operator-deployed. The WLANs are often enterprise-deployed, but sometimes can be operator-deployed as a complement of the operator’s own WWAN. The WPANs can be enterprise-deployed or family-deployed.

When projecting the vertical view into the horizontal view, the satellite network and WWANs absolutely fall into the x-Foreign Network, because they are operator-deployed, relative to any corporate intranet, they are totally external domains. The WLANs can be Home Network, i-Foreign Network, or x-Foreign Network depending on the practical deployment. The same holds for the WPANs.

1.3.1 Handover

When the Mobile Node moves, it has to switch to different networks in order to keep communicating according to various parameters, such as signal-to-noise ratio (SNR), cost, and bandwidth. This procedure is called handover, or handoff.
It is worth noting that there are two distinguishing types of handover. One is the horizontal handover, which means a handover between base stations that are using the same type of wireless network interface (Stemm & Katz, 1998). It typically takes place in the same layer in Figure 1.3. The other is the vertical handover, which means a handover between base stations that are using different wireless network technologies. The vertical handover, in turn, can be divided into two categories: an upward handover is a handover to a wireless network with a larger cell size (and lower bandwidth per unit area), and a downward vertical handover is a handover to a wireless network with a smaller cell size (and higher bandwidth per unit area). Both handover types may probably lead to administrative domain switch.

Another factor, that should be taken into account when deciding on the handover, is the handover overhead. Although smaller coverage generally, but not always, means higher bandwidth, it also means more frequent handover. Handover is a very time-consuming and resource-consuming operation. As a result of the tradeoff between bandwidth and handover overhead, a Mobile Node should connect to the higher layer with larger coverage only where the WLAN is not available. Once the WLAN is available again, the Mobile Node should immediately switch back to the WLAN. This way the Mobile Node can enable seamless handover and keep continual communication. A rule of thumb is that the WLAN serves as a primary access network, and other wireless networks serve as secondary access networks to provide continual communication when the primary network is unavailable.

From the above analysis, it is concluded that the handover will play an important role in the integration of heterogeneous networks. However, based on the above-mentioned rule of thumb, this study only considers handover parameter signal-to-noise ratio and bandwidth.

1.3.2 Routing

The appropriate handover between different networks enables Mobile Nodes to keep continual and ubiquitous communication anywhere. After handover, the packets should be delivered to Mobile Nodes correctly, regardless of which wireless network they attach to. This involves the routing of packets. For the sake of future discussion, four communication paths are identified in this scenario:
1.3. SCENARIO

- i-MN ↔ i-CN
- i-MN ↔ x-CN
- x-MN ↔ i-CN
- x-MN ↔ x-CN

These four communication paths have different representations. Path i-MN ↔ i-CN presents the internal traffic. Path x-MN ↔ i-CN presents the remote access to the internal resource when Mobile Nodes move away from the internal domain. Path i-MN ↔ x-CN and x-MN ↔ x-CN present access to the Internet regardless of whether Mobile Nodes are located inside or outside the internal domain.

1.3.3 Security

To protect against violation of information over the above-mentioned four communication paths, appropriate security services should be provided in place. According to the National Security Telecommunications and Information System Security Committee, the characteristics include availability, accuracy, authenticity, confidentiality, integrity, utility, and possession (Whitman & Mattord, 2003, p. 10). Canavan (2001, p. 9), however, pointed out that in the context of network security, only availability, authenticity, confidentiality, and integrity are relevant. Only these characteristics are therefore involved in this study.

Depending on the domain that the communication paths involve, security services are provided to different degrees. Generally, it is assumed that the internal domain is viewed as a trusted network, and each node in the internal networks trusts each other. Therefore, end-to-end security for path i-MN ↔ i-CN is not necessary, while end-to-end security for x-MN ↔ i-CN is necessary due to an untrusted external domain being introduced.

Although end-to-end security is needed for paths i-MN ↔ x-CN and x-MN ↔ x-CN, due to the lack of a Security Association between them, end-to-end security is difficult to implement between them using a predefined key security mechanism.

Apart from the aforementioned diverse considerations, a common security consideration for all communication paths is the wireless link. Generally, the
wireless link is considered the weakest link along the communication path, and thus it should be secured first.

1.3.4 Mobility

Mobility has been classified in many ways. Where the internal and external domains are concerned, a typical classification, promoted by Adrangi and Levkowetz (2005), is:

- Mobility inside the Intranet, in which Mobile Nodes move inside the Intranet, communicating with the internal Correspondent Nodes.

- Mobility outside the Intranet, in which Mobile Nodes move outside the Intranet, communicating with the internal Correspondent Nodes.

The entities involved in this classification is illustrated in Figure 1.4.

![Figure 1.4: Mobility inside and outside the Intranet](image)

As shown in Figure 1.4, this classification involves only i-MN, i-CN, and x-MN, excluding x-CN. Since all these entities belong to the internal domain, even if x-MN goes outside the internal domain, this study defines mobility between these entities as Internal Mobility. It involves the communication paths: i-MN ↔ i-CN and x-MN ↔ i-CN. On the opposite side, mobility between i-MN, x-MN, and x-CN is defined as External Mobility, because it introduces an external entity, namely x-CN. It involves the communication
paths i-MN ↔ x-CN and x-MN ↔ x-CN, as illustrated in Figure 1.5. It is worth noting that currently most research effort focuses on internal mobility, and relatively little research is done on external mobility.

Figure 1.5: Internal and External Mobility

Apart from the classification in the horizontal direction, a classification in the vertical direction is also made. This includes global mobility, macro-mobility, and micro-mobility. This will be intensively discussed in Section 5.4 on page 120. Figure 1.6 depicts these two classifications.

Figure 1.6: Mobility Classification

1.4 Goal and Scope

After recognizing where the challenge exists and analyzing the above scenario, this study seeks to explore a probable framework for seamlessly integrating
heterogeneous networks with mobility and security support in place, providing the user with Internet access and remote access to the enterprise intranet, as well as pervasive reachability. Considering the wide deployment of GSM and its derivate, UMTS, as well as IEEE 802.11 WLAN, this study chooses UMTS and IEEE 802.11 as the representatives of WWAN and WLAN respectively. It aims to combine GPRS, Mobile IP, and IPSec to realize seamless handover between GPRS and WLAN, as well as seamless global roaming, while ensuring security.

With reference to Figure 1.5, internal mobility is protected under IPSec. However, IPSec is unsuitable to ensure security for external mobility, since the external entity generally does not have security associations with the internal domain. Other upper layer security mechanisms, for example, the Secure Socket Layer (SSL) (Freier, Karlton, & Kocher, 1996) and Transport Layer Security (TLS) (Dierks & Allen, 1999), may be a good choice for external mobility, but to limit the scope are purposefully omitted from this study. In addition, there is an existing standard for wireless link security, that is, Wired Equivalent Privacy (WEP) (Schafer, 2003, p. 303). Likewise, this is not the focus of this study.

This study will not involve physical and link layer technology. Thus the Software-Defined Radio (SDR) will be excluded from this study. The study rather focuses on the network layer. It involves IEEE 802.11 and UMTS networks.

From a system view, the UMTS can be divided crudely into three parts: the air interface, the radio access network and the core network (Wisely et al., 2002, p. 22). That is the Wideband Code Division Multiple Access (WCDMA), the Universal Terrestrial Radio Access Network (UTRAN) and the Core Network, as shown in Figure 2.1 on page 26.

This study will focus on the Core Network and Radio Access Network of the UMTS system, especially the packet-switched domain, that is, the GPRS domain, because that is “where IP can make a difference to the performance and architecture of a 3G network” (Wisely et al., 2002, p. 22). Air interface and hardware are not this study’s concern.

As before mentioned, global roaming will be one of this study’s objectives, but only the technological aspect will be involved in this study; the business aspect on how roaming agreements are signed is not relevant to this study.
Although IPv6 is considered revolutionary for the new Internet model, changing to IPv6 must be an evolutionary process (SPIRENT, 2004). As reported by ADTRAN (2005), there is a limited demand for IPv6 in the U.S. market. Both IPv6 and IPv4 will need to co-exist and inter-operate for an extended transitional period. IPv4 thus still has large business value. Due to IPv4’s wide deployment and high maturity, all IP-relative protocols and standards in this study are based on IPv4. However, the impact of IPv6 on the framework proposed by this study, will be evaluated in the conclusion chapter.

This study attempts to propose a high level framework. As such it is not a practical project, and does not produce implementable specifications. The next section summarizes the objectives.

1.5 Objectives

As mentioned in Section 1.4, the goal of this study is to develop a framework that combine GPRS, Mobile IP, and IPSec, which realizes seamless handover between GPRS and WLAN, while ensuring security.

To successfully achieve this goal, the following objectives need to be addressed:

- An effective IP mobility approach should be selected depending on different geographical ranges.
- A mechanism should be found to ensure seamless handover between GPRS and WLAN.
- Mobile IP services should be offered in the GPRS network.
- Mobile IP should cooperate with IPSec to provide mobility and security simultaneously.

These objectives will be realized systematically.

1.6 Methodology

The methodology followed in this study was primarily a literature study. The emphasis of which was on the review of existing RFC documents and
Internet drafts from the IETF, as well as the specifications and technical reports from the international telecommunication research organization, such as 3GPP, 3GPP2, ETSI and so forth. Although few books were at our disposal, academic papers from appropriate international conferences and referred journals were reasonably available through digital libraries and personal websites on the Internet.

The synthesis of the framework relied on argumentation and drawing analogies to similar problems. The framework was presented in the form of a dissertation, as well as an academic paper. Verification of the framework would be attempted in terms of best practices and, possibly, network simulations.

1.7 Related Work

Some of the aforementioned issues were addressed partially by other researchers. Guardini et al. (2000) pointed out that IP would play a key role in an interworking wireless network. Wisely et al. (2002) further proved this standpoint from the IP design principles perspective.

Das, Misra, and Agrawal (2000) classified IP mobility as Micro-mobility, Macro-mobility, and Global mobility in a hierarchical manner. Their solution, TeleMIP, chooses the most effective method according to the current scope.

Stemm and Katz (1998) presented the concept of the vertical handover, which theorized seamless handover between heterogeneous networks. A lot of practical work has been done based on this concept. OmniCon, for example, created a virtual network device to enable interworking between IEEE802.11 and GPRS (Sharma, Baek, Dodia, & Chiueh, 2004). Recently, the LCE-CL testbed was also deployed to examine the performance of handover between IEEE802.11 and GPRS (Vidales, Mapp, Stajano, Crowcroft, & Bernardos, 2005), and got some experimental results.

As to the integration of MIPv4 and GPRS, ETSI introduced a three-step approach toward this target (European Telecommunications Standards Institute, 2001).

The research community addresses the problem of combining Mobile IP and IPSec in two ways:
1.7. RELATED WORK

- add security to Mobile IP, or
- add mobility to IPSec.

Since currently the firewall and IPSec have much wider deployments than Mobile IP, this study intends to add mobility to IPSec rather than add security to Mobile IP.

The idea of combining Mobile IP and IPSec has the root in the Mobile IP Security system (MoIPS, 1995 – 1997) (Zao et al., 1997). Zao and Condell (1997) summarized this idea in the form of an Internet draft. Another early attempt of combining Mobile IP and IPSec was given in the Secure Mobile Networking project (SMN) of the Portland State University in July 1995 (Binkley, 2001). The comments from this project formed the Internet draft, “Security Considerations for Mobility and Firewalls” (Binkley & Richardson, 1998). SecMIP proposed a solution that the Home Agent was hidden behind a de-militarized zone (DMZ) (Braun & Danzeisen, 2001). This solution required that the IPSec security association had to be renegotiated every time when the Mobile Node obtained a new Care-of Address (COA), which led to a significant delay. For addressing this problem, the concept of dual home agents was introduced. The Secure Universal Mobility project (SUM) proposed this kind of architecture (Dutta et al., 2004). In this proposal, an internal Home Agent tunnels the packet destined for a Mobile Node to the IPSec gateway, which in turn tunnel it to an external Home Agent. Finally, the external Home Agent tunnel it to the Mobile Node as if the external Home Agent expands the IPSec tunnel. Since the outer end of IPSec is in the external Home Agent, which hides the Care-of Address changes from IPSec gateway, the IPSec security association does not need to be renegotiated every time when the Care-of Address changes. However, this introduced the overhead of triple encapsulation, which may not be suitable for real-time communication. A recent research study was done on providing low-latency secure mobile communications based on the previous SUM architecture (Choyi & Barbeau, 2005).

Recently, there has been much activity within the Internet Engineering Task Force (IETF) to develop solutions to maintain IPSec connectivity while a mobile device changes its IP address. An Internet draft has been accepted as the RFC document in August 2005, which pointed out the existing prob-
lems in integrating Mobile IP and IPSec (Adrangi & Levkowetz, 2005). A solution has been proposed in the form of an Internet draft to address these problems (Vaarala & Klovning, 2005). This solution involved both the enterprise and the Internet environments.

Rantala (2004) specially discussed the enterprise use of Mobile IP and IPSec. EURESCOM’s FIT-MIP project addressed the global mobility problem (Morand & Tessier, 2002).

However, none of them provided a whole framework to achieve this study’s goal. Thus this study seeks to stand on the shoulders of these contributors to propose a complete framework towards this study’s goal.

1.8 Overview of This Study

The layout of the dissertation is illustrated in Figure 1.7 and is divided into three parts:

**Part I** introduces knowledge that the reader needs to understand the proposed framework and consists of six chapters. **Chapter 1** provides some background and defines the scope and problems of the topic. **Chapter 2** provides a brief introduction of GPRS network as well as 3G systems. **Chapter 3** introduces the concept of wireless overlay network and network models of implementing this concept. **Chapter 4** describes IP’s design principles and examines its drawbacks. **Chapter 5** describes Mobile IP and discusses the effective IP mobility solution. **Chapter 6** provides an overall description of IPSec.

**Part II** presents the proposed framework and consists of three chapters. **Chapter 7** introduces offering Mobile IP services to GPRS networks. **Chapter 8** investigates the various scenarios of combining Mobile IP and IPSec and introduces a dual home agent architecture for the integration of IPSec and Mobile IP. **Chapter 9** proposes a complete framework for seamless handover between GPRS and WLAN without losing security characteristics.

**Part III** is the epilogue and contains only **Chapter 10**. It concludes this dissertation, evaluates the proposed framework and indicates future work.
1.8. OVERVIEW OF THIS STUDY

Figure 1.7: Layout of Dissertation
Chapter 2

GPRS — A Key Milestone towards 3G

As mentioned in the introduction, one of the objectives of this study is to find a mechanism to realize seamless handover between GPRS networks and 802.11 WLAN. To understand such an mechanism, this chapter will examine the various aspects of the GPRS network.

General Packet Radio Service (GPRS) is a key milestone for GSM data service (Bates, 2002, p. 52), because it introduces a brand-new, packet-switched and IP-based domain into GSM network (European Telecommunications Standards Institute, 2000a), which is more suited to data services. Furthermore, GPRS also provides an added step towards third-generation (3G) networks, as predicted by Regis J.(Bud) Bates, President of TC International consulting, Inc.:

“GPRS will enable the network operators to implement IP-based core architecture for data applications. This will continue to proliferate new services and mark the steps to 3G services for integrated voice and data applications.” (Bates, 2002, p. 54)

Nowadays the broad deployment of GPRS world wide has justified this prediction. With the deployment of GPRS, 3G has progressed from concept to realization. Therefore, regardless of from a theoretical or practical perspective, GPRS is worth proper attention at this point in time.

Due to the significant relationship between 3G and GPRS, an introduction of 3G will be provided first in this chapter. Then an overview of GPRS
will be given. After that, the various aspects of GPRS will be described in proper detail in terms of routing, addressing scheme, procedures and mobility management. Specific attention will be paid to handovers.

2.1 Introduction to 3G

Probably the best description of the original concept of 3G is captured in Alan Clapton’s quote — head of British Telecom’s 3G development:

“3G ... The evolution of mobile communications towards the goal of universal personal communications, a range of services that can be anticipated being introduced early in the next century to provide customers with wireless access to the information super highway and meeting the ‘Martini’ vision of communications with anyone, anywhere and in any medium” (Clapton & Groves, 1996)

A classic graphical description for this scenario has been shown in Figure 1.1 on page 5.

However, reviewing 3G’s history, this scenario was not really realized in practice. 3G has progressed from its ‘Martini’ vision — ‘anytime, anyplace, anywhere’, to a system much closer, in many respects, to the existing 2G networks. The major reason for this was the desire by the existing 2G operators and manufacturers to reuse as much existing equipment, development effort, and services as possible (Wisely et al., 2002, p. 29). This is an important factor in understanding the development of the 3G system below.

2.1.1 Overview of 3G Standards

Although 3G was originally conceived as being a single world-wide standard, by the time it was born, there were two proposed 3G systems and five air interface standards. It is the result of a tradeoff between realism and idealism. The whole project was termed as the IMT-2000 family of standards (Schiller, 2003, p. 136). IMT-2000 is a total “umbrella specification” of all 3G standards, which covers all aspects of 3G systems. From a system view, 3G systems can be divided very crudely into three parts: the Air Interface, the Radio Access Network (RAN) and the Core Network (CN) (Wisely et
2.1. INTRODUCTION TO 3G

al., 2002, p. 22). All 3G standards will be investigated in a comprehensive
and brief way according to this division.

The two proposed 3G systems are Universal Mobile Telecommunications
System (UMTS) (European Telecommunications Standards Institute, 2000b,
2002), developed and promoted by Europe and Japan, and CDMA2000 (3rd
Generation Partnership Project 2, 2001), developed and promoted by North
America. There are also two standardization bodies to support them respec-
tively, 3GPP (www.3gpp.org) for UMTS, and 3GPP2 (www.3gpp2.org) for
CDMA2000.

The five air interface standards concerned are Wideband CDMA, CDMA2000,
TD/CDMA, UMC-136 (EDGE), and DECT, as summarized in Table 2.1.

Table 2.1: Air Interface Standards of IMT2000 Family

<table>
<thead>
<tr>
<th>IMT2000 designation</th>
<th>Common Term</th>
<th>Duplex type</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMT-DS Direct Spread CDMA</td>
<td>Wideband CDMA</td>
<td>FDD</td>
</tr>
<tr>
<td>IMT-MC Multi Carrier CDMA</td>
<td>CDMA2000</td>
<td>FDD</td>
</tr>
<tr>
<td>IMT-TC Time Division CDMA</td>
<td>TD-CDMA</td>
<td>TDD</td>
</tr>
<tr>
<td>IMT-SC Single Carrier</td>
<td>UMC-136</td>
<td>FDD</td>
</tr>
<tr>
<td>IMT-FT Frequency Time</td>
<td>DECT</td>
<td>TDD</td>
</tr>
</tbody>
</table>

As to the core network, for the purpose of reusing, they are chosen based
on the existing 2G Core Network: in the case of UMTS an evolved GSM
Core Network is used; and in the case of CDMA2000 an evolved ANSI-41
Core Network is used.

Since new air interfaces were adopted, the Radio Access Network, as the
 glu linking the Core Network and Base Stations, had to be new for both
Terrestrial Radio Access Network (UTRAN), was produced to adapt to the
new air interface. In the CDMA2000, upgrades to the Base Transceiver
Station (BTS) and Base Station Controller (BSC) were required (Smith &
Collins, 2002, p. 159), due to the introduction of packet data services and
the new air interface.

Due to the wide deployment of GSM, its descendant, UMTS, received the
most attention. Since GPRS is a part of both GSM and UMTS, the rest of
this section will focus on UMTS. As will be seen later, GSM technology played
a remarkable role in the background to UMTS (Kaaranen et al., 2001, p. 9).
Knowledge of GSM is therefore helpful for understanding UMTS. However,
GSM will not be mentioned in this study. Interested readers are referred to a plethora of literature available on GSM. In particular, a good description can be found in the book “GSM Switching, Services and Protocols” (Eberspacher et al., 2001).

UMTS is updated regularly under the control of 3GPP. Until now, 3GPP has released four versions of UMTS: 3GPP Release 99 (R99), sometimes called Release 3 (R3), 3GPP Release 4 (R4), 3GPP Release 5 (R5) and 3GPP Release 6 (R6). Due to 3GPP R99’s wide deployment and high maturity, the description of UMTS in this study is based on 3GPP R99 unless otherwise stated.

2.1.2 UMTS Network Architecture

A complete UMTS architecture consists of the user terminal and network entities. In the context of 3G systems, the user terminal is called User Equipment (UE) and contains two separate parts: Mobile Equipment (ME) and the UMTS Service Identity Module (USIM) (Kaaranen et al., 2001, p. 9). The network entities in turn are divided into two parts: the Radio Access Network (in UMTS it is called UTRAN) and the Core Network. A high level architecture is depicted in Figure 2.1.

The interfaces, between the User Equipment and the UTRAN (Uu interface), and between the UTRAN and the Core Network (Iu interface), are open multi-vendor interfaces. The Uu interface is the place where the WCDMA technology can take effect. However, it is not this study’s focus. This study will focus on UTRAN and the Core Network (CN) as well as the Iu interface between them, because “that is where IP could make a difference to the performance and architecture of a 3G network” (Wisely et al., 2002, p. 22).
Figure 2.2: UMTS Network Architecture (Kaaranen et al., 2001, p. 10)

Figure 2.2 provides much more details of the UMTS architecture. This will be described in detail in the subsequent subsections. For now, observe that the UMTS infrastructures are constructed in a hierarchical fashion, as shown in Figure 2.3. This leads to hierarchical Mobility Management.

A UMTS network consists of at least one administrative region, which is assigned to a Mobile Switching Center (MSC). Each administrative region is made up of at least one Location Area (LA). Sometimes the LA is also called the visited area. A LA consists of several cell groups. Each cell group is assigned to a Radio Network Controller (RNC). Therefore for each LA there exists at least one Base Station (BS), but cells of one RNC may belong to
different Location Areas.

Radio Access Network

As shown in Figure 2.2, there are two types of Radio Access Networks, the UMTS Terrestrial Radio Access Network (UTRAN) and GSM/EDGE Radio Access Network (GERAN). GERAN derives from the GSM with Enhanced Data for GSM Evolution (EDGE) technology, which is introduced into UMTS in 3GPP R4/5 as an alternative to building a UMTS mobile network. However, this study does not intend to handle it in detail. A good description of it can be found in Halonen, Romero, and Melero (2003, p. 585).

The other Radio Access Network, UTRAN, forms the focus of this study. It consists of Radio Network Subsystems, and each Radio Network Subsystem (RNS) in turn contains a different number of Base Stations (BS or, officially, Node B) and one Radio Network Controller (RNC).

The main task of UTRAN is to create and maintain the Radio Access Bearer (RAB) for communication between User Equipment (UE) and Core Network (CN) (Kaaranen et al., 2001, p. 99). This task is distributed to the Base Station and Radio Network Controller.

The Base Station can be considered as the radio edge of the UTRAN. Thus its underlying task is to perform radio signal receiving and transmitting, signal filtering and amplifying, and signal modulation and demodulation. Although the Base Station’s main task is the underlying signal process, nevertheless, there are also other UTRAN control functions in which the Base Station is involved. On the one hand, the Base Station plays a partly supportive role in some functions, including fulfilling, collecting and filtering
radio measurements, and providing them to the Radio Network Controller to execute its control functions. On the other hand, the Base Station plays a central role in code generating, power control executing and Operation & Maintenance, particularly in the network element or at the cell level.

The Radio Network Controller is the switching and controlling element of the UTRAN. On the whole, the functions of the Radio Network Controller can be divided into two parts: UTRAN Radio Resource Management (RRM) and control functions (Kaaranen et al., 2001, p. 111). Radio Resource Management uses a collection of algorithms to ensure the stability of the radio path and the Quality of Service (QoS) of radio connection by the effective sharing and managing of radio resources. UTRAN control functions are those relevant to setup, maintenance and release of the Radio Access Bearer.

Core Network

The Core Network (CN) can be seen as the basic platform for all communication services provided to UMTS subscribers, including switching of circuit-switched calls and routing of packet data (Kaaranen et al., 2001, p. 143). As illustrated in Figure 2.2, the UMTS Core Network primarily consists of two basic domains: Circuit-Switched (CS) domain and Packet-Switched (PS) domain. One of the reasons for such a division is to reflect the traffic characteristics in each domain. These two domains are described briefly below:

- Circuit-Switched Domain: this domain delivers circuit-switched traffic, and is based on the traditional telecommunication network technologies, such as Integrated Service Digital Network (ISDN) and Signalling System #7 (SS7). This domain contains two major elements: the Mobile Switching Center (MSC) and the Gateway Mobile Switching Center (GMSC).

- Packet-Switched Domain: this domain delivers the packet data traffic, and is based on IP-based network technologies. This is where GPRS comes in, which will be discussed later in Section 2.2. Its two major counterparts to the Circuit-Switched Domain are the Serving GPRS Support Node (SGSN) and the Gateway GPRS Support Node (GGSN).

In addition to the above mentioned basic domains, there are two extra elements in the Core Network, which are described in brief below:
• IP Multimedia Subsystem (IMS): this subsystem was introduced in 3GPP Release 5. What makes a difference between a subsystem and the above domains is the following: the Core Network domains have an direct interface to one or more access networks, as shown in Figure 2.2, while the Core Network subsystem does not have such a direct interface with access networks. Instead, it connects to one or more Core Network domains to perform its functionalities (Kaaranen et al., 2001, p. 145). This subsystem’s primary purpose is to enable applications in mobile devices to establish peer-to-peer connections.

• Broadcast (BC) Domain: although it has been defined to be part of the Core Network, its implementation with the 3G network is for further study. For the consideration of integrity, it is mentioned here.

Besides the domains and subsystems mentioned above, some functionalities that are common to all Core Network domains and subsystems are mainly collected in a series of registers as illustrated in Figure 2.2. These functionalities are distributed into the separate elements: the Home Location Register (HLR), the Visitor Location Register (VLR), the Authentication Center (AuC) and the Equipment Identity Register (EIR). Physically, the Visitor Location Register is often implemented at the Mobile Switch Center, while other registers are located in a central site, as shown in Figure 2.2.

The details of these elements, including the above mentioned Circuit-Switched elements, are not handled in this dissertation. A good description of them can be found in Korhonen (2001, pp. 208 – 213). The Packet-Switched elements, that is the GPRS specific elements, will be discussed in detail in Section 2.2. However, it is necessary to highlight the functionality provided by these elements here:

• Mobility Management (MM): this functionality is performed primarily through the Home Location Register and the Visitor Location Register. The former serves as a central database for the subscriber information, for example, addressing information. This information can help pinpoint the user/terminal location within the Mobility Management hierarchy shown in Figure 2.3. The latter contains pretty much the same information, but, only for the visiting subscribers on a temporary basis. That is why it is often co-located with the Mobile Switch
2.1. INTRODUCTION TO 3G

Center. In this way, user mobility is supported through the Circuit-Switched domain, the Packet-Switched domain and the IP Multimedia Subsystem.

- User security information generation, user security support and access authorization: these functionalities are mainly maintained by the Authentication Center, which sends signals to the Core Network domains and subsystems through the Home Location Register.

- Identification handling: these elements, in corporation with each other, provide the appropriate relations among all the identifiers uniquely determining the user in the system. These identifiers will be examined in more detail in Section 2.1.4.

It is worth noting that all those register elements, except for the Visitor Location Register and the Equipment Identity Register, are integrated into a new entity called the Home Subscriber Server (HSS) in the 3GPP R5. Nevertheless the same functionalities are offered via the Home Subscriber Server.

2.1.3 Interfaces and Protocols

The various entities in UMTS are connected together via various interfaces regardless of being wired or wireless. These interfaces are also shown in Figure 2.4 utilizing the GSM/GPRS naming convention, where applicable. They can be roughly divided into two levels: the interfaces (Uu and Iu) splitting the UMTS into three parts, and the interfaces between the entities inside a part. The former are truly open, which means they are well-specified, and the specification is such that the equipment on different ends of the interface can be acquired from different manufacturers (Korhonen, 2001, p. 221). The latter are often, but not always, proprietary. A trend is that they are becoming more and more open.

The following is a brief description of these interfaces:

**Uu interface** is situated between UE and UTRAN. It is the air interface of UMTS. In other words, it is where the WCDMA technologies come in. Some new features that WCDMA brings to this interface will be introduced later.
**Iu interface** is situated between UTRAN and CN. Depending on the domain it attaches to, it is often denoted as:

- $Iu_{CS}$, when it attaches to the CS domain and delivers CS traffic,
- $Iu_{PS}$, when it attaches to the PS domain and delivers PS traffic.

**MAP interface** stands for the Mobile Application Part (MAP) interface, which was originally designed for control signalling in the GSM Circuit-Switched services. It is a umbrella name for all interfaces between Core Network entities. With the introduction of Packet-Switched services by the GPRS, it became necessary to extend the MAP protocol to the interfaces between GPRS support nodes and the home network nodes (Kaaranen et al., 2001, p. 293). Thus it contains the following two categories of interfaces:

- Interface C, D, E, F and G, which originate from the Circuit-Switched domain,
- Interface Gc, Gr, Gf and Gs, which originate from the Packet-Switched domain.
Iub and Iur interface connect different Radio Network Controllers together, which make soft handover possible.

2.1.4 Addresses and Identifiers

One of the major tasks of a communication system is to decide what scheme is used to identify an entity in the system. Thus an identifier for each entity is needed for the purpose of identifying, addressing, or both. Generally, an equipment entity should be both identifiable and addressable while the entity, like a user, requires identifiability only. GSM distinguishes explicitly between user and equipment, and deals with their identification separately (Eberspacher et al., 2001, p. 30). The user equipment and user each have their own internationally unique identifiers separately. This concept is also adopted by UMTS.

In UMTS, the user subscription-specific information is stored in the Universal Subscriber Identity Module (USIM), the Subscriber Identity Module (SIM) in the case of GSM. This USIM usually comes in the form of a chip card, which is transferable between mobile stations. This allows for distinguishing between the terminal mobility and personal mobility (Wisely et al., 2002, p. 144).

Many identifiers have been defined. They are needed for the management of subscriber mobility and for addressing all the other network elements. The most important addresses and identifiers are presented in the following subsections.

International Mobile Station Equipment Identity (IMEI)

The IMEI uniquely identifies mobile stations globally. It is a kind of serial number. The IMEI is allocated by the equipment manufacturer and registered by the network operator, who stores it in the Equipment Identity Register (EIR). By means of the IMEI one recognizes obsolete, stolen, or nonfunctional equipment and, for example, can deny service. This purpose is achieved by means of three lists within the EIR: white list, black list and grey list. A detailed description of how it works can be found in the book “Introduction to 3G Mobile Communications” (Korhonen, 2001, p. 209).
International Mobile Subscriber Identity (IMSI)

When registering for service with a mobile network operator, each subscriber gets a unique identifier, the International Mobile Subscriber Identity (IMSI). This IMSI is stored in the USIM. User Equipment can only be operated if a USIM with a valid IMSI is inserted into equipment with a valid IMEI. This allows the associated subscriber to be billed. The IMSI consists of several parts:

- Mobile Network Code (MNC): 2 decimal places, for unique identification of mobile networks within a country.
- Mobile Subscriber Identification Number (MSIN): maximum 10 decimal places, identification number of the subscriber in his or her mobile home network.

According to this addressing scheme, subscriber identification utilizes a maximum of 15 decimal digits, and IMSI = MCC + MNC + MSIN. While the MCC is defined internationally, the National Mobile Subscriber Identity (NMSI = MNC + MSIN) is assigned by the operator of the home Public Land Mobile Network (PLMN).

Mobile Subscriber ISDN Number (MSISDN)

Simply speaking, the MSISDN is the “real telephone number” of user equipment. It is assigned to the subscriber in his or her USIM, such that a mobile station can have several MSISDNs depending on the USIM. The separation of call number (MSISDN) and subscriber identity (IMSI) primarily serves to protect the confidentiality of the IMSI. In contrast to the MSISDN, the IMSI need not be made public. With this separation, one cannot derive the subscriber identity from the MSISDN, unless the association of IMSI and MSISDN, stored in the Home Location Register, has been disclosed. A rule of thumb is that the IMSI used for subscriber identification should not be known publicly, to make it more difficult to assume a false identity (Eberspacher et al., 2001, p. 32).
The MSISDN categories follow the international Integrated Service Digital Network (ISDN) numbering plan and therefore have the following structure:

- Country Code (CC): up to 3 decimal places, internationally standardized.
- Subscriber Number (SN): maximum 10 decimal places.

The subscriber number is the concatenation MSISDN = CC + NDC + SN and thus has a maximum of 15 decimal digits. It is stored centrally in the Home Location Register.

**Location Area Identity (LAI)**

Each LA of a Public Land Mobile Network has its own identifier. The Location Area Identifier (LAI) is also structured hierarchically and is internationally unique since the Location Area itself is hierarchical, as introduced in Section 2.1.2. LAI consists of an internationally standardized part and an operator-dependent part:

- Country Code (CC): 3 decimal places
- Mobile Network Code (MNC): 2 decimal places
- Location Area Code (LAC): maximum 5 decimal places, or maximum twice 8 bits

This LAI is broadcasted regularly by the base station on the Broadcast Control Channel (BCCH). Thus, each cell is identified uniquely on the radio channel as belonging to a LA, and each MS can determine its current location through the LAI. If the LAI that is “heard” by the MS changes, the MS notices this LA change and requests the updating of its location information in the Visitor Location Register and Home Location Register.
CHAPTER 2. GPRS — A KEY MILESTONE TOWARDS 3G

Temporary Mobile Subscriber Identity (TMSI)

The Visitor Location Register (VLR), being responsible for the current location of a subscriber, can assign a Temporary Mobile Subscriber Identity (TMSI), which has only local significance in the area handled by the VLR. It is used in place of the IMSI for the definite identification and addressing of the mobile station. In this way nobody can derive the identity of the subscriber by listening to the radio channel, since this TMSI is only assigned during the mobile station’s presence in the area of one VLR, and can even be changed during this period (10 hopping).

The mobile station stores the TMSI on the USIM card. The TMSI is stored on the network side only in the VLR and is not passed to the HLR. A TMSI may therefore be assigned in an operator-specific way.

Together with the current location area, a TMSI allows a subscriber to be identified uniquely, i.e. for the ongoing communication the IMSI is replaced by the 2-tuple (TMSI, LAI).

Other Identifiers

Beside the above primary identifiers, there are also several other identifiers, which are listed here for the purpose of completeness:

- Mobile Station Roaming Number (MSRN): the MSRN is a temporary location-dependent ISDN number. It is assigned by the locally responsible VLR to each mobile station in its area. Calls are routed to the MS by using the MSRN. On request, the MSRN is passed from the HLR to the GMSC. The MSRN has the same structure as the MSISDN.

- Local Mobile Subscriber Identity (LMSI): the LMSI serves as a searching key to each visiting mobile station to accelerate database access. It is assigned by VLR when the mobile station registers with the VLR and is also sent to the HLR.

- Cell Identifier (CI): within a LA, the individual cells are uniquely identified with a Cell Identifier (CI), maximum $2 \times 8$ bits. Together with the Global Cell Identity (LAI + CI), cells are thus also internationally defined in a unique way.
• Identification of MSCs and Location Registers: MSCs and location registers (HLR, VLR) are addressed with ISDN numbers. In addition, they may have a Signalling Point Code (SPC) within a PLMN, which can be used to address them uniquely within the SS7 network.

2.1.5 Handover

Mobile phones can maintain their connections in cellular networks when moving from one cell area to another. This benefits from a procedure, called handover or handoff, which switches a connection from one Base Station (BS) to another (Korhonen, 2001, p. 38).

Handovers in UMTS are fundamentally different from handovers in GSM. This is owed to the following two characteristics:

• In WCDMA radio access technology, employed by UMTS, all base stations use the same frequency while in TDMA, employed by GSM, different frequencies are used to multiplex the radio resource.

• Due to using the same frequency, the UMTS User Equipment can utilize a multi-path receiving technology, called RAKE receiver (Korhonen, 2001, p. 32). A RAKE receiver is made of correlators, also known as RAKE fingers, each receiving a multi-path signal. This makes it possible for a User Equipment to simultaneously connect to more than one base station.

A User Equipment (UE) communicating with multiple Base Stations can spend a large part of the connection time in a Soft Handover state (Korhonen, 2001, p. 38). A special type of Soft Handover is Softer Handover, which is a soft handover between cells under the same Base Station. According to a hierarchical structure, there are other types of handovers, such as Internal Handover and External Handover with the latter in turn divided into Intra-MSC Handover and Inter-MSC Handover. This hierarchical handover structure is shown in Figure 2.5. The concept of Soft Handover is applicable through the entire hierarchical structure. There are also other kinds of handovers that are not included in this hierarchical structure, for example, Hard Handover and Inter-system Handover. The following subsections will discuss handovers in brief.
Softer Handover

As aforementioned, a User Equipment can simultaneously connect to multiple Base Stations. In the uplink direction, the signals collected by the cells are combined in the Base Station if all the cells attached by the User Equipment belong to that Base Station. In the downlink direction, that Base Station selects a cell as primary cell to send signal to the User Equipment. In this way when the User Equipment moves between these cells, cell handover is handled seamlessly, and no Radio Network Controller (RNC) is involved. This handover is called Softer Handover, for example, the handover identified by number 1 in Figure 2.5.

When the handover identified by number 2 takes place, in the uplink direction the signal is collected by the BS1 and BS2 first, respectively. Then the signals are combined in the RNC1. In the downlink direction, according to the result of signal measurement, the RNC1 selects the appropriate cell, for example, Cell3, as the primary cell to send the downlink signal. This ensures the connection to the User Equipment is not broken before the User Equipment loses the connection with Cell2. No Mobile Switching Center (MSC) is involved in this situation. This kind of handover, which occurs between the cells controlled by the same RNC, is called Internal Handover. It also includes Softer Handover.
All other handovers require participation of at least one MSC and are called external handovers. When handover 3 takes place, this involves two Radio Network Controllers (RNCs), RNC1 and RNC2. At this time, the RNC1 plays a role of Serving Radio Networking Controller (SRNC). It is in charge of all control between the User Equipment and the UTRAN. The RNC2 plays a role of Drift Radio Network Controller (DRNC). It relays the signal, collected from the Cell5, to the RNC1 via the Iur interface shown in Figure 2.4. The RNC1 serves as an anchor to combine all signals and send them to the MSC1 even if the User Equipment leaves all cells controlled by the RNC1. Obviously it is not effective, because this wastes the resource of the RNC1 and loads up the Iur interface. A procedure, Relocation, needs to occur, which moves the SRNC role from the RNC1 to the RNC2. Relocation is conducted by the MSC of that two RNCs. In this case, it is MSC1. Thus it is called Intra-MSC handover.

When the User Equipment moves over a cell boundary and enters the area of responsibility of a new MSC, the Inter-MSC handover occurs, for example, Handover 4 in Figure 2.5. In this situation, the signals collected by the new MSC, for example, MSC2, are always relayed to the MSC that the User Equipment attached to at the beginning of the connection, for example, MSC1. In this case, the MSC1 constantly keeps an anchor role, combining signals and sending them to the destination, for the entire life time of the connection. This is different from the case of SRNC and DRNC where their roles constantly change with the movement of the User Equipment.

**Hard Handover**

Hard handover is also known as an inter-frequency handover, during which the used radio frequency of the User Equipment changes. It is so called break-before-make handover. It is not seamless. It is not handled in detail in this study. Some description of it can be found in (Korhonen, 2001, pp. 44 – 45).

**Inter-system Handover**

Inter-system Handover is the handover between two different radio access technologies. 3GPP has specified inter-system handover between GSM and UTRAN systems. This is not discussed here. In contrast, the Inter-system
2.2 GPRS Overview

The GPRS is originally designed to allow the service subscriber of GSM to send data to and receive data from external Packet Data Networks (PDNs) in an end-to-end packet transfer mode, without utilizing network resources in Circuit Switched mode (European Telecommunications Standards Institute, 2000a). This introduces a new Packet-Switched domain into GSM, which is inherited by GSM’s descendant, UMTS, with some modification in protocol stack. It is worth noting that here the PDN means general packet data networks, including not only IP networks, but also Point-to-Point Protocol (PPP) networks and X.25 networks. That is why it is called “General Packet Data Service”. However, in practice the GPRS is commonly used to provide access primarily to IP PDNs, especially to the Internet (Salkintzis, 2004, p. 3-2). Thus in this study the focus is on IP-based networks only. In this sense, from a view of the external IP networks, a GPRS network can be seen as a special IP network, which offers IP connectivity to the mobile station (MS, in GPRS terminology) on the move (Salkintzis, 2004, p. 3-2). In other words, the GPRS network resembles a typical IP network in the sense that it provides typical IP routing and interfaces to the external world through one or more IP routers. A high level conceptual reference model is illustrated in Figure 2.6.

Starting with this conceptual model, the rest of this section will describe GPRS from a conceptual view first. Then GPRS system architecture, which implements this concept, will be given. Finally, the interfaces and protocols involved in this architecture will be examined in brief.

2.2.1 Conceptual View of GPRS

Mobile users can get access to the remote IP networks via a special access router, which in GPRS terminology is called Gateway GPRS Support Node (GGSN). From the perspective of the external IP network, for example, the Internet, the GPRS network can be seen as an access network to the Internet as a typical dial-up access network (Salkintzis, 2004, p. 3-2). In this
sense, the GGSN serves as an interface to the external Packet Data Network (PDN) where, on the other end, another network element, the Serving GPRS Support Node (SGSN) offers access points to the mobile stations. When the mobile station moves, it attaches to the closest SGSN to gain access to the external PDN. The GGSN serves as an anchor, which forwards packets destined for the mobile station to its SGSN (Eberspacher et al., 2001, p. 70). In this way the GPRS hides the movement of the mobile station from the external PDN.

With such a conceptual understanding of what the GPRS does, there are several essential issues that need to be clarified:

- IP network and UMTS use different addressing schemes. In an IP network a terminal is identified by its IP address while in the case of an UMTS network, a terminal is addressed by the IMSI of its user. As a result, a Mobile Station has two identifiers depending on the context in which it is used. How is an identifier converted to another identifier when necessary? A one-to-one association between a Mobile Station’s IP address and its IMSI should be established.

- After knowing the corresponding relationship between IP address and IMSI, the question arises how the IP packet from the external IP net-
work is delivered to the destination mobile station identified by its IMSI, and vice versa. That is a matter of routing.

- Another issue is when the Mobile Station moves to a new place, how the IP data destined to this Mobile Station is delivered to it correctly. That is a matter of Mobility Management.

These issues will be addressed later in this chapter. Before doing that, a complete GPRS system architecture is introduced for the purpose of reference in future discussion.

### 2.2.2 GPRS System Architecture

The previous section has introduced the conceptual model of GPRS. This section further describes a detailed system architecture of GPRS, and notes how this architecture implements the above mentioned conceptual model. The GPRS system architecture, as a part of the UMTS, has been presented in Figure 2.2 on page 27. Nevertheless, a dedicated architecture diagram of GPRS is shown in Figure 2.7 for the sake of discussion. In the figure, the
solid lines present the data flow while the dashed lines present the signalling flow.

As seen in Figure 2.2 on page 27, the support of GPRS does not represent a major upgrade to the existing CS domain infrastructures. The greatest impact is the addition of two new elements:

- the Serving GPRS Support Node (SGSN), and
- the Gateway GPRS Support Node (GGSN).

SGSN delivers data packets from and to the mobile stations within its service area. Its tasks include packet routing and transfer, functions for attach/detach of mobile stations and their authentication, and logical link management. The location register of the SGSN stores location information, for example, current cell, current Visitor Location Register, and user profiles of all GPRS users registered with this SGSN (Bates, 2002, pp. 108 – 110).

GGSN acts as an interface to external packet data networks. It converts IP packets coming from the SGSN into the appropriate Packet Data Protocol (PDP) format (IP or X.25) and sends them out on the corresponding external network. In the other direction, the PDP address of incoming data packets is converted to the GPRS address of the destination user. The readdressed packets are sent to the responsible SGSN. For this purpose the GGSN stores the current SGSN addresses and profiles of registered users in its location register (Bates, 2002, pp. 110 – 111).

In general, there is a many-to-many relationship between the SGSNs and GGSNs: a GGSN is the interface to an external network for several SGSNs; a SGSN may route its packets to different GGSNs (Eberspacher et al., 2001, p. 243).

In addition to the above two new elements, the GPRS also needs access to the common database functional elements introduced in Section 2.1.2, such as VLR, HLR, EIR, and AuC. However, some software upgrades need to be done to contain some new data relevant to the GPRS. The access to these elements is performed via an extension of the Mobile Application Part (MAP) protocol (Korhonen, 2001, p. 231). Each MAP interface in Circuit-Switched domain has its counterpart in the Packet-Switched domain, as shown in Figure 2.4 on page 32.
It is worth noting that although the GPRS supports general packet data protocol transfer, the GPRS backbone is built on an IP-based network. Thus the communication between GSNs uses the IP protocol. However, the communication between GSNs and the database elements (VLR, HLR, EIR, and AuC) uses Signalling System #7 (SS7) network (Bates, 2002, pp. 114 – 115).

2.2.3 Interfaces and Protocols

In UTRAN, GPRS adopts the same interfaces as the CS domain. In the core network, there are several new GPRS-specific interfaces, which are described below:

- **Gn**: the interface between GSNs in the same PLMN,
- **Gi**: the interface between GGSN and the external PDNs,
- **Gp**: the interface between SGSN and the GGSN in the external PLMN,
- the GPRS extension of the MAP protocol.

A complete protocol stack in the interfaces along the communication from the UE to the external application server is illustrated in Figure 2.8.

This protocol stack is based on UMTS. However, due to the lack of the information about GPRS in UMTS, the successive discussion is based on GPRS in GSM. To establish a common reference frame for the future discussion, a protocol stack based on GSM is shown in Figure 2.9.
2.3 GPRS Data Routing and Address Conversion

As mentioned above, a GPRS network can be seen as an access network to the external IP network. The GGSN serves as an anchor while the SGSN serves as an access point. Data transfer between the SGSN and GGSN is performed by means of a tunnelling mechanism. SGSNs tunnel data to GGSN using the GPRS Tunnelling Protocol (GTP), shown in Figure 2.9, and vice versa (Eberspacher et al., 2001, p. 70). Tunnelling is only a part of the whole routing along the communication path between the external IP network and the mobile station. The routing of the whole path is performed via a series of identifiers stored in each node along the communication path, including Tunnel ID (TID), which identifies a tunnel between a SGSN and a GGSN. Simply speaking, this is a matter of address conversion.

To facilitate the discussion of address conversion and routing, the rest of this section will expand the conceptual model in Figure 2.6 into a realistic scenario. Based on this scenario, we will discuss how a packet identified by an IP address is routed to a mobile station identified by an IMSI. This includes both data transfer from the MS and data transfer to the MS.

Note that due to the lack of resource about GPRS in UMTS and inconsistency of terms between UMTS and GSM, the following discussion is based on GPRS interfaces and protocols in GSM shown in Figure 2.9 unless otherwise stated. Despite this, it still enhances understanding as the concepts are the
CHAPTER 2. GPRS — A KEY MILESTONE TOWARDS 3G

same.

2.3.1 Scenario

Based on Figure 2.6, a scenario is assumed here where two Mobile Stations, MS1 and MS2, gain access to the external Packet Data Networks (PDNs) via SGSN1 and GGSN1. In the case of MS1, it tries to gain access to three different PDNs simultaneously, namely a PPP network, an X.25 network, and the Internet. The MS2 tries to access to the Internet and a private Intranet. A detailed stack, with various key identifiers, is illustrated in Figure 2.10. Before discussing this scenario, the definition of the APN is introduced.

Access Point Name (APN) identifies each connection between the GPRS network and an external PDN (Salkintzis, 2004, p. 3-9). It is similar to a domain name used in the Internet. It is represented as $\text{PDN}\_\text{name}.\text{PLMN}\_\text{name}.\text{grps}$. PDN\_name is a sequence of labels in the form $\text{label1.label2}...$ and identifies an external PDN. PLMN\_name identifies the Public Land Mobile Network that is used to provide access to the external PDN. The encoding of PLMN\_name depends on the Mobile Country Code (MCC) and the Mobile Network Code (MNC) allocated to the given PLMN. For instance, for a PLMN with MCC = 10 and MNC = 202, PLMN\_name is $\text{mnc202.mcc010}$. To simplify the illustration, APN in Figure 2.10 is shown as APN\_n, where n is a decimal number.

2.3.2 Identifiers

As mentioned above, the conversion between the International Mobile Subscriber Identity (IMSI) and the IP address is performed by using a series of identifiers in each node. As shown in Figure 2.10, a table is assumed to contain these identifiers in each node, that is, the Mobile Station, SGSN, and GGSN. The implementation of the tables is vendor-specific, but the concept is the same. It is worth noting that the MS is identified by Packet Temporary Mobile Subscriber Identity (P-TMSI) (Kaaranen et al., 2001, p. 157) in the SGSN table while it is identified by an IP address in the GGSN table.
At the Mobile Station (MS) end, the table holds the mapping information of the NSAPI to the upper network layer. In the case of MS1, NSAPI1 corresponds to the PPP network layer, NSAPI2 to the IP network layer, and NSAPI3 to the X.25 network layer.

At the SGSN end, the SGSN does not know the IP address of the MS. It uses a Link Layer identifier to distinguish different Mobile Stations. This identifier is called the Temporary Logical Link Identifier (TLLI). A TLLI uniquely identifies a link to a Mobile Station. Thus a combination of TLLI and NSAPI (TLLI, NSAPI) uniquely determines access to an external PDN, identified by an APN. The SGSN also uses the Tunnel Identifier (TID) to uniquely identify a tunnel between the SGSN and the GGSN. Each TID has an one-to-one relationship with each duplet (TLLI, NSAPI). This means each TID uniquely corresponds to an APN. The GGSN’s IP address is also held in the table so that the SGSN can know which GGSN it will tunnel to. The table below the SGSN1 in Figure 2.10 holds all these mapping information.

At the GGSN end, opposite to the SGSN, the IMSI of the Mobile Station, and any telecommunication-specific identifier, except for the IP address, is
unknown. Its table only holds the map of the MS’s IP address and TID as well as the IP address of TID’s peer, that is the SGSN’s IP address.

This subsection has described various identifiers in brief. The following subsections will describe the data transfer in two directions to explain how the address conversion is performed using these identifiers.

### 2.3.3 Data Transfer from MS

When an application, for example a browser in MS1, prepares to send data, for example a HTTP request, to the external Packet Data Network, for example the Internet, it will go through the following steps:

- The browser delivers the request to the IP network layer as normal.
- The IP network layer encapsulates it into an IP packet with MS1’s IP address as the source address and delivers it to the SNDCP via the NSAPI2.
- The SNDCP prefixes a header to this IP packet, which contains NSAPI2. The SNDCP then delivers the data to the link layer.
- SGSN1 receives the data from the link layer, which is identified by TLLI1 in this scenario. It strips off the header, extracts the IP packet, and sends it to GGSN1 via a tunnel identified by TID4.
- GGSN1 receives the IP packet, and sends it to the Internet identified by APN2.

### 2.3.4 Data Transfer to MS

When the HTTP response in the above example is routed to GGSN1 according to MS1’s IP address, this response will be delivered to MS1 through the following steps:

- GGSN1 looks up the tables to determine the particular SGSN address and TID for the Mobile Station that is the intended recipient of the packet. In this case, it finds SGSN1 and TID4.
- GGSN1 tunnels this original packet to the SGSN1 with TID4.
• SGSN1 maps TID4 to the corresponding TLLI1 and NSAPI2 according to its table. At this point, SGSN1 knows where MS1 is and to which network application it must route the packets.

• SGSN1 takes the original packet, adds a header with TLLI1 and NSAPI2, and forwards it to the Mobile Station.

• The SNDCP layer at MS1 strips off the header and delivers the original packet to the IP network layer via NSAPI2.

• The IP network layer extracts the response and delivers it to the browser, which is waiting for this response.

2.3.5 Summary
Taking a scenario as an example, this section described how the address conversion between the IMSI and IP address is performed using a series of identifiers held by each node. An assumption in this scenario is that these identifiers have been established already. Therefore, the question arises how these identifiers are established. In GPRS, this is performed by a series of procedures, which will be discussed in the following section.

2.4 GPRS Procedures
Simply speaking, the association between the mobile station’s IP address and its IMSI is established by setting up a Packet Data Protocol (PDP) context data structure. However, before this procedure can be performed, Mobile Stations have to register with the GPRS network. In the GPRS terminology, it is called the attach procedure (European Telecommunications Standards Institute, 2000a). The opposite procedure is the detach procedure, in which the PDP context, established in the attach procedure, is deleted. The detached mobile station is not IP addressable until a new attach and PDP context set-up procedure has taken place.

2.4.1 Attach Procedure
Attach procedure is performed when Mobile Stations inform the PLMN of their presence for the purpose of using the GPRS services (European
Telecommunications Standards Institute, 2000a). This can be performed immediately after Mobile Stations have been switched on or later as users decide to use the GPRS services. This procedure is initiated by sending an attach request to a SGSN. In response, the SGSN authenticates the Mobile Station, retrieves its subscription data, and checks whether it is authorized to have access to the GPRS network from its current routing area. If none of the checks fails, the SGSN accepts the attach request of the mobile and it returns an accept message. At this point, this SGSN becomes the serving SGSN of that special Mobile Station.

The detailed attach procedure is illustrated in Figure 2.11 in the form of a message sequence diagram. There are a total of five steps to perform this procedure, which are described in detail hereafter.

**Step 1**

In this step, the Mobile Station sends an attach request message to the SGSN, which stores the routing data where the mobile station is located. The attach
request message typically includes the following information (Bates, 2002, p. 135):

- **Packet Temporary Mobile Station Identity (P-TMSI):** as described in Section 2.1.4, this is the Packet-Switched domain counterpart of TMSI. It serves as a temporary identifier for the user identity confidentiality. It is allocated by a visited Routing Area’s Visitor Location Register via a SGSN, possibly another SGSN and, possibly, in another Routing Area. Since it is stored in the nonvolatile memory (USIM) of the Mobile Station, it is used as a temporary Mobile Station identity as long as it is valid.

- **Routing Area Identity (RAI):** the identity of the Routing Area where the above P-TMSI, if present, was allocated.

- **Network Service Access Point Identifier (NSAPI):** NSAPI is specific to a particular network application at the Mobile Station. The Subnetwork-Dependent Convergence Protocol (SNDCP) layer uses this NSAPI to communicate with network applications (Bates, 2002, p. 136).

- **Information related to the Mobile Station capabilities:** supported frequency bands, multi-slot capabilities, and ciphering capabilities.

**Step 2**

Upon receiving the attach request, the new SGSN tries to acquire the permanent Mobile Station identity — that is, its IMSI. It is possible that the P-TMSI included in the attach request message has previously been allocated by the new SGSN. Then the new SGSN also knows the IMSI of the Mobile Station. Thus the new SGSN does not need to acquire this Mobile Station’s IMSI again. Otherwise, the new SGSN can acquire the IMSI according to the following sequence:

- First attempt to contact the old SGSN and request the IMSI value that corresponds to the P-TMSI included in the attach request message. This is performed in Message 2a in Figure 2.11. The address of the old SGSN can be derived by the new SGSN using the RAI included in the attach request.
• If the previous step fails, then the new SGSN acquires the IMSI from the Mobile Station itself, as illustrated in step 2b in Figure 2.11. The obvious drawbacks of this step are that it introduces additional signalling over the radio interface and that it compromises the user identity confidentiality, which is the intention of introducing P-IMSI (Salkintzis, 2004, p. 3-8).

**Step 3**

With the Mobile Station’s IMSI acquired in Step 2, the new SGSN verifies whether the user is authenticated and authorized for that particular service by checking with the Home Location Register (HLR) entry for the Mobile Station.

Typically after the authentication and key agreement procedure, ciphering is enabled on the radio interface, and therefore further messages transmitted on this interface are enciphered (Salkintzis, 2004, p. 3-9).

**Step 4**

This step is initiated by an update location message, which is sent by the new SGSN to update the HLR database with the new location of the MS. This message typically includes the following information:

- the new SGSN’s own IP address,
- its own SS7 address,
- and the IMSI value of the MS.

Subsequently, the HLR informs the old SGSN that it can now release any information stored for this Mobile Station. This is done with the cancel location message, as illustrated in Message 4b in Figure 2.11.

After releasing the resource of the old SGSN, the HLR sends the GPRS subscription data of the Mobile Station to the new SGSN, as illustrated in Message 4c. At this point, the new SGSN may perform several examinations. For example, it may check whether the Mobile Station is allowed to roam in its current Routing Area. If none of the checks fails, then the new SGSN builds up a GPRS Mobility Management (GMM) context for this Mobile
Station and returns a positive acknowledgment to the HLR. Otherwise, if an inspection routine fails, the new SGSN sends a negative response to the HLR and subsequently an attach reject message is sent to the Mobile Station. The specific reason for rejecting the attach request is included herewith.

**Step 5**

In this step, the new SGSN sends an attach accept message to the Mobile Station to indicate that the Mobile Station has been successfully registered for GPRS services. Typically, with the attach accept message, the new SGSN assigns a new P-TMSI value to the Mobile Station. A Temporary Logical Link Identity (TLLI) is also allocated to the Mobile Station as a temporary ID used by the Logical Link Control (LLC) layer. At the final step, the Mobile Station responds with an attach complete message, which acknowledges the correct reception of the new P-TMSI value. As mentioned above, at this stage the messages transmitted via radio interface are typically enciphered, and therefore, the new P-TMSI value can not be eavesdropped on (Salkintzis, 2004, p. 3-9).

A database is maintained at the SGSN that maps the Mobile Identity, IMSI, with the TLLI assigned to it and the NSAPI specific to a network application. This is like the front part of the table below SGSN1 shown Figure 2.10.

### 2.4.2 Detach Procedure

The results of the GPRS detach function is that the SGSN may delete the MM and the PDP contexts; the PDP contexts are actually deleted in the GGSN. The mobile station detaches by sending a detach request message to the SGSN. There are three detach types: GPRS-Only, IMSI-Only, and Combined. A detailed description of the detach procedure can be found in (Bates, 2002, p. 127).

### 2.4.3 Activate PDP Context

Upon completion of a successful GPRS attach procedure, a Mobile Station has been assigned a TLLI that the GPRS network knows. However, external Packet Data Networks (PDNs), for example IP networks, do not yet know
this Mobile Station. Therefore, further actions are needed for accessing an external PDNs. That is where the Packet Data Protocol (PDP) Context Activate procedure comes in.

The message flow sequence for activating a new PDP context is illustrated in Figure 2.12. When a Mobile Station wants to establish a new PDP context, it sends an Activate PDP context Request message to its serving SGSN. This message specifies all characteristics of the requested PDP context, such as (Bates, 2002, p. 142):

- specific PDP type, for example, IPv4, IPv6, X.25, or PPP, which specifies the type of the payload transferred on the PDP context,
- specific APN, which represents an external PDN,
- specific GPRS bearer, which specifies the required QoS properties.

The SGSN checks the requested APN and identifies the IP address of the GGSN that provides access to that APN by using the Domain Name System (DNS). This GGSN may be located either in the serving GPRS network or in the home GPRS network.

After that, the SGSN sends a GPRS Tunnelling Protocol (GTP) signalling a message to that GGSN to request the activation of the requested PDP context. Typically, the GGSN checks whether the Mobile Station is authorized to access the requested APN, and if so, it allocates a new IPv4 address to this PDP context.

Figure 2.12: Message Sequence Diagram in Establishing a PDP Context (Salkintzis, 2004, p. 3-11)
Note that the GGSN may request a new IPv4 address either from an internal Dynamic Host Configuration Protocol (DHCP) server or from an external DHCP server located in the requested network. In the first case, the Mobile Station is allocated an IPv4 address from the address space of the serving or home GPRS network, and the Mobile Station becomes a new IPv4 node within this network. In the latter case, however, the Mobile Station is allocated an IPv4 address from the address space of the external network, and effectively it becomes a new IPv4 node inside this PDN. It is typically used when the external network is an intranet, which may use private IP addresses.

The GGSN accepts the request to create the new PDP context and returns a positive GTP response to the SGSN with TID information. It also updates its database, like the table below GGSN1 in Figure 2.10 on page 47, where it maps the TID and the corresponding SGSN IP address.

Subsequently, the SGSN returns an accept message to the Mobile Station, which includes the IPv4 address allocated to the new PDP context. The SGSN also updates its database with the TID and the corresponding GGSN IP address, like the back part of the table in Figure 2.10 on page 47.

At this point, a new PDP context has been established. Note that the establishment of a PDP context does not involve the reservation of dedicated communication resources in the GPRS network. This applies to both the radio interface and the wired part of the GPRS network. The establishment of a PDP context involves only the storage of new information in the GPRS nodes (Salkintzis, 2004, p. 3-12). A series of associations between different identifiers is maintained in the Mobile Station, SGSN, and GGSN (Bates, 2002, p. 139) respectively, as shown in Figure 2.10 on page 47. This information is subsequently used to route the packets correlated with that PDP context, as described in Section 2.3.

### 2.4.4 Deactivate PDP Context

This procedure removes the MS’s PDP context from the GGSN. To initiate this procedure the MS sends a Deactive PDP context Request message to the SGSN. The SGSN, in turn, sends a Delete PDP context Request message to the GGSN, which removes the PDP context and returns a Delete PDP Context Response message to the SGSN. A detailed description of this
procedure can be found in (Bates, 2002, p. 148)’s book.

2.5 GPRS Handover

The previous section described the GPRS procedures in GSM context. With the conceptual understanding of these procedures in mind, which is the same as in UMTS, this section describes GPRS handovers in the context of UMTS.

The introduction of GPRS makes no difference in the UTRAN part of UMTS (Kaaranen et al., 2001, pp. 288 – 289). Referring to Figure 2.5, the Internal Handover of GPRS is the same as in the Circuit-Switched (CS) domain. As to the External Handover, the involved Mobile Switching Conter (MSC) is replaced by its counterpart in the Packet-Switched (PS) domain, SGSN. Intra-SGSN handover is similar to the Intra-SGSN handover in the sense that their procedures are the same, and only the participating entities are different: one is MSC; one is SGSN. The significant difference is presented in the case of the Inter-SGSN handover. This will be discussed in detail below.

Regarding the above mentioned difference between the Circuit-Switched domain handover and GPRS handover, Figure 2.13 shows a modified hand-
2.5. GPRS HANDOVER

Figure 2.14: Inter-SGSN Handover (Eberspacher et al., 2001, p. 251)

over diagram corresponding to the GPRS.

When the Mobile Station moves from Cell5 to Cell6 in Figure 2.13, which are controlled by SGSN1 and SGSN2, respectively, an Inter-SGSN handover occurs. The whole procedure is illustrated in Figure 2.14 in the form of a message sequence diagram.

The Mobile Station sends a Routing Area Update (RAU) Request message. At this point in time, the new Routing Area is controlled by SGSN2. SGSN2 sends a SGSN Context Request message to SGSN1, asking for the required information of the Mobile Station. The address of SGSN1 is effectively derived from the Routing Area Identity parameter included in the RAU Request message.

Upon the receipt of this SGSN Context Request Message, SGSN1 realizes that the Mobile Station has moved to another Routing Area and stops sending downlink packets to the Mobile Station.

Note that between the instant when the handover took place and the receipt of this message, an interruption period exists. During that period, SGSN1 could have been transmitting downlink packets to the Mobile Station. These packets would be unacknowledged and would need to remain buffered at SGSN1. At this point, the GGSN still believes that the Mobile Station is reachable through SGSN1 and could be transmitting new downlink packets.
to SGSN1. The latter would need to buffer these packets as well (Salkintzis, 2004, p. 3-16).

In response to the above mentioned SGSN Context Request Message, SGSN1 replies with an SGSN Context Response message carrying the requested information to SGSN2. The latter sends a SGSN Context Acknowledgement message, which confirms that it has received the requested information and is ready to receive any buffered packets for the Mobile Station that are still unacknowledged. At this point, SGSN1 forwards all the buffered packets for the Mobile Station to the SGSN2 within a new tunnel. At the same time, SGSN2 sends an Update PDP Context Request to the GGSN to inform it that any further downlink packets for the Mobile Station should be tunnelled to SGSN2. In response, the GGSN replies with an Update PDP Context Response. Now the communication path of the PDP context changes to reflect the location change of the Mobile Station. Furthermore, the location update procedures are also conducted in the HLR to reflect the current location of the Mobile Station.

Note that in the case of inter-SGSN handover, all the Logical Link Control (LLC) connections with SGSN1 in the Mobile Station are released and new LLC connections are established with SGSN2. After the short interruption required for modifying the PDP context and reestablishing the new LLC connections, the communication continues.

2.6 Conclusion

This chapter investigated various aspects of GPRS in the context of UMTS. Special focus was on the handovers in GPRS. These handovers have been categorized into several levels, according to a hierarchical structure:

- Softer Handover,
- Internal Handover,
- Intra-SGSN Handover,
- Inter-SGSN Handover.
Benefit from the same frequency characteristic of WCDMA technology, the Soft Handover, is applied through the whole hierarchical structure of the handover. This enables seamless handover in a GPRS network.

In addition to the handovers in a homogeneous GPRS network, a more sophisticated handover is called the Inter-system handover. In the context of UMTS, the Inter-system handover is between UTRAN and GERAN (Kaaranen et al., 2001, p. 369). It is a result of 2G legacy in the sense that many 3G operators provide their wide-area coverage with the GSM network and use 3G in relatively small traffic hot spots (Korhonen, 2001, p. 332). However, this concept can be expanded to include all wireless access network technologies. This introduces the concept of wireless overlay networks. If the diversity of the various wireless networks are exploited in a effective way, seamless handover between GPRS and WLAN are made possible. The next chapter will examine this concept in terms of its handover procedures and network models.
Chapter 3

Wireless Overlay Networks —
A Practical Approach towards 4G

Chapter 2 investigated the various aspects of the GPRS. As a kind of Wireless Wide Area Network (WWAN), the GPRS network renders wide coverage. However, only low-bandwidth services are offered over this wide coverage region. On the other hand, another broadly deployed wireless network, 802.11 Wireless Local Area Network (WLAN), provides a high bandwidth service but over a narrow coverage region. Although it would be desirable to provide a high-bandwidth service over a wide coverage region to mobile users all the time, no single wireless network technology simultaneously provides a low-latency, high-bandwidth, wide-area data service to a large number of mobile users (Stemm & Katz, 1998).

As pointed out by Stemm and Katz (1998), the solution to the above-mentioned goal is:

“Use a combination of wireless networks to provide the best possible coverage over a range of geographic areas.”

In a more general sense, this combination, spanning in-room, in-building, campus, metropolitan, and regional cell sizes, fits into a hierarchy of network interfaces which is called a wireless overlay network structure. To make this wireless overlay network structure look like a single network, seamless handover between heterogenous networks is required.
This chapter provides a brief examination of the wireless overlay network in terms of its handovers, and network models. Finally, since 802.11 and GPRS offer characteristics that complement each other perfectly, the integration of the GPRS network and 802.11 WLAN is discussed in terms of its implementation architectures.

3.1 Handover in Wireless Overlay Network

Figure 3.1 illustrates a scenario of a wireless overlay network. Lower layers consist of wireless cells with higher bandwidth that cover a relatively small area. Higher layers in the hierarchy provide a lower bandwidth over a larger geographic area.

When a mobile node moves in such a wireless overlay network structure, it needs to switch to different wireless subnetworks to keep communication continuous. This switching procedure is called handover. Relative to this wireless overlay network structure, there are two kinds of handovers that need to be distinguished: horizontal and vertical handover, as illustrated in Figure 3.2.
3.1. HANDOVER IN WIRELESS OVERLAY NETWORK

3.1.1 Horizontal Handover

A horizontal handover is defined as a handover between base stations that are using the same type of wireless network interface, as shown in Figure 3.2(a). It is also called intra-technology handover (Vidales et al., 2005). This is the traditional definition of handover for homogeneous cellular systems such as cellular telephony systems, wide-area data systems, and wireless local area networks. Section 2.5 (see page 56) intensively discussed this kind of handover in the GPRS network. In the case of 802.11 WLAN, many other articles focus on the process involved in this kind of handover (Mishra, Shin, & Arbaugh, 2003).

3.1.2 Vertical Handover

A new type of handover, which is significantly different from the above-mentioned handover, is defined as the vertical handover. This is a handover between base stations that are using different wireless network technologies (also called an inter-technology handover). The term vertical derived from the overlay network structure that has networks with increasing cell sizes at higher levels in the hierarchy, as shown in Figure 3.2(b). This hierarchical
CHAPTER 3. WIRELESS OVERLAY NETWORKS

structure corresponds to the levels shown in Figure 1.3 on page 10. In general, the Wireless Personal Area Network (WPAN) covers an in-room size cell; the WLAN covers a in-building or campus size cell; the WWAN covers the metropolitan-area cell; and the satellite wireless communication system covers the largest region-area cell.

Depending on the direction in which the handover occurs, the vertical handovers can further be divided into two categories: an upward vertical handover is a handover to a wireless overlay with a larger cell size (and lower bandwidth per unit area), and a downward vertical handover is a handover to a wireless overlay with a smaller cell size (and higher bandwidth per unit area). A vertical handover may be to an immediately higher or lower overlay, or the mobile host may “skip” an overlay. For example, a mobile may hand over from an in-room network directly to a wide-area network, or vice versa.

3.1.3 Comparison between Horizontal and Vertical Handover

There are some important differences between the horizontal handover and the vertical handover that might affect the strategy for implementing vertical handovers:

• In horizontal handover systems, a mobile host performs a handover from cell A to cell B while moving out of the coverage area of cell A into the coverage area of cell B. This is not necessarily the case in vertical handover. For example, when a user performs an upward vertical handover from an in-room cell A to an in-building cell B, the user is moving out of the coverage of cell A. However, when a user performs a downward vertical handover from cell B to cell A, the user is not moving out of the coverage of cell B. This implies that downward vertical handovers are less time-critical, because a mobile can always stay connected to a upper overlay while handing over to a lower overlay.

• Many network interfaces have an inherent diversity that arises because they operate at different frequencies. For example, a room-size overlay may use infrared frequencies, a building-size overlay network may use one set of radio frequencies, and a wide-area data system may use another set of radio frequencies. Another way in which diversity exists is
3.1. **HANDOVER IN WIRELESS OVERLAY NETWORK**

in the spread spectrum techniques of different devices. Some devices use Direct Sequence Spread Spectrum (DSSS), while others use Frequency Hopping Spread Spectrum (FHSS). Taking advantage of this diversity between network interfaces, some enhancements can be done in order to reduce the discovery time.

- In a network of homogeneous base stations, the choice of “best” base station is usually obvious: the mobile chooses the base station with the highest signal strength after incorporating some threshold and hysteresis. In a multiple-overlay network, the choice of the “best” network cannot usually be determined by channel-specific factors such as signal strength because different overlay levels may have widely varying characteristics. For example, an in-building RF network with a low signal strength may yield better performance than a wide-area data network with a high signal strength. There are also considerations of monetary cost (some networks charge per minute, others per byte) that do not arise in a homogeneous handover system.

Depending on the above-mentioned specific characteristics of the horizontal and vertical handovers, utilizing them in a strategic way leads to a scenario of seamless handover. This scenario is discussed next.

### 3.1.4 Seamless Handover

Figure 3.3 depicts a scenario with various handovers, each numbered uniquely. Number 1 is a horizontal handover; numbers 2 and 4 are upward vertical handovers; numbers 3, 5, and 6 are downward vertical handovers. The decision state for each handover is identified by the different patterns in the figure.

As mentioned above, it is not necessarily the case that a vertical handover is due to move out of the coverage of a wireless network. An intersection state exists where the mobile node has physical links to both the upper and lower networks, or the same level networks in the case of horizontal handover, for example, handover 1 in Figure 3.3. This results in a make-before-break handover mechanism becoming possible (Dutta et al., 2004), which presents a seamless handover.

The idea behind the make-before-break mechanism is that the mobile node connects to the upper larger coverage network in advance when it de-
tects that it is losing connection to the lower smaller coverage network. For a short period the mobile node simultaneously connects to both networks. When the mobile node completely loses the connection to the lower network, it does not suffer from data loss because it has had a connection to the upper network ready. In this way, the seamless handover is achieved when the mobile node spans the gap between the lower networks, like the space between numbers 2 and 3, as well as between numbers 4 and 5 in Figure 3.3.

This section described the concept of seamless handover in wireless overlay networks. The next section will discuss several different architectural models that can be implemented toward a wireless overlay network.

### 3.2 Models of Wireless Overlay Network

Several basic models are implemented towards a wireless overlay network, using multiple wireless networks. This is illustrated in Figure 3.4 using two wireless networks as an example, network A and network B. These models adopt the concept of a hybrid network. A hybrid core network and several wireless access networks, each with its own service optimization, are used
3.2. MODELS OF WIRELESS OVERLAY NETWORK

The hybrid core network processes the interaction between different wireless access networks while the selected wireless access network delivers the service to end users.

The distinguishing factor between these models is the layer on which the wireless networks communicate (Wu & Havinga, 2001). Many derivatives of these models can be found (Walsh et al., 2000).

The rest of this section will describe the three basic models depicted in Figure 3.4.

3.2.1 Tunnelled Networks

Figure 3.4(a) illustrates this network model. In this model, a user has independent service agreements with operators of several access systems. Upper layers access the different technologies independently. According to a certain policy, the best network is selected and the hybrid core tunnels the traffic across the Internet and the chosen radio access network to the mobile node. Thus, no modifications are required to the existing network stacks.

The hybrid functionality is provided with the hybrid core, which resides in the home network of the mobile node. For example, the home agent based on Mobile IP can be used to support the hybrid functionality. The two main disadvantages of this are that the hybrid core will often duplicate functions of the access networks, such as authentication, and all data transactions involve multiple round trips across the Internet, increasing network load and service
3.2.2 Hybrid Networks

Figure 3.4(b) illustrates this model. In this model, the individual wireless access networks implement the three bottom layers: physical, link, and network layers. There is a hybrid core that interfaces between the Internet and the different wireless access networks. In this case, the agreements between the wireless access network operators and the hybrid core operator must be more permanent. An advantage of the hybrid network is that it is able to offer advanced transport services for the network layer. For example, the hybrid core may implement smooth hand-over support for inter access network mobility. The main drawback of this model is that the networking activities are duplicated. Despite this, the service latency is reduced relative to the tunneled network model because there is not as much duplicated functionalities as in tunneled networks.

3.2.3 Heterogeneous Networks

In this model there is a common core layer that deals with all network functionality and operates as a single network to the upper layers. Thus, different wireless access networks implement only the physical and link layers, which are specifically relevant to each technology. Communication between wireless access networks belonging to the common core is based on lower layers. This reduces the overhead, and improves performance. A major obstacle of this model is that the different access networks must converge, which requires a huge standardization effort and operator commitment. Figure 3.4(c) illustrates this model.

When implementing a wireless overlay network, the integration of 802.11 WLAN and 3G cellular networks, especially the UMTS network, warrants special attention due to their wide deployment. Efforts made in this regard are described in the next section.


3.3 Integration of WLAN and 3G Cellular Network

A feasibility study on the integration of the 3G system and 802.11 has been done by the 3GPP TSG working group. They identified a total of six scenarios (European Telecommunications Standards Institute, 2003), which can be divided into four levels of integration (Vidales et al., 2005). The levels are open coupling, loose coupling, tight coupling, and fully integrated. Of the four levels, most research was done on tight coupling and loose coupling, as shown in Figure 3.5. These levels will be respectively discussed in detail below.

3.3.1 Tight Coupling Interworking

The tight coupling approach aims to make the 802.11 network appear to the 3G core network as another 3G access network (Buddhikot et al., 2003). Thus the 802.11 gateway appears to play the role of an SGSN, for example, in the case of UMTS. It means this gateway would implement the interface Gb and emulate functions which are natively available in the UTRAN, as shown in Figure 3.5. This hides the details of the 802.11 network to the 3G
core network. Similar to the above-mentioned heterogeneous network model, this makes these two different networks appear to be a single network for the outside. They would share the same authentication, signalling, transport and billing infrastructures, independent from the protocols used at the physical layer on the radio interface. In this sense, the approach can be seen as an example of the heterogeneous network model.

However, the mobile nodes in this approach are required to implement the corresponding 3G protocol stack on top of their standard 802.11 network cards, and switch from one physical layer to the next as needed. All the traffic generated by the mobile nodes in the 802.11 network is injected into the 3G core network.

Furthermore, this approach renders several disadvantages both on the network and terminal side (Buddhikot et al., 2003). On the network side, the 3G core network directly exposes its interfaces to the 802.11 network, and the same operator must own both the 802.11 and the 3G parts of the network. In fact, in this case, independently operated 802.11 networks could not be integrated with 3G networks. Current 3G networks are being deployed using carefully engineered network-planning tools, and the capacity and configuration of each network element are calculated using mechanisms which are specific to the technology utilized over the air interface. By introducing the 802.11 traffic directly into the 3G core, the setup of the entire network, as well as the configuration and the design of network elements, have to be modified to support the increased load.

On the terminal side, this approach presents the following issues:

- As mentioned above, the 802.11 network cards would need to implement the 3G protocol stack.

- It would constrain the use of 3G-specific authentication mechanisms, based on Universal Subscriber Identity Module (USIM) or Removable User Identity Module (R-UIM) cards, for authentication on WLANs, forcing 802.11 providers to interconnect to the 3G carriers’ SS7 network to perform authentication procedures.

- This would also imply the use of 802.11 network interface cards with built-in USIM or R-UIM slots or external cards plugged separately into the subscriber devices.
For the reasons described above, special care should be taken when choosing this architecture to implement the integration of the 802.11 WLAN and 3G cellular network. The complexity and the high cost of the reconfiguration of the 3G core networks and of the 802.11 gateways could force operators that chose the tightly-coupled approach to become uncompetitive (Buddhikot et al., 2003).

### 3.3.2 Loose Coupling Interworking

This approach can be seen as an example of the tunneled network models mentioned previously. This architecture also introduces a gateway into the 802.11 network, as shown in Figure 3.5. However, unlike the tight coupling architecture, this gateway connects to the Internet and does not have any direct link to 3G network elements such as PDSNs, GGSNs or 3G core network switches. The user population that accesses services of the 802.11 gateway may include users that have locally signed on, as well as mobile users visiting from other networks. This approach is called loose coupling interworking because it completely separates the data paths in the 802.11 and the 3G networks. The high speed 802.11 data traffic is never injected into the 3G core network, but the end user still achieves seamless access.

In this approach, different mechanisms and protocols can handle authentication, billing and mobility management in the 3G and 802.11 portions of the network. However, for seamless operation to be possible, they have to inter-operate. In the case of inter-operation with CDMA2000, this requires that the 802.11 gateway supports Mobile IP functionalities to handle mobility across networks.

Since the UMTS standards do not yet include Mobile IP, more adaptation is required to integrate with UMTS networks. Chapter 7 will discuss the issues relevant to offering a Mobile IP service in GPRS. Mobile IP services would need to be tailored to the GGSNs to enable seamless mobility between 802.11 and UMTS. Common subscriber databases would need to interface with Home Location Registers (HLR) for authentication and billing on the UMTS side of the network. These problems will be addressed in Chapter 7.

Advantages of the loose coupling integration approach are as follows:

- It allows the independent deployment and traffic engineering of 802.11
and 3G networks. 3G carriers can benefit from other providers’ 802.11 deployments without extensive capital investments. At the same time, they can continue to deploy 3G networks using well-established engineering techniques and tools.

- While roaming agreements with many partners can result in widespread coverage, including key hot-spot areas, subscribers benefit from having just one service provider for all network access. They no longer need to establish separate accounts with providers in different regions, or with providers who cover different access technologies.

- Finally, unlike the tight coupling approach, this architecture allows an organization to provide its own public 802.11 hot-spot, inter-operate through roaming agreements with public 802.11 and 3G service providers, or manage a privately installed enterprise WLAN.

From the aforementioned it becomes clear that the loose coupling approach offers several architectural advantages over the tight coupling approach, with almost no drawbacks. Therefore, this study promotes the loose coupling approach as the preferred architecture for the integration of WLAN with UMTS networks, in particular, with GPRS networks.

### 3.4 Conclusion

This chapter introduced the concept of the wireless overlay network, and investigated the basic models for implementing this concept. The advantages and disadvantages of each model were also discussed. As a practical implementation, the integration of the 802.11 WLAN and the GPRS network received more attention. It has been argued that a loose coupling architecture was more suitable for this integration.

Now, the appropriate architecture has been chosen. The question arises which mechanisms will fulfill the role of glue in such a loose coupling architecture to enable the various networks to work together and enable seamless handover between them. The answer is dominantly prone to be the IP, which has proved its huge success in the wired network. The next chapter will examine the various aspects of the IP, demonstrate why it is justified that the IP fulfills this role, and what problems exist towards fulfilling such a role.
Chapter 4

IP — An Interworking Role in Overlay Networks

In Chapter 2, GPRS has been introduced, a technology which provides data service in mobile networks. In Chapter 3, the concept of a wireless overlay network has been introduced, which is considered as an appropriate architecture for seamless handover in heterogenous networks. To achieve seamless handover, a mechanism that integrates all heterogeneous network access technologies is needed. As pointed out by Guardini et al. (2000):

“A key role of IP in future mobile data systems will probably be the provision of efficient and cost effective interworking between overlay networks.”

The vitality of the IP technologies has been demonstrated by the exponential growth of the Internet. This chapter attempts to answer which characteristics justify IP to fulfil an interworking role in future mobile systems. It also investigates the existing problems towards fulfilling this role.

For the purposes mentioned above, the rest of this chapter is organized as follows: firstly, an architectural model of IP networks section will be introduced. Thereafter the design principles of IP will be discussed. This is followed by the introduction of several IP’s functionalities highly relevant to this study, including the addressing and routing of IP, and mapping IP addresses to physical addresses. The existing IP drawbacks will be investigated next. Finally, the IP network security will be discussed.
4.1 Architectural Model

The advent of the Internet is primarily for the purpose of interconnecting many disparate physical networks and making them function as a coordinated unit (Comer, 1991, p. 1). As such, an interconnection scheme, or internetwork (also called internet), is introduced to hide the low-level details and make the collections of networks appear to be a single large network. Actually, an internet is a logical network, as opposed to the low-level physical network, built on a collection of physical networks (Peterson & Davie, 2000, p. 248). It has its own data packet format and addressing scheme. A new node, also called a router, or called a gateway, is introduced to interconnect two or more different networks, which shapes the architectural model, as illustrated in Figure 4.1.

A boundary can be distinguished between the hosts and the entities that only provide network functionalities. This division separates the hosts from the network subsystem. This leads to the question of which functionalities should be put where. The next section seeks to answer this question from an IP design principle perspective.
4.2 IP Design Principles

Perhaps the most important distinction between the Internet and traditional telecommunication networks is how the systems are designed. This distinction is captured in the following statement:

“The traditional telecomms approach is to design everything as part of a single process, leading to what is conceptually a single standard (in reality, a tightly coupled set of standards). Building a new system will thus involve the design of everything from top to bottom from scratch (and thus it is often called the ‘Stovepipe Approach’). By contrast, the IP approach is to design a ‘small’ protocol that does one particular task, and to combine it with other protocols (which may already exist) in order to build a system.” (Wisely et al., 2002, p. 5 – 6)

As opposed to ‘the Stovepipe Approach’ of the traditional telecommunication network, we define the approach used to design the Internet as ‘the IP approach’ in the context of this study. The above mentioned distinction of the IP approach can be captured by an IP design principle, namely ‘always keep layer transparency’, or by the phrase, “IP over everything and everything over IP” (Wisely et al., 2002, p. 6). An hourglass depicts this concept (Deering, 1998), as illustrated in Figure 4.2.
Another distinction between the Internet and traditional telecommunication networks is where the functionality is placed (Wisely et al., 2002, p. 6). Traditional telecommunication networks put much functionality within the network while, in contrast, the Internet places functionality as close to the edge of the network as possible, to keep the network as simple as possible. This is captured by another IP design principle: ‘always think end to end’.

Two key principles, layering and the end-to-end principle, have been mentioned above. However, the next subsection begins with what is probably the more fundamental principle: connectivity.

### 4.2.1 Connectivity

Providing connectivity is the key goal of the Internet (Wisely et al., 2002, p. 77). The Internet’s original intention was that the network would provide connectivity, even if large parts of the network were destroyed by enemy actions. This connectivity is built on the low-level physical network’s connectivity that is structured in such a way that each of these networks carries IP data packets. Routers connect these networks together and route the IP packets to their destination address. Here, the routers employ a connectionless packet route. IP packets, independent of the physical network type, have the same common format and common addressing scheme. Thus, it is easy to take a packet from one type of network and send it over another network.

In such a connectionless network, there is no need to establish a path for the data through the network before data transmission (Wisely et al., 2002, p. 80). There is no state information stored within the network about particular communications. Instead, each packet of data carries the destination address and can be routed to that destination independently of the other packets that might make up the transmission. Thus if any node is destroyed, packets would still be able to find alternative routes through the network. No state information about the data transmission could be lost, as all the required information is carried with each data packet.

Since the Internet concentrates on connectivity, it supports not just a single service like telephony but a whole host of applications all using the same connectivity. This principle promotes that the network layer should focus on transport only, and leave other functionalities on the edge of the network, as previously mentioned. It also promotes a thin IP network, as
4.2. IP DESIGN PRINCIPLES

illustrated in Figure 4.2.

Next, consider the other two key principles: first the end-to-end principle, then the principle of layering.

4.2.2 End-to-End Principle

This principle is first stated by Saltzer, Reed, and Clark (1984):

“The function in question can completely and correctly be im-
plemented only with the knowledge and help of the application
standing at the endpoints of the communication system. There-
fore, providing that questioned function as a feature of the com-
munication system itself is not possible. (Sometimes an incom-
plete version of the function provided by the communication sys-
tem may be useful as a performance enhancement.)”

This statement proclaims that only end systems can correctly perform functions that are required from end-to-end, such as security and reliability. Therefore, these functions should be left to the end systems (Wisely et al., 2002, p. 81). End systems are the hosts that actually communicate, such as a PC or mobile phone. This end-to-end approach removes much of the complexity from the network, and prevents unnecessary processing, as the network does not need to provide functions that the terminal will need to perform for itself.

This end-to-end principle is often reduced to the concept of the ‘stupid’ network, as opposed to the telecommunications concept of an ‘intelligent net-
work’ (Wisely et al., 2002, p. 83). The end-to-end principle means that the basic network deals only with IP packets and is independent of the transport layer protocol. This confirms the above mentioned connectivity principle from another side, that is, the network layer should focus on transport only.

4.2.3 Layering and Modularity

The layering principle can be summarized as follows:

“Layered protocols are designed so that layer n at the destination
receives the same object sent by layer n at the source.” (Comer,
1991, p. 149)
This allows each layer to focus on its own functionalities at a time, without worrying about how lower layers work. The protocols are easy to implement and understand. This is a structured way of dividing functionalities to reduce or hide complexity. Each layer offers specific services to upper layers, while hiding the implementation detail from the upper layers. Ideally, there should be a clean interface between each layer. This simplifies programming and makes it easier to change implementation of any individual layer.

When this concept is applied in the vertical direction, it is called modularity. Any protocol performs one well-defined function at a specific layer. These modular protocols can be reused. Ideally protocols should be reused wherever possible, and functionality should not be duplicated. Avoiding duplication also makes it easier for users and programmers to understand the system (Wisely et al., 2002, p. 83).

According to the above mentioned layering concept, a layering model of the Internet is shown in Figure 4.3. This model is often referred to as the TCP/IP Internet Layering Model. It will be referred to frequently later when describing a solution in IP networks.

Besides the above mentioned network layering model specific to IP networks, there is also another general model promoted by the International Organization for Standardization (ISO), namely the Open Systems Interconnection (OSI) reference model (Stallings, 2004, p. 27). This model will be referred to frequently when describing non-IP networks.

Many books provide information on these two models. This study does not intend to repeat that information here. A good description of the OSI model can be found in (Stallings, 2004, pp. 27 – 38), and of the TCP/IP model in (Comer, 1991, pp. 146 – 147). However, for the purpose of compar-
4.2. IP DESIGN PRINCIPLES

Figure 4.4: A Comparison of the OSI and TCP/IP Models (Stallings, 2004, p. 40)

ison, Figure 4.4 roughly presents correspondences in functionality between two models.

4.2.4 Summary

The previous subsections described the three principles of IP, respectively. They are highly interrelated. Achieving connectivity is the goal while the end-to-end principle is a consequence of progressing towards this goal. The network thus focuses on providing connectivity only, all other application-specific functionalities are thus left to the end systems to process. The principle of layering and modularity is the mechanism to align with the previous two principles. When applying the TCP/IP model, in Figure 4.3, to the Internet architecture in Figure 4.1, we can get a better understanding of this concept. This is shown in Figure 4.5.

The network boundary is clearly identified in the figure. Inside the network boundary, there are routers and subnetworks, which cooperate with each other to offer connectivity. Outside the network boundary, there are hosts which process the application specific functionalities, based on the connectivity offered by the networks.

In this architecture, the great advantage of the layering principle is that it renders transparency between different layers. This allows changes to protocol components without having to do a complete update of all protocols. This is particularly important when dealing with the heterogeneity of networking.
technologies. There is a big range network types with different capabilities, and different application types with various capabilities and requirements. By providing the linchpin — the internet layer — it is possible to hide the complexities of the networking infrastructure from users and concentrate on purely providing connectivity (Wisely et al., 2002, p. 86). This has led to the catchphrase ‘IP over Everything and Every-thing over IP’.

Therefore, these three principles will be fundamentally followed in this study. However, they are not absolute. Perhaps the best description of the philosophy about the principles is as follows:

“Principles that seemed inviolable a few years ago are deprecated today. Principles that seem sacred today will be deprecated to-morrow. The principle of constant change is perhaps the only principle of the Internet that should survive indefinitely.” (Carpenter, 1996)

Thus, “engineering feedback from real implementations is more important than any architectural principles” (Carpenter, 1996). This is seen as ‘the principle of the principles’ in discussing IP networks.

The previous sections looked at IP from a conceptual perspective. The next section describes IP addressing and routing, while Section 4.4 will explain how IP addresses are mapped to physical addresses.
Figure 4.6: Formats of Internet Protocol (IP) Datagram and Address

4.3 IP Addressing and Routing

As mentioned in Section 4.1, an internet can be seen as a logical network, built on a collection of physical networks. It has its own packet format and addressing scheme, as illustrated in Figure 4.6.

This study does not intend to describe each field, included in the figure above, in detail here. A good description of the IP datagram format can be found in (Comer, 1991, pp. 91 – 106), and of the IP address format in (Comer, 1991, pp. 62 – 70). To understand the problem that arises with IP addresses when considering mobility, some concepts about IP addresses are nevertheless worth discussing.

IP addresses are hierarchical, which means that they are made up of several parts that correspond to some sort of hierarchy in the internetwork (Peterson & Davie, 2000, p. 262), as shown in Figure 4.6(b). Specifically, IP addresses consist of two parts, a network part and a host part. The network part of an IP address identifies the network to which the host is attached; all hosts attached to the same network have the same network part in their IP addresses. The host part then identifies each host uniquely in that particular network.
The hierarchical structure of IP addresses facilitates the routing of an IP datagram. The routing algorithm determines the next hop address based on the network part of the destination IP address of the IP datagram. If the destination host and sender lie on the same physical network, the sender directly delivers the IP datagram to the destination’s physical address via the network interface. To see if the destination lies on the same physical network, the sender extracts the network part of the destination IP address and compares it to the network part of its own IP address. A match means they are in the same physical network. This situation is called direct routing in IP terminology. As to how the senders know the physical address of the destination, it is a matter of address resolution, which will be addressed in the next section (Comer, 1991, p. 111).

The opposite of direct routing is indirect routing in which the destination is not on a directly attached physical network with the sender. This forces the sender to pass the datagram to a gateway, or router, for delivery. In this case, the next hop is a router. The sender directly delivers the original datagram to that router’s physical address. It is important to understand that the datagram source and destination addresses remain unaltered. Upon the receipt of that datagram, the router thus can repeat the above mentioned comparison and calculate out the next hop address based on its routing table. This process is repeated until the datagram reaches the router which is on the same physical network as the destination. At this point, the router performs direct routing and directly delivers the datagram to the destination (Comer, 1991, p. 116).

Typically, a routing table contains pairs (N, G), where N is the IP address of a destination network, and G is the IP address of the next router along the path to the network N (Comer, 1991, p. 113). It is important to know that the routing table contains only network addresses, not the individual host IP addresses. This makes routing efficient and keeps routing tables small. However, it also renders a constraint on the IP network:

“If a host moves from one network to another, its IP address must change” (Comer, 1991, p. 65).

Although host-specific routing can solve this problem, it causes the routing table to expand rapidly. This is especially in the case of the Internet,
4.4 Mapping IP Address to Physical Addresses

As mentioned above, before a sender can directly deliver the IP datagram to the next hop, it needs to know the next hop’s physical address. This is a matter of mapping high-level addresses to physical addresses, also known as the address resolution problem (Comer, 1991, p. 74). There are several solutions to this problem. The Address Resolution Protocol (ARP) is the easiest and most efficient solution.

Figure 4.7 demonstrates the idea behind the ARP, which is explained as follows: when host A wants to resolve IP address $I_B$ to a physical address, it broadcasts a special data frame that asks the host with IP address $I_B$ to respond with its physical address, $P_B$. All hosts in the same physical network, including B, receive the request, but only host B recognizes its IP address and sends a reply that contains its physical address. When A receives the reply, it uses the physical address to send the IP datagram directly to B. In this way, the IP addresses are dynamically resolved without the requirement of a centralized database (Comer, 1991, p. 75).
4.4.1 Proxy ARP

Proxy ARP lets a proxy agent answer ARP requests on behalf of a host on another network. This makes the sender believe that the proxy agent is the destination host, while in fact the destination host is on another network. The router often plays the role of proxy agent for the destination host, relaying packets to it from other hosts.

4.4.2 Gratuitous ARP

Another feature of ARP is called gratuitous ARP. It occurs when a host sends an ARP request looking for its own IP address. This is usually done when the interface is configured at bootstrap time.

While IP allows for advantages in terms of connectivity, end-to-end and modular, it also has some drawbacks. These are discussed in the next section.

4.5 IP Drawbacks

As mentioned early, the birth of the Internet was due to the need for global connectivity. Moreover, the Internet community embraced a realistic philosophy in the sense that “it is better to adopt an almost complete solution now, rather than to wait until a perfect solution can be found” (Carpenter, 1996). Consequently, the Internet was born with just a set of protocols, not a complete architecture or a network design (Wisely et al., 2002, p. 7). Providing global connectivity to the users is its major goal. To become a complete architecture, other key elements are needed. These elements are examined below. This includes billing and management, quality of service, address Exhaustion, IP mobility, session management, and security issues.

4.5.1 Billing and Management

With the exponential growth of the user traffic entering the network, it is obvious that a mechanism should exist to limit the user traffic and manage user activities. A reasonable approach is by means of charging. The users pay for what they occupy. Thus a billing engine and management platform are needed. Here the following three problems need attention:
4.5. IP DRAWBACKS

- **Authentication:** a validation of a user’s identity, in other words, confirm you are whom you claim to be (Whitman & Mattord, 2003, p. 312).

- **Authorization:** the granting of specific types of service (including “no service”) to a user, based on their authentication, what services they are requesting, and the current system state. Authorization may be based on restrictions, for example time-of-day restrictions, or physical location restrictions, or restrictions against multiple logins by the same user. Authorization determines the nature of the service which is granted to a user. Examples of types of service include, but are not limited to: IP address filtering, address assignment, route assignment, QoS/differential services, bandwidth control/traffic management, compulsory tunnelling to a specific endpoint, and encryption.

- **Accounting:** the tracking of the consumption of network resources by users. This information may be used for management, planning, billing, or other purposes. Real-time accounting refers to accounting information that is delivered concurrently with the consumption of the resources. Batch accounting refers to accounting information that is saved until it is delivered at a later time. Typical information that is gathered in accounting is the identity of the user, the nature of the service delivered, when the service began, and when it ended.

These three problems are often termed as the Authentication Authorization Accounting (AAA) architecture (Laat, Gross, Gommans, Vollbrecht, & Spence, 2000). A lot of IETF effort has been placed on this area. A AAA authorization framework is proposed in RFC 2904 (Vollbrecht et al., 2000b). Some application examples of this framework are provided in RFC 2905 (Vollbrecht et al., 2000a). Moreover, the requirements that AAA protocols must meet are specified in RFC 2906 (Farrell et al., 2000).

4.5.2 Quality of Service

The Internet provides a “best effort” transport service. There is, however, no guarantee that the packets definitely reach their destinations without delay and in the correct order. Nevertheless, due to diversity of the application services, there are different constrains in terms of quality of service that the
IP layer provides. Some require absolute correctness with no concern for time limits, for example, file download service, while some require arrival within the limited time even if some packets are lost or wrong, for example, voice services. All these elements are defined as Quality of Service (QoS), which determines the degree of satisfaction of the user of a service (Wisely et al., 2002, p. 201). In fact, IP is fundamentally unsuited for delivering packets within a time limit. Adding this functionality into the IP is a very active topic in the research community.

Several QoS mechanisms have been developed within the IETF, which are listed below,

- Integrated Services (IntServ) in RFC 1633 (Braden, Clark, & Shenker, 1994),
- Multi-Protocol Label Switching (MPLS) in RFC 3031 (Rosen, Callon, & Viswanathan, 2001),
- Differentiated Services (DiffServ) in RFC 2475 (Blake et al., 1998),
- Integrated Services over Specific Link Layers (ISSLL) in RFC 2816 (Ghanwani, Pace, Srinivasan, Smith, & Seaman, 2000), and
- Resource Reservation Protocol (RSVP) in RFC 2205 (Braden, Zhang, Berson, Herzog, & Jamin, 1997), RFC 2208 (Mankin et al., 1997).

4.5.3 Address Exhaustion

Global addressing is one of the Internet’s goals. Unfortunately, there are not enough IP addresses to allocate to every host, since the address field is limited to 32 bits. To address this problem, a new version, IPv6 (Deering & Hinden, 1998), is introduced to expand the address space to 128 bits.

4.5.4 IP Mobility

Traditionally, IP-based networks have been designed with the assumption that they are fixed. Although the concept works well in a fixed environment, when it comes to mobility, a major problem arises from the fact that an IP address has two contradictory semantics associated with it. On the one
hand, the IP address serves as an identifier of the mobile terminal for the upper protocols — this should be the same regardless of the sub-network it attaches to. On the other hand, the IP address also serves as a locator of the mobile terminal for routing the datagram to its destination — this requires change so as to reflect its current position of sub-network that it has moved to, as mentioned in Section 4.3.

A lot of effort has been made to solve this problem. This will be discussed in detail in Chapter 5.

4.5.5 Session Management

Because IP connectivity is just a socket on a computer, it is quite often the case that applications on different terminals are incompatible in some way. For example, when trying to set up something like a real time voice call, it requires quite a lot of negotiation on coding rates and formats, etc. In addition, the user’s IP address will change at each login, which means that it is almost impossible to locate individuals instantly for setting up a voice session. What is needed is a way of identifying fixed users, binding them more rapidly to one changing IP address, and then negotiating sessions (Wisely et al., 2002, p. 8).

For the purpose of identifying users, a standardized method, Network Access Identifier (NAI), is proposed in RFC 2486 (Aboba & Beadles, 1999). For the purpose of negotiation, an early protocol suite, H.323, is developed by ITU (Wisely et al., 2002, p. 128). Another much more recent proposal by IETF is the Session Initiation Protocol (SIP) (Handley, Schulzrinne, Schooler, & Rosenberg, 1999; Handley & Jacobson, 1998).

4.5.6 Security Issues

There are many serious security flaws inherent in the TCP/IP protocols suite. Some of these flaws exist because hosts rely on IP source address for authentication. Others exist because network control mechanisms, and in particular routing protocols, have minimal or non-existent authentication (Bellovin, 1989).

For defending these security issues, security mechanisms exist at each layer of the TCP/IP model shown in Figure 4.3. This is not a case of re-
peating functionalities. The word ‘security’ actually contains many different functions, and different types of security are employed to solve different security issues, as shown in Table 4.1. Generally, true security requires what is known as ‘security in depth’, that is, if one wants security, one should not rely on just on one mechanism to achieve it (Wisely et al., 2002, p. 107).

Table 4.1: Security Mechanisms at Different Layers of the TCP/IP Model (Wisely et al., 2002, p. 108)

<table>
<thead>
<tr>
<th>Layer</th>
<th>Example Security Mechanisms</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application</td>
<td>SET Digital Signature</td>
<td>Enables private, authenticated transactions. Relies on certificate infrastructure.</td>
</tr>
<tr>
<td>Transport</td>
<td>SSL</td>
<td>Enables data encryption. Relies on certificate infrastructure.</td>
</tr>
<tr>
<td>Network</td>
<td>IPSec &amp; AAA protocols Firewalls</td>
<td>Protects the network. Protects data across the network.</td>
</tr>
<tr>
<td>Link</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical</td>
<td>Coding schemes, Physical Isolation</td>
<td>Especially useful for wireless links which are easily tapped</td>
</tr>
</tbody>
</table>

The next section will comprehensively investigate security in IP networks while Chapter 6 will focus on a specific IP Security Architecture, IPSec (Kent & Atkinson, 1998c).

4.5.7 Summary

This section has investigated the drawbacks of the IP approach in a comprehensive manner. Also some proposed methods have been introduced in brief. It is worth noting that some approaches to solving these drawbacks involves ‘weakening’ the basic IP design principles mentioned in Section 4.2. For example, adding QoS to some routers fattens the waist of the hourglass depicted in Figure 4.2 on page 75, that is, the IP layer, which weakens the end-to-end principle. IPSec has the same consequence. As stated by Deering (1998), this will lead to a much fatter waist, as illustrated in Figure 4.8.
Thus care should be taken when seeking approaches to solve these problems. With these basic principles in mind, this study focuses on two of the above mentioned problems: mobility and security, while the other problems are not addressed. Furthermore, the problems that arise when solutions to these two problems are combined will also be addressed. Chapter 5 will focus on mobility while Chapter 6 will focus on security. Then next section investigates security in IP networks in a comprehensive way.

4.6 IP Network Security

The previous section has investigated the weaknesses of the IP from a system perspective. This section will focus on the IP network security issues only. In a sense, IP network security issues arise from the inherent weakness in the design, configuration, implementation, or management of a network or system (Canavan, 2001, p. 25). In the context of information security, this weakness is called as a vulnerability. Any deliberate act that exploits vulnerabilities is called an attack (Whitman & Mattord, 2003, p. 64), which presents a threat against the operation, functioning, integrity, or availability of a network or system. To defend any threat, the controls, also called safeguard, or countermeasure, should be in place. This includes security mechanisms, policies, procedures, or technologies which can successfully counter attacks, reduce risk, and resolve vulnerabilities (Whitman & Mattord, 2003, p. 27).
4.6.1 Information Security

When looking at the essence of network security, we discover that it is the information, and our ability to access that information that we want to protect, and not the computers and networks (Canavan, 2001, p. 9). Thus a more appropriate term to use is information security. Information Security is defined as the protection of information, and the systems and hardware that use, store, and transmit that information according to the concept developed by the National Security Telecommunications and Information System Security Committee (NSTISSC) (Whitman & Mattord, 2003, p. 9).

From an information security perspective, therefore, the purpose of the security controls mentioned above is to protect the value of the information over the IP network from being compromised in terms of availability, accuracy, authenticity, confidentiality, integrity, utility, and possession (Whitman & Mattord, 2003, p. 10). However, in the context of the network security, only availability, authenticity, confidentiality, and integrity characteristics make sense (Canavan, 2001, p. 9). Thus, a threat is anything that could lead to the violation of the above mentioned characteristics of the information (Schafer, 2003, p. 6), while an attack is seen as the actual implementation of a threat.

IP networks are often divided into two parts: the outside world, known as the untrusted network, for example, the Internet, and the inside world, known as the trusted network, for example, the intranet of an organization (Whitman & Mattord, 2003, p. 276). Most controls defend against attacks from the outside. However, it is worth noting that often threats come from inside an organization: “We have found the enemy and it is us” (Canavan, 2001, p. 10). This section, however, still focuses on defense against outside attacks.

This study attempts to propose a framework in a mobile environment, where it is often the case that inside users move to the outside. The question arises how, in this situation, can users be provided with continuous access to inside information over an untrusted network? In other words, how is the availability of the information ensured? Furthermore, availability should be ensured without the violation of any of the other characteristics, namely, authenticity, confidentiality and integrity.

After all technology countermeasures have been deployed, we have not
necessarily provided a secure network system. It is important to remember that information security is not simply a matter of technology. It is concerned with software, hardware, data, people, and procedures. There are many management side elements involved, such as policy, awareness, training and education (Whitman & Mattord, 2003, p. 10). Generally, controls against threats are provided in three levels: managerial, operational, and technical (Whitman & Mattord, 2003, p. 221). However, this study focuses on the technical aspect only. A more comprehensive introduction to all aspects of information security can be found in the book, “Principles of Information Security” (Whitman & Mattord, 2003).

Coming back to the attacks mentioned previously, intensive research efforts have been made in this area. Bellovin (1989) widely investigated the security problems in the TCP/IP protocol suite. Chakrabarti and Manimaran (2002) focused on the network infrastructure, and presented a taxonomy of existing attacks. This section does not attempt to repeat this work in a comprehensive manner, but will introduce some attacks relevant to Mobile IP and IPSec. Thereafter some countermeasures against these attacks will be introduced. The effect of these countermeasures might be reduced due to the introduction of Mobile IP and IPSec, or might even be incompatible with them.

### 4.6.2 Attacks

As mentioned above, an attack leads towards a threat to the characteristics of the information. In reality, a specific attack often leads to a combination of threats towards several characteristics of the information (Schafer, 2003, p. 8). An actual hacker will often combine several attacks to construct a threat towards the characteristics of the information. There are two general categories of attacks (Schafer, 2003, p. 9):

- **Passive attack:** there is no obvious activity that can be monitored or detected in this type of attack. Attackers stealthily eavesdrop on protocol data units. It is often used to monitor and record the traffic on the network, usually in order to gather information that can be used later in active attacks (Canavan, 2001, p. 26). Examples of this attack type include packet sniffing and traffic analysis.
• Active attack: attackers delay, replay, delete, modify, or insert data into the protocol data units. As the name implies, this kind of attack employ more overt actions on the network. Examples of this type of attacks are denial of service attack (DoS) or active probing of systems and networks.

A hacker often uses passive attacks to obtain information to launch an active attack. Thus, the discovery of passive attacks can be regarded as a premonition of an active attack. This feature is often exploited by the Intrusion Detection System (IDS) (Whitman & Mattord, 2003, p. 286).

In this section, a passive attack, namely sniffer, and two active attacks, spoofing and DoS, will be introduced, as well as general countermeasures against them.

**Sniffer**

A sniffer is a program or device that can monitor data travelling over a network (Whitman & Mattord, 2003, p. 69). Sniffers can be used for legitimate network management functions as well as for stealing information from a network. When it is exploited by illegitimate users to gain internal information, the confidentiality of the information is compromised. Sniffers can be extremely dangerous to a network’s security, as they are virtually impossible to detect and can be inserted almost anywhere.

A measure against this attack type is employing cryptology to encrypt traffic over the network. Many approaches exist, such as Transport Layer Security (TLS) (Dierks & Allen, 1999), Secure Socket Layer (SSL) (Freier et al., 1996), Virtual Private Network (VPN), etc.

**Spoofing**

Spoofing is a technique used to falsify one’s identity or masquerade as some other individual or entity to gain unauthorized access to a system or network (Canavan, 2001, p. 30). A broad category of spoofing has been identified, including IP address spoofing, session highjacking, DNS spoofing, sequence number spoofing, and replay attacks.

Among them, IP address spoofing receives particular attention in this study. To engage in IP address spoofing, a hacker must first use a variety
of techniques to find an IP address of a trusted host and then modify the packet headers so that it appears that the packets are coming from that host. Spoofing attacks often compromise the authenticity and integrity of information transmitted over the networks. Some spoofing attacks are denial of service oriented (Binkley & Richardson, 1998), which leads to the loss of availability of the attacked systems.

One example of IP address spoofing is “smurf” attacks. In such an attack, the attacker sends one or many ping packets to an IP directed-broadcast address with an IP source address that may also be at the destination site. For example, if a site had a site specific class C address along the lines of 192.168.1.0, the attacking IP destination might be 192.168.1.255 or 192.168.1.0 (0 broadcast addresses may be used as well) with an IP source of 192.168.1.1. The result is that two systems may be attacked. The IP source itself is bombarded with ping reflections from all the systems at the directed broadcast address. Further the smurf vehicle could also be used for single packet “ping of death” attacks (Binkley & Richardson, 1998). The defense against this kind of attacks will be addressed in Section 4.6.3.

**Denial of Service**

Denial of Service (DoS) attacks are designed to shut down a system or network, or render it inoperable (Canavan, 2001, p. 38). The goal of DoS is not to gain access to information but to make a system or network unavailable for legitimate users, which compromises the availability of the information. There are many different types of DoS attacks. Some are listed below:

- Ping of death,
- Synchronize sequence number (SYN) flooding,
- Spam,
- Smurf.

This dissertation will discuss two types of countermeasures against these threats. The next section describes firewalls, while Section 4.6.4 introduces virtual private networks (VPNs) as a countermeasure.
4.6.3 Firewall

To protect internal systems from the attacks mentioned above, firewalls are often deployed on the border of networks to prevent a specific type of information from moving between the untrusted network and trusted network, and vice versa (Whitman & Mattord, 2003, p. 276).

Firewalls have made significant advances since their earliest implementations. The first generation of firewall devices were routers that perform only simple packet filtering operations. More recent generations of firewalls offer increasingly complex capabilities. At present, firewalls have evolved into the fifth generation, with many types available. These types of firewalls can be implemented in a wide variety of architectures. This section will firstly introduce various types of firewalls. Thereafter, deployment architectures will be introduced.

Types of Firewalls

Firewalls can be categorized in different ways, for example, by the layer of the OSI model at which they operate, by the technology that they implement, or by the general approach they employ (Canavan, 2001, p. 214). This subsection will introduce them according to the time that they were born, that it, generation by generation.

Packet filtering firewall is the earliest firewall. It simply filters packets based on their headers as they travel to and from the organization’s networks.

Application-level firewall is also called proxy server, since it runs special software designed to serve as a proxy for a service request. The application-level firewall is often a dedicated computer separate from the filtering router and very commonly used in cooperation with a filtering routers.

Stateful inspection firewall, like the packet filtering firewalls, also performs packet filtering, but it takes it a step further. It keeps track of each connection established between internal and external networks using a state table. These state tables track the state and context of each packet in the session, by recording which station sent what packet and when. A stateful firewall can restrict incoming packets by denying access to packets that are responses to internal requests. If the stateful firewall receives an incoming packet that cannot be matched in its state table, then it defaults to its Access
Control List (ACL) to determine whether to allow the packet to pass.

**Dynamic packet filtering firewall** takes a further step than stateful inspection firewalls. It allows only a particular packet with a particular source, destination, and port address to enter through the firewall. It does this by understanding how the protocol functions, and by opening and closing “doors” in the firewall, based on the information contained in the packet header. With this functionality, dynamic packet filters are intermediate, between traditional static packet filters and application proxies.

**Kernel proxy** is a specialized form that works under the Windows NT Executive, which is the kernel of Windows NT. It evaluates packets at multiple layers of the protocol stack, by checking security in the kernel as data is passed up and down the stack.

### Firewall Architectures

Each of these firewalls can be configured in a number of network connection architectures. These approaches are sometimes mutually exclusive and sometimes can be combined. Each is examined in more detail below.

**Packet filtering routers** are the border routers between the organization’s internal networks and the external service provider, which are equipped with packet filtering firewall function (Whitman & Mattord, 2003, p. 279). They filter packets that the organization does not allow into the network. They often use the source IP address to make the distinction between allowed and unallowed packets. This is a simple but effective means to lower the organization’s risk to external attack. Drawbacks to this type of system are a lack of auditing and strong authentication. IP header information is furthermore unreliable as it can be set by the source, or any intermediary to any arbitrary value (Binkley & Richardson, 1998). Thus it can not solely defend the attacks mentioned previously. In addition, the complexity of the access control lists, used to filter the packets, can grow and degrade network performance. Figure 4.9 shows this configuration.

To defend against the IP address spoofing attacks described in Section 4.6.2, packet filtering routers often enforce a security policy that packets arriving at an outside interface must have an outside source address, or must not have an inside source address. Likewise, packets leaving from inside must have an inside source address, or must not have an outside source address. This way,
one may prevent spoofing attacks from originating from the inside of a network (Binkley & Richardson, 1998). This is called network ingress filtering (Ferguson & Senie, 1998). Mobile IP is specifically affected by ingress traffic filtering, which will be intensively discussed in Section 5.2 on page 113.

**Screened host firewall systems** combine the packet filtering router with a separate, dedicated firewall, such as an application proxy server. This approach allows the router to prescreen packets to minimize network traffic and load on the internal proxy. The application proxy examines an application layer protocol, such as HTTP, and performs the proxy services. This separate host is often referred to as a bastion host, as it represents a single, rich target for external attacks, and should be very thoroughly secured. An advantage is that the proxy requires the external attack to compromise two separate systems, before the attack can access internal data. In this way the bastion host provides more protection for the data than the router alone. Figure 4.10 shows this configuration.

**Dual-homed host firewalls** contain two Network Interface Cards (NICs), rather than the one contained in the bastion host configuration, as illustrated in Figure 4.11. One NIC is connected to the external network, and one is connected to the internal network, providing an additional layer of protection. With two NICs all traffic must physically go through the firewall to move between the internal and external networks.

**Screened subnet firewalls** consist of two or more internal bastion hosts, behind a packet filtering router, with each host protecting the trusted network, as illustrated in Figure 4.12. There are many variants of the screened
subnet architecture. The first general model consists of two filtering routers, with one or more dual-homed bastion hosts between them. The screened subnet is an entire network segment that performs two functions: it protects the DMZ systems and information from outside threats; and it protects the internal networks by limiting how external connections can gain access to internal systems. Although extremely secure, the screened subnet can be expensive to implement and complex to configure and manage. To apply the architecture, the value of the information it protects must justify the cost.

The architecture of a screened subnet firewall provides a DMZ. The DMZ can be a dedicated port on the firewall device linking a single bastion host, or it can be connected to a screened subnet. Until recently, servers providing services through an untrusted network were commonly placed in the DMZ.
Examples include Web, file transfer protocol (FTP), and certain database servers. More recent strategies using proxy servers have provided much more secure solutions.

4.6.4 Virtual Private Network

Firewalls protect the internal network from the outside attacks. However, when users move to the outside, they will not be under protection of their firewalls. A mechanism is required to ensure the availability of the internal information for them in such a way that authenticity, confidentiality and integrity are not compromised. Firewalls solely can not satisfy this demand. That is where the Virtual Private Network (VPN) can play a role.

A VPN is a means of transporting traffic in a secure manner over an unsecured network (Canavan, 2001, p. 205). A VPN often gains this ability by using some combination of encryption, authentication, and tunnelling. The following is the four most commonly used protocols to construct VPN:

- Point-to-Point Tunnelling Protocol (PPTP),
- Layer 2 Tunnelling Protocol (L2TP),
- Internet Protocol Security (IPSec), and
4.7 Conclusion

IP is considered as a mechanism to interwork heterogenous networks. This chapter demonstrated it from a design principle perspective and then investigated its drawbacks. Two of these drawbacks are the focus of this study, namely mobility and security. IP mobility will be discussed in Chapter 5. With the solving of IP mobility problem, it is possible to enable seamless handover in the selected GPRS and WLAN interworking architecture as discussed in Chapter 3. The remaining problem toward this study’s goal is security.

Security is a very wide-scope problem. This chapter briefly examined IP network security in terms of attacks and countermeasures against these attacks, for example, firewalls. Of popular firewall deployment architectures, screened subnet firewall architecture was chosen as the base of the proposed framework. In addition, to protect traffic between the internal domain and the Mobile Node, visiting external domains in such an architecture, a VPN is considered to be a popular mechanism. IPSec, as one of the most popular VPN mechanisms, will be introduced in detail in Chapter 6.
Chapter 5

Mobile IP

Chapter 4 has concluded that IP would play an interworking role in the future mobile network. In that chapter, it has been pointed out that a major problem towards this role is that IP was originally designed to internetwork fixed networks, with the assumption that the hosts are fixed. The hosts in IP networks are identified and located based on IP addresses. On the one hand, as an identifier, the IP address should never be changed, otherwise the upper session will break. On the other hand, as a locator, the IP address should change to reflect the host’s new point of attachment when it moves. Many solutions attempt to address this problem. Mobile IP (MIP), proposed by the IETF, is the best-known.

This chapter will give a broad overview to enhance understanding of Mobile IP in Section 5.1. Furthermore, this chapter will expand to discuss Mobile IP in terms of reverse tunnelling and route optimization. Finally, a hierarchical approach to IP mobility will be discussed.

5.1 Overview of Mobile IP

To solve the above-mentioned IP addressing problem, the fundamental idea behind Mobile IP is that a Mobile Node (MN) should have two IP addresses (Perkins, 1996b). They are:

- Home Address, which is permanently assigned to a Mobile Node, and acts as its identifier.
- Care-of Address (CoA), which is temporarily assigned to a Mobile Node when it moves.
Node, and acts as its locator reflecting its actual location.

Thus, in basic Mobile IP operation, a Correspondent Node (CN), which is going to communicate with a Mobile Node, addresses the Mobile Node via its Home Address (HoA), while the Mobile IP handles the Mobile Node’s movement and enables the routing of packets to its actual location so that the mobility of the Mobile Node is hidden from the Correspondent Node. This operation requires the introduction of two extra entities into the IP network infrastructure, namely:

- **Home Agent (HA)**, which maintains location information for Mobile Nodes. It resides in the Mobile Node’s home network and stands for the Mobile Node, when that Mobile Node is away from home to intercept packets addressed to the Mobile Node’s Home Address and tunnels them to its CoA.

- **Foreign Agent (FA)**, which resides in the Mobile Node’s visited network, namely the foreign network. It provides routing services to the Mobile Nodes that have registered with it. The Foreign Agent detunnels and delivers the packets, which were tunnelled by the Mobile Node’s Home Agent, to the Mobile Node. For the packets sent by a Mobile Node to a Correspondent Node, the Foreign Agent may serve as a default router for registered Mobile Nodes.

There are two Mobile IP versions corresponding to the two versions of IP: Mobile IPv4 (MIPv4) (Perkins, 2002) and Mobile IPv6 (MIPv6) (Johnson, Perkins, & Arkko, 2004). MIPv6 is based on the same principles as MIPv4 but is more powerful (Dimopoulos & Venieris, 2004, p. 5-15). It is well integrated into the new IPv6 protocol and exploits its facilities for overcoming several drawbacks encountered in MIPv4. However, as mentioned in Chapter 1, this study builds the framework on IPv4 networks. Thus this chapter concerns Mobile IPv4 only. In the context of this study, the notations, Mobile IP and MIP, refers to the general Mobile IP functionalities common to both MIPv4 and MIPv6.

A typical network configuration with Mobile IP support is illustrated in Figure 5.1. It is assumed that the Mobile Node in this configuration is equipped with 802.11 WLAN access ability. This configuration will be
5.1. OVERVIEW OF MOBILE IP

When the Mobile Node moves to a foreign network, the CoA can be acquired by the Foreign Agent in that foreign network, or by alternative mechanisms, such as the Dynamic Host Configuration Protocol (DHCP) (Droms, 1997). They are respectively denoted as:

- Foreign Agent Care-of Address (FA-CoA), which comprises of one of the Foreign Agent’s addresses. In this way, all visiting Mobile Nodes to a foreign network can share a few addresses, which saves the scarce IP address resource. This approach is represented in light gray in Figure 5.1.

- Co-located Care-of Address (Co-CoA), which comes from the IP address pool of the foreign network. Each Mobile Node has a dedicated IP address. This approach does not require the presentence of the Foreign Agent in the foreign network. It is represented in dark gray in Figure 5.1.

As shown in Figure 5.1, there is a remarkable triangle along the communication path $CN \rightarrow HA \rightarrow MN \rightarrow CN$. In Mobile IP’s terminologies, it is referred to as Triangle Routing, which contributes to one primary disadvantage of Mobile IP. This is that unnecessary load is placed on the network due to

Figure 5.1: A Typical Mobile IP Network Configuration (Dimopoulou & Venieris, 2004, p. 5-3)
to indirect routing (Dimopoulou & Venieris, 2004, p. 5-10). This problem will be intensively discussed in Section 5.3. This overview section focuses on the basic Mobile IP procedures only.

In the basic Mobile IP operation, three phrases can be distinguished: agent discovery, registration, and routing (Dimopoulou & Venieris, 2004, p. 5-3). The rest of this section will discuss them in sequence.

5.1.1 Agent Discovery

Agent Discovery is a procedure where the mobility agents, Home Agent and Foreign Agent, advertise their availability on each link so that a newly attached Mobile Node can determine the current point of attachment, and identify the services provided by the agents (Perkins, 1996b). More specific, the mobility agents periodically broadcast an Agent Advertisement message on the link, or this message is sent in response to an Agent Solicitation message sent by a Mobile Node for advancing the Agent Discovery procedure. Through the information provided by the advertisements, a Mobile Node is able to detect if it is in the same network or has moved to a new network. This is referred to as Move Detection. If the Mobile Node determines that it has migrated to a new network, it can acquire a new Care-of Address (CoA). The new CoA can be obtained from the CoA address list following the Agent Advertisement message, called Foreign Agent CoA (FA-CoA), or through other mechanisms, for example, DHCP, called Co-located CoA (Co-CoA) in this case. Besides facilitating the move detection and CoA acquisition, the Agent Advertisement also indicates the services offered by the mobility agents via some bit setting.

The Agent Advertisement is an ICMP Router Advertisement message (Deering, 1991), that is followed by a Mobility Agent Advertisement Extension, and optionally, by other extensions. Its format is shown in Figure 5.2.

Based on the format offered here, the rest of this subsection will explain how it facilitates the move detection, and CoA acquisition. Other services provided will also be indicated.
5.1. OVERVIEW OF MOBILE IP

Move Detection

Two primary algorithms are used by Mobile Nodes to detect whether they have migrated from one subnetwork to another. The first is based on the Lifetime field within the main body of the ICMP Router Advertisement portion of the Agent Advertisement, as shown in Figure 5.2. A Mobile Node records the Lifetime received in any Agent Advertisement. If an Agent Advertisement is absent from the same agent for longer than the specified Lifetime, it can be deduced that the Mobile Node has moved to a new network. If, at the same time, an Advertisement is received from another agent, it registers with this new agent. Otherwise, it immediately sends an Agent Solicitation message that leads to the timely discovery of the agent (Perkins, 1996b).

The second method for move detection uses network prefixes. This is based on the Prefix-Lengths Extension. The prefix length is applied to the router addresses, which are found in the router advertisement part of the message, and can serve as the identification of the subnetwork. If the agent prefix differs from the prefix of the Mobile Node’s CoA (FA-CoA or Co-CoA), the Mobile Node assumes that it has moved, and then registers with the new agent. Otherwise, it assumes that it is still located in the same network.

Care-of Address Acquisition

When a Mobile Node recognizes that it has moved to a new foreign network, through the above-mentioned Move Detection algorithms, it acquires a Care-of Address (CoA) in the foreign network. This address can be obtained.
from the CoA address list appending the Agent Advertisement message, as illustrated in Figure 5.2. In this case, it is FA-CoA. It can also be obtained from other mechanisms, for example, DHCP. In this case, it is called Co-CoA.

Other Services

Besides facilitating Move Detection and CoA acquisition, the Agent Advertisement, through its flag bits depicted in Figure 5.2, indicates other services supported by the mobility agents:

- F and H indicate the agent’s role on the network, Foreign Agent, Home Agent or both.

- M and G indicate the type of encapsulations supported by the Foreign Agent. M corresponds to the Minimal encapsulation (Perkins, 1996c) and G to the Generic Routing Encapsulation (GRE) (Farinacci, Li, Hanks, Meyer, & Traina, 2000). Note that these two encapsulation types are optional while IP in IP encapsulation (Perkins, 1996a) is mandatory for all mobility agents.

- T indicates the Foreign Agent supports the reverse tunnelling. This will be discussed in Section 5.2.

- R indicates that registration with a Foreign Agent is required, even if in the Co-CoA mode. This is needed for the purpose of authentication.

5.1.2 Registration

After a Mobile Node (MN) detects that it has moved to a new network and has acquired a new Care-of Address (CoA), it must inform its Home Agent (HA) of this CoA, so that the latter can tunnel the packets destined for its Home Address to the Mobile Node. This can be achieved through the registration procedure. In this procedure, authentication is conducted to establish mobility security association. This might take place between a Mobile Node and Foreign Agent, between a Mobile Node and Home Agent, as well as between a Foreign Agent and Home Agent (Dimopoulou & Venieris, 2004, p. 5-6).

In addition, a registration procedure is also launched when the Mobile Node needs to renew its soon-to-expire registration with its HA. Therefore,
regardless of whether the Mobile Node migrates to a new network or not, it is engaged in sending renewal registrations to its Home Agent for resetting the lifetime of its registration (Perkins, 1996b).

The registration procedure is performed via two messages exchanged between the Mobile Node and its Home Agent: a Registration Request (from Mobile Node to Home Agent) and a Registration Reply (from Home Agent to Mobile Node). Depending on the CoA mode, the Foreign Agent might be located between Home Agent and Mobile Node. The Foreign Agent will process the information of interest and relay messages to their actual destinations. However, the Foreign Agent might be not involved in this procedure if the Mobile Node is using a Co-located CoA (Co-CoA), unless the Foreign Agent has set the R bit of its Agent Advertisement, as mentioned in Section 5.1.1.

In the rest of this section, it is assumed that the registration is performed through the Foreign Agent, unless otherwise stated. The complete description of direct registration with the Home Agent can be found in RFC 2002 (Perkins, 1996b).

For a clear understanding of how the registration procedure establishes an association between a Mobile Node’s Home Address and its CoA, two conceptual data structures are introduced here, as shown in Figure 5.3. One is a visitor list, which resides in the Foreign Agent, and accommodates the associations between the Mobile Node’s Home Address and its physical address. The other is a mobility binding list, which resides in the Home Agent, and accommodates the association between the Mobile Node’s Home Address and its CoA.

The Registration Request message format is illustrated in Figure 5.4. A complete description of each field can be found in RFC 2002 (Perkins, 1996b). The following discussion will directly reference these terms without providing a detailed explanation.

Roughly speaking, the registration procedure can be divided into the establishment of the association in the above-mentioned data structure, authentication between communicating entities, and the determination of expire time, which will be discussed hereafter, respectively.
Association Establishment

Upon receipt of a Registration Request from a Mobile Node, that is, message 1a in Figure 5.3, one of the main tasks of the Foreign Agent is to establish an association between the Mobile Node’s physical address and its the Home Address. Because this request packet is sent with the Mobile Node’s Home Address as the IP source address, the Home Address can be extracted from the IP header. Otherwise the Home Address can be directly determined from the registration request message’s Home Address field shown in Figure 5.4. The physical address is determined from the link layer data frame. An association between the Mobile Node’s physical address and its Home Address is thus established in the visitor list of the Foreign Agent. In this case, an entry is inserted into the visitor list, mapping the Mobile Node’s Home Ad-

![Figure 5.3: Registration Procedure in Mobile IP](image1)

![Figure 5.4: Registration Request Message Format (Perkins, 1996b)](image2)
dress, 172.17.1.19, to its physical address, “89 E3 A2 73 2B A8”. However, the actual update of the Foreign Agent’s visitor list takes place when it receives a Registration Reply from the Home Agent indicating the success of the procedure, that is, message 2a.

The Foreign Agent relays the Registration Request message to the Home Agent using the address from the Home Agent field in Figure 5.4, as demonstrated by message 1b in Figure 5.3. The Home Agent updates its mobility binding list for the Mobile Node. It adds a new entry, and either removes the existing ones if the S bit is not set, or maintains them in conjunction with the new one if the S bit is set. If the Mobile Node was previously located in its home network, a mobility binding is created because the Home Agent did not maintain any location information for the Mobile Node until that moment. In this case, the new entry maps the Mobile Node’s Home Address to its CoA, 207.102.6.15.

The deregistration procedure follows a similar message format and flow when the Mobile Node returns to its home network. The exception is that the lifetime is set to zero in the registration request message. This will result in the deletion of all mobility binding entries, for the Mobile Node in the home agent, if the Care-of Address field is set to its Home Address, or the deletion of only the CoA indicated by the Care-of Address field, if this field is set to that particular CoA.

Apart from the registration and deregistration procedures performed when the Mobile Node leaves or returns to its home network, it is also necessary that appropriate Address Resolution Protocol (ARP) procedures should be performed to enable other hosts in the home network to contact the Mobile Node.

When the Mobile Node moves away from its home network, hosts in the home network will not have valid ARP cache entries for that Mobile Node from that moment on. Once the Home Agent receives and accepts the Registration Request message, it starts gratuitous ARP, which was discussed in Section 4.4.2 on page 84. Hosts on the Home Network will be notified that packets destined to the Mobile Node’s Home Address should be delivered to its own physical address, that is, the Home Agent. In this way, the Home Agent will be able to intercept such packets and tunnel them to the Mobile Node. Accordingly, the Home Agent also starts proxy ARP, which
was discussed in Section 4.4.1 on page 84, for replying to ARP requests for
the Mobile Node (Dimopoulou & Venieris, 2004, p. 5-8 ).

Similarly, when the Mobile Node returns to its home network, it should
perform gratuitous ARP for updating the ARP caches of hosts on the Home
Network. This associates the Mobile Node’s own physical address with its
home IP address to enable the receipt of packets destined to its Home Ad-
dress. The Mobile Node then sends the Deregistration Request to its Home
Agent. Upon receipt and successful processing of the message, the Home
Agent stops using proxy ARP for that Mobile Node and once again performs
gratuitous ARP (Dimopoulou & Venieris, 2004, p. 5-8 ).

**Authentication**

Authentication is performed via various authentication extensions, includ-
ing Mobile-Home authentication extension, Mobile-Foreign authentication
extension and Foreign-Home extension (Perkins, 1996b). These extensions
append to the fixed portion of the registration request message. Among
them, one extension should always be present in both request message and
reply message, that is, Mobile-Home authentication extension. This exten-
sion requires that a mobility security association exists between the Mobile
Node and the Home Agent and is used for authenticating registration data
exchanged between these entities.

Certainly, when these authentication extensions are present at the same
time, a specified order should be obeyed. For example, a Mobile Node, after
securing the Registration Request message with a Mobile-Home Authentica-
tion Extension, may add other extensions to be processed by the Foreign
Agent and can secure these extensions with a Mobile-Foreign Authentication
Extension. The Foreign Agent, when receiving the message, should process
and remove any extension after the Mobile-Home Authentication Extension,
before relaying it to the Home Agent. Furthermore, it can add any informa-
tion to be processed by the Home Agent, which is secured with a Foreign-
Home Authentication Extension. The same logic also applies to the Reply
messages sent in the reverse direction. It is important, however, that the
Foreign Agent should not modify any of the fields up to the Mobile-Home
Authentication Extension because it is likely that the Home Agent will not
be able to authenticate this part of the message (Dimopoulou & Venieris,
5.1. OVERVIEW OF MOBILE IP

2004, p. 5-7).

**Expire Time Determination**

The Lifetime field in Figure 5.4 indicates the number of seconds remaining before the registration is considered expired. This value is initially set by the Mobile Node when sending a registration request, taking into account the value of the registration lifetime field set by the Foreign Agent in its Agent Advertisement as illustrated in Figure 5.2. However, the registration lifetime is decided not only by the Mobile Node, but needs the approval of the Home Agent as well. If the Home Agent specifies a smaller lifetime value in the Registration Reply, the Foreign Agent will, upon the receipt of this registration replay, reset its timer to align to the lifetime indicated by the Home Agent, and then relay this reply to the Mobile Node. Rather than resetting its timer, the Mobile Node simply will decrease the registration’s remaining lifetime by the difference of the two lifetime values. By so doing, it will ensure that the registration lifetime expires at the Mobile Node before it does at the Home Agent and Foreign Agent. Consequently, a renewal registration will be transmitted early enough for updating the mobility bindings of the Foreign Agent and Home Agent (Dimopoulou & Venieris, 2004, p. 5-8).

5.1.3 Routing

After the Mobile Node (MN) has registered with its Home Agent (HA), it is able to receive routing services from the Home Agent with, when needed, the cooperation of the Foreign Agent (FA) for packets addressed to its Home Address (HoA). The Home Agent, in particular, intercepts packets destined to the Mobile Node’s Home Address and tunnels them to the registered Care-of Address (CoA). The end of the tunnel might be a Foreign Agent or the Mobile Node itself, depending on the CoA mode employed. If the Home Agent supports the simultaneous binding feature, and the Mobile Node has registered more than one CoA, the Home Agent tunnels a copy to each one of them. The default mechanism used for tunnelling is IP in IP encapsulation (Perkins, 1996a), which is mandatorily supported by all Mobile IP entities, Home Agents, Foreign Agents, and Mobile Nodes. It is also possible that an alternative type of encapsulation, such as Minimal encapsulation (Perkins,
1996c) or GRE encapsulation (Farinacci et al., 2000), can be requested by the Mobile Node and agreed upon during the registration procedure. Figure 5.5 takes IP in IP encapsulation as example to demonstrate the message sequence and encapsulation format. As mentioned previously, there is a remarkable triangle routing along the communication path between the Correspondent Node (CN) and the Mobile Node. Depending on the CoA mode employed, either Foreign Agent CoA (FA-CoA) or Co-located CoA (Co-CoA), there are two possible communication loops:

- \( A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow B_3 \) if FA-CoA is used, and
- \( A_1 \rightarrow A_2'' \rightarrow B_3 \) if Co-CoA is used.

In the case that a Foreign Agent is present, the encapsulated packet, \( A_2 \), reaches the Foreign Agent first, which checks its visitor list for an entry matching the inner packet’s destination address, that is, the Mobile Node’s Home Address (HoA). If such an entry is found, the Foreign Agent delivers the inner packet, \( A_3 \), to the Mobile Node by addressing the Mobile Node’s physical address. This physical address becomes known to the Foreign Agent during either the agent discovery from either an Agent Solicitation, or during the registration procedure from the Registration Request message transmitted by the Mobile Node. This entry is created during the registration procedure, as described in Section 5.1.2. If the Foreign Agent has no entry for the Mobile Node’s Home Address in its visitors list, then it simply discards this packet.

In the reverse direction, standard IP routing is used for packets transmitted by the Mobile Node. As shown in Figure 5.5(a), \( B_1 \) is directly sent by the Mobile Node to the Correspondent Node (CN), with the CN’s address as destination address, and its own Home Address (HoA) as source address.

Note that the Mobile Node must be configured with a default router to serve the next hop for outgoing packets. If it has registered a FA-CoA, it simply uses its Foreign Agent as the first-hop router for outgoing packets. The Foreign Agent’s physical address is known from the agent’s advertisements. Alternatively, the Mobile Node can choose its default router from the Router Addresses advertised in the ICMP Router Advertisement portion of the Agent Advertisement. These two methods are also used by Mobile Nodes that use a Co-CoA (Dimopoulou & Venieris, 2004, p. 5-9).
Mobile IP routing services are used when the Mobile Node is away from home. However, if the Mobile Node returns to its home network, it operates exactly as any other stationary host. Packets addressed to the Mobile Node’s Home Address reach the Mobile Node through standard IP routing. The Home Agent maintains no mobility binding and thus is not involved in the routing of datagrams.

This section provided an overview of Mobile IP. The next section will further discuss reverse tunnelling in Mobile IP while Section 5.3 will pay attention to route optimization in Mobile IP’s triangle routing.

## 5.2 Reverse Tunnelling

As mentioned in Section 4.6.3 on page 95, ingress filtering takes place in most networks' border routers to interrupt packets with a source IP address that appears topologically incorrect. In this way, malicious users are not able to impersonate other legitimate users when attached to ingress-filtering domains (Ferguson & Senie, 1998). However, ingress filtering prevents the standard operation of Mobile IP. Mobile Nodes residing in a foreign network...
CHAPTER 5. MOBILE IP

are likely to send packets that have their Home Addresses as the source IP address.

Reverse tunnelling seeks to address this problem by introducing the use of a reverse tunnel, having the Care-of Address (CoA) of the Mobile Node (MN) as a starting point, and the Home Agent (HA) as an endpoint (Montenegro, 1998). In this way the topologically incorrect Home Address of the Mobile Node is hidden from the foreign network’s ingress filtering firewall. Mobile Nodes are aware that Foreign Agents (FA) support reverse tunnelling by detecting if the T flag is set in their Agent Advertisements. Consequently, they request this service during their registration through the selected Foreign Agent by setting the corresponding flag in the Registration Request message. Likewise, Home Agents must accept the reverse tunnel support and should be ready to decapsulate packets originating from the Mobile Node’s CoA.

In Foreign Agent CoA mode, there are two styles of reverse tunnelling, as illustrated in Figure 5.6, namely:

- Direct Delivery Style,
- Encapsulating Delivery Style.

In Direct Delivery Style, the Mobile Node specifies the Foreign Agent as its default router for outgoing IP packets. From the Mobile Node’s point of view, packets to be reverse tunnelled are processed in the same way as any other packet. The destination address of such packets is the Correspondent Node’s IP address and the source address is the Mobile Node’s Home

Figure 5.6: Reverse Tunnelling
5.2. **REVERSE TUNNELLING**

Address. The Foreign Agent, in turn, is responsible for examining these packets by checking their source IP address, and for tunnelling them to the Mobile Node’s Home Agent. Compared with Encapsulating Delivery Style, the benefit of this style is that no extra processing is required at the Mobile Node and that packets consume less bandwidth on the wireless link between the Mobile Node and the Foreign Agent because they are not encapsulated. However, it lacks flexibility because all packets will be directed to the Home Agent and there is no way for the Mobile Node to avoid reverse tunnelling at runtime (Dimopoulou & Venieris, 2004, p. 5-13).

In the Encapsulating Delivery Style, the Mobile Node does not send packets directly to the Correspondent Nodes, but rather encapsulates them to the Foreign Agent. The Foreign Agent is responsible for checking that the packets have been sent by a Mobile Node that has requested this service. Then the packets are decapsulated and finally re-encapsulated to the Mobile Node’s Home Agent. The Home Agent, in both schemes, will decapsulate the packets and forward them based on the destination address of the original packet, that is, the Correspondent Node’s address.

The benefit of this style is that it allows the Mobile Node to indicate at runtime whether reverse tunnelling should be used by the Foreign Agent. In this way, if the Mobile Node, which has requested the Encapsulating Delivery Style service, sends packets to its Correspondent Node without encapsulating them first to the Foreign Agent, the latter does not perform reverse tunnelling to these packets. As a consequence, it will forward them based on standard IP routing mechanisms.

A Mobile Node requests this service by adding an Encapsulating Delivery Style Extension (Montenegro, 1998) to the Registration Request message. This extension is added after the Mobile-Home Authentication Extension and is removed by the Foreign Agent after being processed, because it affects only the Foreign Agent’s operation. In all cases, encapsulation is performed according to the forward tunnel configuration. This means that the same encapsulation type and the same tunnel endpoints are used for both forward and reverse tunnels.

If the Mobile Node uses Co-located CoA (Co-CoA) mode, the Mobile Node itself tunnels the original packets to its Home Agent, with its CoA as the source address and its Home Agent as destination address.
5.3 Optimization

The base Mobile IP protocol renders transparency to the operation of Correspondent Nodes (CN) by making them believe that Mobile Nodes (MN) always reside in their home networks. Although this feature is desirable because it supports mobility without the requirement of any changes to the Correspondent Node’s protocol stack, it suffers from several performance problems.

Firstly, Mobile IP’s triangle routing problem leads to sub-optimal routing of packets. Secondly, packets in transit during a handover are often lost because they are tunnelled based on out-of-date location information. Thirdly, base stations with small cells result in frequent handovers, and require a registration with a distant Home Agent (HA) for each such local handover. This leads to higher overhead and further aggravates packet loss (Perkins & Wang, 1999).

Mobile IP route optimization alleviates triangle routing by informing the Correspondent Node of the Mobile Node’s Care-of Address (CoA), while smooth handover reduces packet loss during handovers by informing the previous Foreign Agent (FA) of the Mobile Node’s Care-of Address (Perkins & Johnson, 1999).

For the third problem, a hierarchical Foreign Agent management scheme is proposed to alleviate the frequent local handover (Gustafsson, Jonsson, & Perkins, 2005).

The rest of this section will address these three problems especially in cooperation with a corresponding proposal by IETF.

5.3.1 Route Optimization

The idea behind route optimization is that similar to Home Agent, Correspondent Nodes also need to maintain mobility binding for the Mobile Nodes they communicate with, so as to directly send packets (tunnelled packets) to the Mobile Node’s actual point of attachment (Dimopoulou & Venieris, 2004, p. 5-10). However, the fact that mobility information for the Mobile Node is now available to all possible communicating Correspondent Nodes, makes the timely update of this information more difficult when the Mobile Node changes its point of attachment. In the standard protocol, only the
Home Agent had to be notified. Now, in addition to the Home Agent, all Correspondent Nodes need to update their mobility binding caches as well. Until this happens, packets forwarded to the old Foreign Agent are lost. For this reason, optimization extensions are also provided to allow the forwarding of packets, which were already on their way to the old Foreign Agent, to the Mobile Node’s new CoA.

Four new messages are defined for route optimization (Perkins & Johnson, 1999):

- A Binding Warning message is used to transmit advice that one or more correspondent nodes or foreign agents are likely to have either no binding cache entry or an out-of-date binding cache entry for some mobile node.
- A Binding Request message is used by a Correspondent Node to request a Mobile Node’s current mobility binding from the Mobile Node’s home agent.
- A Binding Update message is used for notification of a Mobile Node’s current mobility binding.
- A Binding Acknowledge message is used to acknowledge receipt of a Binding Update message.

When the Home Agent (HA) receives data packets from a Correspondent Node (CN) destined to a Mobile Node (MN) away from home, it deduces that the Correspondent Node is not aware of the Mobile Node’s actual location because packets are addressed to the Mobile Node’s Home Address. Then it sends a Binding Update (BU) message to notify the Correspondent Node of the Mobile Node’s Care-of Address (CoA). Upon the receipt of this Binding Update message, the Correspondent Node is able to know that the Mobile Node has moved to a new point of attachment. In subsequent communication, it directly tunnels a packet to the Correspondent Node without the Home Agent (HA) involved, as illustrated in Figure 5.7. In the rest of session, the Corresponding Node periodically sends the Home Agent the Binding Request message to update its binding cache for that Mobile Node. In response to this request message, the Home Agent send a Binding Update message to the Correspondent Node as well.
In this basic operation, a problem might arise when the Mobile Node moves to a new Foreign Agent, but the Correspondent Node has not yet been informed of the Mobile Node’s change of CoA and keeps sending packets to the old one. These packets will be lost if no other mechanisms is in place. This is where the Binding Warning message comes in, which enables smooth handover.

5.3.2 Smooth Handover

Smooth handover has been defined for two cases. Firstly, it is defined for data being sent, based on out-of-date binding caches maintained at Correspondent Nodes (CN). Secondly, it is defined for data packets in transmit while the Mobile Node (MN) changes its point of attachment, but before it registers with its Home Agent (HA). In both cases, either the Correspondent Nodes or the Home Agent maintain “wrong” information for the Mobile Node with respect to its Care-of Address (CoA). Consequently, the previous Foreign Agent (FA) needs to be informed of the Mobile Node’s new location in order to redirect packets addressed to the Mobile Node. This is performed by making the Mobile Node instruct its new Foreign Agent, via a Previous Foreign Agent Notification Extension (PFANE) (Perkins & Johnson, 1999), to send a Binding Update message to its previous Foreign Agent, as illustrated in Figure 5.8.

Upon receipt of this message, the previous Foreign Agent deletes the
visitors list entry for the Mobile Node and adds a binding cache entry for it for forwarding packets to its new location. Then, it replies to the Mobile Node with a Binding Acknowledge message, because the Mobile Node needs to be sure that the update has been carried out. In the absence of this message, the new Foreign Agent retransmits the Binding Update message to the previous Foreign Agent. In the event that the Mobile Node returns to its home network, smooth handover functionality is not supported, and the Binding Update message causes only the previous Foreign Agent to delete its visitor list entry (Perkins & Johnson, 1999).

From now on, if the previous Foreign Agent receives encapsulated packets for that Mobile Node, which it is not currently serving but for which it maintains a binding cache, it deduces that the Correspondent Node needs to update its binding cache for the Mobile Node and sends a Binding Warning message to the Mobile Node’s Home Agent, which carries the Correspondent Node to be updated. The Home Agent then sends a Binding Update message to the indicated Correspondent Node to inform it of the Mobile Node’s new CoA. A Binding Warning message can also be sent by the Mobile Node to its...
Home Agent at the time it registers a new CoA. This way the Home Agent is informed in a timely manner of the Correspondent Node to be updated during the registration procedure (Perkins & Johnson, 1999).

5.3.3 Hierarchical Foreign Agent Scheme

Direct routing from Correspondent Nodes to Mobile Nodes and smooth handover reduce unnecessary traffic and packet loss during handover. However, they cannot reduce the frequent local handover. That is where the MIPv4 Regional Registration (Gustafsson et al., 2005) comes in, which promotes a hierarchical Foreign Agent (FA) scheme.

The idea is to hide the movement of Mobile Nodes within the visited domain from the Home Agent under one globally routable entity, denoted as the Gateway Foreign Agent (GFA) (Dimopoulou & Venieris, 2004, p. 5-38). This means that time required for signalling messages to the home network, in conjunction with the time needed for the host to update the path to its current location, is reduced. This is because a change in the Mobile Node's path, within a visited domain, is not propagated to the Home Agent but is handled locally.

The regional registration proposal defines a tree-like hierarchy of Foreign Agents, where the FA, located at the root of the tree, is referred to as the GFA, as illustrated in Figure 5.9.

5.4 A Hierarchical Approach to IP mobility

Mobile IP has been widely accepted as the most appropriate protocol for IP mobility management in future wireless mobile networks. However, it suffers from several well-known weaknesses, such as the increasing signaling load placed on the network and the latency in restoring the communication path to the host's new point of attachment (Dimopoulou & Venieris, 2004, p. 5-22). This latency, in turn, leads to the loss of packets while the end-to-end path is being restored. Consequently, although providing a robust and simple solution, Mobile IP does not guarantee seamless mobility, and thus Mobile Nodes participating in active sessions experience disruption in their service when changing their point of attachment. These drawbacks have led to the
investigation of other solutions that complement Mobile IP. Section 5.3.3 described one of these solutions. This section takes that concept a step further, and promotes a hierarchical approach to mobility management. In this hierarchical structure, IP mobility is further divided into the following three levels (Das et al., 2000):

- **Micro-mobility** is the movement of a Mobile Node within or across different base stations within a subnetwork and occurs very rapidly. Currently, management of micro-mobility is accomplished using link-layer support (layer 2 protocol).

- **Macro-mobility** (or intradomain mobility) is the movement of a Mobile Node across different subnetworks within a single domain or region, and occurs relatively less frequently. This is currently handled by Internet mobility protocols (layer 3), such as Mobile IP.

- **Global mobility** (or interdomain mobility) is the movement of a Mobile Node among different administrative domains or geographical regions. At present, this is also handled by layer 3 techniques such as Mobile IP.
With such a hierarchical approach, the macro-mobility protocols often operate in the core network while micro-mobility protocols cover the region of an access network and hide the Mobile Node’s movement within the same local region from the upper macro-mobility protocols. The global mobility protocols deal with the movement between the different administrative domains and operate at a global level. With three level mobility protocols cooperating with each other, a global mobility architecture is enabled. This architecture allows for low latency and as little signalling load as possible, as illustrated in Figure 5.10.

Mobile IP is widely accepted as the macro-mobility and global mobility protocol (Das et al., 2000). Otherwise, several proposals have been made for handling the micro-mobility. Most of them consist of extensions to Mobile IP or interact with Mobile IP signaling for providing global mobility to users. A classification of these protocols has been made by researchers based on the forwarding mechanisms used for the routing of packets to the Mobile Nodes. Protocols may be categorized as those that use hierarchical tunnelling techniques as their forwarding mechanism and those that use host-based forwarding entries along the path to the Mobile Nodes (Dimopoulou &
Venieris, 2004, p. 5-24). The former includes MIPv4 Regional Registration and Hierarchical MIPv6 while the latter includes Cellular IP and HAWAII. Their primary difference lies in where location information is maintained within the network. In the first category, agents, often placed in a hierarchical manner, hold this information and use the standard IP routing of the “mobile-unaware” fixed infrastructure for the forwarding of packets. In contrast, host-based forwarding schemes establish location information for the hosts in all nodes pertaining to the downstream forwarding path.

The detailed description of these protocols are not the concern of this study. A comprehensive comparison of various micro-mobility protocols, in terms of architecture, scalability, reliability, and philosophy, can be found in the book “IP for 3G” (Wisely et al., 2002, p. 176).

5.5 Conclusion

This chapter described the basic operation of Mobile IP, and investigated the various extensions to it as well, such as reverse tunnelling, route optimization, smooth handover, and regional registration. With solving the IP mobility problem using Mobile IP, based on the aforementioned GPRS and WLAN interworking architecture in Chapter 3, a seamless handover mechanism between GPRS and WLAN has come into shape. To ultimately make this work requires offering Mobile IP services in GPRS networks, which will be discussed in Chapter 7.

After constructing the seamless handover mechanism, this chapter further introduced a hierarchical architecture for IP mobility. This makes IP mobility more effective, which meets this study’s other objective, namely effective IP mobility.

Before presenting the proposed framework, the next chapter will introduce another part of the framework: IPSec.
Chapter 6

IPSec

Chapter 4 has comprehensively investigated the existing problems for IP towards a complete network design. Among them, the IP mobility and security issues are the focus of this study. Chapter 5 has thoroughly discussed the IP mobility, while a brief introduction of IP network security has been done in Section 4.6 on page 89. At that time, it was argued that the availability of the information in the intranet should be ensured without the violation of any other characteristic, that is, authenticity, confidentiality and integrity when the users of the intranet move to the outside. More specifically, this involves the authentication of the user, authentication and confidentiality of the data, and the integrity of the data. We have stated that IPSec (Kent & Atkinson, 1998c) is such a mechanism that provides the solution in an effective and elegant way.

This chapter will describe IPSec to demonstrate how it offers the above mentioned security services. The first section will provide an overview of IPSec Architecture. The second section will provide a brief examination of the existing IPSec scenarios. The next two sections will describe two of three primary IPSec components, Authentication Header (AH) and Encapsulation Security Payload (ESP) protocols, respectively. The fifth section will briefly introduce the third primary component, namely key management. The sixth section will evaluate IPSec architecture design in accordance with IP design principles, mentioned in Section 4.2 on page 75, and attempt to answer why it is an effective and elegant solution.
6.1 Overview of IPSec

IPSec primarily consists of three components, the Authentication Header (AH) protocol (Kent & Atkinson, 1998b), the Encapsulating Security Payload (ESP) protocol (Kent & Atkinson, 1998a), and key management (Maughan, Schertler, Schneider, & Turner, 1998; Harkins & Carrel, 1998). The security services of IPSec are achieved by means of certain cryptographic mechanisms employed by AH and ESP. These cryptographic mechanisms, in turn, require a separate set of mechanisms for putting cryptographic keys in place, which forms the third part of IPSec, Key Management. In addition, there are many other supporting components, such as encryption algorithms, authentication algorithms, and domain of interpretation (DOI) (Piper, 1998). An overview of IPSec architecture is shown in Figure 6.1.

6.1.1 Security Services

IPSec is designed to provide inter-operable, high quality, cryptographically-based security for IPv4 and IPv6 (Kent & Atkinson, 1998c). The set of security services offered includes access control, connectionless integrity, data origin authentication, protection against replays (a form of partial sequence integrity), confidentiality, and limited traffic flow confidentiality. These services are offered through the use of the AH protocol and ESP protocol as well as the key management procedures and protocols. The services that each protocol provides are summarized in Table 6.1.
6.1. OVERVIEW OF IPSEC

Table 6.1: Security Service Matrix

<table>
<thead>
<tr>
<th>Service</th>
<th>AH</th>
<th>ESP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connectionless integrity</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Data origin authentication</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Protection against replays</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Confidentiality</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Traffic flow confidentiality</td>
<td></td>
<td>√</td>
</tr>
</tbody>
</table>

6.1.2 Security Association

A Security Association (SA) is a simplex ‘connection’ that provides security services to the traffic carried by it (Kent & Atkinson, 1998c). The term ‘connection’ is used with quotation marks because the IP realizes a connectionless and thus stateless service for all communication entities. However, the respective peer entities need certain state information to implement security services. This includes security mechanisms, algorithms and keys used. This state is maintained through Security Associations established, deployed and dissolved between peer entities (Schafer, 2003, p. 200). The concept of a Security Association is fundamental to IPSec. Both AH and ESP make use of Security Associations and one of the major functions of key management is the establishment and maintenance of Security Associations. The rest of this subsection will describe the various aspects of the Security Association.

Security Association Entities

As mentioned above, a Security Association is a secure ‘connection’ between communication entities. Here, the communication entity can be a host or gateway. In the case of a gateway, this gateway serves as a security proxy on behalf of the hosts in the network behind this gateway. In the context of this study, it is referred to as Security Gateway or, more specific, IPSec Gateway. The following are the three possible combinations of entity connections:

- Host ↔ Host,
- Host ↔ Gateway, and
- Gateway ↔ Gateway.

These combinations correspond to three topology structures of IPSec:
• Host-to-Host,

• Host-to-Network, and

• Network-to-Network.

A comparison of these three topologies is shown in Figure 6.2. In host-to-host mode, security is truly end-to-end while in modes wherever the gateway is involved, that is, the host-to-network and network-to-network modes, the communication from the host to its gateway is protected by IPSec. This problem will be further addressed later when discussing the combination of Security Associations.
6.1. OVERVIEW OF IPSEC

Security Association Identifier

It is worth noting that a Security Association is used only in one communication direction. To secure a typical bi-direction communication between two peer entities, two Security Associations are thus required. Each Security Association, in each direction, should be uniquely identified. The following definition meets this requirement:

“A Security Association is uniquely identified by a triple consisting of a Security Parameter Index (SPI), an IP destination address, and a security Protocol (AH or ESP) identifier.” (Kent & Atkinson, 1998c)

Since the IP destination address is different in each communication direction, it solely distinguishes the Security Associations in each direction.

The SPI is a 32-bit pseudo-random number included in the header of each of the AH and ESP protocols. To function correctly, SPIs must be synchronized between two endpoints of a Security Association.

Security Association Modes

A Security Association is basically operated in one of the following two modes:

- Transport Mode: this mode can only be used between two hosts.

- Tunnel Mode: this mode can be used between hosts or gateways. However, whenever either end of a Security Association is a security gateway, the Security Association must be tunnel mode.

To well understand the Security Association modes well, especially tunnel mode, two terms need to be introduced first:

- Communication Endpoints, which denote the source and destination system of the IP packet exchange (Schafer, 2003, p. 206).

- Cryptographic Endpoints, which denote the systems that generate and process the AH or ESP protocol headers of IP packet exchanges within the framework of a Security Association.
With these two terms explained, the transport and tunnel modes become easier to understand when described as follows:

- Transport mode is used when the cryptographic endpoints are the same as the communication endpoints.
- Tunnel mode is used when at least one cryptographic endpoint is not a communication endpoint.

As we know, the endpoint is identified by the IP address in IP networks. Since the cryptographic endpoint and communication endpoint are the same in the transport mode, only an IP header is required to represent two roles of the same endpoint, as illustrated in Figure 6.3(a).

In contrast, these two endpoints are different in the tunnel mode. Consequently, two IP headers are required to represent them respectively. An outer IP header represents the cryptographic endpoints while an inner IP header represents the communication endpoints, as illustrated in Figure 6.3(b).

Hosts can play a role of sole communication endpoint or both communication endpoint and cryptographic endpoint, but not sole cryptographic endpoint. Gateways, in most cases, play a role of sole cryptographic endpoint, but not absolutely. For example, for network management a gateway system can be managed by a management station via the Simple Network Management Protocol (SNMP). At this time, it is acting as a host.

Thinking about IPSec topologies, mentioned earlier, and Security Association modes, it is observed that constraint relationships exist between them. In host-to-network and network-to-network modes only tunnel mode can be applied, because the gateway is involved in these two modes. In host-to-host mode, both transport and tunnel modes can be applied. Actually, the transport mode can be seen as a simplified form of tunnel mode in host-to-host mode. Since the communication ends and cryptographic endpoints are the
same in this situation, an extra IP header is not needed so that the tunnel mode is evolved into the transport mode.

Table 6.2 summarizes the relationships between IPSec topologies and Security Association modes.

<table>
<thead>
<tr>
<th>Security Association Combination</th>
<th>Tunnel Mode</th>
<th>Transport Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host-to-Host</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Host-to-Network</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Network-to-Network</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Security Association Combination**

The IP traffic transmitted over a single Security Association is offered protection by exactly one security protocol, either AH or ESP, but not both. Sometimes a security policy may call for a combination of security services for a particular traffic flow that can not be achieved using a single Security Association. In such a situation it is necessary to use multiple Security Associations to implement the required security policy. The term “Security Association bundle” is applied to a sequence of Security Associations through which traffic must be processed to meet a security policy (Kent & Atkinson, 1998c). It is worth noting that the Security Associations that construct a bundle may terminate at different cryptographic endpoints. For example, one Security Association may extend between a mobile host and a security gateway and a second, nested Security Association may extend to a host behind the gateway. Security associations may be combined into bundles in two ways: Transport Adjacency and Iterated Tunnelling.

- **Transport Adjacency** refers to applying more than one security protocol to the common IP datagram, without using tunnelling. This approach of combining AH and ESP allows for only one level of combination, and further nesting yields no extra benefit since the processing is performed at one IPsec entity at the destination. As mentioned above, transport mode is applicable in host-to-host topology only. Likewise, this approach is only applicable in host-to-host topology.
Iterated Tunnelling refers to the application of multiple layers of security protocols effected through IP tunnelling. This approach allows for multiple levels of nesting, since each tunnel can originate or terminate at a different cryptographic endpoint along the path. There are three basic cases of iterated tunnelling:

1. Both endpoints for the Security Associations are the same.
2. One endpoint of the Security Associations is the same.
3. Neither endpoint is the same.

These two approaches can also be combined, for example, a Security Association bundle could be constructed from one tunnel mode Security Association and one or two transport mode Security Associations, applied in sequence. Note that nested tunnels can also occur where neither the source nor the destination endpoints of any of the tunnels are the same. In such a case, there will be no host or security gateway with a bundle corresponding to the nested tunnels.

Security Association Databases

There are two conceptual databases for the administration and specification of Security Associations: the Security Policy Database (SPD) and the Security Association Database (SAD). The former specifies the policies that determine which security services should be applied to which IP packets and how this application should be executed. The latter database contains parameters that are associated with each Security Association. This also defines the concept of a Selector, a set of IP and upper layer protocol field values that are used by the SPD to map traffic to a policy, that is, a Security Association (Kent & Atkinson, 1998c).

Three basic IPSec topologies have been identified in Section 6.1.2, namely host-to-host, network-to-network and host-to-network, and two basic combining methods of Security Associations have been introduced in Section 6.1.2. When these combining methods are applied to the topologies, it leads to many IPSec scenarios. Some of these scenarios will be discussed in the next section.
6.2 IPSec Scenarios

Among the consideration of IPSec topologies and security associations, three basic scenarios must be mandatory in the compliant IPSec hosts or security gateways (Kent & Atkinson, 1998c). The rest of this section will introduce these three scenarios, according to the categories of topologies, respectively.

6.2.1 Host-to-Host

In this case, both communication endpoints and cryptographic endpoints are hosts. Thus both transport adjacency and iterated tunnelling methods can be used. This scenario provides true end-to-end security.

When applying transport adjacency in host-to-host topology, which actually, is the only topology to which transport adjacency can be applied, a possible packet format is shown in Figure 6.4.

It is worth noting that there is no situation where Encapsulating Security Payload (ESP) appears prior to the Authentication Header (AH). This is because the protection range of the AH is bigger than that of the ESP. The AH is applied to both the upper layer protocols and parts of the IP header, but the ESP is not. Thus, in order to protect as big a range as possible, it seems appropriate that the AH should appear as the first header after IP.

It is possible to apply iterated tunnelling to the host-to-host topology, but there is no requirement to support general nesting in tunnel mode (Kent & Atkinson, 1998c). Thus, this study does not intend to discuss it here.
6.2.2 Network-to-Network

In this configuration, the traffic between two networks is protected by IPsec. Its basic form presents a scenario with simple virtual private network support, as shown in Figure 6.2(b). Only tunnel mode can be applied here. The packet format is shown in Figure 6.3(b).

One of the problems associated with the above configuration is the fact that traffic is protected between the security gateways, and traffic between the hosts and security gateways are left unprotected (Held, 2004, p. 148). If true end-to-end security between the hosts are required, applying iterating tunnelling to this topology can meet this requirement, which presents another scenario, as illustrated in Figure 6.5.

6.2.3 Host-to-Network

This covers the situation where a remote host uses the Internet to reach an organization’s IPsec gateway and then gains access to some server or other machine. If applying iterated tunnelling to this configuration, the two Security Associations originate from the same endpoint, namely the host, but terminate at different endpoints. One being the peer host, and another one the security gateway, as shown in Figure 6.6.

This is the so-called ‘road warrior scenario’ (Schafer, 2003, p. 207). It is worth noting that this scenario is suitable to a mobile environment. Assume that the host in Figure 6.6(a) is a wireless mobile station. When the user
moves to the outside, it connects to the Internet via the external wireless network. Then it crosses the Internet to the home network’s security gateway. After establishing the Security Association with the security gateway, the traffic between the security gateway and this mobile station is protected by IPSec. Furthermore, if another Security Association between this mobile station and another host in the home network is established, the traffic along the whole path is protected, and end-to-end security is offered.

Due to its high relevance to the mobile environment, the most focus will be placed on this scenario in this study.

6.3 IPSec Standards

According to the layering and modularity principle, discussed in Section 4.2.3, all the components or functionalities mentioned above are implemented in such a way that any protocol implements a functionality only so that it is easy to reuse, and to be combined without the problem of functionality duplications. IPSec is designed to be algorithm-independent, which intends to provide a framework to populate various algorithms for required security services. This modularity permits the selection of different sets of algorithms without affecting the other parts of the implementation. As a result, many subsequent RFC documents are issued to specify these protocols and algorithms, and RFC 2411 serves as a guideline for specifications describing encryption and authentication algorithms used in this system. A structural
diagram of these documents is shown in Figure 6.7.

The rest of this chapter will describe specifics regarding the three major components of the IPsec architecture. Section 6.4 will discuss the AH protocol, followed by a discussion of the ESP protocol in Section 6.5. Section 6.6 then discusses key management.

6.4 AH Protocol

The Authentication Header (AH) protocol specified, in RFC 2402 by IETF, provides connectionless integrity and data origin authentication, and optional anti-replay services (Kent & Atkinson, 1998b). In general, it is divided into two parts: a definition of the base protocol and the use of cryptographic algorithms with AH.

This section provides an overview of the AH protocol. Its protocol format is described first. Then a description is provided of how the security services, mentioned above, are achieved.

6.4.1 AH Protocol Format

The AH header consists of six fields: namely next header, payload length, reserved, security parameter index (SPI), sequence number, and authentication
data. After these fields, comes payload, as shown in Figure 6.8. The first two fields perform normal protocol functionalities, such as demultiplexing and boundary positions. The remaining fields perform the security-relative functionalities. They are introduced in brief below.

**Next Header:** this 8-bit field identifies the type of payload after the AH header. The Internet Assigned Numbers Authority (IANA) defines the assigned numbers that can be used (Reynolds & Postel, 1994). For example, the value 6 for TCP packet, 17 for UDP packet, as the Protocol field of IP header does. This assists the receiver to demultiplex the payload to an appropriate higher level protocol processor.

**Payload Length:** this 8-bit field contains the length of the whole AH header expressed in 32-bit words, minus two. Note that here the payload does not mean payload data in Figure 6.8, but the two integrity-securing fields: sequence number and authentication data. The length of the whole AH header and payload, as the payload of the IP datagram, can be calculated from the HLEN field and total length field of the IP header. From this length, and the payload length, the authentication data’s length can be calculated. Since the payload length is measured in 32-bit words, padding is required where the authentication algorithm is not a multiple of a 32-bit word. So the effective length of the authentication data becomes a problem. It can be determined by the selected algorithm.

**Reserved:** this 16-bit field is reserved for future use. It is currently set to zero.
Security parameter index: this is a 32-bit field. As mentioned earlier, this field uniquely identifies the Security Association to be used for processing the packet in conjunction with the IP destination address and the being used security protocol.

Sequence number: this 32-bit field is used for anti-replay service. This service, also referred to as replay protection, is optional although the field is always included by the sender. It is up to the receiver to determine whether or not to process the data in this field.

Authentication Data: this field is variable. It is used to hold the output of a cryptographic algorithm for providing integrity and authentication service. This will be described in detail later.

6.4.2 Connectionless Integrity and Data Origin Authentication

Integrity is a security service that ensures that modifications to data are detectable. It is called connectionless integrity here in that the AH detects modification of only an individual IP datagram, without regard to the ordering of the datagram in a stream of traffic (Kent & Atkinson, 1998c). Thus it does not ensure the integrity of the whole IP stream. Other mechanisms should exist to ensure that. Data origin authentication is defined as:

“providing to one party that receives a message assurance of the identity of the party which originated the message.” (Menezes, Van Oorschot, & Vanstone, 1997, p. 25)

These two security services are usually bundled together.

Authentication Algorithm Procedures

The AH uses the same mechanism to achieve both of the above two security services. It is performed primarily through the calculation and validation of an Integrity Check Value (ICV), also called the Message Authentication Code (MAC) in the case of the data origin authentication.

The sender employs a certain keyed cryptographic algorithm to calculate a digest of the whole packet, that is, an ICV. The ICV is calculated in such a
way that nobody can deduce the original message from the ICV and nobody can calculate the ICV of a message without knowing the key. The sender places the calculated ICV in the Authentication Data field and sends out the whole packet. Upon receipt of this packet, the receiver uses the same algorithm and key to calculate the ICV of the received packet. Then the receiver compares the calculated ICV with the ICV in the packet received’s authentication data field. If it is equal, the receiver can deduce that the packet originally comes from the sender without modification in the middle. This deduction is based on the fact that the key is secret between the sender and the receiver and the selected algorithms ensure that nobody can fake and modify a packet to pass the above validation without the key. In this way, the connectionless integrity and data origin authentication are ensured.

Both the selected algorithms and the key are the parameters of an Security Association, they are negotiated during the Security Association initialization process. This is the task of the key management part, which will be introduced later. What is highlighted here is that the AH offers the high flexibility in this way. It is not limited to a certain algorithm and does not limit length of the ICV. As a result, there is a broad range of algorithms available. Once again, this benefits from the IP’s modularity principle.

For point-to-point communication, suitable algorithms include keyed Message Authentication Codes (MACs) based on symmetric encryption algorithms, such as DES (Menezes et al., 1997, p. 250) or on one-way hash functions, such as MD5 (Rivest, 1992) or SHA-1 (Eastlake & Jones, 2001). Furthermore, these algorithms have been enhanced with the special consideration of the IPSec, such as HMAC-MD5-96 (Madson & Glenn, 1998a) and HMAC-SHA-1-96 (Madson & Glenn, 1998b). For multi-cast communication, one-way hash algorithms, combined with asymmetric signature algorithms are appropriate, though performance and space considerations currently preclude use of such algorithms.

**ICV Calculation**

There are some aspects worth noting with respect to the calculation of the Integrity Check Value (ICV).

The first aspect is about mutable fields. IP header fields can be classified into three categories: immutable, mutable but predictable, and mutable. For
the first two categories, their values are inserted into the field for the purpose of ICV calculation. For the last category, such as the TTL field, it will be modified during the transit. Its value will be set to zero for the purpose of the ICV computation. The value of authentication data field is also set to zero.

The second aspect is about padding. Some encryption algorithms are based on block size. Sometimes it is not a multiple of 32 bits. So padding is required to ensure the length of the whole AH is a multiple of 32 bits. The content of the padding is arbitrarily selected by the sender. These padding bytes are included in the authentication data calculation, counted as part of the payload length, and transmitted at the end of the authentication data field to enable the receiver to perform the ICV calculation.

From the fields the ICV covers, we can see the extent of the protection offered by the AH. It protects the whole IP packet, including the original IP header, and even the new IP header if in the tunnel mode, as shown in Figure 6.9.

It is worth noting that this mechanism solely can not absolutely ensure the data origin authentication. If only the above mechanism adopted, it cannot provide defence against a replay attack, where a third part can intercept the above processed message from the sender, store it and resend it later. When this resent message arrives at the receiver, it can pass all validation. Integrity is not comprised, but data origin authentication is violated. This can be prevented by another mechanism, called replay protection, which will be introduced below.
6.4.3 Replay Protection

The replay protection service is achieved using a sequence number. It is initially set to zero when a Security Association is established. Thereafter, before each datagram is transmitted, the value of this field is incremented by one. A new Security Association must be established, because sequence numbers associated with a Security Association are not allowed to repeat when the sequence number value of $2^{32} - 1$ is reached. As this sequence number is included in the computation of the ICV, any modification of it by an attacker will be detected.

The receiver of a packet always verifies if the sequence number contained in the packet is within a range of acceptable numbers. This range is called a ‘sliding window’. The reason why an entire window is used, is that in the Internet the order of IP packets can change during transmission of IP packets via different routes, even during normal operation. Therefore, packets sent later may possibly arrive at the receiver sooner than packets that were sent earlier.

RFC 2401 specifies a minimum window size of 32 with the value 64 recommended as the default. Upon receipt of an IP packet, the receiving system performs the following actions depending on the sequence number:

- If the sequence number is to the left of the receiving window, the receiver discards the packet.
- If the sequence number is inside the receiving window, the receiver verifies the ICV and accepts the packet if the verification is successful.
- If the sequence number is to the right of the receiving window, the receiver verifies the ICV and, if the verification is successful, accepts the packet and advances the window to the right.

6.5 ESP Protocol

As described above, the AH protocol does not provide a confidentiality service. This is achieved by the Encapsulating Security Payload (ESP) protocol. Optionally, it can also offer an authentication service and a replay protection
service. Similar to the AH protocol, it uses the same mechanism to implement the authentication and replay protection service. This section focuses on how confidentiality is achieved in the ESP.

The rest of this section will describe the individual field first. Thereafter a description of how the confidentiality of information is achieved, is provided.

6.5.1 ESP Protocol Format

The ESP protocol header consists of seven fields: security parameter index (SPI), sequence number, payload, pad, pad length, next header, and authentication data, as shown in Figure 6.10.

Security Parameter Index: this field is 32-bits in length. Similar to the SPI in an AH header, this field is used to identify different Security Associations with the same destination address and security protocol.

Sequence Number: this 32-bit field functions in the same manner as its counterpart in the AH header.

Initialization Vector: if the cryptographic algorithm being used requires an initialization vector (IV), the IV is transported in plain-text in each IP packet so that each packet can be processed independently of other packets.

Payload: this variable length field contains the actual data being transported. This field, as well as any padding and the Pad Length and Next Header fields to be described shortly, are encrypted, with the algorithm used for encryption selected when the Security Association was established. The actual type of data carried in the payload, such
6.5. **ESP PROTOCOL**

as a TCP segment, is identified by the Next Header field contained in the trailer.

**Pad:** this field ensures that the payload being encrypted is padded to a length that is equivalent to an integer multiple of the block size of the algorithm used and that the two following fields end up in the higher-order 16 bits of a 32-bit word.

**Pad Length:** this field indicates the number of octets added.

**Next Header:** this field functions in the same way as it does in the AH.

**Authentication Data:** the optional authentication data field contains an ICV, if available.

### 6.5.2 Confidentiality

The ESP achieves confidentiality by means of the symmetric encryption algorithms. This is specified by the Security Association, and negotiated during the initialization process. Many other parameters also need to be decided at that time, such as, cipher block length if block encryption algorithms is used, and the key.

The mode of the ESP decides the extent of confidentiality offered. For the transport mode, just the original upper layer protocol is encrypted, and only data confidentiality is offered. For the tunnel mode, the entire original IP packet is encrypted. As a result, the original source and destination IP address is not visible to the third party. This offers partial traffic confidentiality, in addition to the data confidentiality.

### 6.5.3 Other Security Services

The replay protection is offered similar to the Authentication Header (AH) protocol. Also, the same algorithms used by the AH are applied in the ESP connectionless integrity and data origin authentication services. What is different though, is the extent of coverage. The ESP does not count the original IP header to the calculation of the ICV, while the AH does. Thus, the ESP does not protect the integrity of the original IP header, as shown
When both confidentiality and integrity are selected in the ESP, a notable point is the order in which the encryption algorithm and authentication algorithm are performed. Kent and Atkinson (1998a) argued that encryption should be performed first before authentication. It is because encryption does not encompass the Authentication Data field while, by contrast, the calculation of ICV involves the payload. In addition, this order of processing also facilitates rapid detection and rejection of replayed or bogus packets by the receiver, prior to decrypting the packet, hence potentially reducing the impact of denial of service attacks. Note that since the Authentication Data is not protected by encryption, a keyed authentication algorithm must be employed to compute the ICV.

### 6.6 Key Management

The above sections described how security services are achieved using the Authentication Header (AH) and Encapsulating Security Payload (ESP) protocols. Before data exchange between two systems can be protected by these IPSec protocols, the Security Association between them must firstly be set up. There are two options for the establishment of the Security Association: manual or dynamic establishment. As the manual method is not only time-consuming but also error prone, it should only be used in very manageable configurations.

IPSec defines a standard method for the dynamic establishment of Security Associations. The Internet Security Association and Key Management
Protocol (ISAKMP) defines generic protocol formats and procedures for negotiation (Maughan et al., 1998). The actual applications of this protocol for the negotiation of parameters for IPSec Security Associations is presented in detail in the IPSec Domain of Interpretation (IPSec DOI) (Piper, 1998).

Based on the message formats and procedures provided by the ISAKMP, the Oakley protocol defines the actual key exchange procedure (Orman, 1998). In conjunction with ISAKMP, it is often referred to as ISAKMP/Oakley, as it represents a key exchange protocol that supports the ISAKMP framework.

Another protocol, the Internet Key Exchange (IKE), can be seen as a hybrid protocol that combines portions of ISAKMP and Oakley to provide a key management capability (Harkins & Carrel, 1998).

This study does not intend to explain them in detail, but a good description can be found in the corresponding RFC documents, as indicated above.

6.7 Evaluation with IP Design Principles

The previous sections have introduced various aspects of IPSec. In doing so, it has been mentioned many times that the layering and modularity principles have been obeyed in the design of the IPSec. However, it seems to violate the principle of the end-to-end principle, as the IPSec operates on the network layer. According to Wisely et al. (2002, p. 82), the security services are a matter of the end-to-end principle, and placing it in network layer fattens the network layer and makes the network more complicated. Only the endpoints that communicate with each other know if they need to hide the information from other users. Thus, a matter of course is that the security functionalities should not be placed in the network layer. Despite this, this section still attempts to argue that IPSec does comply with the end-to-end principle of IP. Furthermore, a more thorough discussion of end-to-end and node-to-node encryption is conducted.

The argument of the end-to-end principle is conducted in terms of all three IPSec topologies: host-to-host, network-to-network, and host-to-network. In the case of host-to-host, it is obvious that any solution can be seen as end-to-end wherever the security functionalities are placed, as the hosts are just
the endpoints of the communication. This does not add any complexity into
the networks. The IP packets that have been processed by IPSec are routed
in the networks as the normal IP packet does. Any intermediate routers do
not need to have any acknowledge of IPSec, and are left unchanged. Only
the communicating hosts need to have IPSec support. It is really what the
end-to-end principle requires.

In the case of network-to-network, IPSec processes are performed in the
security gateways. It is seen as end-to-end in the sense that it is the security
gateways that are seen as the endpoints, but not the individual hosts in the
networks behind those gateways. Only the security gateways need to have
IPSec support. Similar to the host-to-host case, no acknowledge of IPSec is
required in the intermediate routers, or even in the communicating hosts. In
this sense, it is seen as end-to-end.

The same deduction is applicable in the case of host-to-network where
the endpoints are a host and a security gateway, respectively. Thus, in any
case, IPSec is indeed an end-to-end solution.

Another solution for security is node-to-node encryption (Canavan, 2001,
p. 202), as shown in Figure 6.12. Node-to-node encryption often operates in
the link layer. That is why it is also referred to as link-to-link encryption.
The source host puts encrypted data into the data link layer frame, and
sends it to the next node, that is, a host or router in the network. This
node decrypts the received data frame, and then, if in the case of the router,
encrypts it into another data frame, which is sent to the next node. Otherwise
the encrypted data frame is delivered to the upper layer. Each node along
the path between source host and destination host then repeats the same
decryption and encryption procedure again and again until it reaches the
destination host.

An advantage of node-to-node encryption is that it is strong encryption.
As mentioned above, it operates at the link layer, and encrypts all the upper
protocols data, which exposes as little information to the outside as possible.
However, it has some inevitable disadvantages which include the following:

- It requires that the router devices along the whole path must be com-
patible, which means they must know how to encrypt and decrypt the
data frame. As a result of the above requirement, a key management
process is also required between any involved node. Moreover, with
the increment of the involved routers along the path, it will become incredibly complex, especially in a large network like the Internet.

- In each router the data is decrypted and exposed to the outside. This is a security gap, which could be exploited by hackers to obtain valuable information. The solution to this issue is to establish a trust relationship between the host and all the routers along the whole path. However, it seems an impossible mission in the Internet due to its high autonomy.

- It is inefficient that the same encryption and decryption processes are repeated in intermediate routers again and again.

When the above disadvantages are taken into account, IPSec provides a much better solution, which renders effective and elegant. The only drawback of IPSec, compared with the node-to-node encryption, is that IPSec is not as strongly encrypted as the node-to-node, because it exposes more information to the outside. Fortunately, this can be rectified by employing tunnel mode. In this way, the whole IP packet is protected under IPSec, as mentioned earlier. In this sense, IPSec, by comparison, is better than the other end-to-end solutions operating in the upper layer, such as SSL and TLS, as they expose even more information to the outside.
6.8 Conclusion

This chapter examined various aspects of IPSec. The Authentication Header (AH) protocol protects the authenticity and data integrity of traffic while the Encapsulating Secure Payload (ESP) protocol protects the confidentiality of traffic, and optionally, the authenticity and data integrity as well. Of various IPSec scenarios, the so-called “road warrior scenario”, using host-to-network topology, is considered most suitable to the mobile environment. This scenario will be intensively discussed in the rest of this thesis in terms of the proposed framework. The incorporation of Mobile IP and IPSec will be discussed in Chapter 8.

At this stage, we have the necessary background knowledge to understand the proposed framework. Hereafter, we will go to the next part of this thesis, namely the Framework, and present the proposed framework.
Part II

The Framework
Chapter 7

Integration of GPRS and Mobile IP

Part I introduced necessary background knowledge and motivated some direction for the proposed framework. A loose coupling GPRS and WLAN interworking architecture was suggested in Chapter 3. IP was selected to fulfill the interworking role in this architecture. After describing IP in Chapter 4, we further described Mobile IP in Chapter 5. Moreover, a hierarchical architecture for IP mobility was suggested for effective IP mobility. IPSec, in corporation with a screened subnet firewall architecture, was selected to protect traffic’s authenticity, integrity and confidentiality when Mobile Nodes move away from the internal domain.

With the necessary background knowledge in mind, hereafter, the proposed framework is presented. This will be performed through three steps:

1. Offering Mobile IP service in GPRS, which is discussed in this chapter.

2. Integration of Mobile IP and IPSec, which will be discussed in Chapter 8.

3. Combining the above discussions to propose a complete framework, which will be presented in Chapter 9.

The rest of this chapter will introduce an approach to introducing Mobile IP service into GPRS networks as proposed by the the 3rd Generation Partnership Project (3GPP). Then, based on “Step 2” of their approach, we will present a model for offering Mobile IP service into GPRS networks. After
that, we will describe this model in terms of its initial stage, routing and handover in correspondent sections.

7.1 Introduction of Mobile IP into GPRS

3GPP has proposed a “stepwise” approach to introducing Mobile IP into the GPRS core network (European Telecommunications Standards Institute, 2001). The essential idea is to enhance the Gateway GPRS Support Node (GGSN) with Foreign Agent (FA) functionality, which introduces a new infrastructure entity, GGSN/FA, into the GPRS network. The development of a GPRS network towards a full Mobile IP-based network can be performed in three steps, all backwards compatible with networks and mobile terminals that are not handling Mobile IP.

**Step 1:** This represents a minimum configuration for an operator, who wishes to offer the Mobile IP service. The current GPRS structure is kept and handles the mobility within the Public Land Mobile Network (PLMN), while Mobile IP offered by GGSN/FA allows the user to roam between other systems, such as WLAN, without losing an ongoing session.

**Step 2:** A more efficient routing could be obtained after inter-SGSN handovers by changing the GGSN/FA, to which the Mobile Node is attached, to a more optimal one. By keeping the tunnels from the new SGSN to both the old and new GGSN/FA, for a short period of time, potential problems with packet loss are minimized.

**Step 3:** This step combines the Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN) into one node, the Internet GPRS Support Node (IGSN), and to let Mobile IP handle inter-IGSN handover, that is, mobility within the PLMN core network and between networks. It renders a full Mobile IP based mobile network.

An operator may implement “Step 2” or “Step 3” without first implementing the previous one. The rest of this section will introduce these three steps in brief, respectively.

In Figure 7.1 – 7.3, the Border Gateway (BG) denotes the functionality to avoid unwanted traffic between GPRS Public Land Mobile Networks
(PLMNs). The Border Gateway falls outside the scope of GPRS specifications.

### 7.1.1 Step 1 — Offering Mobile IP Service

This step assumes a minimal impact on the GPRS standard and on the networks when introducing Mobile IP into GPRS. In this step, the Mobile Node must be able to find a Foreign Agent, preferably the nearest one. The underlying assumption is that the Foreign Agent is located at the GGSN and that not all GGSNs may have Foreign Agents. One Foreign Agent in a Public Land Mobile Network is sufficient for offering Mobile IP service, however for capacity and efficiency reasons, more than one may be desired. This means that the Mobile Node must request a Packet Data Protocol (PDP) context to be set up with a GGSN that offers Foreign Agent functionality. The solution is to define an Access Point Name (APN), for example “MOBILEIPv4FA”. This APN is used to connect to the correct GGSN with a Foreign Agent.

While setting up the PDP context, the Mobile Node must be informed of the network parameters of the Foreign Agent (FA), for example, the care-of address. Depending on the capabilities of a visited network, two roaming schemes can be identified: GPRS roaming and Mobile IP roaming. With GPRS roaming, it means roaming via the Gp interface, which is necessary when the visited network does not offer any Foreign Agent. In those cases where the visited network offers a Foreign Agent, Mobile IP roaming is utilized. It is assumed that the Mobile Node stays with the same GGSN/FA as long as the PDP context is activated. A typical network architecture after this step is shown in Figure 7.1.

### 7.1.2 Step 2 — Intermediate GPRS-MIP System

In “Step 2”, the routing is improved by performing a Mobile IP based streamlining after an inter-SGSN handover. A “very mobile” Mobile Node might perform several inter-SGSN handovers during a long session which may cause inefficient routing. The inefficient routing takes place when the GGSN/FA that is closest to the new SGSN is different from the closest one to the old SGSN. The routing could be improved by changing the GGSN/FA for the mobile during an UMTS/GPRS session. During such a change two tunnels...
are maintained, for a short period of time, between the new SGSN and the old and the new GGSN/FA. This will minimize problems with packet loss.

The possibility of optimizing the route is especially desirable in those cases where there are several GGSN/FAs in the Public Land Mobile Network, and the GGSN/FA and the SGSNs are co-located, as illustrated in Figure 7.2.

The Mobile Node will get a new care-of address with the same procedure as is defined in step 1 for giving the Mobile Node a care-of address. As in the previous step, the GPRS interfaces (Gn and Gp) need to be deployed for roaming customers, since there might be networks which not yet support Mobile IP. Roaming between Public Land Mobile Networks (PLMN) can be
handled either with Mobile IP or with GPRS.

### 7.1.3 Step 3 — Using Mobile IP for Intra System Mobility

The third step is to let Mobile IP handle intra Core Network mobility, intra-PLMN mobility as well as inter-system mobility in the packet domain. The functionality of the SGSN and the GGSN are combined into one node, the Internet GPRS Support Node (IGSN), and functionality is added to utilize Mobile IP for handling inter-IGSN mobility. The IGSN/FA will be the node that marks the end of the UMTS specific part of the PLMN. Figure 7.3 depicts a logical view of the core network architecture. To allow compatibility with UMTS/GPRS networks, an option to let the IGSN also act as a SGSN or GGSN will be necessary during a transformation period.

The basic functionalities of the IGSN includes:

- Support of UMTS/GPRS mobility management across the UTRAN/BSS, that is, what the SGSN does today.

- Support of MAP (Mobile Application Part), which is used to communicate with UMTS/GPRS specific nodes, such as Home Location Register (HLR), Equipment Identity Registry (EIR), SMS-C and the functionality needed to handle the information to and from these nodes, such
as SIM based authentication and handling of keys for encryption over the radio interface.

- Interaction with the Home Location Register (HLR) and via the Foreign Agent with the Authentication Authorization Accounting (AAA) infrastructure.

- Charging data collection and formatting according to UMTS/GSM specifications. IETF specifications may be used for Foreign Agent accounting.

- Support of Mobile IP with the necessary functionality to be compliant with Mobile IP deployment in non-UMTS networks around the world.

- Support of inter-IGSN handovers, either by Mobile IP or GTP. In the control plane, the PDP context for a Mobile Node might need to be transferred from the old to the new IGSN by GTP.

### 7.2 Overview of the Model

Based on the above description of the three steps, this study chooses the second step as a model of integrating Mobile IP into GPRS, because it has the most minimal impact on the current network infrastructure, and optimized mechanisms for handovers. It, furthermore, has been widely put into practice (Sharma et al., 2004; Buddhikot et al., 2003).

Two points need to be taken into account when designing this model:

- Due to scarcity of radio resources and IPv4 addresses, they should be used with care;

- The impact on the current GPRS signalling messages as well as on the Mobile Node and SGSN functionality should be minimized to ensure that this model can be implemented for Release 99 (R99) version of UMTS, as the R99 is a widely deployed version.

The first point leads to the choice of using the FA-CoA mode, because it saves the scarce IP address resources, as explained in Chapter 5. The second point leads to the choice of using the APN to find the desired GGSN instead
of introducing a new PDP type and the choice of transporting all Mobile IP messages in the UMTS/GPRS user plane.

Based on “Step 2”, a model of integrating Mobile IP into the GPRS is shown in Figure 7.4. The following subsections will describe this model in terms of its handover process, protocol architecture, and procedures.

### 7.2.1 Hierarchical Handover

This model handles handover at four levels: Base Station (BS) handover, Radio Network Controller (RNC) handover, SGSN handover, and Foreign Agent handover, as illustrated in Figure 7.5. All handovers, excluding the Foreign Agent handover, are UMTS-specific procedures, and the Mobile IP is transparent for these procedures. Chapter 2 described these procedures, and this chapter will not repeat them again. The Foreign Agent handover is Mobile IP specific. This handover is often triggered by a SGSN when this SGSN finds a closer GGSN/FA to it.
7.2.2 Protocol Architecture

Multiple protocols and entities are involved along the communication path from the corresponding node (CN) to the Mobile Node (MN), as illustrated in Figure 7.6. It is easy to observe that the GGSN/FA is where the Mobile IP and GPRS interact with each other. Two sides of this point are the Mobile IP specific domain and the GPRS specific domain, respectively. The end-to-end principle, promoted by the Internet community, is still obeyed by the upper layer protocols, that is, the transport layer and the application layer.
7.2.3 Procedures

For a Mobile Node attaching to a GPRS network, the following basic procedures are performed to obtain Mobile IP service over that GPRS network, as illustrated in Figure 7.7.

- GPRS attachment
- PDP context activation
- Mobile IP Agent advertisement
- Mobile IP Registration
- IP packets routing

The former four procedures can be seen as the initial stage. After that, the Mobile Node is attached to the GPRS network, and known by its Home Agent. At this point, the Correspondent Node can send packets to the Mobile Node’s home address. These packets are routed to the Mobile Node via a Mobile IP tunnel and a GPRS Tunnelling Protocol (GTP) tunnel.

The rest of the sections will introduce the initial stage first, and then introduce the routing stage in this model.

7.3 Initial Stage

The initial stage is divided into two parts: GPRS specific procedures (attachment and PDP context activation), and Mobile IP specific procedures.
(agent advertisement and registration). An interaction point between the GPRS and Mobile IP is when the Mobile Node finishes PDP context activation procedure and the GGSN/FA is going to advertise its presence to the Mobile Node. At this point the association between the GPRS specific address and Mobile IP specific is established. The PDP context activation assigns an IP address to the Mobile Node, and the Mobile Node also acquires an IP address in the agent advertisement procedure. The first question when combining the GPRS and Mobile IP is which address is determined as the IP address of the Mobile Node.

Relative to the two procedures present at the interaction point, the other two procedures concerned are more straightforward. The GPRS attachment functions exactly as described in Section 2.4.1 on page 49, and will, therefore, not receive further attention in this section. As to the Mobile IP registration procedure, the messages relevant to this procedure are delivered via the GTP tunnel established in the PDP context activation procedure.

To enhance understanding of the discussion to follow, a sequence diagram of the whole initial stage is provided in Figure 7.8. Based on this figure, the rest of this section will describe the initial stage in detail, and seek to answer the question mentioned above.

The PDP Context Activation procedure functions normally, as described in Section 2.4.3 on page 53. One exception is that the Mobile Node should omit the IP address, which the GGSN sends to it in response to the Activate PDP Context Request. Thus, at this point in time the Mobile Node still does not have an IP address. Instead, it is allocated a local link layer address, that is, TID. Via the GTP tunnel, identified by that TID, the Foreign Agent advertises its presence to the Mobile Node. As the Mobile Node not yet has an IP address, a limited broadcast address (255.255.255.255) needs to be used as the destination address in the Agent Advertisement message.

As the normal Agent Advertisement message does, this message carries the necessary parameters regarding the agent, such as Foreign Agent addresses. When the Mobile Node receives the advertisement, it can pick up one of the Foreign Agent address lists as its care-of address.

After that, the Mobile Node registers the new Care-of Address with its Home Agent. The format of the registration was given in Figure 5.4. This message is delivered to the GGSN/FA via the GTP tunnel established above.
The Foreign Agent relays this message to the Mobile Node’s home agent. To map the reply from the home network with the correct Mobile Node, the GGSN/FA needs to store the home address of the Mobile Node and the local link address of the Mobile Node, that is, TID. When the GGSN/FA receives the registration reply, it must, besides the normal process, also insert the Mobile Node’s home address into its GGSN PDP Context as that Mobile Node’s PDP address.

At this point, the initial stage is finished. The care-of address of the Mobile Node, that is, the address of the GGSN/FA, is known by the Mobile Node’s home agent. No extra PDP address is allocated to the Mobile Node. Its home address is set as its PDP address in the GGSN/FA when relaying the Mobile IP registration reply message to the Mobile Node. In this way, the association between the Mobile Node’s home address and its TID is established. When the Foreign Agent receives the encapsulated packets by the home agent, the Foreign Agent extracts the original packet, and delivers it to the GGSN part. The GGSN part tunnels this packet to the corresponding SGSN via the GTP tunnel according to the association established above.
This will be described in more detail in the next section.

7.4 Routing

After the initial stage, the associations between a series of identifiers in different participating entities are established. Relying on these associations, the routing of the packets are conducted. Figure 7.9 illustrates all the possible routes. As the FA-CoA mode is mandatory in this model, only one parameter impacts on the routing of the packets, that is, the reverse tunnels. Depending on whether the reverse tunnel is employed or not, there are the following communication loops:

- Reverse tunnel: $A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow A_4 \rightarrow B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow B_4$
- No reverse tunnel: $A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow A_4 \rightarrow B_1 \rightarrow B_2 \rightarrow B_3''$

7.4.1 CN → MN

In this direction, the corresponding node sends a packet $A_1$ to the Mobile Node’s home address. This packet is intercepted by the home agent. The home agent encapsulates $A_1$ into $A_2$ with the Mobile Node’s CoA as the destination address. As the FA-CoA mode is employed in this model, the packet $A_2$ is received by the GGSN/FA that the Mobile Node attaches to. The Foreign Agent functionality part of the GGSN/FA extracts the original packet $A_1$. Then it delivers the $A_1$ to the GGSN functionality part for further processing. The GGSN part looks up its database for this packet’s destination address, that is, the Mobile Node’s home address. As mentioned previously, in Mobile IP registration, the PDP context address of the Mobile Node was updated using its home address, and an association between the home address and TID, which identify the tunnel to the corresponding SGSN, was established. According to this association, the GGSN puts the packet $A_1$ into the GTP tunnel identified by that TID. The SGSN receives the packet $A_1$ from the GTP tunnel. According to the association between TID and TLLI, which was established in the GPRS attachment procedure, packet $A_1$ is delivered to the Mobile Node.
7.4. ROUTING

7.4.2 MN → CN

When the Mobile Node sends packet B1 to the corresponding node, it delivers B1 to the SGSN first. Then B1 is tunnelled to the GGSN/FA by the SGSN. The GGSN part of GGSN/FA extracts the original packet, and delivers it to the Foreign Agent part. If the reverse tunnel is employed, the Foreign Agent encapsulates B1 into the packet B3. The home agent receives B3, extracts B1, and relays it to the corresponding node. Otherwise, if no reverse tunnel is employed, the Foreign Agent part directly delivers packet B1 to the public IP network. Finally, B1 is routed to the corresponding node by the normal IP routing mechanism.
7.5 Handover

This section focuses on the situation where a GGSN/FA handover is required when the Mobile Node attaches to a new SGSN. The GGSN/FA handover is controlled by the SGSN. The GGSN/FA handover would naturally be done after the SGSN handover, but it could also be used for load balancing between two GGSN/FAs.

Figure 7.10 illustrates the message flow when the GGSN/FA handover takes place. After a SGSN handover, the SGSN has the option to change the GGSN/FA. The decision is based on the knowledge that SGSNs have of the GGSNs. If the decision is negative, the PDP Context is kept as normal and the old GGSN is kept. On the other hand, if a GGSN/FA handover must be performed, the following procedure is followed:

1. The new SGSN sends a Create PDP Context Request to the new GGSN with the information that the PDP Context is a Mobile IP PDP Context. The information of the type of the context is put in the APN field as described in Section 7.3.

2. The new GGSN answers with a Create PDP Context Response and creates the connection between the Foreign Agent and the new PDP Context.

3. The new SGSN sends a Delete PDP Context Request to the old GGSN.
4. The old GGSN deletes the PDP Context and responds to the request with a Delete PDP Context Response.

5. The Foreign Agent sends the Mobile Node an Agent Advertisement.

6. Agent registration is performed as described in Section 7.8.

After successful creation of the new PDP Context, a timer can be set. The timer counts down the time until the old PDP Context is deleted. This allows the packets that arrive at the old GGSN/FA to be forwarded to the Mobile Node. The timer can also be set to zero to indicate the absence of a timer. Hence, the PDP Context to the old GGSN is deleted immediately after the new PDP Context is created.

7.6 Conclusion

This chapter described a stepwise approach to offering Mobile IP services in a GPRS network as proposed by the 3GPP. Based on “Step 2” in this proposal, we proposed a model for offering Mobile IP service into GPRS networks. This model requires minimal changes to the GPRS infrastructure. Only a combined entity, GGSN/FA, is introduced to serve as a gateway between GPRS and Mobile IP. Micro-mobility is handled by the GPRS part, while the Mobile IP part is responsible for macro-mobility. This way, GGSN/FA hides the movement of Mobile Nodes in the GPRS network to the Home Agent. This effectively reduces signalling payload and latency.

The next chapter will discuss the integration of Mobile IP and IPSec.
Chapter 8

Integration of IPSec and Mobile IP

Chapter 7 has described the integration of Mobile IP and GPRS. Now Mobile Nodes present a unified Mobile IP interface to Correspondent Nodes, regardless of whether they are located in GPRS networks or 802.11 WLANs. As mentioned in Chapter 6, IPSec is used to protect the communication between Mobile Nodes and their internal domain when they move away. This chapter seeks to address the integration of Mobile IP and IPSec to provide mobility and security at the same time.

The rest of this chapter is organized as follows. The first section will describe the network environment where Mobile IP and IPSec will be deployed. Thereafter the two encapsulation modes for combining the Mobile IP tunnel and IPSec tunnel are discussed. The third section will investigate the existing deployment scenarios where the two encapsulation modes are applied to combine Mobile IP and IPSec tunnels. The fourth section will propose a dual home agent architecture for combining Mobile IP and IPSec, which provides both internal and external mobility.

8.1 Background

Typical corporate networks consist of three different domains: the Internet (untrusted external domain), the intranet (trusted internal domain) where private addresses (Rekhter, Moskowitz, Karrenberg, Groot, & Lear, 1996) are typically used, and the Demilitarized Zone (DMZ), which connects the
two domains (Vaarala & Klovning, 2005), as was illustrated in Figure 4.12 on page 98. Often the firewall and VPN gateway are deployed in the DMZ to guard the access to the internal domain. Access is only allowed if both the firewall and VPN security policies are respected.

In this study when we refer to a VPN we refer to an IPSec-based remote access VPN. Among the three topologies of IPSec, described in Chapter 6, the Host-to-Network topology, shown in Figure 6.6 on page 135, is considered the most suitable for the individual remote user to access the internal domain (Schafer, 2003, p. 207). This deployment works well when an IPSec peer is stationary. However, when an IPSec peer moves, which is called a Mobile Node (MN) in Mobile IP terminology, a problem arises.

The problem is owed to a limitation in the current IPSec version, whereby a new Internet Key Exchange (IKE) negotiation (Harkins & Carrel, 1998) must be done whenever a Mobile Node moves, which means its Care-of Address (CoA) changes. The main reason is that a Security Association is unidirectional and identified by a triplet consisting of (Vaarala & Klovning, 2005):

- The destination address, which is the outer address when tunnel mode is used,
- The security protocol (ESP or AH), and
- The Security Parameter Index (SPI).

Although an implementation is not required to use all of the above for its own Security Associations, it cannot be assumed that another implementation also will not use them. When a Mobile Node sends packets to a stationary IPSec Node, there is no problem. This is because the Security Association is owned by the stationary IPSec Node, and therefore the destination address does not need to change. The source address is often simply ignored by the stationary node although some implementations do check the source address as well.

The problem arises when packets are sent from the stationary node to the Mobile Node. As the Security Associations are unidirectional, the Security Association owned by the Mobile Node is different from the above one. The destination address of this Security Association is established during IKE
8.1. BACKGROUND

negotiation, and is the Care-of Address of the Mobile Node at time of negotiation. Therefore the packets will be sent to the original Care-of Address, not a changed Care-of Address.

A new version of Internet Key Exchange (IKE), IKE2, is supplemented with the mobility extension (MOBIKE) that may solve this mobility problem (Kivinen & Tschofenig, 2005), but it is still in the early stages of development. Thus a demand still exists that mobility should be provided in conjunction with the existing protocol. For this reason Mobile IP needs to be combined with IPSec.

In combining the two, there are two primary questions that needs to be answered:

• What is the order of tunnels? Thus, should the Mobile IP tunnel be first and the IPSec tunnel second, or should it be the other way round?
• Where should the Home Agent be deployed? Should it be behind the IPSec gateway, in front of IPSec gateway, or co-located with it?

As will be seen later, these two questions are not independent of each other, instead, they are somewhat intertwined.

Before attempting to answer these two questions, some notations are introduced first, for the sake of future discussion. Some of the notations were introduced in Chapter 1, and are repeated here for convenience:

• i-MN refers to the Mobile Node (MN) when it is located in the internal domain.
• x-MN refers to the Mobile Node when it is located in the external domain.
• i-CN refers to the Correspondent Node (CN) in the internal domain.
• x-CN refers to the Correspondent Node (CN) in the external domain.
• i-Foreign Network refers to the foreign network inside the internal domain.
• x-Foreign Network refers to the foreign network outside the internal domain.
Besides the above notations, there are others, which uses the same naming convention:

- **i-MIP** refers to an instance of Mobile IP that runs inside the internal domain, which uses public address space.

- **x-MIP** refers to an instance of Mobile IP that runs outside the internal domain, which uses private address space.

- **i-HA** refers to the Home Agent (HA) that resides in the internal network.

- **x-HA** refers to the Home Agent (HA) that resides outside the internal network.

- **i-FA** refers to the Foreign Agent (FA) that resides in the internal network.

- **x-FA** refers to the Foreign Agent (FA) that resides outside the internal network.

- **i-CoA** refers to the Care-of Address (CoA) that is registered with i-HA.

- **x-CoA** refers to the Care-of Address (CoA) that is registered with x-HA.

- **i-HoA** refers to the home address of the Mobile Node in its i-HA.

- **x-HoA** refers to the home address of the Mobile Node in its x-HA.

Finally, some points about these notations need to be highlighted. Firstly, note that i-MN and x-MN represent different roles of the same entity, rather than different entities. When a Mobile Node moves inside the internal domain, it is referred to as i-MN. When that Mobile Node moves outside the internal domain, it is referred to as x-MN: although it is the same entity, it is fulfilling a different role. From the view of the administrative domain, the Mobile Node always belongs to the internal domain, regardless of whether it is located inside or outside. This is represented as a Security Association between the Mobile Node and the internal domain. In contrast, x-CN always represents an external domain node, while i-CN always represents an internal domain node.
Secondly, i-CoA and x-CoA have the same effect on i-MN and x-MN. At any point in time, a Mobile Node (MN), regardless of playing a role of i-MN or x-MN, attaches to one foreign network only. Its CoA is always acquired from the foreign network currently attached, regardless of it being the i-Foreign Network or the x-Foreign Network. When this CoA is registered with its i-HA, it is referred to as i-CoA, and when it is registered with its x-HA, as x-CoA. Although it is the same address, it has a different name, depending on the network.

Thirdly, there is only one home network for a Mobile Node, which, in the context of this study, always resides in the internal domain.

8.2 Two Modes of Combining Mobile IP and IPSec Tunnels

When combining the Mobile IP tunnel and IPSec tunnel, the original packet has to be encapsulated into a new IP packet twice. Depending on the order of the encapsulation, there are two combination modes (Adrangi & Levkowetz, 2005):

- Mobile IP tunnel inside IPSec tunnel, and
- IPSec tunnel inside Mobile IP tunnel.

This section will discuss these two combination modes from the perspective of their packet encapsulation. As will be seen later, the order of encapsulation regarding Mobile IP and IPSec tunnels has a significant impact on these two modes’ characteristics. There are also many other parameters that impact on the characteristics of two modes, for example, Foreign Agent Care-of Address (FA-CoA) or Co-located Care-of Address (Co-CoA), reverse tunnel or non-reverse tunnel. These will be discussed after introducing the essentials of the two modes.

8.2.1 Mobile IP Tunnel inside IPSec Tunnel

In this mode, the original packet is encapsulated into the Mobile IP tunnel first, and thereafter encapsulated into the IPSec tunnel, as illustrated in Figure 8.1(a). The whole operation flow is described hereafter:
The original packet destined for the Mobile Node (MN) is intercepted by its Home Agent (HA).

The Home Agent encapsulates this packet into a Mobile IP tunnel between the Home Agent’s address and the Mobile Node’s Care-of Address (CoA).

This encapsulated packet is routed to the IPSec gateway to go outside.

The gateway inspects this packet and checks its Security Association database. When a Security Association is found between the Mobile Node’s CoA and the gateway, the gateway encapsulates this packet once again, but this time into an IPSec tunnel between the Mobile Node’s CoA and the gateway.

When this packet arrives at the Mobile Node, the Mobile Node’s IPSec software strips off the outer IP header. The Mobile Node’s mobile IP software then strips off the inner IP header. Finally, the packet sent by the Correspondent Node (CN) to the Mobile Node is delivered to the upper layer.
8.2. TWO MODES OF COMBINING MOBILE IP AND IPSEC TUNNELS

Figure 8.2: IPSec Tunnel inside Mobile IP Tunnel

The problem with this mode is that the IPSec tunnel’s end is the Mobile Node’s CoA address, and, therefore, the IPSec tunnel has to be renegotiated every time when the Mobile Node changes its point of attachment (Adrangi & Levkowetz, 2005). Since the tunnel renegotiation is a very time-consuming and resource-consuming operation, this combination mode in practice does not provide efficient mobility support, especially for real time applications.

As such, the reverse traffic from the Mobile Node to the Correspondent Node has to traverse the IPSec gateway to reach the inside Correspondent Node, as illustrated in Figure 8.1(b). It is worth noting that when considering compatibility with ingress filtering, described in Section 5.2 on page 113, no reverse tunnel is required here. The IPSec tunnel alone can have the same effect as what the reverse tunnel does. Thus no Home Agent is involved in reverse tunnelling, which can be seen as a side effect of the IPSec tunnel.

8.2.2 IPSec Tunnel inside Mobile IP Tunnel

In this mode, the original packet is encapsulated into the IPSec tunnel first, and in turn encapsulated into the Mobile IP tunnel, as illustrated in Figure 8.2. The whole operational flow is described below:

- The original packet destined for the Mobile Node (MN) is routed to the IPSec gateway to go outside.

- The gateway looks up its Security Association database. It can find a Security Association between the Mobile Node’s home address and the gateway. It is worth noting that this is significantly different from the Mobile IP tunnel inside IPSec tunnel mode where an end of the IPSec tunnel is the Mobile Node’s Care-of Address (CoA).
CHAPTER 8. INTEGRATION OF IPSEC AND MOBILE IP

- The gateway encapsulates the original packet into an IPSec tunnel with the gateway’s IP as source address and the Mobile Node’s home address as the destination address.

- This packet is routed to the Home Agent (HA) as per normal.

- The Home Agent encapsulates this packet once again, but this time into an IPSec tunnel between Mobile Node’s CoA and the Home Agent’s address as the normal Mobile IP protocol does.

- At the Mobile Node end, the Mobile IP software strips off the outer IP header first. Then IPSec software strips off the inner IP header. The original packet is delivered to the upper layer.

As another end of the IPSec tunnel is the Mobile Node’s home address, the IPSec tunnel does not need to be renegotiated when the Mobile Node changes its point of network attachment. This effectively solves the problem of the previous mode. However, this is achieved at the expense of leaving the Home Agent unprotected by the IPSec gateway. Thus the Home Agent should have its own authentication mechanism in place to defend against spoofing attacks.

When considering the reverse traffic from the Mobile Node to the Correspondent Node (CN), the IPSec tunnel has the same effect as the reverse tunnel without Home Agent involvement, as in the previous mode. This can be seen as an optimization for the combination. Normally, where the reverse tunnel is required, there should be two tunnels. By making the IPSec tunnel perform the functionality of the reverse tunnel, this eliminates the necessity for an extra Mobile IP reverse tunnel.

The outer tunnel plays a different role in each of the modes. In a Mobile IP tunnel inside IPSec tunnel mode, the outer tunnel (IPSec tunnel) acts like a hider, hiding the inner tunnel (Mobile IP tunnel) from the outside. In an IPSec tunnel inside Mobile IP tunnel, the outer tunnel (Mobile IP tunnel) acts like an expander, expanding the inner tunnel (IPSec tunnel) to the Mobile Node.
8.3. Five Scenarios of Deploying HA and IPSec Gateway

Besides the above mentioned two modes of combining Mobile IP and IPSec tunnels, there are also five scenarios of deploying the Home Agent (HA) and IPSec gateway, identified by IETF (Vaarala & Klovning, 2005), where the above modes can be applied. These scenarios are based on the network environment shown in Figure 8.3. As this chapter’s focus is on the inter-operation between Mobile IP and IPSec VPN, firewalls are purposely omitted from the following scenarios in order to keep things simple.

![Figure 8.3: Network Environment](image)

8.3.1 Home Agent Inside the Intranet behind a VPN Gateway

In this scenario, the Home Agent is deployed inside the intranet protected by an IPSec gateway, thus denoted as i-HA, and is not directly reachable by the external Mobile Node (x-MN) outside the intranet. Figure 8.4 illustrates this scenario.

Direct application of Mobile IP standards, denoted as i-MIP in this scenario, is successfully used to provide mobility for the internal Mobile Node
Figure 8.4: Home Agent behind the VPN Gateway (Adragi & Levkowetz, 2005)

(i-MN). This corresponds to the communication path $i$-MN $\leftrightarrow$ $i$-CN. However, when considering another part of the internal mobility, $x$-MN $\leftrightarrow$ $i$-CN, a problem arises. The $x$-MN can only access the intranet resources through the IPSec gateway, which will allow only authenticated IPSec traffic to go inside. This implies that the Mobile IP traffic has to run inside IPSec, which leads to two distinct problems:

1. When the $x$-Foreign Network uses the FA Care-of Address mode, Mobile IP registration becomes impossible. This is because the Mobile IP traffic between $x$-MN and the IPSec gateway is encrypted, and the $x$-FA, which is likely to be in a different administrative domain, cannot inspect the Mobile IP headers needed for relaying the Mobile IP packets.

2. In Co-located Care-of Address mode, successful registration is possible, but the IPSec tunnel has to be re-negotiated every time that the $x$-MN changes its point of network attachment. This is a very time-consuming and resource-consuming operation, which in effect makes mobility impossible, as discussed in Section 8.2.1.

Recalling the 3G scenario in Figure 1.3 on page 10, the above analysis only involves the communication path $i$-MN $\leftrightarrow$ $i$-CN and $x$-MN $\leftrightarrow$ $i$-CN. In other words the analysis involves the access to the intranet resource regardless of the Mobile Node being inside or outside the intranet. The Mobile Node, regardless of being inside or outside the intranet, and $i$-CN, are both entities in the internal domain, which has a Security Association with the IPSec
8.3. FIVE SCENARIOS OF DEPLOYING HA AND IPSEC GATEWAY

It is therefore not a problem for x-MN or i-CN to traverse the IPSec gateway. It is referred to as ‘internal mobility’ because all traffic takes place between the internal entities, as mentioned in Section 1.3.4 on page 14.

Note that this scenario does not provide ‘external mobility’, which covers communication path i-MN ↔ x-CN and x-MN ↔ x-CN, as described in Section 1.3.4. This is due to two reasons. Firstly, the Mobile Node uses a private address. It is impossible for x-CN to address the x-MN. Secondly, due to the lack of a Security Association between x-CN and the IPSec gateway, it is impossible for x-CN to traverse the IPSec gateway to address i-MN.

8.3.2 Home Agent and IPSec gateway on the VPN Domain Border

In this scenario, the Home Agent (HA) is deployed on the VPN domain border, that is, DMZ, together with the IPSec gateway, and it is directly reachable by both the x-MN and i-MN. Figure 8.5 illustrates this scenario.

The Home Agent has a public interface to the external domain, and a private interface to the internal domain. Similar to the previous scenario, the standard Mobile IP successfully provides mobility to the internal Mobile Node (i-MN). In the case of x-MN, besides the “Mobile IP tunnel inside IPSec tunnel” mode, which only the previous scenario can use, the Home Agent’s public interface makes another mode possible, that is, “IPSec tunnel inside Mobile IP tunnel”. The former has similar problems to the previous scenario. However, the Home Agent’s public interface makes a small difference, namely that in the Foreign Agent Care-of Address (FA-CoA) mode Mobile IP regis-
tration can be directly delivered to the Home Agent via its public interface without being encapsulated in IPSec tunnel, so that it can pass by the IPSec gateway. This requires that an authentication mechanism must be in place at registration to protect against spoofing attacks. The registration is thus not a problem regardless of FA-CoA mode or Co-located Care-of Address (Co-CoA) mode. However, the problem of renegotiation every time that the Mobile Node changes its point of attachment still exists.

The IPSec tunnel inside the Mobile IP tunnel does not present these problems. However, it does require some modifications to the routing logic of the Home Agent or the IPSec gateway.

The above discussion considers internal mobility only. External mobility are made possible by the Home Agent’s public interface. Besides an internal home address (i-HoA), which enables addressability for i-CN, an external home address (x-HoA) can be assigned to the Mobile Node so that an x-CN can address it via its x-HoA. In the case of x-MN, this presents no problem. However, in the case of i-MN, due to the lack of Security Association between x-CN and i-MN, it is difficult for the x-CN to traverse the IPSec gateway to address i-MN. Some policy should be put in place so that the communication between x-CN and i-MN does not compromise the security of the internal domain.

In this scenario, it is difficult to define the role of a Home Agent, as it runs in both public and private address space. In this sense, it can been seen as an i-HA when it distinguishes a Mobile Node using that Mobile Node’s i-HoA, which presents internal mobility. It can also been seen as an x-HA when it distinguishes a Mobile Node using that Mobile Node’s x-HoA, which presents external mobility. For this reason, it is just denoted as HA in Figure 8.5.

### 8.3.3 Combined VPN Gateway and Home Agent

This scenario is similar to the scenario described in Section 8.3.2, with the exception that the IPSec gateway and Mobile IP Home Agent are running on the same physical machine, as illustrated in Figure 8.6.

Running the Mobile IP Home Agent and VPN on the same machine resolves routing-related issues that existed in Section 8.3.2 when an “IPSec tunnel inside Mobile IP tunnel” configuration is used. However, it does not satisfy multi-vendor inter-operability in environments where Mobile IP Home
8.3. FIVE SCENARIOS OF DEPLOYING HA AND IPSEC GATEWAY

Agent and VPN technologies must be acquired from different vendors.

8.3.4 Home Agent outside the VPN Domain

In this scenario, the Mobile IP Home Agent is deployed outside the internal domain, as illustrated in Figure 8.7.

The Home Agent has a public interface only. Though, in corporation with the IPSec gateway, it nevertheless provides partial internal mobility. The condition is that it must run in an IPSec tunnel inside Mobile IP tunnel mode. In the case of x-MN, the x-MN registers its Care-of Address with its Home Agent first. Then the x-MN negotiates with the IPSec gateway and establishes an IPSec tunnel, with its x-HoA as outer endpoint and its
i-HoA as inner endpoint. In this way, when i-CN sends a packet to its i-HoA, the packet is routed to the IPSec gateway. The IPSec gateway tunnels this packet to x-MN’s x-HoA. The x-MN’s Home Agent intercepts this tunnelled packet, and further tunnels it to x-MN’s CoA address.

However, in the case of i-MN, it is impossible for it to register with its Home Agent, because the latter does not have a private interface. Similarly, it is impossible for an x-CN to address i-MN due to the lack of the Security Association between them. This scenario thus does not support mobility inside the internal domain.

It is still possible for an x-CN to address x-MN via its x-HoA. This scenario thus also provides partial external mobility.

### 8.3.5 Combined VPN Gateway and Home Agent on the Local Link

This is similar to the deployment scenario described in Section 8.6, with the difference that the IPSec gateway/Home Agent is sitting on the local link. The IPSec gateway/Home Agent is assumed to be reachable from the external network. In other words, it is assumed to have a public IP address. The firewall is assumed to be configured to allow direct access to the VPN/Home Agent from the external network. Figure 8.8 depicts this scenario.

This deployment can work without any technical problems with IPSec tunnel inside Mobile IP tunnel mode. In Mobile IP inside the IPSec tunnel mode, it presents the same problems as reported in Section 8.3.1. This option is not practical for large deployments because of the extent of the distributed...
8.3.6 Evaluation of Scenarios

Adrangi and Levkowetz (2005) have evaluated the above mentioned deployment scenarios to identify those most in need of solving. The evaluation was based on the following two main criteria:

1. Is the deployment scenario common and practical?
2. Does the deployment scenario reveal any problems resulting from Mobile IP and VPN coexistence?

Adrangi and Levkowetz (2005) believed the scenario in Section 8.3.1 to be the most important and practical because of a rising need for providing corporate remote users with continuous access to their Intranet resources. The problems occurring in scenarios in Sections 8.3.2 and 8.3.4 are either the same as those in the scenario in Section 8.3.1 or a subset of them. Therefore, solving the scenario in Section 8.3.1 will also solve the scenarios in Sections 8.3.2 and 8.3.4. The scenarios in Sections 8.3.3 and 8.3.5 do not introduce extra functional problems resulting from Mobile IP and IPSec coexistence than the scenario in Section 8.3.1, and thus there is no need to seek a solution. A solution for the scenario in Section 8.3.1 is therefore seen as essential, and can in turn also be applied to solve problems in other scenarios.

Based on this evaluation, the next section will focus on the scenario in Section 8.3.1 and examine the existing challenges of this scenario towards a complete solution to Mobile IP and IPSec coexistence.

8.4 Dual Home Agent Architecture

As mentioned above, the scenario shown in Figure 8.4 is seen as the essential scenario. It provides mobility for the i-MN perfectly using standard Mobile IP. However, in the case of x-MN, problems do arise, which are repeated below:

- Registration in Foreign Agent Care-of Address (FA-CoA) mode
- Renegotiation when the MM changes its attachment of point
The first problem has an alternative solution, namely using only the Co-located Care-of Address mode. The second problem is seen as an inherent problem in this scenario due to the nature of the Mobile IP tunnel inside the IPSec tunnel, namely that an end of the IPSec tunnel is the Care-of Address (CoA) of the Mobile Node (MN), which will change every time when the Mobile Node changes its point of attachment.

As discussed in Section 8.2.2, this is not a problem for the IPSec tunnel inside Mobile IP tunnel mode, because in that case an end of the IPSec tunnel is the Home Address (HoA) of the Mobile Node, which never changes. Based on this characteristic of the IPSec tunnel inside Mobile IP tunnel mode, an extra Home Agent is introduced to encapsulate the IPSec tunnel, which has already encapsulated a Mobile IP tunnel, into another most outer Mobile IP tunnel (Vaarala & Klovning, 2005), as illustrated in Figure 8.9. This architecture has been implemented in a multimedia test-bed by Dutta et al. (2004).

8.4.1 Network Environment

For the above architecture, there are other parameters that have an impact on it, for example, Foreign Agent Care-of Address (FA-CoA) or Co-located Care-of Address (Co-CoA) mode. The IETF has thoroughly investigated various possible situations, and presented a complete reference network environment (Vaarala & Klovning, 2005), as illustrated in Figure 8.3 on page 175.

Every subnetwork in Figure 8.3 demonstrates a situation where the Mobile Node (MN) can establish a connection to the corresponding Home Agent (HA) by using an appropriate access mode. According to Vaarala and Klovn-
ing (2005), an access mode is defined as consisting of:

1. A composition of the Mobile Node networking stack: i-MIP or x-MIP/VPN/i-MIP.

2. Registration mode of i-MIP and x-MIP (if used): FA-CoA or Co-CoA.

Each possible access mode is denoted as “xyz”. Its meaning is described below:

- “x” indicates whether the x-MIP layer is used. If used, “f” indicates FA-CoA, “c” indicates Co-CoA. Otherwise, absence indicates not used.
- “y” indicates whether the VPN layer is used. “v” indicates VPN used, and absence indicates not used.
- “z” indicates mode of i-MIP layer (it is always used). “f” indicates FA-CoA, and “c” indicates Co-CoA.

When combined, they result in four access modes. When the Mobile Node is inside the Intranet (i-MN):

- c: i-MIP with Co-CoA
- f: i-MIP with FA-CoA

When the Mobile Node is outside the Intranet (x-MN):

- cvc: x-MIP with Co-CoA, VPN-TIA as i-MIP Co-CoA
- fvc: x-MIP with FA-CoA, VPN-TIA as i-MIP Co-CoA

The former two modes are covered by mobility inside the Intranet while the latter two modes is about mobility outside the Intranet.

### 8.4.2 Registration

Depending on where the Mobile Node is located, the registration procedure varies. In the case of i-MN, the normal registration procedure as described in Section 5.1.2 (see page 106) is performed. However, the i-MN needs to register twice, once with i-HA and once with x-HA via its private interface.
When the i-MN registers with i-HA, the Home Address field in the registration request message, shown in Figure 5.4 (see page 108), is filled with its i-HoA and the Care-of Address (CoA) field is filled with its CoA, which is called i-CoA in this case. Note that at this time the CoA is a private address. When the i-MN registers with x-HA, the Home Address field is filled with its x-HoA. The Care-of Address (CoA) field is also filled with its CoA, but it is called x-CoA in this case. As such, the associations between an i-MN’s i-HoA and its CoA, and between its x-HoA and its CoA are established so that the packets destined to its i-HoA from the i-CN and destined to its x-HoA from the i-CN can be relayed to its CoA, as shown in Figure 8.10. In the figure, it is assumed that the Mobile Node has physical address, “89 E3 A2 73 2B A8”.

In the case of x-MN, the registration procedure becomes much more complicated. It does not only involve the Mobile IP registration procedure, but also IPSec tunnel establishment if needed. It involves the following main steps:

- Registers its new Care-of Address, acquired from an x-Foreign network, with its x-HA. It is called x-CoA.
8.4. DUAL HOME AGENT ARCHITECTURE

- Establishes an IPSec tunnel between its x-HoA and the IPSec gateway inside the DMZ of its enterprise network.

- Registers the tunnel inner address (TIA) of the IPSec tunnel with the i-HA. It is called i-CoA. This will cause the i-HA to tunnel packets sent to the Mobile Node’s i-HoA to the IPSec gateway, which will then tunnel the packets through the IPSec tunnel and the x-MIP tunnel to the Mobile Node.

- When the Mobile Node moves back to the internal domain, the IPSec and the x-MIP tunnels will be torn down.

Figure 8.11 illustrates how the Mobile IP and IPSec tunnels are set up during the Mobile Node’s movement from the internal domain to the external domain.

8.4.3 Handover

When the Mobile Node is located in the internal domain, the handover procedure is as normal, which was discussed in Section 5.3.2 on page 118. When the Mobile Node hands over from the internal domain to the external domain, the procedure depicted in Figure 8.11 takes place. However, when the Mobile Node moves from an external network to another external network, this procedure does not need to take place again. The Mobile Node only
needs to perform the normal Mobile IP registration procedure to register its new x-CoA with its x-HA. Since the Mobile Node’s x-HoA remains the same, the Mobile Node’s IPSec tunnel does not need to renegotiate. It looks as if the x-MIP tunnel curved the IPSec tunnel and the i-MIP tunnel within it so that the IPSec packets can reach their new destinations without breaking IPSec tunnelling, as illustrated in Figure 8.12

### 8.4.4 Routing

Besides the normal triangle routing rendered by Mobile IP, in the most complicated case routing in this dual Home Agent architecture requires tri-tunnelling, an i-MIP tunnel, IPSec tunnel and x-MIP tunnel. This takes place in the case of x-MN ↔ i-CN. The packets destined to x-MN from i-CN are tunnelled to the IPSec gateway first by i-HA. Then the IPSec gateway tunnels tunnelled packets to the x-HA. Upon receipt of these packets, twice tunnelled, the x-HA tunnels them to the Mobile Node’s CoA.

In the case of x-MN ↔ x-CN, normal triangle routing takes place between x-CN, x-HA and x-MN. Similarly, in the case of i-MN ↔ i-CN, it takes place between i-CN, i-HA and i-MN; and in the case of i-MN ↔ x-CN, between x-CN, x-HA and i-MN.

### 8.5 Conclusion

This chapter investigated the various scenarios of combining Mobile IP and IPSec. Based on an essential scenario, a dual Home Agent architecture was
proposed to solve the existing problems, namely registration in Foreign Agent Care-of Address mode and IPSec renegotiation when the Mobile Node’s Care-of Address changes. This is performed by introducing an extra external Home Agent. This leads to triple tunnelling (i-MIP, IPSec and x-MIP). Although it adds encapsulation overhead, it indeed solves the above mentioned problems. This achieved one of this study’s objectives, namely incorporating Mobile IP and IPSec.

At this point, the two parts of the framework, the integration of GPRS and Mobile IP and the integration of IPSec and Mobile IP, have been discussed separately. The next chapter will combine these two parts to propose a complete framework for achieving this study’s goal.
Chapter 9

A Complete Framework

As mentioned in Chapter 1, this study attempts to produce a framework for enterprise users that realize seamless handoff between GPRS and WLAN, as well as seamless global roaming with security in place. Towards this goal, Chapter 7 discussed one part of the framework: offering Mobile IP service to GPRS. Chapter 8 discussed another part of the framework: the integration of IPSec and Mobile IP. This chapter seeks to propose a complete framework based on the discussion of the previous two chapters, which complies with the goal of this study.

This framework addresses both internal and external mobility issues. The following sections will describe them respectively in terms of the mobile agent discovery, Mobile IP registration, and IP routing.

9.1 Overview of the Framework

A typical DMZ network topology is adopted in this framework. It consists of the intranet, DMZ, and the external network, as illustrated in Figure 9.1. Similar to the typical configuration in this kind of deployment, the intranet uses the private address space while the external network belongs to the public address space. The x-HA and IPSec gateway are located in the DMZ. Both of them have two network interfaces: one connects to the external network (public address space); another one connects to the intranet (private address space). As seen later, this subtle difference enables the i-MN’s reachability from outside. Except for this aspect, nothing new is involved. In the external network, the combination of GPRS and WLAN described in Chapter 7 are
CHAPTER 9. A COMPLETE FRAMEWORK

Figure 9.1: Framework
entirely adopted here. The bottom of Figure 9.1 shows the combination of all necessary tunnels in the most complicated case where an i-CN communicates with an x-MN. Note that the combination is different in two directions. The combination above shows the case of i-CN $\rightarrow$ x-MN while the combination below shows the case of x-MN $\rightarrow$ i-CN. These tunnels will be discussed in detail later.

9.1.1 Address Scheme

As aforementioned, the private address scheme is adopted in the intranet. The DMZ connects the intranet to the Internet with security in place. Generally, the GPRS backbone also uses the private address space except for the GGSN, which in addition to the private address interface, also has a public address interface to the Internet. According to this address scheme, a possible implementation of the framework is illustrated in Figure 9.2.

In this figure, the GGSN/x-FA, IPSec, and x-HA have both a public address and a private address, which serve as the gateway between the internal world and external world.

The MN’s configuration is much more complex than others. It needs two home addresses, i-HoA and x-HoA, as discussed in Section 8.4 on page 181. Correspondingly, it has two care-of addresses, i-CoA and x-CoA. Depending on where the MN is located, they have different significance. When the MN is inside the intranet where it plays a role of i-MN, it registers i-HA with the address of the attached i-FA’s address, denoted as i-CoA. For example, if the MN in Figure 9.2 attaches to the foreign network 192.168.1.0, its i-CoA is that network’s i-FA’s address, 192.168.1.1. It also registers x-HA with this address via x-HA’s private address interface, 192.168.3.5. In other words, its x-CoA also is 192.168.1.1.

When the MN is outside the intranet, denoted as x-MN, it registers x-HA with the address of the attached x-FA’s address, denoted as x-CoA. For example, if the MN in Figure 9.2 attaches to the above GGSN/x-FA (57.23.4.1), its x-CoA is 57.23.4.1. It also registers i-HA with its TIA. That is 45.2.1.7. In other words, its i-CoA also is its TIA, namely 45.2.1.7.
9.1.2 Protocol Stack

This framework involves a large number of communicating entities and protocols, which varies in different communication pathes. All possible paths include:

- i-CN ↔ i-MN
- i-CN ↔ x-MN
- x-CN ↔ i-MN
- x-CN ↔ x-MN

The lower layer network transport technologies also vary, but they all serve as a bearer for the upper IP network layer. Here, omitting diversification between them, they are totally abstracted as a link layer bearer.
Figure 9.3 shows the interaction of various communicating entities and protocols in each communication path. Note that the figure shows the cases in the CN → MN direction only. In the reverse direction the interaction is somewhat different. This will be discussed in detail later in the form of the packet encapsulation format in Section 9.3 (See page 200) and Section 9.4 (See page 200).

### 9.1.3 Security Services

In this framework, the security services are offered by IPSec. Consulting Figure 9.3, among all communication paths only i-CN ↔ x-MN involves the IPSec gateway. This means that security services offered by IPSec is only applicable to this communication path. These services includes:
• Data origin authentication
• Data integrity
• Confidentiality
• Replay Protection

Depending on the security association agreed to by the communicating entities, the range of security services varies. The detail of these security services have been discussed in Chapter 6.

Although the intranet is generally considered trusted and regarded as a safe domain, this is applicable to the wired network only. The wireless link is considered much less safe than the wired link. Thus security mechanisms for the wireless link are necessary, especially in the cases where no IPSec is involved, such as i-CN ↔ i-MN, x-CN ↔ i-MN, and x-CN ↔ x-MN. There is an existing standard for wireless link security, Wired Equivalent Privacy (WEP) (Schafer, 2003, p. 303). It is a node-to-node security mechanism. Since the entities in the intranet are considered the trusted nodes, the deployment of WEP in the intranet offers some degree of security in communication path i-CN ↔ i-MN. But for two other cases, due to the introduction of the external untrusted entities, WEP alone does not render the necessary security guarantee. A detailed discussion of the WEP falls outside of the scope of this study. Many books, though, focus on this topic. One of them is “Security in Fixed and Wireless Networks: an Introduction to Securing Data Communications” (Schafer, 2003).

The paths, x-CN ↔ i-MN and x-CN ↔ x-MN, are not under protection of the IPSec. This is because there is no security association between the x-CN and Mobile Node, and it is difficult to establish a security association between them in IPSec. This study suggests that other mechanisms, such as SSL (Freier et al., 1996) and TLS (Dierks & Allen, 1999), be deployed to protect the communication in these two paths. However, this study is not going to focus on these mechanisms.

9.1.4 Hierarchical Mobility

This framework deals with mobility in a hierarchical fashion, including two levels: macro-mobility and micro-mobility, as illustrated in Figure 9.4. All
WLAN Foreign Agents fall in the level of macro-mobility. The GPRS network, GPRS-relative parts, GGSN and SGSN, deal with micro-mobility while the Foreign Agents, next to the GGSNs, deal with macro-mobility.

### 9.1.5 Authentication Agreement

When the Mobile Nodes roam between different wireless networks, this involves multiple administrative domains. The intranet, generally, is owned by an organization. The GPRS network is owned by an operator, which is mostly not the owner of the intranet as well. Furthermore, the external foreign networks may be owned by a third party or in some cases it may belong to the GPRS operator. Authentication agreements need to exist between different administrative domains to allow the Mobile Nodes access to the wireless networks. This is a precondition before seamless handover can be conducted.

This, however, falls outside of the scope of this study. Except for technical issues, there is also commercial and legal issues to be dealt with.

### 9.1.6 Handover

For seamless handover between the GPRS and WLAN, the Mobile Nodes must be able to receive and send the signal of both the WLAN and GPRS base station. This can be achieved by using a dual network interface card.
(NIC). This card can receive both signals simultaneously and measure their signal-to-noise ratio (SNR). When the Mobile Node moves to the boundary of the network, depending on the quality of the signal, the Mobile Node determines the point of the attachment network.

If a handover is required as a result of determination, it takes place in the link layer first. If this handover leads to a switch from the WLAN to the WLAN or from the GPRS to the GPRS, it is identified as a horizontal handover. If it leads to a switch from the WLAN to the GPRS, it is identified as a vertical handover, and vice versa.

After the link layer handover, a network layer handover is performed. This is where Mobile IP comes in. The normal Mobile IP initial procedures are performed, that is, agent discovery and registration.

Figure 9.5 illustrates the concept of the handover. When the MN moves from left to right, it constantly measures the signal-to-noise ratio (SNR) of all possible wireless networks. Based on the measured SNR, it switches from WLAN1 to WLAN2 to GPRS to WLAN3. Corresponding to the link layer switch, its network layer switches from FA1 to FA2 to GGSN/FA to FA3.
9.2 Network Layer Handover

We expanded Figure 9.5 into more detail and an illustration of a more detailed handover scenario is found in Figure 9.6. This figure focuses on the network layer handover only.

Each handover that occurs when the Mobile Node moves from left to right is numbered. Among them, handovers 2 and 4 lead to an administrative domain switch from inside to outside. Handover 3 leads to an opposite switch, that is, from outside to inside. Other handovers only switch access networks in the same administrative domain. This can be seen as another approach to categorizing the handovers, namely the intra-administrative domain handover and the inter-administrative domain handover.

9.2.1 Handover from Inside to Outside

As discussed previously, when handovers 2 and 4 in Figure 9.6 occur, the Mobile Node attaches to the external domain. At this point, the Mobile Node is denoted as x-MN. After link layer handover has been performed, the x-MN needs to register with its x-HA, establish a security association with IPSec Gateway, and then register with its i-HA. This was introduced in Section 7.2. What is new to this framework is that the GPRS attachment and PDP context activation procedures must be performed first, when switching to the GPRS network before the above mentioned Mobile IP procedure can start.
A detailed description of the GPRS procedures was given in Section 2.4. A sequence diagram of the combination of the GPRS, Mobile IP, and IPSec procedures is shown in Figure 9.7.

- **GPRS Attachment**: its major purpose is to establish link layer association between the Mobile Node and the SGSN. Each Mobile Node is allocated a Temporary Logical Link Identifier (TLLI) to identify them in the SGSN. The detailed description can be consulted in Section 2.4.1 on page 49.

- **PDP Context Activate**: the previous GPRS attachment procedure assigns an identifier that the GPRS network knows while the PDP Context Activate procedure assigns an identifier to the Mobile Node that the external IP network nodes knows, that is, the IP address. A GTP tunnel is established between the GGSN and SGSN. It is identified by a specific tunnel identifier (TID) for each Mobile Node. Section 2.4.3 on page 53 provides more details.

- **x-FA Agent Discovery**: after the IP-based communication tunnel has been established, the x-FA can advertise its presence to the Mobile Node via this tunnel. Upon the receipt of this advertisement, an Agent discovery procedure can be conducted in the Mobile Node, as described in Section 5.1.1 on page 104.
9.2. NETWORK LAYER HANDOVER

- Mobile IP registration with x-HA: in this procedure, x-MN register its x-CoA with the x-HA. At this point a mobile IP tunnel is established between x-HA and the x-MN’s x-CoA, that is, x-FA’s IP address.

- IKE negotiation: x-MN issues an IKE negotiation with the IPSec gateway. Note that x-MN directly communicates with the IPSec gateway without the x-HA being involved. At this point, an IPSec tunnel is established between IPSec gateway and the x-MN’s x-HoA. Its TIA is the x-MN’s x-HoA.

- Mobile IP registration with i-HA: the x-MN registers its TIA as its i-CoA with the i-HA via the IPSec tunnel without x-HA being involved. Hereafter, another Mobile IP tunnel is established between i-HA and the x-MN.

With these procedures completed, the x-MN is attached to the external domain. For each handover where the administrative domain does not change, regardless of it being horizontal handover, like handover 6, or a vertical handover, like handover 5, 7 and 8, the x-MN only needs to register x-HA with its new care-of address, or in addition launch GPRS procedures in the case of GPRS network being attached. No extra IPSec security association establishment procedure is required. In this way, the IPSec tunnel is not broken when x-MN hand over outside the intranet. A detailed discussion of this was given in Section 8.4 on page 181.

9.2.2 Handover from Outside to Inside

When handover 3 in Figure 9.6 occurs, it means the x-MN returns to its intranet. At this point, it is denoted as i-MN. No IPSec is needed in this situation. The i-MN tears down the IPSec tunnel that was established before. It then receives the Agent Advertisement, acquires the care-of address, and registers i-HA with its new care-of address, which is denoted as i-CoA in this case.

Then i-MN also registers x-HA with its new care-of address via x-HA’s private interface, which is denoted as x-CoA, although it is the same as the i-CoA.
CHAPTER 9. A COMPLETE FRAMEWORK

When the i-MN roams inside the intranet, for each new i-FA that it attaches to, besides registering with i-HA, the i-MN also registers its new i-CoA with the x-HA via the latter’s private interface. Consulting the registration format in Figure 5.4 on page 108, the Home address field is x-HoA, and the CoA field is i-CoA, which is different from the case of registration with i-HA where the Home address field is i-HoA, and the CoA field is also i-CoA. This establishes the association between the i-MN’s x-HoA and its i-CoA, which allows i-MN reachability from the x-CN.

9.2.3 Smooth Handover

To support seamless mobility, it is important to reduce handoff delay and transient data loss during handoff. Establishing VPN tunnels or establishing connectivity to a cellular data network could introduce excessive delays that are intolerable to real-time applications. This framework adopts make-before-break handoff mechanisms to reduce handoff delay and data loss during handoff.

9.3 Internal Mobility

As described in Section 1.3.4 on page 14, this study limits the scope of the internal mobility to the communicating entities pertaining to the internal administrative domain, including the i-CN and Mobile Nodes, both inside the intranet and outside the intranet. Consequently, there are two possible communication paths:

- i-CN ↔ i-MN, also known as mobility inside the intranet
- i-CN ↔ x-MN, also known as mobility outside the intranet

The rest of this section will discuss the routing in these two situations, respectively.

9.3.1 Routing between i-CN and i-MN

Besides two communicating entities (i-CN and i-MN), this scenario also involves other intermediate entities, such as i-HA and i-FA. Since all entities
are inside the intranet, no traffic goes beyond the IPSec gateway. Thus the IPSec Gateway is not involved in this situation. All the mobility procedures, such as agent discovery, registration and routing, are performed similar to how the normal Mobile IP does, which were described in Chapter 5. Note that the i-MN not only registers i-HA, but also x-HA for the i-MN’s reachability from outside, as mentioned in Section 9.2.2. Routing still adopts a typical triangle routing as with normal Mobile IP. For completeness, Figure 9.8 shows all possible routing. Depending on the parameters of Mobile IP, there are the following routing loops:
• FA-CoA, no reverse tunnel: A1 → A2 → A3 → B1 → B2
• FA-CoA, reverse tunnel: A1 → A2 → A3 → C1 → C2 → C3
• Co-CoA, no reverse tunnel: A1 → A2" → B1
• Co-CoA, reverse tunnel: A1 → A2" → D1 → D2

Figure 9.8(b) shows all possible packet formats in the i-CN → i-MN direction while Figure 9.8(c) shows the i-MN → i-CN direction.

9.3.2 Routing between i-CN and x-MN

In this situation, the Mobile Node moves outside the intranet, which is thus denoted as x-MN. Traffic between the x-MN and the internal entity, i-CN, spans through the IPSec gateway. Other entities involved includes x-HA, i-HA, x-FA, as well as GGSN/x-FA and SGSN, depending on whether GPRS is the attached wireless network. The case where no GPRS is involved was discussed in Section 8.4 on page 181. This subsection only focuses on the case where GPRS is involved.

A series of initial procedures relative to this case were introduced in Section 9.2.1. After these initial procedures all the necessary tunnels are established and ready. All traffic between x-MN and i-CN is transported via these tunnels. It is worth noting that the routing is somewhat different in two directions, as illustrated in Figure 9.9. Also note that there is no alternative communication, and the FA-CoA mode is mandatory in the case of the GPRS. A unique communication path is:

A1 → A2 → A3 → A4 → A5 → A6 → B1 → B2 → B3 → B4

The rest of this subsection will describe the routing procedures in both the i-CN → x-MN and x-MN → i-CN direction.

i-CN → x-MN

In this direction, the packets are routed to the x-MN along the path A1 → A2 → A3 → A4 → A5 → A6. Their packet formats are shown in Figure 9.9(b).

The i-CN sends a packet A1 to the x-MN with x-MN’s i-HoA as destination address. The i-HA intercepts this packet because the x-MN has
registered it with its TIA as i-CoA, as described in Section 9.2.1. The i-HA encapsulates A1 into A2 with TIA as destination address. A2 is routed to the IPSec gateway for going outside because the x-MN’s x-HoA is selected as its TIA in this framework, which is a public address.

The IPSec gateway consults its Security Association Database, and find a security association identified by a triple SPI, x-HoA, AH or ESP, which has been established during the IKE negotiation stage, as described in Section 9.2.1. The IPSec gateway encapsulates A2 into A3 according to the selected IPSec protocol format (AH, ESP, or both), with x-HoA as the destination address.

A3 is intercepted by the x-HA, because x-MN has registered it with its x-CoA before. The x-HA in turn encapsulates A3 into A4 with the x-CoA
as the destination address as per normal Mobile IP procedure.

Since the FA-CoA mode is adopted in the GPRS network in this framework, the x-CoA is equal to the x-FA’s address that the x-MN attaches to. Thus, A4 is routed to the GGSN/x-FA via the Internet as normal IP routing does. The FA functionality part of GGSN/x-FA then decapsulates A4, gets packet A3, and delivers it to GGSN functionality part. The GGSN consults its database, and looks for the TID corresponding to A3’s destination address, that is, x-HoA. As mentioned in Section 9.2.2, the GGSN has associated this x-HoA with a TID, which identified a GTP tunnel to the SGSN attached by the x-MN. The GGSN put A3 into that GTP tunnel, which forms the packet A5.

A5 is delivered to the corresponding SGSN via the GPRS backbone. The SGSN strips off the GTP header, and gets packet A3. According to the association between x-HoA and a TLLI that was established in GPRS attachment stage, the SGSN delivers A3, which is denoted as A6 at this point, to the x-MN.

The x-MN’s IPSec part validates all protection examination for A6. If this packet passes examination, the IPSec header is striped off. The extracted packet is A2. A2, in turn, is delivered to the x-MN’s Mobile IP part. Mobile IP strips off the outer IP header, and delivers the inner packet, which is the original packet sent by the i-CN, A1, to the IP network layer for further processing.

$x$-MN → i-CN

In this direction, the packets are routed back to i-CN from x-MN along with the path $B_1 \rightarrow B_2 \rightarrow B_3 \rightarrow B_4$. Their packet formats are shown in Figure 9.9(c).

The x-MN adds an IPSec header to the original IP packet destined for i-CN, and forms packet $B_1$. $B_1$ is routed to the IPSec gateway via the GPRS network and the Internet. The IPSec gateway examines the packet and validates security requirements. If the packet passes the examination, the IPSec gateway strips off the IPSec header, and gets the original packet, $B_4$. $B_4$ is routed to i-CN using the normal IP routing because $B_4$’s address is i-CN’s IP address, which was set by x-MN.

It is worth noting that no reverse tunnel is required to pass ingress filtering
9.4 External Mobility

In the context of this study, external mobility focuses on mobility management when the Mobile Node communicates with the external entity, x-CN. This is regardless of whether the Mobile Node is inside the intranet, where it is denoted as i-MN, or outside the intranet, where it is denoted as x-MN. Consequently, there are two possible communication paths:

- x-CN ↔ i-MN
- x-CN ↔ x-MN

The rest of this section will discuss the routing in these two situations, respectively.

9.4.1 Routing between x-CN and i-MN

In this situation, the i-MN is located by its i-CoA from the private address space, which is not reachable from the public address space where x-CN only comes from. As mentioned in Section 9.1.3, no IPSec is involved in this path. Consequently, the IPSec gateway can not serve as a bridge between two incompatible address spaces. Thus, a question of what can play the role of such a bridge arises. A private interface is added to the HA to fulfil this demand. Via the x-HA’s private interface, the i-MN registers x-HA with its i-CoA as its x-CoA, as described in Section 9.2.2. The x-HA intercepts each packet destined to i-MN’s x-HoA, and relays this packet to the i-MN’s i-CoA via its private interface. In this way, the i-MN is enabled reachable from outside.

Figure 9.10 shows all possible routes in this case. It is worth noting that the reverse tunnel must be employed in the return direction. Otherwise, the i-MN can not send the packets to the x-CN.

Because the reverse tunnel is mandatory in this case, only the CoA mode varies. Consequently, there are the following two communication loops:

- FA-CoA: A1 → A2 → A3 → B1 → B2 → B3
CHAPTER 9. A COMPLETE FRAMEWORK

![Routing Sequence Diagram](image)

(a) Routing Sequence Diagram

<table>
<thead>
<tr>
<th>Packet Sequence</th>
<th>Destination Address</th>
<th>Source Address</th>
<th>x-HoA</th>
<th>x-CN</th>
<th>Payload</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1:</td>
<td></td>
<td></td>
<td>x-HoA</td>
<td>x-CN</td>
<td></td>
</tr>
<tr>
<td>A2:</td>
<td>i-CoA (i-FA)</td>
<td>x-HA (Private Address)</td>
<td>x-HoA</td>
<td>x-CN</td>
<td></td>
</tr>
<tr>
<td>A3:</td>
<td></td>
<td></td>
<td>x-HoA</td>
<td>x-CN</td>
<td></td>
</tr>
<tr>
<td>A2''</td>
<td>i-CoA</td>
<td>x-HA (Private Address)</td>
<td>x-HoA</td>
<td>x-CN</td>
<td></td>
</tr>
</tbody>
</table>

(b) Packet Encapsulation in x-CN → i-MN direction

<table>
<thead>
<tr>
<th>Packet Sequence</th>
<th>x-CN</th>
<th>x-HoA</th>
<th>Payload</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2:</td>
<td>x-HA (Private Address)</td>
<td>i-CoA (i-FA)</td>
<td>x-HoA</td>
</tr>
<tr>
<td>B3:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B1''</td>
<td>x-HA (Private Address)</td>
<td>i-CoA</td>
<td>x-CN</td>
</tr>
</tbody>
</table>

(c) Packet Encapsulation in i-MN → x-CN direction

Figure 9.10: Routing between x-CN and i-MN

• Co-CoA: A1 → A2'' → B1'' → B3

The rest of the section will describe the routing of the packets in two directions, respectively.

**x-CN → i-MN**

The x-CN sends i-MN the packet A1 with the i-MN’s x-HoA as the destination address. The x-HA intercepts this packet, and looks up its binding cache. It finds an association between the x-HoA and i-CoA. The x-HA adds a new IP header to the A1 with the i-MN’s i-CoA as the destination address. In the case of FA-CoA, the i-CoA is its i-FA’s IP address. The A2 is generated, and sent to the i-FA. In the case of Co-CoA, the i-CoA is a dedicated address. The A2'' is generated, and directly sent to the i-MN. Finally, the
i-MN receives the original packet A1.

**i-MN → x-CN**

In the reverse direction, i-MN must deliver a B1-like packet so that it can be routed to the x-CN. B1’s destination address is a public address. However, this is not routable in the intranet due to the use of a private address space. Thus, this original packet must be encapsulated into a new IP packet, which uses a private address. For this reason, the reverse tunnel must be employed. The end of tunnel is the x-HA’s private address. Depending on the CoA mode, the encapsulated packet could be B2 if using FA-CoA, or $B1''$ if using Co-CoA. The x-HA then receives the encapsulated packet, extracts the original packet B1 and delivers it to the public network. Finally, the packet B1 is routed to the x-CN as per normal IP routing.

Note that in addition to the purpose mentioned above, the reverse tunnel maintains its functionality, that is, to pass ingress filtering examination.

### 9.4.2 Routing between x-CN and x-MN

In this case, both two communicating entities are in the external public domain. Thus, no private address is involved. All Mobile IP procedures and the parameters’ impact continue as normal Mobile IP does, which was introduced in Chapter 5. This subsection places emphasis on the situation where the x-MN moves to the GPRS network.

The FA-CoA mode is mandatory in the case of the GPRS networked involved. Thus, there is only a parameter, namely the reverse tunnel option, to impact on the routing. All possible routes are shown in Figure 9.11. There are the following communication loops:

- **Reverse Tunnel:** $A1 \rightarrow A2 \rightarrow A3 \rightarrow A4 \rightarrow B1 \rightarrow B2 \rightarrow B3 \rightarrow B4$

- **No Reverse Tunnel:** $A1 \rightarrow A2 \rightarrow A3 \rightarrow A4 \rightarrow B1 \rightarrow B2 \rightarrow B3''$

There is no special points that need to be highlighted. The normal Mobile IP procedures are applicable in the IP network segment, and the normal GPRS procedures are applicable in the GPRS segment. The readers can consult aforementioned relevant description to understand Figure 9.11 in both the x-CN → x-MN and x-MN → x-CN direction.
9.5 Conclusion

This chapter presented the complete framework for seamless handover between GPRS and WLAN with security protection. The framework adopts a screened subnet firewall architecture. An IPSec gateway is deployed in the DMZ to separate the internal domain and external domain. Besides an internal Home Agent (i-HA) residing in the internal domain, an external Home Agent (x-HA) is deployed in the DMZ to eliminate IPSec’s renegotiation problem when Mobile Nodes handover in the external domain. As such a secure Mobile IP framework has been constructed. To make GPRS users enjoy services provided by this framework, a GGSN/FA is introduced to the GPRS infrastructure instead of the original GGSN, as discussed in Chapter 7. This
extends Mobile IP service to GPRS networks. This way, seamless handover between GPRS and WLAN also is achieved.

The next chapter will evaluate this framework against this study’s objectives and indicate some future work.
Part III

Epilogue
Chapter 10

Conclusion

This study presented a security framework for seamless handover between GPRS and WLAN. To develop this framework, this study demonstrated a scenario in Section 1.3 on page 9. Based on the discussion of this scenario, this study defined the scope, the existing problems and objectives for addressing these problems.

To achieve these objectives, this study introduced GPRS in Chapter 2. Then, the concept of wireless overlay network was introduced in Chapter 3, as well as the architectures for implementing this concept. Due to the significant role of IP in these architectures, Chapter 4 provided an examination of IP’s various aspects relevant to this study. The deficiencies of IP mobility and security aspects are addressed in Chapters 5 and 6, which introduced Mobile IP and IPSec, respectively.

After the necessary background knowledge for understanding the proposed framework was provided, this study started to present its framework. This was performed through three steps. Firstly, the integration of Mobile IP and GPRS was discussed in Chapter 7. Then, the integration of Mobile IP and IPSec was discussed in Chapter 8. Finally, a complete framework towards this study’s goal was proposed in Chapter 9, based on the previous two chapters’ discussion.

The rest of this chapter will evaluate whether this framework achieved the predefined objectives and indicate future work on this framework.
10.1 Revisiting the Objectives

Section 1.5 (see page 17) presented four objectives of the proposed framework towards this study’s goal. The following subsections will evaluate to what degree the framework achieved these objectives.

10.1.1 Effective IP Mobility

To make IP mobility more effective, this framework adopted a hierarchical mobility structure. In particular, it deals with mobility on two levels: micro-mobility and macro-mobility. In the case of GPRS, micro-mobility is performed through SGSN and GGSN while macro-mobility is performed using the Mobile IP protocol. In WLAN, micro-mobility and macro-mobility are performed through Mobile IP regional registration. This way, the lower level hides the movement of Mobile Nodes from the upper level. This effectively reduces the signalling load and latency.

10.1.2 Seamless Handover

Seamless Handover is achieved through exploiting the coverage diversity of GPRS and WLAN. Generally, the GPRS has a much larger coverage than WLAN. In the situation where WLAN becomes unavailable, Mobile Nodes in advance switch to GPRS networks in the form of a vertical handover using a make-before-break mechanism and vice versa. This allows Mobile Nodes to always attach to at least an access network and to render seamless handover.

10.1.3 Offering Mobile IP Service in GPRS

Mobile IP service is offered to GPRS via a combined entity, GGSN/FA. This entity combines both GPRS and FA functionalities. Though a series of associations between various domain-specific identifiers, it performs the translation between the Mobile Node’s IP address and its International Mobile Subscriber Identity (IMSI). As such, packets sent to a Mobile Node by Mobile IP are delivered to that Mobile Node and vice versa.
10.1.4 Incorporation of Mobile IP and IPSec

To solve the problem of IPSec mobility, an additional external Mobile IP instance is used to tunnel IPSec packets to its current point of attachment. An internal Mobile IP instance is nevertheless required to function as normal. This forms a dual Home Agent Architecture. This way, IPSec protects the traffic between the internal domain and external domain without losing mobility, in corporation with Mobile IP.

10.2 Future Work

The proposed framework achieved the goal of this study, but future work is required. The feasibility of the work needs to be verified in practical terms. An obvious future step would be to implement this as a prototype or through a simulation exercise.

However, on a conceptual this proposal still suffers from some deficiencies. Firstly, it suffers from a double triangle routing problem, which comes from the internal Mobile IP (i-MIP) and the external Mobile IP (x-MIP), respectively. This adds extra payload to the traffic and calls for a certain degree of optimization in future work. Initial efforts could be directed to the work of Choyi and Barbeau (2005), who investigated similar issues.

Another point worth noting is that in the case of the communication path x-MN ↔ x-CN, when the x-CN is another x-MN, belonging to the internal domain, both of them should be considered as internal nodes. Thus a Security Association between them exists already. However, the communication between them should also be protected under IPSec. This presents another avenue for future work.

As mentioned in Chapter 7, this study adopted “Step 2” as the base of offering Mobile IP in a GPRS network. Future work can attempt to adopt “Step 3”, which expands Mobile IP service to the level of micro-mobility. Consequently, the framework will be all Mobile IP-based.

For future research it would also be beneficial to constantly keep track of the development of Internet Key Exchange version 2 (IKEv2) (Kivinen & Tschofenig, 2005), also called MOBIKE, where the Mobile Node does not need to tear down its existing Security Association and renegotiate it. Instead, it can modify the existing security association when it changes its
IP address. This will help mitigate heavy tunnelling overhead due to triple encapsulation (i-MIP, IPSec and x-MIP).

Finally, this framework should be upgraded to IPv6. The final framework will benefit from many new features of the new version. For example, in Mobile IPv6, a Foreign Agent is no longer required. In addition, the design of Mobile IPv6 incorporates security features that differ significantly from Mobile IPv4 (Kempf, Arkko, & Nikander, 2004). This will facilitate the security design of the framework.

10.3 Final Word

In conclusion, this study has developed a framework for security mobility between GPRS and WLAN with guarantee of security characteristics. However, possibility exists for further research, that was out of the scope of this study as discussed in Section 10.2.

Nevertheless, it is believed that this study contributes to the understanding of the wireless overlay networks in terms of mobility and security. The author also wants to express the hope that this study will fuel further interest in this domain of discourse.
Part IV

Appendices
Appendix A

Academic Paper

An academic paper based on the research contained in this dissertation has been written.

The paper entitled, “Issues When Securely Integrating UMTS, Mobile IP & IPSec”, will be submitted to a suitable conference.
References


References

Mobility and Firewalls. Internet draft. IETF. (Retrieved on Sept. 18, 2005 from http://www.cs.pdx.edu/research/SMN/papers/mobifirewall.txt)


References


**References**


Francisco, USA: Morgan Kaufmann.


Hill.


