INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6” x 9” black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700  800/521-0600
EFFICIENT TECHNIQUES FOR SUPPORTING MOBILE HOSTS IN WIRELESS NETWORKS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Gopal K. Dommety, B.Tech, M.S.

* * * * *

The Ohio State University 1999

Dissertation Committee:
Dr. Mukesh Singhal, Adviser
Dr. Ten H. Lai
Dr. Sadayappan Ponnuswamy

Approved by

[Signature]
Adviser
Department of Computer and Information Science
© Copyright by

Gopal Dommety

1999
ABSTRACT

Traditionally, network protocols were designed with the objective of supporting fixed network elements and end-hosts. This dissertation examines network layer aspects of location management, handoff management, route optimization, and transport of signaling messages to support mobile users in hierarchically routed networks such as IP (Internet Protocol)-based or ATM (Asynchronous Transfer Mode)-based networks.

Integrated, one-phase crankback, LR (Location Register), and flat location management schemes are developed to track mobiles and route data/connections. The integrated scheme uses features of current dynamic routing and signaling protocols with minimal modifications to support mobility, but data could follow a suboptimally routed path. The one-phase crankback scheme uses topology information and the technique of crankback to set up optimally routed connections. While the integrated and the one-phase crankback schemes integrate location management (Integrated approach) with routing and signaling protocols, the LR scheme and the flat scheme use an overlay (Overlay approach) of external location registers to perform location management. LR scheme is a generalized overlay-approach scheme. Effect of user and network characteristics on design of optimal location management architectures is studied. Results show that at low Call-to-Mobility Ratio (CMR < 0.025), the overlay approach performs better than the integrated approach.
Handoff management procedures reroute connections when a user moves. One-phase dynamic handoff scheme which results in an optimally routed path and incurs low latency, and signaling and buffering alternatives to perform lossless and in-sequence delivery of data are developed.

A route optimization procedure to reroute suboptimal connections that result during connection setup or during handoffs is developed. Amount of resources required for a suboptimally routed connection can be significantly reduced by performing this route optimization.

A method to efficiently transport short messages in a connection-oriented network, by utilizing existing routing information is developed. Connectionless ATM protocol (CL-ATM) uses this method to transport short messages in ATM Networks. CL-ATM protocol provides an efficient method to transport IP traffic over ATM backbone.

These techniques are implemented in a prototype with mobile users, to demonstrate effective working of data-transfer, real-time, and interactive applications.
To my parents
ACKNOWLEDGMENTS

First, I would like to thank my advisers, Professor Mukesh Singhal and Dr. Malathi Veeraraghavan for their support and guidance. They both have followed my work closely, and their guidance has been invaluable. I also thank Professor Ten-Hwang Lai, Professor P. Sadayappan, and Professor Raj Jain for many valuable comments, constructive suggestions, and advises without which this thesis would not be possible.

I am grateful to members of the Broadband Systems Research Department at Bell Laboratories, Holmdel, NJ, and especially Kye Eng, Pramod Pancha, Mark Karol, and S. Srinivasan for numerous discussions and helpful insights. I would also like to thank Kent Leung of IOS Network Protocols, Cisco Systems for his numerous discussions which helped me considerably.

Many friends have shared and shaped my life during the past four years. I would like to thank Alpesh Patel, Ananth Dommety, Anil Mathew, Ashutosh Sabrawal, Bama Ramarathanam, Bhargavi Sarma, Bindu Pillai, Bobby Vandalore, Debasish Basak, Gayatri Srinivasan, Jay Iyer, Kiran Uppuluri, Manish Prabhu, Manpreet Singh, Naveen Kakani, Prabhu Thayakal, Prashant Sawant, Prashun Raha, Praveen Pauskar, Rajeshwari Pechetty, Raja Chatterjee, Ram Kesevan, Ravi Prakash, Rohit Goyal, Sandeep Kulkarni, Shaji Ullatil, Shivkumar Kalayanaraman, Sonia Fahmy, Srikanth Kilaru, Srinivas Phani,

Finally, and most importantly, I wish to thank my parents, Nagaveni and Venkata Swamy, my brother, Srikanth, my sister, Vasundara, and brother-in-law, Balakishore for their unconditional love and unwavering support.
VITA

March 26, 1972 .......................... Amalapuram, India.


1995 ................................. M.S. Computer and Information Science, The Ohio State University.

1994-96 .............................. Graduate Teaching and Research Associate, The Ohio State University.

1996-97 .............................. Graduate Research Associate, The Ohio State University and Visiting Member of Technical Staff, Bell Laboratories, Lucent Technologies.

1998-Present ........................ Ameritech Presidential Fellow, The Ohio State University.

PUBLICATIONS


FIELDS OF STUDY

Major Field: Computer and Information Science

Studies in:

<table>
<thead>
<tr>
<th>Field</th>
<th>Professor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Networking</td>
<td>Prof. Mukesh Singhal</td>
</tr>
<tr>
<td>Theory and Algorithms</td>
<td>Prof. Rephael Wenger</td>
</tr>
<tr>
<td>Computer Architecture</td>
<td>Prof. Ponnuswamy Sadayappan</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>DEDICATION</td>
<td>iv</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>v</td>
</tr>
<tr>
<td>VITA</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xx</td>
</tr>
<tr>
<td>CHAPTERS:</td>
<td></td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Problem statement</td>
<td>3</td>
</tr>
<tr>
<td>1.1.1 Location management</td>
<td>6</td>
</tr>
<tr>
<td>1.1.2 Handoff management</td>
<td>8</td>
</tr>
<tr>
<td>1.1.3 Route optimization</td>
<td>10</td>
</tr>
<tr>
<td>1.1.4 Transport of signalling messages</td>
<td>11</td>
</tr>
<tr>
<td>1.2 Important design considerations</td>
<td>12</td>
</tr>
<tr>
<td>1.3 Motivation and scope</td>
<td>16</td>
</tr>
<tr>
<td>1.4 Dissertation outline</td>
<td>17</td>
</tr>
<tr>
<td>2. BACKGROUND</td>
<td>19</td>
</tr>
<tr>
<td>2.1 Routing protocols</td>
<td>19</td>
</tr>
<tr>
<td>2.1.1 Distance-vector routing protocols</td>
<td>20</td>
</tr>
<tr>
<td>2.1.2 Link-state routing protocols</td>
<td>21</td>
</tr>
<tr>
<td>2.2 Hierarchically organized networks</td>
<td>21</td>
</tr>
<tr>
<td>2.2.1 ATM networks and ATM routing architecture</td>
<td>22</td>
</tr>
<tr>
<td>2.2.2 Internet routing architecture</td>
<td>27</td>
</tr>
</tbody>
</table>
3. LOCATION MANAGEMENT (ROUTING FOR MOBILE HOSTS) ............................... 30
  3.1 Requirements ................................................................................................... 31
  3.2 Approaches to performing location management ........................................ 32
    3.2.1 Overlay approach .............................................................................. 33
      3.2.1.1 Cellular location management standards .......................... 34
      3.2.1.2 Forwarding scheme ....................................................................... 37
      3.2.1.3 Anchor scheme ........................................................................ 38
      3.2.1.4 Flat tracking/locating scheme ............................................. 39
      3.2.1.5 Hierarchical location management scheme ...................... 40
      3.2.1.6 Discussion ............................................................................... 41
    3.2.2 Integrated approach ............................................................................ 41
      3.2.2.1 IETF mobile IP ..................................................................... 42
      3.2.2.2 IBM mobile host protocol .................................................... 42
      3.2.2.3 Integrated mobile ATM methods ....................................... 43
  3.3 Organization of the solutions ......................................................................... 46

4. INTEGRATED SCHEME (MOBILE PNNI SCHEME) ............................................... 47
  4.1 Architecture to support mobile endpoints ...................................................... 47
  4.2 Mobile tracking ................................................................................................. 51
  4.3 Connection Setup/Mobile locating ................................................................... 58
  4.4 Discussion .......................................................................................................... 61
    4.4.1 Determining Scope ............................................................................. 61
    4.4.2 Route optimization ............................................................................. 61
    4.4.3 IP mobility support ............................................................................. 62
    4.4.4 Fault tolerance ..................................................................................... 63
    4.4.5 Security and authentication .................................................................. 64
    4.4.6 Mobile network or mobile router/switch ........................................ 65
  4.5 Summary ............................................................................................................ 65

5. ONE-PHASE CRANKBACK SCHEME .................................................................. 66
  5.1 Problem statement ............................................................................................. 68
  5.2 Crossover node determination ....................................................................... 69
    5.2.1 Optimal crossover node determination ............................................ 71
    5.2.2 Minimal crossover node determination .......................................... 75
  5.3 One-phase crankback scheme ......................................................................... 83
  5.4 Performance analysis ...................................................................................... 86
    5.4.1 Analytical model ................................................................................. 87
      5.4.1.1 Notation ...................................................................................... 87
      5.4.1.2 Property .................................................................................... 88
      5.4.1.3 Location management schemes ........................................... 89
      5.4.1.4 Numerical results ......................................................................... 96
7.3 Qualitative comparisons ................................................................. 172
7.4 Summary ........................................................................................ 173
7.5 Probability Distributions ................................................................. 174

8. ROUTE OPTIMIZATION ................................................................. 178

8.1 Background .................................................................................... 179
  8.1.1 Prior work on handoff management ........................................... 179
  8.1.2 Prior work on route optimization .............................................. 181
8.2 Problem statement ........................................................................ 182
8.3 Route optimization algorithm ....................................................... 187
8.4 Application to handoff management .............................................. 192
  8.4.1 Second phase of two-phase handoff schemes ......................... 192
  8.4.2 Multiple handoffs ................................................................... 198
  8.4.3 Second phase of two-phase location management schemes ... 199
8.5 Results .......................................................................................... 203
8.6 Summary ....................................................................................... 205

9. HANDOFF REROUTING ............................................................... 206

9.1 Problem statement ....................................................................... 207
9.2 One-phase dynamic COS search handoff scheme ....................... 208
9.3 Analysis and discussion ............................................................... 213
  9.3.1 Analytical model ...................................................................... 214
    9.3.1.1 Notations .......................................................................... 214
    9.3.1.2 Property .......................................................................... 215
    9.3.1.3 Handoff management schemes ....................................... 216
  9.3.2 Numerical results ................................................................. 223
9.4 Summary ...................................................................................... 227
9.5 Probability Distributions .............................................................. 228

10. TRANSPORT OF SIGNALING MESSAGES .................................. 231

10.1 Mobility applications ................................................................. 232
  10.1.1 Locating mobile endpoints .................................................... 232
  10.1.2 Handoff in wireless ATM networks ..................................... 233
10.2 Potential solutions ....................................................................... 234
10.3 Proposed solution ....................................................................... 237
  10.3.1 CL-ATM network layer protocol and ATM router .............. 238
  10.3.2 Applications ......................................................................... 242
    10.3.2.1 Mobility management messages .................................. 243
    10.3.2.2 IP-over-ATM usage ....................................................... 244
10.4 Summary .................................................................................... 246
11. A PROTOTYPE IMPLEMENTATION ................................................. 248
   11.1 Network architecture ......................................................... 248
   11.2 Mobility management ....................................................... 251
   11.3 Summary ................................................................. 254

12. CONCLUSION AND DIRECTIONS FOR FUTURE WORK ............... 255
   12.1 Research contributions ................................................... 255
   12.2 Future work ..................................................................... 258
       12.2.1 Security and authentication ..................................... 258
       12.2.2 Broadcast and multicast support ................................. 259
       12.2.3 Fault tolerance ..................................................... 259
       12.2.4 Quality of Service (QoS) support ............................... 260
       12.2.5 Ad-Hoc networks and mobile networks ......................... 260
       12.2.6 Physical, link, transport, and application layer issues ......... 261

13. BIBLIOGRAPHY ................................................................. 262
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Architecture of a Mobile Communication Network</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Handoff Procedure</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>Path Extension Handoff Procedure</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>Sending of Registration Messages</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>ATM User plan protocol reference model</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>ATM Control plan protocol reference model</td>
<td>23</td>
</tr>
<tr>
<td>7</td>
<td>ATM Interfaces</td>
<td>23</td>
</tr>
<tr>
<td>8</td>
<td>ATM Connection setup signaling</td>
<td>24</td>
</tr>
<tr>
<td>9</td>
<td>PNNI-based hierarchical ATM network</td>
<td>25</td>
</tr>
<tr>
<td>10</td>
<td>Internet Routing Hierarchy</td>
<td>28</td>
</tr>
<tr>
<td>11</td>
<td>Overlay location management approach</td>
<td>32</td>
</tr>
<tr>
<td>12</td>
<td>Spectrum of location management schemes</td>
<td>33</td>
</tr>
<tr>
<td>13</td>
<td>Mobile location tracking</td>
<td>35</td>
</tr>
<tr>
<td>14</td>
<td>Locating a mobile for incoming call delivery</td>
<td>36</td>
</tr>
<tr>
<td>15</td>
<td>Two schemes to improve mobile tracking costs</td>
<td>38</td>
</tr>
<tr>
<td>16</td>
<td>Hierarchical location databases</td>
<td>40</td>
</tr>
<tr>
<td>17</td>
<td>Integrated mobile ATM methods</td>
<td>43</td>
</tr>
<tr>
<td>Page</td>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------</td>
<td>-------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>18</td>
<td>Spectrum of mobile connection setup schemes</td>
<td>45</td>
</tr>
<tr>
<td>19</td>
<td>Terminal mobility scenarios</td>
<td>48</td>
</tr>
<tr>
<td>20</td>
<td>Mobility Scenarios</td>
<td>49</td>
</tr>
<tr>
<td>21</td>
<td>Configuration of a zone</td>
<td>49</td>
</tr>
<tr>
<td>22</td>
<td>Supporting mobility in an ATM network</td>
<td>50</td>
</tr>
<tr>
<td>23</td>
<td>Wireless ATM Protocol architecture</td>
<td>51</td>
</tr>
<tr>
<td>24</td>
<td>Propagation of reachability information</td>
<td>52</td>
</tr>
<tr>
<td>25</td>
<td>Flowchart representing how reachability updates are propagated</td>
<td>54</td>
</tr>
<tr>
<td>26</td>
<td>Flowchart representing how forwarding pointers are set</td>
<td>56</td>
</tr>
<tr>
<td>27</td>
<td>Registration messages</td>
<td>56</td>
</tr>
<tr>
<td>28</td>
<td>Flowchart representing how connections are routed</td>
<td>58</td>
</tr>
<tr>
<td>29</td>
<td>Connection setup</td>
<td>59</td>
</tr>
<tr>
<td>30</td>
<td>Internet Routing Hierarchy</td>
<td>62</td>
</tr>
<tr>
<td>31</td>
<td>Crankback and call forwarding schemes</td>
<td>67</td>
</tr>
<tr>
<td>32</td>
<td>Rerouting of connection f-o to form a connection f-t</td>
<td>69</td>
</tr>
<tr>
<td>34</td>
<td>Relative positions of nodes f, o and n</td>
<td>72</td>
</tr>
<tr>
<td>33</td>
<td>Determination of crossover nodes</td>
<td>73</td>
</tr>
<tr>
<td>35</td>
<td>Rerouting connections</td>
<td>74</td>
</tr>
<tr>
<td>36</td>
<td>Minimal crossover node determination</td>
<td>76</td>
</tr>
<tr>
<td>37</td>
<td>Minimal crossover node determination</td>
<td>80</td>
</tr>
<tr>
<td>38</td>
<td>Rerouting connections</td>
<td>82</td>
</tr>
<tr>
<td>39</td>
<td>Relative positions of x, y, and z</td>
<td>88</td>
</tr>
</tbody>
</table>

xvi
40 Relative positions of nodes c, h and v ......................................................... 94
41 Estimation of number of crankback nodes ......................................................... 94
42 Relative positions of nodes c, h and v .............................................................. 96
43 Delay and resource comparisons .................................................................... 98
44 Switch model .................................................................................................... 101
45 Comparison of average call setup delays ......................................................... 104
46 Maximum call setup delay ................................................................................ 105
47 Number of hops in a connection ....................................................................... 106
48 Crankback nodes involved in a connection setup ............................................. 107
49 Mobile location procedures in a flat tracking scheme ..................................... 114
50 Comparison of mobile tracking approaches ................................................. 117
51 Reference signaling architecture. .................................................................... 120
52 Mobile tracking procedures in the flat scheme ............................................. 122
53 Comparison of computation costs ................................................................... 129
54 Comparison of the communication costs ....................................................... 131
55 Comparison of the average total cost .............................................................. 133
56 Comparison of call setup delay ........................................................................ 135
57 Location registers arranged hierarchically ..................................................... 141
58 Overlay of LR architecture on a PNNI-based network ................................. 141
59 Flowchart representing how location registers are updated in the LR scheme . 143
60 Mobile locating procedure in the LR scheme (worst-case) ............................ 146
61 Illustration of the hierarchy numbering notation .......................................... 148
62 Relative positions of ancestors-are-siblings levels for nodes x, y and z ........ 150
63 Distances between calling party, and home and visiting locations ............ 159
64 Comparison of the average total costs of the two schemes ..................... 165
65 Effect of S on the average total costs in the two schemes ....................... 166
66 Contrasting behavior of the two schemes with varying S and CMR ............ 167
67 Effect of h on the average costs in the two schemes ......................... 168
68 Effect of change in S on the average costs in the mobile PNNI scheme ....... 170
69 Effect of change in value of S on average costs in the LR scheme .......... 171
70 Generalized handoff scheme ............................................................. 180
71 Spectrum of handoff schemes ............................................................ 180
72 Rerouting of connection f-o to form a connection f-t ............................... 183
72 Suboptimal paths after a two-phase handoff is performed .................... 183
73 Suboptimal paths after the connection setup in the call forwarding scheme .. 185
74 Route optimization algorithm .............................................................. 187
75 Determination of crossover nodes and new segment setup .................... 190
76 Connection after the handoff is completed ........................................ 194
77 Buffering of upstream cells at COS and downstream cells at target node ... 196
78 Switching of user cells by buffering at the target node .......................... 197
79 Suboptimal paths after multiple handoffs ......................................... 198
80 Buffering of upstream cells at COS and downstream cells at target node ... 202
81 Switching of user cells from the old to the new path by buffering at the Target node\textsuperscript{203} ............................................................. 203
82 Number of hops in a connection after connection setup ........................ 204
83 Number of hops in a connection after a handoff .................................................... 205
84 Base rerouting problem .......................................................................................... 207
85 Old and new path during soft handoff ..................................................................... 211
86 Switching of down-cells by buffering at the new switch ....................................... 212
87 Switching of down-cells by buffering at the COS .................................................. 213
88 A signaling alternative for backward handoff ......................................................... 213
89 A signaling alternative for forward handoff ............................................................ 214
90 Relative positions of x, y, and z ................................................................................. 216
91 Estimation of the number of nodes on the common path ....................................... 219
92 Relative positions of nodes c, o and n ...................................................................... 221
93 Delay and resource comparisons of one- and two- phase handoff methods ....... 224
94 Delay in one-phase handoff with and without detailed path information ............. 226
95 Mobile location using HLRs ..................................................................................... 232
96 Handoff in a wireless ATM network ........................................................................ 233
97 IP overlay network using PVCs ............................................................................... 236
98 Connectionless ATM datagram transport ............................................................... 239
99 CL-ATM packet format ............................................................................................ 239
100 IP-over-ATM using the CL-ATM protocol ............................................................ 245
101 Wireless Network Prototype .................................................................................. 249
102 Architecture of ATM Switch Router ....................................................................... 250
103 CL-ATM IP Gateway Architecture ........................................................................ 251
104 Mobile Node Architecture ...................................................................................... 251
105 Path of the data traffic ............................................................................................ 25
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Notation ................................................................. 78</td>
</tr>
<tr>
<td>2</td>
<td>Notation ................................................................. 87</td>
</tr>
<tr>
<td>3</td>
<td>Input Values ............................................................ 97</td>
</tr>
<tr>
<td>4</td>
<td>Network configurations .............................................. 103</td>
</tr>
<tr>
<td>5</td>
<td>Input parameters ..................................................... 104</td>
</tr>
<tr>
<td>6</td>
<td>Notation ................................................................. 121</td>
</tr>
<tr>
<td>7</td>
<td>Input Values ............................................................ 128</td>
</tr>
<tr>
<td>8</td>
<td>Tracking and locating costs ....................................... 132</td>
</tr>
<tr>
<td>9</td>
<td>Comparison of Flat and Hierarchical Schemes ................. 139</td>
</tr>
<tr>
<td>10</td>
<td>Notation ................................................................. 149</td>
</tr>
<tr>
<td>11</td>
<td>Tracking costs in the mobile PNNI scheme .................... 152</td>
</tr>
<tr>
<td>12</td>
<td>Tracking costs in the LR scheme .................................. 155</td>
</tr>
<tr>
<td>13</td>
<td>Input values ............................................................ 164</td>
</tr>
<tr>
<td>14</td>
<td>Mapping to the common rerouting problem ..................... 184</td>
</tr>
<tr>
<td>15</td>
<td>Mapping to the common rerouting problem during connection setup ............... 186</td>
</tr>
<tr>
<td>16</td>
<td>Notation ................................................................. 189</td>
</tr>
<tr>
<td>17</td>
<td>Analysis notation ..................................................... 215</td>
</tr>
<tr>
<td>18</td>
<td>Input values ............................................................ 223</td>
</tr>
</tbody>
</table>

xx
CHAPTER 1

INTRODUCTION

Advances and the widespread usage of wireless technology, portable computing, and high-speed networks has facilitated the support of several communication services to mobile users. Specifically, advancements in VLSI (Very Large Scale Integration) and surface mount technologies allow for scaling down in size and power consumption of the computing devices, thus leading to the proliferation of portable devices such as laptop computers, cellular phones, PDAs (Personal Digital Assistants), and Palm PC’s. Advancements in wireless technology, improvements in digital and RF circuit fabrication, new large-scale circuit integration, and other miniaturization technologies which make portable radio equipment smaller, chapter and more reliable. Advances in digital switching techniques, networking technology, and networked applications have facilitated the large scale deployment high-speed communication networks.

Due to these technological advancements wireless communication networks have emerged as an important field of activity. During the 80’s, mobile systems provided voice services and were analog cellular communication systems such as AMPS (Advance Mobile Phone Service)[104], NMT (Nordic Mobile Telecommunications), TACS (Total Access Cellular Systems), etc. There were also systems that provided paging and cordless
services. During the 90's, mobile systems provided enhanced voice and paging services and used digital cellular communication systems such as IS-54, IS-95, GSM (Global System for Mobile Communications)[9], DCS1800 (Digital Communication Systems in the 1800 MHz band), DFCT (Digital European Cordless Telecommunications), CPDP [107][108], and TETRA (Trans European Trunked Radio). In the third generation the aim is to provide ubiquitous communication to a wide range of mobile terminals supporting wide range of services. These mobile terminals can be simple cellular phones, dual devices such as cell phone and web browser, or complex computers. Some of the services are voice and video services; multimedia services requiring quality of service guarantees; access to e-mail, databases and on-line shopping services; sending faxes and alpha numeric pages; browsing the web, etc. Recently, we are seeing wireless campus area networks (e.g., Ricochet service offered by Metricom [105][106]) and in-building local area networks. Several efforts such as IMT 2000 (International Mobile Telecommunications 2000), Wireless ATM, Mobile IP [13], IEEE802.11, General Packet Radio service (GPRS) are underway to realize the vision of ubiquitous communication to a wide range of mobile terminals supporting wide range of services.

The aim of third generation wireless networks is to provide ubiquitous communication to a wide range of mobile terminals supporting wide range of services. To achieve this objective, new network architectures and techniques to support mobile users providing cost effective support for a variety of cellular and wireless data technologies. This dissertation focuses on developing efficient techniques to manage mobile users in wireless networks. A typical wireless network architecture to support mobile users and some of the issues in such networks are described in Section 1.1.
1.1 Problem statement

Architecture of a Mobile Communication Network (MCN)\(^1\) or a Wireless Network is as shown in Figure 1 and consists of a wireline network, mechanism for wireline-wireless interface, and mobile terminals. The mobile terminal is a portable device such as laptop computer or PDA equipped with a wireless radio equipment, or cellular phone (mobility scenarios are discussed in detail later). Scope of the wireless network is limited to the mobile users within the wireless coverage area of the network and to the mobile users that can access this network via other access mechanisms such as dialup connections. The entire coverage area of the network is divided into large number of smaller areas called cells (see Figure 1). Each cell has a base station or a wireless access point. Each base station has a wireless interface through which it provides network access\(^2\) for all the mobile terminals within the cell and a wireline interface through which it is attached to an MSC (Mobile Switching Center), a switch, or a router. Cells are grouped into "zones" or registration areas. Base stations (Access points) periodically emit “beacons” containing information such as the identity of the base station and the zone to which it belongs and other configuration information. The mobile unit receives these beacons and determines the cell or the zone in which it currently is. The mobile units within a cell communicate with the wireline network through a base station. Base stations belonging to a zone are connected to an MSC. The wireline network or the backbone network connects these MSCs or switches to the each other and to the rest of the network. Typically, the wireline network also supports fixed terminals as shown in Figure 1. The wireline network can be

---

1. The Ad-hoc network model is not considered.
2. The access part of the network implements radio related functions required to connect the mobile terminal to the wireline network. The access could be via radio access using techniques such as TDMA (Time Division Multiple Access) or CDMA (Code Division Multiple Access).
the Internet, telephone network (as in the case of cellular\textsuperscript{3} telephone network), a LAN as in the case of a wireless LAN, ATM network as in the case of wireless ATM network, or a combination of several of these networks (Each of these networks may use different protocols and techniques to manage mobility).

![Diagram of Mobile Communication Network](image)

**Figure 1**: Architecture of a Mobile Communication Network

The objective of mobile communication networks is to allow communicating devices (mobile terminals) to autonomously move and still stay connected to the network, thereby being able to access services as if it was stationary. In other words, the mobile unit can receive a connection/data or establish a connection/send data to other terminals in the network while moving. Since a mobile moves independent of time, location, and access arrangement the point of attachment or the access point through which the mobile is connected to the network/ or accesses the network changes. For example, in Figure 1 the mobile terminal could be in different cells at different points of time, there by having access to the network via different base stations.

\textsuperscript{3} In a more narrow sense, cellular refers to wireless telephone networks operating in the 800 MHz or 900 MHz portions of the radio spectrum.
Since the mobile units are free to move between cells, a mobility management mechanism is needed to manage the effects of the movement. Mobility management algorithms enable networks to support mobile users allowing them to move, while simultaneously offering them continuous network connectivity to receive incoming calls, data packets, and other services. Mobility management consists of location management, handoff management. The objective of location management is to route connections or data to the mobile. Location management deals with tracking the location of a mobile, locating the mobile, and setting up a connection or delivering data to a mobile. Since the access point through which the mobile can be accessed changes, the network in order to establish a connection or deliver data packets needs to keep track of the access point (location) through which the mobile can be accessed. Mechanisms are required to track mobiles, determine the current location of the mobile, and to route connections or data packets to the mobile. These mechanisms are called location management mechanisms. The access point through which the mobile can be accessed could also change while the mobile is actively communicating. As the mobile unit moves while communicating, rerouting of connections or data is required. This rerouting process is called handoff. Other mechanisms that are required for efficient mobility management are route optimization techniques and techniques to transport signalling messages. The main focus of this research is to develop techniques to manage mobile terminals efficiently. In this thesis, the above mentioned issues relating to mobility management are addressed. These issues are described next.
1.1.1 Location management

Location management procedures are needed in order to route connections or data to the current location of the mobile. In traditional networks, a fixed terminal is identified by a network address, which served two purposes: 1) as a location indicator, i.e., the network address also implicitly identified the network route to the terminal and 2) as a terminal identifier, i.e., higher level protocols (e.g., transport protocols) used this address to establish and maintain connections. With mobile terminals, it may not be possible for both the functions to be served by the same identifier. For example, in mobile communication network shown in Figure 1, since the mobile terminal could be anywhere within the coverage areas of the network, i.e., in any of the cells, the access point or the base station through which the mobile can be accessed could be different at different points of time. In other words, the location indicator of the mobile will change often. The terminal identifier cannot change as it is used by other terminals to identify the mobile terminal and also by higher layer protocols to maintain connections. In order to establish a connection or deliver data packets, information regarding the access point (location indicator) through which the mobile can be accessed is needed. Location management provides mechanisms to track the current location (indicator) of the mobile and locate the current position of the mobile before or during delivering the connection or data. Location management is an important problem, whose efficient solution is central to the realization of efficient mobile communication networks.

4. Location of the mobile terminal refers to the access point, base station, MSC, cell, or the zone in which the mobile terminal is currently residing.
There are two aspects to the problem of location management, mobile tracking and mobile location. *Mobile tracking* is the procedure by which the network elements update information about the location of a mobile. Mobile tracking is initiated when the mobile registers with network. A mobile registers with the network by sending a registration message. Typically mobiles generate registration messages when they move into a new zone, when powered on, powered off, or periodically. These registration messages are used to track the current location of the mobile, by updating the information in the various network elements. The exact updating procedures depend on the location management architecture and the scheme used (discussed in Chapters 3-7).

*Mobile location* is the procedure by which the network finds the exact location of a mobile and routes the connection or data. This procedure is initiated when a connection is to be established or data is to be routed to the mobile. Both the Mobile tracking and the Mobile locating procedures involve the exchange of signaling messages among the various network elements. The exchange of these messages consume significant wireless, wireline and network resources. The information acquired during the tracking phase is used during the locating phase.

There is a trade-off between the resources used during the mobile tracking and mobile locating phases. An aggressive tracking strategy can result in a simple locating strategy. This trade-off is illustrated by the following example. If the mobile terminal registers every time it moves into a new cell, the network has information regarding the exact location of the mobile. In this case, during connection establishment, the connection can be established directly since the exact location of the mobile is available. With this strategy, the mobile tracking is resource intensive but the mobile location requires less resources and incurs less delay. If the mobile never registers, then the current location of
the mobile may not be available. In this case during connection establishment, the current location of the mobile needs to be determined by broadcasting query messages. With this strategy, the mobile tracking does not incur any resources but the mobile location is resource intensive and incurs higher delay.

The problem of location management, location management architectures and strategies are discussed in Chapters 3 to 7. Chapter 3, presents details of the existing solutions to the problem of location management. Chapters 4 to 7, present the details of the proposed architectures and strategies. These chapters also present the results obtained from comparative performance analysis. Next the problem of handoff management is described.

1.1.2 Handoff management

In a network with fixed terminals, after a connection is established to a terminal, the connection does not change under normal circumstances, i.e., except due to failures. But in a network with mobile hosts, as the mobile terminal moves from one location to another, the connection is to be rerouted. This rerouting procedure is called Handoff management. Handoff procedures, typically involve functions such as identification of the new base station, rerouting of the wireline connection between the far end and the old base station to the new base station, radio link establishment between the mobile and the new base station, etc. In this work, we are primarily concerned with the connection rerouting aspect of the handoff.

As mentioned earlier, handoff management includes procedures to reroute connections on which a mobile user is communicating while moving. Consider a mobile with an active connection to it and currently residing in cell 1 as shown in Figures 2. In this scenario.
there is an existing wireline connection between the other-end and base station I and a wireless connection between the base station I and the mobile terminal. Figures 2 shows a mobile moving from base station I to base station II. When the mobile moves from one base station to another, wireless resources are allocated at the new bases station, and the connection needs to be rerouted to the new base station as shown in Figures 2. The allocation of resources and the rerouting procedure should be done with minimal disruption of the connection. In other words, time required to perform the handoff procedures (handoff latency) should be minimized. Also, some applications require that during handoff procedure there should be no loss of data or minimal loss of data. The problem of handoff management is discussed in detail in Chapter 9. A new solution to the problem of handoff management is described in Chapter 9. This chapter also discusses the various signalling and buffering alternatives to achieve lossless handoffs (handoff procedure without loss of data).
1.1.3 Route optimization

In mobile communication networks, during handoff procedures and location management procedures paths taken by connections could become "suboptimal." For example, it is essential to minimized the handoff latency as increase in handoff latency leads to degradation of the quality of service of the connection (longer disruption in service). Most handoff schemes propose performing a local connection reroute rather than an end-to-end connection reroute to keep handoff latencies low. An example of such a handoff procedure is the path extension scheme. In the path extension scheme, rerouting of the connection is achieved by setting up a connection from the old switch (Switch I in Figure 2) to the new switch (Base station I in Figure 2). Such reroute operations could result in making the connection path suboptimal. The connection paths could become longer after multiple handoffs. Similarly, location management schemes that propose setting up the connection to the "home" location of a mobile and then rerouting the

---

5. The path of a connection is classified as being "suboptimal" if it is not the shortest path between the two endpoints of the connection.
connection to the mobile's current location based on location data provided by the home node could result in suboptimal connection paths. In other words, suboptimality that is introduced during connection setup is primarily due to the lack of exact information about the location of the mobile at the call originating switch, and suboptimality after connection setup occurs because of user movements by a communicating mobile. It is important to optimize routes of connections since suboptimal paths imply an inefficient usage of network resources. In order to optimize the route of existing connections route optimization procedures are required. The problem of route optimization is discussed in detail in Chapter 8. A route optimization procedure that can be used to reroute connections that become suboptimal due to operations such as handoffs and location-based reroutes is described in Chapter 8.

1.1.4 Transport of signalling messages

In mobile networks, there is a need to exchange short mobility management and authentication messages between network entities. For example, mobility related messages such as registration messages, handoff messages, and location request messages are sent between various network entities. Previous studies [80][81] indicate that the signalling traffic and database queries associated with PCS due to user mobility are likely to grow to levels well excess of that associated with a conventional call. Hence, it is extremely important to reduce the amount of signalling and also the overhead incurred during this signaling. In order to efficiently support mobile endpoints in a network, a mechanism to efficiently transport mobility related messages is required. In networks supporting connectionless transport of messages, the mobility related signaling messages can be sent efficiently. In connection-oriented networks such as ATM networks, a
connection needs to be established first and then the mobility related messages are to be sent. For example, in order to send a registration message from the visiting switch to the home switch in an ATM network (see Figure 4) a potential solution is to first establish a connection between the visiting switch and the home switch, then send the message, and later release the connection. This procedure incurs the overhead of establishing a connection and then releasing it. The problem of transporting signalling messages is discussed in detail in Chapter 10. An efficient solution to this problem and its application to other motivating applications is described in Chapter 10.

1.2 Important design considerations

In order to develop efficient solutions to these problems of mobility management, one needs to work with the limitations imposed by the mobile environment and the artifacts of existing technology and infrastructure. The main limitations imposed by the mobile environment are limited computation resources (power of the hand-held devices, limited processing power of network elements), limited communication resources (limited
wireless and wireline bandwidth), and limited storage resources. Some of the *important design issues* to be considered are:

1. Communication resources: Communication resources primarily consist of wireless bandwidth and wireline bandwidth. *Wireless bandwidth:* One of scarce resources is the wireless bandwidth. The availability of the usable frequencies is limited. The demand of such resources is growing at a fast rate. Protocols should be designed to optimize the use of this scarce resource. *Wireline bandwidth:* Wireline bandwidth is also limited. Excessive signalling unduly effects the performance of the system (both in-band signalling and out-of-band signalling). In particular, effort should be made to reduce the "long distance signalling".

2. Computation resources: Mobile units are usually hand-held devices with very minimal computational resources. Therefore, it is infeasible to implement computationally intensive algorithms in a mobile unit. Computational power of various network elements could also become a bottleneck. Different network elements have different computational capabilities and computational requirements. The solutions being devised should avoid excessive computational load on the various network elements.

3. Power (energy supply): Portable devices have very limited energy supply. Since energy supply is required for computation as well as communication, these activities need to be optimized.

4. Delay: The location management procedures and the handoff procedures incur delays/latencies during their execution. In certain situations these delays are unacceptable. The solutions being devised should minimize these delays.

13
5. Existing infrastructure and protocols: Artifacts of the existing infrastructure can be a limiting factor to the design of efficient architectures. To support mobile endpoints in an existing wireline network, enhancements to the existing infrastructure are required. Changes to existing protocols should be minimized and mechanisms to support these changes should be provided. It is necessary to limit these changes to few of the network elements. We refer to the switches with enhancements to support mobility as mobile-enhanced switches. In order to facilitate the easy integration of mobile endpoints into the existing networks, an approach of incremental deployment of mobile-enhanced switches needs to be taken. Therefore, it is important to consider isolating mobility-related features to mobile-enhanced switches.

6. Scalability: These protocols should scale to handle large number of users and in large networks. Protocols should also support local area mobility with the given technology.

7. User characteristics: The aim of next generation mobile communication networks is to provide ubiquitous communication to a wide range of mobile terminals supporting wide range of services. Different users have different call arrival and mobility patterns. Use of these parameters can help to optimize the resource utilization. Therefore the protocols should be adaptable to user characteristics.

8. Security and authentication: Wireless environment has inherent limitations in terms of security due to the broadcast nature of the communications. Mobile users also break the traditional model of security and authentication. While designing techniques to support mobile users, an important consideration is to incorporate security and authentication mechanisms such that no new security risks are introduced compared
to supporting static and wired hosts. Issues such as denial of service attacks, unauthorized access, unauthorized redirection of traffic, and firewall traversal should be considered.

9. Fault tolerance: Most of the current mobility management techniques that are implemented in the first and second generation systems depend on certain central network elements to provide service to mobile users. An example of a central network element in the case of mobile IP is the Home Agent [13] and HLR in the case of IS-41. The new protocols developed should have built-in fault tolerance such that even if some of the network elements fail the mobile users will still be connected.

10. Broadcast and multicast support: Ideally, the mobile user should be able to access services as if the mobile user was stationary. Currently, only unicast data delivery is supported to/from mobile terminals. Other important modes of data delivery that need to be considered while designing mobility support are broadcast and multicast support.

11. Application Support: Different applications require different Quality of Service (QoS) requirements. It is also very important to consider the behaviour of applications that are being supported while designing protocols for mobility support.

Other issues to be considered are techniques to support other mobility scenarios such as mobile networks (networks in which a part of the network moves: mobile routers) and ad-hoc networking scenarios, and extensions to support additional services such as Virtual Private Network access to mobile users. Also, most networks do not have homogenous, i.e., do not support the same protocols throughout the entire network, in this scenario it is important to consider the interworking between the different protocols.
1.3 Motivation and scope

Extensive research has been performed for supporting mobility and currently, there exists a wide range of technologies supporting mobile users. Examples of such technologies are Cellular Telephony and PCS. In the next generation technologies supporting mobile users, the aim is to provide ubiquitous communication to a wide range of mobile terminals supporting wide range of services. In order to achieve this goal, new network architectures are needed to provide generic cost effective support for a variety of cellular, PCS, and wireless data technologies. This gives rise to two architectural choices, IP (Internet Protocol)-based architecture and ATM (Asynchronous Transfer Mode) -based architecture. It is important to provide efficient support for mobility in IP-based architecture because of the wide deployment of IP-based infrastructure. The choice of ATM-based switching architecture is considered because it is a de facto transfer mode for the broadband fixed communication networks. The choice of ATM as the backbone network for generic mobility supporting infrastructure is motivated by: 1) ATM’s superior cost/performance when used as a switching technology for large traffic volumes, 2) ATM's transport capabilities which allows different traffic types to be carried over the same network, and 3) Increasing acceptance of ATM-based transmission systems and advantages of common infrastructure for fixed and mobile terminals.

This work focuses on the supporting mobile terminals in wireless networks, primarily ATM Networks. It addresses the network layer aspects of mobility support such as, Location Management, Handoff Management, Route Optimization, and Transport of Signaling Messages. The techniques developed are independent of wireless access technologies, and therefore can be used with any access technology. These techniques can be applied to provide efficient IP-mobility support and can be deployed to enhance
performance in existing wireless networks such as PCS. In this thesis, the application of the developed techniques to ATM networks, IP-based networks, and PCS networks are also discussed.

1.4 Dissertation outline

The outline of the dissertation is as follows:

Chapter 1 provides an introduction to wireless networks and describes the problems that occur in wireless networks.

Chapter 2 provides the required background information. It reviews routing protocols and briefly describes ATM Networks and Protocols and the IP-based networks and protocols.

Chapter 3 describes the problem of location management and provides background information regarding the various approaches to location management and existing solutions.

Chapter 4 proposes an integrated scheme to perform location management in wireless networks. The techniques are explained by applying these techniques to support mobility in ATM Networks (mobile PNNI scheme). It also describes how this scheme can be used to provide efficient IP-mobility support.

Chapter 5 proposes a one-phase crankback scheme to route connections to mobile hosts such that the established connection is an optimal one. It also presents comparative performance results comparing the one-phase crankback scheme to two-phase schemes such as the Integrated Scheme (mobile PNNI Scheme).

Chapter 6 proposes a flat overlay scheme to perform location management. It presents comparative performance results comparing this new scheme with existing schemes.
Chapter 7 proposes a generalized overly scheme called the Location Register scheme (LR scheme). The performance analysis shows the effect of user and network characteristics on the hierarchies of location registers in a location management architecture. It also presents results comparing the integrated scheme with the LR scheme.

Chapter 8 describes the problem of route optimization in wireless networks and proposes a solution to this problem. It also demonstrates how this solution can be applied to the location management and handoff management.

Chapter 9 proposes a one-phase dynamic COS search handoff scheme. It presents comparative performance results comparing the proposed scheme to existing schemes.

Chapter 10 explores the issue of transporting signaling messages in ATM networks and presents an efficient method to transport signaling messages.

Chapter 11 describes the details of the wireless network prototype built to demonstrate the effectiveness of the mobility techniques developed.

Chapter 12 concludes the dissertation with a summary of the presented work and suggestions for future work.
CHAPTER 2

BACKGROUND

This Chapter provides a brief introduction to routing protocols, concentrating on the routing architectures in ATM (Asynchronous Transfer Mode) networks and the Internet.

2.1 Routing protocols

Routing protocols are required to route connections in connection-oriented networks and datagrams (packets) in connectionless networks. Routing protocols can be classified as static and dynamic routing protocols. In static routing protocols, routing information in the network elements (switches/routers) is statically configured and so does not change. In dynamic routing protocols, routing information in the network elements changes with the state of the network. Dynamic routing protocols could be implemented in a centralized fashion or in a distributed fashion. In this work we primarily deal with the aspects of

1. In connection-oriented approach, a connection is established before sending the data by determining the path which the data will take. And additionally resources such as buffer and bandwidth can be reserved along the path to give service guarantees to the connection. Once the connection is formed the data is sent. Examples of networks supporting connection-oriented approach are circuit switched networks such as the telephony network and virtual circuit networks such as ATM networks.

2. In connectionless approach, data is sent in packets and each packet is routed independently of the other. In this mode no prior connection setup is required. An examples of a network supporting connectionless approach is the Internet.
dynamic routing protocols that are implemented in a distributed fashion. There are two types of distributed dynamic routing protocols, Distance vector routing protocols and Link state routing protocols. For detailed description of routing protocols please refer [111].

2.1.1 Distance-vector routing protocols

In distance-vector routing, each network element maintains a distance vector, a list of \(<\text{destination}, \text{cost}>\) tuple for each destination \(x\), where cost is the estimate for the sum of the link costs on the shortest path to that destination. The distances are initialized to a value higher than the expected cost of any route in the network. Each network element periodically sends a copy of its distance vector to all its neighbors. Each network element monitors the cost of its outgoing links, and also sends a copy of its distance vector to all its neighbors if there is a significant change in the characteristics of the link. When a network element receives a distance vector from a neighbor, it compares its current cost to the a destination with the sum of the cost to reach its neighbor and its neighbors cost to reach the destination, to determine whether its cost to reach any destination would decrease if it routed packets to that destination through that neighbor. Distance-vector algorithms suffer from problems of both short and long lived loops, the problem of count-to-infinity, and slower convergence [102]. Implementation of distance-vector routing protocols is usually simple. Examples of distance vector protocols are Routing Information Protocol (RIP), Border Gateway Protocol (BGP), and EIGRP (Enhanced Inter-Gateway Routing Protocol).
2.1.2 Link-state routing protocols

In link-state routing the topology of the network and the cost of every link is distributed to all the network elements. In other words, all the link information such as the state of the link and the nodes it connects is known to all the network elements. Each node maintains a view of the network topology with a cost for each link. Using this information each network element independently computes the optimal paths to all destinations. Typically flooding algorithm is used to spread the link information. Link-state routing suffers from problem of short-lived loops (which usually disappear in the time it takes for a message to traverse the diameter of the network) and that the implementation is complex compared to distance-vector routing. The convergence in link-state routing protocols is faster compared to distance-vector routing. Examples of link-state routing protocols are OSPF (Open Shortest Path First), IS-IS (Intra-Domain Intermediate System to Intermediate System), and PNNI (Private Network-Network Interface) routing protocols.

2.2 Hierarchically organized networks

Consider a network with \( N \) Nodes and \( E \) edges using a link-state routing algorithm. It can be shown that computing the shortest paths takes \( O(E \log E) \) computation, the routing table requires \( O(N) \) storage, the link-state database requires \( O(E) \) storage, and at least \( O(E*N) \) messages to send the link-state information. Clearly, the computation, communication, and the space requirements become excessive as \( N \) becomes large. Because networks are expected to grow to several billion endpoints, it is necessary to use
hierarchical routing. In this thesis networks that use hierarchical routing are referred to as hierarchically organized networks. Next, a brief overview of ATM Networks and the Internet and the routing protocols and signalling protocols used in them are described.

2.2.1 ATM networks and ATM routing architecture

Asynchronous Transfer Mode (ATM) networks aim to combine the flexibility of the Internet with the quality-of-service guarantees of the telephone network. They are designed for high bandwidth, Quality-of-Service, scalability, and manageability. ATM networks are virtual circuit based connection-oriented networks. The user and control plane protocol reference models are shown in Figures 5 and 6. The ATM Adaptation Layer (AAL) segments user information into ATM cells at the source and reassembles cells at the destination. There are different types of AALs defined [6], AAL type 0: Virtually empty for cell relay service, AAL type 1: Continuous Bit Rate traffic, AAL type 2: Real-
time Variable Bit Rate (such as compressed voice or video), AAL type 3/4: Non-real-time Variable Bit Rate (bursty data), AAL type 5: Non-real-time Variable Bit Rate (much simpler protocol).

ATM signalling is a set of protocols used for call/connection establishment and clearing over ATM interfaces. Figure 7 shows the ATM network interfaces. The interface between a End-Host and an ATM switch is referred to as User-Network Interface (UNI)
The interface between two public ATM switches is referred to as Broadband Inter-Carrier Interface (B-ICI). The interface between two private ATM switches is called Private Network-Network Interface (PNNI). Since the model of hierarchically organized networks used in this thesis is based on PNNI-based networks, the PNNI routing and signaling protocols are described later.

The ATM connection setup involves exchange of SETUP, CONNECT, and CONNECT_ACK messages. Connection setup signaling interfaces involved is shown in Figure 8. Connection setup involves call processing and signaling. Call processing involves functions such as route computation and switch resource management. Switch resource management which includes, execution of connection admission control (CAC) algorithm to determine the availability of resources (locally as well as globally), selection of VPI/VCIs on links, set Port/VPI/VCI mapping table (fabric configuration), and set parameters for user-plane algorithms (such as scheduling, priority, etc.). Next, PNNI-based networks and the routing and signaling protocols are described.
PNNI-based ATM networks [3] are arranged in hierarchical peer groups as shown in Figure 9. The lowest level \( l = L \) consists of ATM switches connected in arbitrary topologies. Each peer group has an elected *Peer Group Leader* (PGL). Nodes within a peer group exchange topology, loading and reachability information using the *PNNI routing protocol*. The PGL of each peer group represents all the nodes within its peer group in the higher-level peer group and sends summarized topology/loading/reachability information about its lower-level peer group to its peers in the higher-level peer group. Each higher-level peer node broadcasts this summarized information to all its nodes in the lower-level peer group. Using this technique, each node has topology/loading/reachability data about its own peer group and all its ancestor peer groups. For example, in Figure 9, node A.1.1 has the topology information of the entire peer group A.1, as well as the topologies of peer group A and the top-level \( l = 1 \) peer group.

The *PNNI signaling protocol standard* defines ATM connection setup and release procedures. In order to set up a connection, the first switch receiving the connection setup request determines the hierarchical source route for that connection. The computed
hierarchical source routes are carried as D TL (Designated Transit Lists) parameters in the PNNI signaling SETUP messages [3]. A D TL is list of node identifiers, where a node at the lowest level is an ATM switch, while at higher levels, a node is a logical group node that represents a peer group. A stack of D TLs is used to specify the complete path of a connection from the current node to the destination with one D TL for each level. The exact path (D TLs) within each peer group is computed at the ingress border node of the peer group, which then pushes a D TL specifying the path for that peer group on to the stack of D TLs. While computing the path through a peer group, the ingress border node of a peer group uses the next entry in the next D TL (the one corresponding to the next higher level peer group) as the target for exiting this peer group. In case there is only one D TL remaining (i.e., next D TL does not exist), the destination address is used to determine the route. The egress border node of a peer group pops D TLs that are exhausted from the stack. A call being set up according to a specified stack of D TLs may be blocked at a node due to a lack of sufficient resources or connectivity. In such situations, the call is cranked back (released) to the border node that created the unusable D TL and an alternate route is attempted. "Optimality" of connection paths in PNNI-based networks should be regarded within the context of this hierarchical organization of switches. In other words, because of the hierarchical organization, the "shortest-path" computed by PNNI may not be the true shortest-path. This penalty is paid in return for network scalability.

As an example, a setup request from an endpoint connected to switch A.1.4 to an endpoint connected to switch B.2.2 (see Figure 9) results in an initial path computation of the stack of D TLs [{A.1.4, A.1.3}, {A.1, A.2}, {A, B}], where each D TL is specified within { }. When the SETUP reaches the egress border node A.1.3, it pops the D TL {A.1.4, A.1.3} from the stack of D TLs since the connection is completely established
through the nodes specified in this DTL. Node A.2.1, the border node of peer group A.2, computes the path through peer group A.2 to reach peer group B (the next entry in the next DTL) as consisting of only node A.2.1. Since it is both the ingress and egress border node of peer group, it effectively "pushes-on and pops-off" a DTL \{A.2.1\} from the stack to DTLs. It also pops-off the DTL \{A.1. A.2\} since this DTL is also exhausted. The SETUP message generated to node B.1.5 carries only one DTL in its stack of DTLs, i.e., \{\{A,B\}\}. At the next node, B.1.5 (the ingress border node of peer group B), the path through peer group B is computed to reach the endpoint connected to B.2.2. The new stack of DTLs is \{\{B.1.5, B.1.4\}, \{B.1, B.2\}, \{A, B\}\}, with each DTL corresponding to a level as viewed by node B.1.5. At B.1.4, the first DTL is popped out of the stack. At node B.2.3, a new DTL is computed \{B.2.3, B.2.2\} and pushed on to the stack of DTLs. Thus, the SETUP from B.2.3 to B.2.2 carries the stack of DTLs, \{\{B.2.3, B.2.2\}, \{B.1, B.2\} \{A, B\}\}. When the SETUP reaches B.2.2, egress switch for the end-to-end connection, all three DTLs are popped off.

### 2.2.2 Internet routing architecture

Internet is a loose collection of networks organized into a multilevel hierarchy using a wide variety of interconnection technologies. The main components of the internet are the routing [111], the Internet Protocol (IP), and the addressing [112]. The Internet routing has
a hierarchical structure as shown in Figure 10. At the highest level is the Internet backbone that connects Autonomous Systems. Routing between autonomous systems uses the exterior gateway protocol. An Autonomous System (AS) is a set of routers under a single administrative control, which appears to other Autonomous System to have a single coherent interior routing plan. AS boundary routers represent the AS to the Internet backbone and advertise AS external routes into the AS. Routing within a AS is called Interior Gateway Protocol. At the lowest level we have routing within a single broadcast LAN such as Ethernet or FDDI. The Border Gateway Protocol (BGP) has been recommended as an Exterior Gateway Protocols and OSPF has been recommended as the Interior Gateway Protocol (IGP). RIP is another widely used IGP.

OSPF is a link state routing protocol. An AS consists of smaller networks called areas. Area is a collection of contiguous networks and hosts, together with the routers having interfaces to any of the included networks. The backbone of an AS consists of all the remaining networks not contained in any area, their attached routers and routers that are attached to multiple areas. Area boundary routers are responsible for representing an area to the AS backbone and advertising external routes into the area. In the case of a broadcast
network with multiple routers a Designated Router is elected to advertise. It is also possible to use any protocol in an autonomous system, possibly a routing protocol that allows from more levels of hierarchy.
CHAPTER 3

LOCATION MANAGEMENT (ROUTING FOR MOBILE HOSTS)

Location management is required to be able to route connections or datagrams to mobile hosts. Location management deals with tracking mobiles and locating them prior to establishing an incoming call. Location management algorithms provide mechanisms to track and locate a mobile. The two aspects of this problem, mobile tracking and mobile locating, are also referred to as MOVE and FIND operations, respectively in the literature. Mobile tracking is the procedure by which the network elements update information about the location of the mobile. Mobile location is the procedure by which the network finds the location of the mobile while delivering incoming calls, data packets or other services. The information acquired during the tracking phase is used in the locating phase. There is a trade-off between the resources used during the mobile tracking and the mobile locating procedures. If the mobile tracking procedure is resource intensive, then typically the mobile locating procedure will be less resource intensive. For example, if the mobile registers with its "home" every time it moves, the mobile tracking is resource intensive but the mobile location requires less resources and incurs less delay.
3.1 Requirements

Some of the requirements of location management are given below. Some of the design considerations are given in Chapter 1.

*Permanent address:* Every mobile host should have a permanent address. This address is also referred to as the home address of the mobile. Corresponding hosts or the calling party uses the permanent address to send packets or initiate connection setup to the mobile host.

*Interoperating with existing infrastructure:* Artifacts of the existing infrastructure can be a limiting factor to the design of efficient architectures. Since there is considerable investment in existing infrastructure, and it is more economical to integrate the support of mobile terminals into existing networks that support only fixed hosts. It is necessary to limit the changes to few network elements.

*No weakness of security:* The schemes should provide sufficient security to ensure the communication to and from mobile host is secure, and no additional security holes are created to the fixed terminal already attached to the network.

*Multicast and Broadcast capability:* It is desirable that the mobile hosts get the similar multicast and broadcast service as the fixed terminals.

---

1. Dynamic configuration protocols can be used to provide a partial solution to the problem of location management for portable terminals. In this scenario, temporary address can be obtained from the remote network. This approach has three problems 1) Higher layer connections do not survive the temporary attachment, 2) the temporary address has to be advertised every time (via DNS or some such service). 3) Location privacy is lost.
Location privacy: The location management scheme should provide techniques by which the current location of the mobile terminal is not known to the corresponding host or the calling party.

3.2 Approaches to performing location management

There are two approaches to preforming location management. In the first approach (overlay approach), a location management architecture is overlaid on the backbone network to manage the location of the mobiles (see Figure 11). In other words, location management is performed with the help of external location registers. The location management architecture consists of location registers (databases) interconnected by a signaling network. In this approach, an explicit location query can be issued to the associated location registers, and after determining the exact location of the mobile the
connection is set up. This approach is being used in the current connection-oriented networks, e.g., cellular networks. In the second approach (integrated approach), the mobility management is integrated into the fixed network. In this approach the routing and signaling protocols are modified to handle mobility. In the second approach, immediately after the call arrival, the connection setup is initiated. In this approach routing data is used to deliver the call to the mobile. This approach is being used in the current connection-less networks, i.e., used to deliver packets rather than calls. This chapter provides a brief overview of the various methods proposed under the two approaches for location management. In Chapters 4 to 7, new solutions to the problem of location management are proposed.

3.2.1 Overlay approach

A classification of the location management schemes under the first approach is presented in Figure 12. A notable feature of the schemes in this approach is that they use a hierarchy of location registers to track the location of the mobile. These schemes can be classified as static or dynamic schemes depending on the variation of hierarchy. In the
dynamic schemes, the number of levels of hierarchy will vary with the current location of
the mobile and the mobility of the user. Whereas in the static schemes, the number of
levels in the hierarchy does not depend on the mobility of the users. Several static
schemes have been proposed. These scheme primarily vary in the number of levels in
hierarchy. On one extreme we have the fully hierarchical schemes and on the other
extreme there are flat schemes. A brief review and comparison of these methods is
provided below.

3.2.1.1 Cellular location management standards

Two standards exist for performing location management in cellular networks. The
standard commonly used in North America is the EIA/TIA interim standard 41 (IS-41) [4]
and in Europe the Global System for Mobile Communications [9]. Both these standards
use a two level hierarchical architecture. Figure 13 shows the cellular network
architecture and the mobile tracking scheme defined in the cellular standards. The
network elements consist of BS (base stations), MSCs (Mobile Switching Centers), VLRs
(Visitor Location Registers) and HLRs (Home Location Registers). Base stations provide
wireless access to mobiles. MSCs are the switching centers through which connections to
and from mobiles are made. HLRs and VLRs track the location of mobiles and answer
queries for mobile locations. The cellular standards allow for two variations of the
network architecture: collocated VLR/MSCs and separated VLRs and MSCs as shown in
Figure 13.

Mobile tracking begins with mobiles generating registrations to the network. In this
work, we assume that zone-change registrations are generated by each mobile. Zone-
change registrations allow the network to track a mobile user on a larger scale than on a
per base station move basis. A zone is defined to include all the base stations under an MSC (Mobile Switching Center). Each base station emits a periodic broadcast beacon carrying the zone identifier of the base station. This allows a listening mobile to identify when it moves into a new zone. An effect of this form of tracking is that a page needs to be generated by the MSC to determine the exact base station at which a mobile is located during call setup. The zone size, which determines the level of granularity at which mobiles are tracked, should be chosen to limit the number of registrations received by the network without increasing the paging load significantly. In Figure 13, when a user moves from base station 1 to base station 2, it does not generate a registration, since both base stations are in the same zone. A zone-change registration is sent by the mobile only when the user moves from base station 2 to base station 3. These registrations are conveyed from the base station to the MSC (Mobile Switching Center) using IS-634 messages [82].
The network response to receiving a registration from a mobile is dependent on the network configuration. As shown in Figure 13, MSCs may be collocated with VLRs (Visitor Location Registers) or they may be separate. In the network configuration where MSC/VLR are collocated, the MSC on receipt of the IS-634 message sends a REGNOT (Registration Notification) message to the HLR of the mobile, which in turn sends a REGCANC (Registration Cancellation) message to the old MSC/VLR. In the network configuration where MSCs and VLRs are separate entities, the MSC generates a REGNOT message to the VLR, which in turn, generates a REGCANC message to the old MSC, if the old MSC is connected to the current VLR. Otherwise, it generates a REGNOT message to the HLR of the mobile. On receiving the REGNOT message, the HLR then initiates a registration cancellation procedure with the old VLR.

Given the two IS-41-based forms of mobile tracking shown in Figure 13, the corresponding mobile location procedures are shown in Figure 14. To deliver an incoming call to a mobile $M$, the calling party's switch (originating switch) sends a LOCREQ (Location Request) message to the HLR of mobile $M$ as shown in Figure 14. Since cellular telephony networks use geographical addressing, the HLR of a mobile is determined from the identity of the mobile. The HLR sends a ROUTEREQ (Route
request) to the VLR/MSC from which it received the last registration message for mobile $M$. In case (a), where the VLR is collocated with the MSC, the VLR/MSC allocates a TLDN (Temporary Location Directory Number) for this mobile and returns this value in the response back to the HLR. The HLR sends this parameter in its response to the originating MSC. This allows the originating MSC to initiate connection setup to the MSC of the called mobile. Routing of the connection through the network is done using the TLDN which indicates the far end MSC (of the called mobile). When the call setup message reaches the MSC of the called mobile, the latter issues a page to locate the exact base station at which the mobile is located for call completion. In case (b), two ROUTEREQ messages are required as shown in Figure 14. While the hierarchical arrangement of separated VLRs and MSCs reduces the signaling traffic generated by mobile tracking, it is clear from Figure 14b that the penalty paid is increased mobile location delay.

3.2.1.2 Forwarding scheme

In the forwarding scheme [7], a chain of forwarding pointers is maintained, as shown in Figure 15a, to limit the registration (tracking) signaling load. As a user moves between base stations connected to different VLR/MSC nodes, registration messages are sent by the mobile to the network. The new VLR/MSC simply communicates with the old VLR/MSC which adds a pointer to the forwarding chain. Thus, registration signaling traffic is localized. However, during call setup, a chain of pointers need to be traced to locate a mobile, potentially leading to long setup delays. This scheme is proposed for networks with low CMRs (Call-to-Mobility Ratios).
3.2.1.3 Anchor scheme

An improvement to this scheme is proposed in [10] where a local anchor VLR is created. As users move, the pointer at the local anchor is updated as shown in Figure 15b. This localizes registration signaling and, at the same time, achieves lower mobile location delays than the forwarding scheme. We note that at low CMRs, the anchor scheme is similar to the IS-41 scheme with separated VLR/MSCs by comparing Figure 14b and Figure 15:b. In effect, by creating a local anchor, the mobile location requires a three-hop message exchange (HLR -> Anchor VLR/MSC -> Current VLR/MSC). Similarly, in the IS-41 model with separated VLRs and MSCs, the mobile location requires a three-hop message exchange (HLR -> VLR -> MSC). One difference between these two schemes is that the assignment of VLRs to MSCs is static in IS-41, leading to automatic updates of the VLR as a user moves, while in [10], the local anchor is changed either on each call arrival (static local anchoring), or in a dynamic manner only if the expected cost is lower.
(dynamic local anchoring). At high CMRs, since the local anchor will be frequently updated to be the current VLR/MSC, the anchor scheme resembles the IS-41 scheme with collocated VLRs/MSCs more closely than the scheme with separated VLRs and MSCs.

The tracking costs of the forwarding scheme are smaller than the tracking costs of the anchor scheme because only two messages are exchanged during a move in the forwarding scheme, while four messages are required in the anchor scheme (to update the local anchor and to cancel records of the mobile in the old VLR). However, the search costs are smaller in the anchor scheme compared to the forwarding scheme. Furthermore in the anchor scheme, if the cancellation of the information at the old VLR is done by a timer-based mechanism, the tracking cost is same as in the forwarding scheme.

3.2.1.4 Flat tracking/locating scheme

In [55] a "flat" scheme is proposed for location management. In this approach, one level of tracking is used for all mobiles. In other words, HLR tracks the current switches of all its mobiles. Therefore, whenever a mobile moves to a new location, the HLR needs to be notified. In Chapter 3, a "hybrid flat" tracking/locating scheme is proposed for location management. In this approach, one level of tracking is used for mobiles in their "home" network, and a two-level tracking approach is used for mobiles visiting in other networks, referred to as "remote" mobiles. An HLR tracks the MSC locations of its home mobiles, and the VLR locations of its remote mobiles. VLRs track the MSC locations of these remote mobiles.
3.2.1.5 Hierarchical location management scheme

[12] proposed a hierarchical location management scheme, which reduces long-distance signaling for both the mobile tracking and locating operations. Figure 16 shows the network architecture assumed in [12]. The topmost level can be considered as "Earth". The next level nodes then represent "countries", with following level representing "states within countries", etc. If the mobile, whose home is node 1, moves to node K, a chain of pointers are set up as shown in Figure 16 from the home node (node 1) to node K. If an endpoint at node L calls the mobile, node L sends a call setup message to node M (in the direction of the home node of the mobile). Since node M has a pointer for the mobile (indicating node K), it sets up a connection to node K. Thus, long-distance signaling is avoided by using this chain of pointers to cut short searches. Similarly, as the mobile moves between nodes, say the mobile shown in Figure 16 moves from node K to node L, then the mobile tracking update only propagates until node M. Node M changes its pointer but does not require the message to be propagated further upstream.
3.2.1.6 Discussion

The flat scheme results in lower \textit{computation costs} but incurs larger \textit{communication costs} than the cellular scheme, while the hierarchical scheme achieves the opposite (lower communication costs, but higher computation costs). By building a rooted tree of location registers, the need for home location registers is eliminated, thus removing the need for "long-distance" signaling messages. On the other hand, to determine the location of a mobile, the location query needs to be stopped and processed at a much larger number of location registers (on the average). This increases the computation costs of the network. In contrast, the flat scheme increases the overall signaling load (communication costs) on the network since registrations and location queries need to be sent "long-distance" to the HLRs of mobiles, but it also decreases the computation costs since processing is needed only at one node for both registrations and location queries. The flat scheme also results in a lower mobile location delay due to the one-hop location query processing, which, in turn, leads to a lower overall call setup delay. Other improved schemes, such as the forwarding scheme of [7] and the anchor scheme of [10], are in between these two extreme schemes in terms of computation and communication costs. In Chapter 7, a generalized overlay scheme, the Location Registers (LR) scheme, is described. It is a hybrid scheme whose parameters can be set to default to one of the two improved schemes, i.e., the flat scheme or the hierarchical scheme.

3.2.2 Integrated approach

In the \textit{second approach}, the mobility management is integrated into the fixed network. This approach is being used in the current connection-less networks, i.e., used to deliver packets rather than to establish connections. Consider the architecture of the network node
shown in Figure 11. In the integrated approach, the signaling and routing protocols in the switch/router are modified to handle mobility. Next, schemes proposed under the integrated approach are briefly described.

3.2.2.1 IETF mobile IP

Mobile IP [13, 14] is an extension to the Internet Protocol (IP), which enables hosts to change their point of attachment to the internet without changing their IP address. In this protocol, a packet for a mobile host is routed to the home network of the mobile as identified by its permanent IP address. The home network tracks the current location of the mobile and tunnels the packet to the current network of the mobile. In order to prevent this “triangle” routing, route optimization extensions have been proposed [14]. These extensions provide a means for communicating-nodes to maintain a binding between the mobile and its current location, and use this binding to tunnel datagrams directly to mobile. Extensions are also provided to allow datagrams in flight when a mobile node moves, and datagrams sent based on an outdated binding information, to be forwarded directly to the mobile.

3.2.2.2 IBM mobile host protocol

In this method, a Mobile Router that resides on the subnet associated with the MH’s home address is used to perform routing. The Mobile Router has the current location information of the mobile terminal (via a registration procedure). In this scheme the normal IP routing delivers a packet destined to a mobile terminal to its home network. If the mobile is currently in its home network, the packet is delivered. Otherwise, the Mobile Router intercepts the packet and using Loose Source Routing (LSR) option delivers the packet to the mobile terminal. Once the correspondent host learns of the location (through
the return packet's LSR option). It can send the later packets via an optimally routed path using LSR option. The major disadvantage of this proposal is that LSR option is not implemented in many existing routers. Several other proposals have been proposed for providing IP-mobility support [113][114].

3.2.2.3 Integrated mobile ATM methods

Recently several methods have been proposed to perform location management in mobile ATM networks. Next, a brief overview of representative proposals is given.

Chapter 4 proposes a mobile PNNI scheme, which integrates the handling of mobility into the ATM network instead of creating an overlay of location registers. In this method limited reachability updates are propagated as mobiles move allowing a small "neighborhood" of switches to know the switch through which the mobile can be reached. Calls originating within this neighborhood will be set up optimally on the shortest paths. To handle calls originating at switches outside this neighborhood, a forwarding pointer is set at the home switch of the mobile indicating the current location of the mobile. On call arrival, originating switches set up connections as per the data indicated in their reachability information. If the call originates within the neighborhood of the mobile, as

Figure 17: Integrated mobile ATM methods
shown in Figure 17 for the called mobile located in position I, or if the mobile is at its home location, the connection is routed on the shortest path. However, if the call originates at a node outside the mobile's current neighborhood, it will be routed toward the home of the called mobile, at which point, it needs to be somehow rerouted to the current location of the mobile. The mobile PNNI scheme proposes forwarding the call from the home switch to the current location. This is illustrated in Figure 17 for the case when the called mobile is in position II. Other options for handling this situation are proposed in [21, 23]. Acharya et al. [23] proposed a complete release of the connection from the called mobile's home switch to the calling party's switch followed by a setup from the calling party's switch to the current location of the mobile. Ayyagari et al. [21] proposed using the crankback feature of PNNI signaling (as shown in Figure 17 for the case when the called mobile is located at position III), which allows for alternative routing of calls that are blocked at a node due to insufficient resources or connectivity. In [21], the authors note that typically crankback will be executed at a local level, which means that the overall end-to-end path taken by the connection could be "suboptimal."
These three schemes can be classified into one-phase or two-phase schemes as shown in Figure 12. In one-phase schemes, such as the complete release scheme of [23], an optimally routed connection is constructed during the call setup procedure, thereby eliminating the need for route optimization. In two-phase schemes, such as [34, 21], the call is delivered to the mobile possibly on a suboptimal route (first phase) and a route optimization (second phase) is performed subsequently to achieve an optimal route.

The complete release method incurs a high connection setup delay, while the call forwarding scheme (in which the crankback node is statically assigned to be the home node) uses the most amount of network resources (albeit for a short period of time until the route optimization phase is complete). The crankback case, in which the node to which the connection is "cranked back" is determined dynamically, is between these two extremes as shown in Figure 12. The mobile PNNI Scheme is described in Chapter 4 and a one-phase crankback scheme in which the connection is cranked back and rerouted such that the resulting connection is routed via an optimal path, is described in Chapter 5.
3.3 Organization of the solutions

In this Chapter an overview of the various location management schemes is presented. Chapter 4 proposes an integrated scheme (mobile PNNI scheme) to perform location management in Mobile ATM Networks. Chapter 5 proposes a one-phase crankback scheme to route connections to mobile hosts such that the established connection is an optimal one. It also presents comparative performance results comparing the one-phase crankback scheme to two-phase schemes such as the Integrated Scheme (mobile PNNI Scheme). Chapter 6 proposes a flat overlay scheme to perform location management, and Chapter 7 proposes a generalized overly scheme called the Location Register scheme (LR scheme). The performance analysis shows the effect of user and network characteristics on the hierarchies of location registers in a location management architecture. It also presents results comparing the integrated scheme with the LR scheme.
CHAPTER 4

INTEGRATED SCHEME (MOBILE PNNI SCHEME)

In this chapter the integrated scheme (mobile PNNI scheme) is described. In the this scheme the mobility management is performed by enhancing the routing and signalling protocols. This scheme is explained in the context of ATM networks and is referred to as the mobile PNNI scheme. There are three aspects to mobile PNNI scheme: Mobile tracking (Move, power-up, and power-down), Routing connections/data (combined with mobile locating), and Route optimization. Though the principles of the integrated scheme are presented in the context of PNNI routing and signaling protocols, they can be easily applied to any hierarchical routing protocols. Application of the integrated scheme to provide IP-mobility support is also discussed.

4.1 Architecture to support mobile endpoints

In this Section the various mobility scenarios are discussed and architectural changes to current networks needed to support mobile terminals are presented. Access to a network can be provided by wired medium or wireless medium and the accessing terminals can be
mobile or fixed. This gives rise to four possible scenarios as shown in Figure 19. In this
work we are primarily concerned with scenarios with portable and mobile terminals. In
Figure 19 only terminal mobility is considered. One could potentially consider network
element mobility and also ad-hoc network configuration. Example of network element
mobility scenario would be a network in a plane or a train where a part or the entire
network moves. In this thesis, we describe the techniques with the assumption that the
terminals are mobile terminal integrated schemes. These techniques can also be used for
other mobility scenarios also.

Mobile endpoints can be supported in an existing network architecture in the following
manner (see Figure 20). Mobiles are located at base stations, which are assumed to be
organized as in cellular networks, with multiple base stations connected to each switch as
shown in Figure 21. Zone-change registrations are used to limit air interface registration
traffic, where a “zone” consists of all the base stations under a single switch. When a call
setup arrives at a switch, it pages all its base stations to determine the exact base station on
which it is located. General configurations, allowing base stations to be connected to
Network with mobility enhancements

Figure 20: Mobility Scenarios

Figure 21: Configuration of a zone

Multiple switches, and/or with different definitions of zones, are possible. The configuration shown in Figure 21 ties in with the architecture of Figure 9 at the lowest level. In other words, some of the switches in Figure 9 are connected to a set of base stations as shown in Figure 21. These base stations offer wireless access to mobile endpoints.
Wireless ATM reference configuration is shown in Figure 22. Radio access to the fixed

ATM network can be provided by adding radio access capability to the ATM switches as an add-on functionality. All supporting radio functionality such as radio MAC (Media Access Control), radio resource management functions, etc. are implemented in the ATM Switch. The interaction between base station and the switch control is internal to the switch and base stations are simple radio transreceivers. It is also possible to provide radio access by adding a new stand-alone device between the ATM switch and the mobile terminal. In this configuration the new device implements the radio specific functionality and the ATM switch implements call and connection control and mobility management aspects. Figure 23 shows the Wireless ATM Protocol Architecture. Modifications needed to the existing protocol architecture are also shown (+M). In this architecture shows the end-to-end ATM scenario. It is also possible to terminate the ATM mode prior to the wireless access.
4.2 Mobile tracking

In the integrated scheme the mobility management functions are performed by enhancing the routing and signaling protocols. In particular we use the feature of sending reachability updates when a mobile moves into a new area, powers on or powers off. The reachability information, which indicates how a specific set of addresses can be reached, can be propagated using the reachability information feature in PNNI networks, or as host specific routes or external routes. For example, Figure 24(a) shows the reachability information corresponding to the mobile A.1.1.2 in the different network elements. A tuple $<x, y>$ is used to denote the reachability information, second element $y$ represents the peergroup though with the first element $x$ is reachable. Nodes in the peergroup A.1 have the reachability information stating that all nodes with the prefix A.1.1 are reachable.
through the switch A.1.1 (A.1.1.* is used to denote nodes whose addresses have a prefix of A.1.1). Node A.2.1 has the reachability information that all nodes with the A.1 as a prefix are reachable through the peergroup A.1. When the mobile moves from A.1.1 to A.1.2, A.1.2 advertises reachability for A.1.1.2. In other words, all the nodes in the peergroup A.1 have reachability information stating that A.1.1.2 is reachable through A.1.2. Note that the reachability information in the nodes outside peergroup A.1 does not change. Once this reachability information is in place, routing can be done based on longest prefix match or using special information table for mobiles. Most current routing is based on longest prefix match routing.

The above technique of sending reachability information gives an integrated solution to the problem of location management. This technique is referred to as fully-integrated scheme. This is an elegant solution to the location management problem for local area mobility, i.e., the mobile is “near” its home location. Home location of a mobile is the switch from which the mobile derives its address. For example A.1.1 is the home location of A.1.1.2. For wide area mobility, i.e., the area to which the reachability information is large, this solution has the following problems:
• Flooding of messages over a large area.
• Large delays in propagation of reachability information.
• Increased number of routing entries, which increases the route lookup latency.
• Increased size of link state database.

In order to avoid these disadvantages, we use the notion of scope which limits the area to which the reachability information. We refer to this area as neighborhood in this text. Before describing how reachability updates propagate, we define three terms, ancestors-are-siblings level, scope, and the neighborhood of a node. The ancestors-are-siblings level \( a_{ij} \) of nodes \( i \) and \( j \) is the level at which the ancestors of the two nodes \( i \) and \( j \) belong to the same peer group. The scope \( S \) is used to set the stopping point for reachability information propagation. If scope is \( S \), reachability data sent by a node \( i \) does not propagate to any node \( j \) for which \( a_{ij} < S \). The neighborhood \( G_i \) of a node \( i \) is defined to include all nodes \( j \) such that \( a_{ij} \geq S \).

As an example, consider the PNNI based network shown in Figure 9. The ancestors-are-siblings level of nodes A.1.1 and A.2.2 is 2. If the scope \( S = 2 \), the neighborhood of node A.1.1 includes all nodes in peer group A but excludes all nodes in peer group B.

Reachability updates are propagated according to the rules shown in Figure 25. If a mobile powers on within its home neighborhood \( (a_{hv} \geq S) \), reachability data needs to be changed at only a few nodes (the exact set of nodes updated is indicated in Figure 25). However, if it powers-on outside this neighborhood, the whole new neighborhood receives a reachability update about the mobile overriding its default summarized reachability, which indicates that the mobile is at its home. The same rules apply for reachability updates sent when a mobile powers off. As mobiles move, if the new location is within the neighborhood of the old location, reachability updates are sent up only to a
subset of nodes whose reachability data changes as a result of the move (the exact set of nodes is defined in Figure 25). On the other hand, if the mobile moves outside its current neighborhood, the entire new neighborhood needs to receive a reachability data update that overrides the default summarized reachability data about the mobile. In addition, the old neighborhood is also updated to cancel the limited reachability update that overrode the default summarized reachability data. In effect, this resets reachability information about the mobile to indicate that the mobile is at its home location. Using this approach, all nodes within a mobile's neighborhood know its exact location, while nodes outside its neighborhood believe that the mobile is at its home location (default reachability). The reason for this arrangement is to allow calls originating from a switch within a mobile's neighborhood to be routed directly to the mobile (without having to be routed to the home switch first). The forwarding pointer at the home switch allows for calls originating at nodes outside the neighborhood to be routed to the home location as per reachability data in these nodes, and then forwarded to the mobile's current location based on the
forwarding pointer. The scope parameter $S$ allows the network provider to use a large neighborhood by setting $S$ to a low value (which will lead to a large number of reachability updates but will result in more calls being routed directly) or vice versa for a small neighborhood.

As an example, consider the PNNI based network of Figure 9 and assume that the scope $S$ is 2. If a mobile A.1.1.5 powers on at A.1.2 (i.e., within its home neighborhood), the reachability update overriding the default reachability data only propagates to nodes in peer group A.1. No PTSP is sent from A.1 to A.2, since there is no change of reachability data stored in A.2 nodes regarding mobile A.1.1.5. If instead it powers on at a base station connected to node B.1.1 (i.e., outside its home neighborhood), reachability updates will be propagated through peer group B.1, and then upwards (i.e., PTSPs carrying reachability updates are sent from B.1 to all nodes of peer group B) and finally, downwards from LGNs other than B.1 in peer group B to their child peer groups. Since $S = 2$, no reachability updates are sent to nodes in peer group A, which all believe that the mobile is at its home A.1.1.

**Setting of forwarding pointers (sending registrations):** Figure 26 shows a flowchart of the actions involved in setting forwarding pointers. In two of the three cases shown in Figure 26, forwarding pointers have to be set at the home in order to route calls generated by nodes outside the neighborhood (defined by scope $S$). In the third case, when the mobile does not change neighborhoods during a move, the forwarding pointer data in the home switch is accurate even after the move, and hence no registration is sent to the home. Forwarding pointers have to be set at the old location in the case of a move for the following reason. The forwarding pointer is needed to handle calls that the home switch may have forwarded to the old location before it receives the registration message.
Registration resulting from a move from node o to new node n

Power on/off registration —> Mobile registers at switch n

Send registration to home switch

Yes

New location outside the neighborhood of the old location?

No

Send registration to home switch; Send registration to old switch to set a pointer at the old switch to forward calls to new location (if home = old, only one registration is needed)

Send registration to old switch to set a pointer at to the old switch to forward calls to new location

Figure 26: Flowchart representing how forwarding pointers are set

Updating its forwarding pointer with the new location of the mobile. Also, calls originated within the old neighborhood will be directed to the old location until the old location issues a reachability update to cancel the limited reachability update override issued earlier for the mobile. Such calls will need to be forwarded from the old location to the new location of the mobile.

As examples, consider the mobile B.1.1.5 located under a base station connected to its home B.1.1 as shown in Figure 27. Let \( S = L \), which means that the neighborhood of a mobile is the set of switches in its level \( L \) peer group as shown in Figure 27. If the mobile
moves to a base station under switch B.1.2 (case of Figure 26: in which the new location is in the neighborhood of the old location), then a Registration message will be sent from B.1.2 to B.1.1 (as shown in Figure 27). Further, if it moves to B.3.1, two Registrations are sent to set forwarding pointers at the home and old locations (B.1.1 and B.1.2, respectively) of the mobile (case of Figure 26 in which the new location is outside the neighborhood of the old location).

Details regarding how registration messages are transported are two-fold. First, each mobile maintains the identifiers of its old and home switches, allowing it to communicate this information when registering at a switch (power-on, power-off or move). This information is used by the new switch to generate the registration message to the home or old switch. Second, we assume the availability of connectionless transport to send location management messages, such as registrations. One such transport mechanism is connectionless ATM (CL-ATM) proposed in Chapter 10. Without this assumption, on-demand connections would need to be set up and released for the transport of every Registration message, which creates a considerable processing and signaling overhead.

In summary, when a mobile powers on or changes locations, the mobile tracking procedure uses a combination of setting forwarding pointers at the home and old (in case of a move) locations of the mobile by sending registration messages, and sending limited reachability updates (with a scope $S$) using the PNNI routing protocol to override the default summarized reachability information, which indicates that the mobile is at its home.
4.3 Connection Setup/Mobile locating

In the mobile PNNI approach, there is no explicit mobile location procedure prior to connection setup. Instead, connection setup proceeds with every switch "believing" its reachability information. Figure 28 shows how incoming connections to mobiles get routed in the mobile PNNI scheme. The path taken depends on the locations of the calling party, and the home and visiting locations of the called mobile. If the calling party is in the mobile's neighborhood or the mobile is in its home neighborhood, the call is routed directly to the mobile. Otherwise, the call is routed to the home switch first, since the switches outside the neighborhood of the mobile have default reachability information. In

Figure 28: Flowchart representing how connections are routed
this scenario, the home switch forwards the call to the current location of the mobile. If the mobile has moved recently, the path may also depend on whether the call arrives after the reachability updates have propagated and forwarding pointers have been set, or whether the call arrives prior to the completion of reachability update propagation and/or the setting of forwarding pointers.

We illustrate some of the cases shown in Figure 28 with examples. Consider that A.1.1.1 issues a call setup to the mobile B.3.1.4, currently located at B.1.2, as shown in Figure 29. If $S = 2$, then the called mobile is in its home neighborhood and the calling party is outside the called mobile’s neighborhood. Call setup proceeds through peer groups A.1 and A.2 and arrives at peer group B (at a node B.1.1, which is in peer group B), since the reachability information in the nodes of peer group A indicate that the mobile is in its home peer group (B). Once the call setup message arrives at peer group B, it is routed efficiently to B.1.2 since all nodes in the neighborhood (which includes all nodes in peer group B) have accurate reachability for mobile B.3.1.4. A second example illustrating the case when the calling party is outside the called mobile’s neighborhood, the called mobile is in its home neighborhood, and the call arrives at a switch before it is updated with the correct reachability information about the mobile. Consider the situation in which the mobile B.3.1.4 has just moved from B.3.1 to B.1.2 and a call arrives at switch B.1.1
before B.1.1 is updated with reachability information for the mobile endpoint B.3.1.4. In this case, switch B.1.1 may choose \{B.1.1, B.1.4, B.2, B.3\} as the shortest path by which to reach peer group B.3 based on its current reachability information for mobile B.3.1.4 (which points toward B.3). The call is then routed from the old location B.3.1 to B.1.2. Thus, the connection route will be inefficient as shown in Figure 29 (with the arrows indicating the connection route). As a third example, consider that \(S = 3\), and that a call to mobile B.3.1.4 is generated by an endpoint attached to switch B.2.1. Since \(S = 3\), the B.2 nodes are not updated about the move of the mobile B.3.1.4 from B.3.1 to B.1.2. This is an example of the case when the calling party is outside the called mobile's neighborhood and the called mobile is not in its home neighborhood. In this case, the connection will be routed to the home and then to the new location (from B.2.1 to B.3.1, back to B.2.1 and then to B.1.2).

For connections being forwarded from the old (or home) location of the mobile to its current location, the call setup message needs to "tunnel" the mobile's home address while using the mobile's current (temporary) address to perform connection routing. In the second example described above (also shown in Figure 29), when the connection is rerouted from B.3.1 to switch B.1.2, if the called party number parameter in the SETUP message indicates the home address of the mobile, then node B.2.1 will again turn the connection setup back toward B.3 (if its reachability information for B.3.1.4 is not yet updated). To avoid this, switch B.3.1 must use a temporary address, such as B.1.2.0 (where the "0" extension indicates "mobile users") in the called party number field. This will allow node B.2.1 to route the connection toward node B.1.2. Upon receiving the setup,

---

1. The first node receiving the call setup message in each peer group determines the route of the connection through that peer group.
B.1.2 will recognize the B.1.2.0 number in the call setup to indicate a mobile. It then looks for the tunneled mobile's home address in the setup message, i.e., B.3.1.4, to page for the mobile and complete connection setup.

4.4 Discussion

4.4.1 Determining Scope

Scope is an important parameter that determines the performance of the mobile PNNI scheme. In Chapter 7 we discuss details regarding ways to determine the optimal scope for a user. It is important to note that the parameter Scope can be set on a per mobile basis and can also be dynamically changed depending on the current environment and the mobility characteristics of the mobile. As mentioned earlier it is not desirable to have a small value for scope (large neighborhood). Therefore some of the connections get routed in triangular path and therefore route optimization is required.

4.4.2 Route optimization

As shown in the examples of Section 4.3, connections may be inefficiently routed due to the lack of correct reachability information. This implies a need for route optimization. A similar need for route optimization exists in mobile IP networks [14]. However, unlike IP networks, ATM networks deliver cells in sequence. The ATM route optimization procedures should maintain cell sequence. The route optimization is performed in two steps. First, a "switchover node" at which the connection is to be rerouted from the old path to the new path is found, and a new segment is set up from the route optimization-initiating switch to the switchover node. Second, user data is switched over from the old
path to the new path using "Tail" signals and buffering to perform this action without loss of cell sequence. Details of the switchover node selection procedure are presented in Chapter 8.

4.4.3 IP mobility support

In this section application of the mobile PNNI scheme to IP mobility support is briefly described. As described earlier, the Internet routing has a hierarchical structure as shown in Figure 30s. At the highest level is the Internet backbone that connects Autonomous Systems (AS). In OSPF, AS is subdivided into smaller networks called areas. Area is a collection of contiguous networks and hosts, together with the routers having interfaces to any of the included networks. At the lowest level we have routing within a single broadcast LAN such as Ethernet or FDDI.

Each mobile host and the agent can be configured as a stub area or the mobile host can be part of an area that has other networks. Using feature of link advertisements and setting policy in current routes, networks can be configured to send reachability information. So, if $S=4$, then the reachability information is sent only in the lowest level, i.e. single
broadcast LAN. If $S=3$, then the reachability information is sent only to the area in which the mobile is currently residing. If $S=2$, then the reachability information is sent to the area in which the mobile is currently residing and also to the area boundary routers in the AS. If $S=1$, then the reachability information is sent additionally to the routers in Internet backbone.

4.4.4 Fault tolerance

Most location management architectures have a single point of failure. For example in the IS-41 standards technique the single point of failure is HLR and for mobile IP it is Home Agent. In the integrated scheme (mobile PNNI scheme) technique of sending reachability information increases the fault tolerance of the system. For example, when the mobile is in its home neighborhood even if a fault occurs in the Home Switch, data can be delivered to the mobile (if $S \leq L$). This technique increases the fault tolerance even when the mobile is not in its home neighborhood, if the failure occurs in the current neighborhood.

When the mobile is not at its home neighborhood or when (if $S > L$), the Home Switch/Visiting Switch are single point of failures. In order to increase the fault tolerance, technique of replication can be used. In addition to the Home Switch the location information can be stored in a set of switches. When a mobile registers, the Home Switch can send the required information to this set of switches or if a mechanism exists the set of switches can learn of this information automatically (this is possible in a broadcast environment). When the Home Switch fails (in dynamic environments, Hello protocol is used to learn of link and system failures), one of the switches in the set can act as the home switch. The selection can be priority based or by an election procedure. The new Home
Switch should advertise the reachability for the mobiles which were the home mobiles of the old Home Switch. By this mechanism the future communication to the mobiles is not disrupted.

In order to take care of the existing connections, all the connections could be reestablished (this might affect the applications) or the correction could be rerouted around the old Home switch and through the new Home Switch. Registration messages could be sent to an any cast address so that it reaches the acting Home Switch irrespective of the who is acting as the home switch. These techniques could be used at the Visiting Switch also.

4.4.5 Security and authentication

Mobile users break the traditional model of security and authentication. In order to prevent denial of service attacks, unauthorized access, unauthorized redirection of traffic, authentication of mobility messages is required. For example, the Visiting Switch and the Home Switch have to ensure that the registration message is sent by the authorized mobile terminal. Authentication Headers can be used similar to the ones used in mobile IP. Secret key mechanisms can be used between the Home Switch and the mobile due to one-to-one nature of the relationship. Public key challenge authentication can be used between the Home Switch and the Visiting Switch, thus authenticating one switch to the other. The Visiting Switch can ensure the identity of the mobile by relying on the Home Switch authentication as a third party authentication. It is also necessary that the routing protocols use the necessary security and authentication mechanisms (these are necessary even in non-mobile environments).
4.4.6 Mobile network or mobile router/switch

A mobile node can be a switch/router that has an entire network which moves with it. This mobile router/switch could be responsible for the mobility of networks moving together. Examples of this scenario are networks in Airplanes, Trains, Ships, and Satellites. It is also possible that each of these networks could consist of multiple mobile networks (mobility-in-mobility), or could support mobile users. The techniques integrated scheme and some of the techniques explained in later chapters can be modified to provide mobility in these scenarios. These techniques can be applied recursively to the mobility-in-mobility scenario also.

4.5 Summary

In this Chapter, the mobile PNNI scheme was described. In the mobile PNNI scheme the mobility management is performed by enhancing the routing and signalling protocols. The PNNI routing protocol is used to convey reachability information about endpoints to ATM switches. The “scope” parameter (set to some number $S$) is used to limit the region of nodes which receive reachability updates as mobiles move. There is no explicit mobile location phase prior to connection setup. Instead, connections are set up to mobiles according to the reachability information at the switches. Application of the mobile PNNI scheme to IP mobility support is also described briefly.
CHAPTER 5

ONE-PHASE CRANKBACK SCHEME

In the previous chapter, we described a mobile location management scheme called the mobile PNNI scheme. In this scheme, reachability updates are propagated with limited scope as mobiles move allowing a small "neighborhood" of switches to know the switch through which the mobile can be reached. In other words, switches within the neighborhood of the mobile have the correct reachability information about the mobile, and the switches outside the neighborhood of the mobile will have default reachability information indicating that the mobile is at its home switch. For example, in Figure 9:, neighborhood of a mobile currently under the switch A.1.1 with scope set to 2 consists of all the switches in the peer group A. Calls originating within the neighborhood of the mobile will be set up optimally on the shortest paths since all the switches on the paths have correct reachability information.

Calls originating outside the neighborhood of the mobile proceed toward the home of the mobile since the reachability information in the switches outside the neighborhood indicate that the mobile is at its home (default reachability information). In this case, the following two scenarios are possible: i) called mobile is within the neighborhood of its home, or ii) mobile is outside the neighborhood of its home. If the called mobile is within
the neighborhood of its home, the default reachability information (indicating that the mobile is in its home neighborhood) in switches outside the neighborhood of the mobile is the correct reachability information. Since the switches within the neighborhood as well as outside the neighborhood have correct reachability information, calls get routed via optimal paths. Therefore, if the call originates within the neighborhood of the mobile, as shown in Figure 31 for the called mobile located in position I, or if the mobile is at its home location (position II in Figure 31), the connection is routed on the shortest path.

To handle calls originating at switches outside this neighborhood when the mobile is not in its home neighborhood, a forwarding pointer is set at the home switch of the mobile indicating the current location of the mobile. On call arrival, originating switches set up connections as per the data indicated in their reachability information. Since the reachability information in switches outside the neighborhood of the mobile indicate that the mobile is reachable through its home, the call will be routed toward the home of the called mobile, at which point, it needs to be somehow rerouted to the current location of the mobile. In Chapter 4, we proposed forwarding the call from the home switch to the current location. This is illustrated in Figure 31 for the case when the called mobile is in position III. Other options for handling this situation are proposed in [21, 23]. Ayyagari et
al. [21] proposed using the crankback feature of PNNI signaling (as shown in Figure 31 for the case when the called mobile is located at position IV), which allows for the call setup procedure to be cranked back (to B in Figure 31) and then rerouted. In [21], the authors note that typically crankback will be executed at a local level, which means that the overall end-to-end path taken by the connection could be “suboptimal.” In this chapter, we propose a one-phase crankback connection setup scheme in which the connection is cranked back to a “right” node and rerouted to the current location of the mobile. In other words, an optimal route can be achieved using the crankback approach in one phase by cranking back to the “right” node. The principle used to determine the “right node” can also be used to provide a solution to the route optimization problem for the second phase of the two phase methods (route optimization after call setup). This is described in Chapter 8.

5.1 Problem statement

In this section, we state the route optimization problem (also referred to as the base rerouting problem) and demonstrate how the rerouting problem of the one-phase crankback scheme can be mapped to the base rerouting problem. The solution to this problem is presented in section 5.2. The base rerouting problem is stated below.

Consider a connection between two end points f and o (where f is used to denote the “far-end” and o is used to denote the “old point”). The problem is to reroute the connection f-o to create a connection from node f to node t (target node). This problem is referred to as base rerouting problem or the base route optimization problem and is illustrated in Figure 32. Next, we demonstrate how the route optimization problems created in the three location management schemes can be mapped to this base problem.
The crankback location management scheme can be converted from a two-phase scheme to a one-phase scheme if the call is cranked back to an optimal crossover node. Such a scheme can be viewed as one in which route optimization is performed during call setup. In this case, the same problem illustrated in Figure 32 needs to be solved wherein node $f$ corresponds to the calling party's switch $c$, node $o$ corresponds to the home switch $h$ and node $t$ corresponds to the visiting location of the mobile. The optimal crossover node $p$ is the switch to which the call is cranked back during call setup to the mobile.

5.2 Crossover node determination

In this section, we propose a solution to the base rerouting problem. In Section 5.3, we apply this algorithm and present the one-phase crankback solution to the location management problem.

There are three steps involved in solving the base rerouting problem, (i) determining a crossover node $p$, (ii) establishing a new segment $p$ to $t$, and (iii) releasing the old segment from $p$ to $o$. Given that standard ATM signaling procedures can be used for the latter two steps, only the crossover node determination step needs a solution. Thus, we only present a solution to the problem of determining a crossover node $p$ for the base rerouting problem illustrated in Figure 32. Ideally, the crossover node $p$ should be such that the path
\( f - p - t \) is the shortest path between \( f \) and \( t \), while at the same time, there should be a maximal overlap in the paths of the old connection \( t - o \) and the new connection \( f - t \). We refer to such a crossover node as the \textit{minimal crossover node} for the two paths \( f - o \) and \( f - t \).

While finding the minimal crossover node has both the advantages of minimizing the resources required by the new connection and minimizing the new segment setup/old segment release overhead, there are certain constraints in the PNNI standard (in its current form) that do not allow for the determination of this node. In the PNNI standard, there is currently

\begin{enumerate}
  \item[(i)] no requirement mandating that all nodes retain the hierarchical path of a connection after it is established, and
  \item[(ii)] a restriction that only nodes that created a DTL (Designated Transit List) can change that DTL.
\end{enumerate}

The minimal crossover node determined under these constraints is defined as the \textit{optimal crossover node} since it is a node \( p \) such that \( f - p - t \) is an optimal path, and as much overlap as is possible between the old and new paths is achieved within the constraints of the PNNI standard.

Thus, we have defined two types of crossover nodes, the optimal crossover node and the minimal crossover node. We present our solution for determining the optimal crossover node in Section 5.2.1. In Section 5.2.2, we present a solution for determining the minimal crossover node under the assumption that it is possible to make minor modifications to the PNNI standard to allow for the relaxation of these constraints. The relationship between the optimal and minimal crossover node determination procedures is
illustrated in Figure 33. It shows that if either hierarchical path information is not available or only the nodes that created the DTL can change the DTL, the procedure terminates after determining the optimal crossover node. Otherwise, the procedure to determine the minimal crossover node is also executed. Figure 33 also shows the details of the two procedures.

5.2.1 Optimal crossover node determination

In this section, procedure to determine the optimal crossover node is described. Two aspects of the PNNI based networks that are fundamental to this procedure are: i) every node has only the summarized information regarding the topology of the network (nodes outside of a peer group do not have information regarding internal structure of that peer group), and ii) connections are routed using hierarchical source routing. The basis of this procedure is that, if a connection is routed using source routing and the nodes outside a peer group do not know the internal details of the peer group, then the connection to any node within the peer group will follow the same route until the first node (ingress border node) of that peer group. The procedure to determine an optimal crossover node is described next. The notation used in this description is defined below.

Definition: Ancestors-are-siblings peer group, denoted by $P_{ij}$, of two nodes $i$ and $j$ is the lowest peer group in the hierarchy at which ancestors of both nodes belong to the same peer group. For example, in Figure 9, ancestors-are-siblings peer group of nodes A.1.1 and A.2.1 is A.

Definition: The ancestors-are-siblings level of two nodes $i$ and $j$, denoted by $a_{ij}$, is the level at which the ancestors of the two nodes $i$ and $j$ belong to the same peer group, i.e.,
the level of their ancestors-are-siblings peer group. For example, in Figure 9, the ancestors-are-siblings level of A.1.1 and A.2.1 is 2, since the level of peer group A is 2.

The location of the optimal crossover node depends on the exact relation between the ancestors-are-siblings levels of the nodes $f$, $o$, and $t$.

If $a_{of} < a_{ot}$ (Case I of Figure 34), the scenario is one in which the target node $(t)$ is closer\(^2\) to the old node $(o)$ than the far-end $(f)$. In this case, the optimal crossover node is the ingress border node of the ancestors-are-siblings peer group of the old and the target nodes ($P_{to}$) through which the connection setup procedure for the $f-o$ connection entered peer group $P_{to}$. Since nodes contain only the summarized topology and reachability information, nodes outside a peer group cannot distinguish between exact locations of the different nodes reachable through that peer group. Therefore, a connection from the far-end to either the target or the old node will follow the same route until it reaches the ingress node $(p)$ of the ancestors-are-siblings peer group of the old and the target nodes ($P_{to}$). Once the connection setup arrives at the ingress node $p$, it computes the source route to the exact location. Therefore, at $p$, the connection to the target node $(t)$

---

1. The numbering of levels in the hierarchy is such that lower level peer groups have higher level values assigned to them, i.e., the level numbers increase from top to bottom in the hierarchy.
2. Closeness is defined with respect to the node relationships defined by the peer groups. Note that closeness according to this criterion is not necessarily close with respect to number of hops or geographical distance.
Figure 33: Determination of crossover nodes
would have a different source route than a connection to the old node \((o)\). Hence, an optimal path can be obtain by adjusting the existing connection from the crossover switch \(P\).

Example: Consider the scenario in which there is an existing connection between A.1.1 (far-end) and B.3.1 (old node), and a new connection between A.1.1 and B.2.1 (target node) is desired (see Figure 35). In this case \(a_{of}, a_{ot}\) and \(P_{to}\) are 1, 2, and B respectively. Since \(a_{of} \lt a_{ot}\), the crossover node is the ingress border node of peer group B, which is B.1.1. Therefore, the optimal path between A.1.1 and B.2.1 is achieved by setting up a new segment from B.1.1 to B.2.1, as shown in Figure 35.

If \(a_{of} > a_{ot}\) (Case II a of Figure 34), the scenario is one in which the old node \((o)\) is closer to the far-end \((f)\) than to the target node \((t)\). In this situation there may not be any segment common between the existing connection and the new desired connection. Therefore, the optimal crossover node is the far-end \((f)\). The new path is optimal, since the entire path between the far-end and the target node is being set up.

If \(a_{of} = a_{ot}\) (Case II b of Figure 34), the scenario is one in which the target node is closer \((a_{ft} > a_{fo}\) to the far-end than to the old node or equidistant \((a_{ft} = a_{fo}\) to the far-end and the old node. In this situation too there may not be any segment common between
the existing connection and the new desired connection. The optimal crossover node is the far-end (f). Figure 33 shows the optimal crossover node determination procedure described above.

The significance of the optimal crossover node is three fold. First, if a new connection is established from this node to the target node, then the resulting connection path between the far-end and the target node will be optimal. Second, the optimal crossover node is determined without the use of the hierarchical path information of the existing connection. Third, the current PNNI standards stipulate that only the node that created a hierarchical DTL can change that DTL. This optimal crossover node is the node that created the DTL that is to be changed. Therefore, rerouting of the connection at this optimal crossover node conforms to the stipulation. Next, the procedure to determine the minimal crossover node is described.

5.2.2 Minimal crossover node determination

As explained earlier, it is possible for a part of the existing segment from the optimal crossover node to the old node to be in common with the new segment from the crossover node to the target node. For example, in Figure 35, while B.1.1 is the optimal crossover node, B.2.3 is the minimal crossover node. In this section, we describe the procedure to determine the minimal crossover point. Before presenting the details of this procedure, the main principle of this procedure is illustrated by an example.

Consider the scenario in which the far-end switch, old switch and the target switch are in the same level-L peer group as shown in Figure 36. In this case, the optimal crossover node is the far-end switch since \( a_{of} = a_{ot} \) (Case IIb of Figure 34). To determine the minimal crossover node initiating the procedure at node o as specified in the problem
definition (see Section 5.1), node $o$ needs to first compute the shortest path from node $f$ to node $t$ and then compare it with the old path $f-o$. Since the PNNI routing protocol is a link-state routing scheme (and not a "distance-vector" scheme), this operation can be easily done. Nodes $f$, $o$ and $t$ maintain the same topological information about the peer group allowing a node (in this case node $o$) to determine the shortest path between two other nodes (in this case, nodes $f$ and $t$).

For the scenario shown in Figure 36, node $o$ computes the shortest path between $f$ and $t$ to be $\{f, c_1, c_2, c_3, c_4, n_1, n_2, t\}$. It knows the path of the old connection as being $\{f, c_1, c_2, c_3, c_4, m_1, m_2, m_3, o\}$. By comparing the old and the new paths, it determines the minimal crossover node to be the point of intersection of the two paths closest to the target node ($t$). For the scenario shown in Figure 36, the minimal crossover node is $c_4$.

Henceforth, we refer to the above-described procedure as the new path determination/new path-old path comparison routine, and is applied on paths through a single peer group.

Next, consider how this minimal crossover node determination procedure can be extended to the more general case in which nodes $f$, $o$ and $t$ belong to different level-$L$ peer groups. The extensions are based on the following observations:
(i) the end-to-end hierarchical route stored at a node on the path consists of a stack of
DTLs, with each DTL representing a path through a peer group at a level of the hierarchy
as viewed by that node, and

(ii) there is always a peer group at which the ancestors of the three nodes $f$, $o$ and $t$
belong to the same peer group.

Given these observations, we begin the minimal crossover node determination by
having node $o$ exercise the *new path determination/new path-old path comparison routine*
for the paths through the peer group at which ancestors of nodes $f$, $o$ and $t$ are siblings.
This leads to the determination of a *minimal crossover peer group* rather than the
identification of the actual minimal crossover node. The minimal crossover peer group
contains the minimal crossover node, but given the hierarchical structure of the network,
node $o$ has no detailed information about the inside structure of this minimal crossover
peer group. This makes it necessary to execute the next step of the procedure at a node that
lies inside this minimal crossover peer group.

For this purpose, the old connection is traced from node $o$ until the egress border node
of the identified minimal crossover peer group on the old connection is reached. This node
can then perform the *new path determination/new path-old path comparison routine* for
the paths through the minimal crossover peer group to determine the next lower-level
minimal crossover peer group. This procedure is applied recursively until a physical ATM
switch is identified as the minimal crossover node.
**Details of the algorithm:** Figure 33: shows the details of the minimal crossover node

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$LGN_t^{phy}$</td>
<td>Logical group node ancestor of the physical ATM switch $phy$ at level $l$. For example, in Figure 9, the $LGN_1^{A1.1}$ is A</td>
</tr>
<tr>
<td>$x_l$</td>
<td>Minimal crossover peergroup of level $l$. $x_L$ (value of $l$ is $L$) is the minimal crossover node.</td>
</tr>
<tr>
<td>$e_l$</td>
<td>Egress border node of the peergroup $x_{l-1}$ on the old path. This node executes the new path determination/new path-old path comparison routine at level $l$, i.e., computes the minimal crossover peergroup at level $l$ ($x_l$). The value of $e_{old}$ is $o$ (old node).</td>
</tr>
<tr>
<td>$E_{old}^l$</td>
<td>Logical group node ancestor of $e_l$ at level $l$, i.e., $LGN_l^{e_l}$.</td>
</tr>
<tr>
<td>$i_l$</td>
<td>Ingress border node of $x_{l-1}$ on the old path. If $l$ is equal to $a_{old}$, $i_l$ is the optimal crossover node $p$.</td>
</tr>
<tr>
<td>$I_l$</td>
<td>Logical group node ancestor of $i_l$ at level $l$, i.e., $LGN_l^{i_l}$.</td>
</tr>
<tr>
<td>$E_{new}^l$</td>
<td>Logical group node adjacent to $x_{l-1}$ in the new path $P_{l-1}^{new}$. If there is no node adjacent to $x_{l-1}$ in the new path or the value of $l$ is $a_{old}$, $E_{new}^l$ is the logical group node ancestor of $t$ at level $l$, i.e., $LGN_l^t$.</td>
</tr>
<tr>
<td>$P_{old}^l$</td>
<td>Hierarchical path of the existing connection segment between the logical group nodes $I_l$ and $E_{old}^l$. This path is specified in terms of the logical groups nodes of level $l$.</td>
</tr>
<tr>
<td>$P_{new}^l$</td>
<td>Shortest path between the logical group nodes $I_l$ and $E_{new}^l$. This path is specified in terms of the logical groups nodes of level $l$.</td>
</tr>
</tbody>
</table>

Table 1: Notation

determination procedure. The notation used in the description of this algorithm is listed in Table 1.
First, the optimal crossover node is determined using the procedures described in section 5.2.1. Next, in order to determine the minimal crossover node, the procedure of identifying the minimal crossover peergroup at level $l$ is repeated iteratively at levels $a_{ot}$ through $L$. The minimal crossover peergroup at level $l$ ($x_l$) is identified by executing the new path determination/old path comparison routine at the egress border node of the minimal crossover peergroup at level $l-1$ on the existing path ($e_l$). Initially, i.e., for level $a_{ot}$, $e_l$ is node $o$. The node $e_l$ determines $x_l$ by computing the following parameters (see flow chart given in Figure 33):

19. $E_l^{new}$: Logical group node adjacent to $x_{l-1}$ in the new path $P_{l-1}^{new}$. This logical group node is used as the destination node while computing the path of the new segment ($P_l^{new}$) through this peergroup ($x_{l-1}$). If there is no node adjacent to $x_{l-1}$ in the new path or the value of $l$ is $a_{ot}$, $E_l^{new}$ is the logical group node ancestor of $t$ at level $l$.

i.e., $LGN_l^t$ (see Figure 37).

20. $I_l$: Logical group node ancestor at level $l$ of the ingress border node of $x_{l-1}$ on the old path. If $l$ is equal to $a_{ot}$, $I_l$ is $LGN_l^o$ (see Figure 37:). $I_l$ is the first LGN (Logical Group Node) of level $l$ in the peergroup $x_{l-1}$ that is common to both the old and the new segment. Therefore, in the worst case $I_l$ is equal to $x_l$.

21. $E_l^{old}$: Logical group node ancestor of $e_l$ at level $l$, i.e., $LGN_l^{e_l}$. This is the last LGN of level $l$ in the peergroup $x_{l-1}$ on the existing path.
22. $P_{l}^{new}$: Path of the new connection segment through the peergroup $x_{l-1}$ is determined by computing the shortest path between the logical group nodes $I_l$ and $E_l^{new}$. Node $e_l$ has the summarized topology of the peer group $r_{l-1}$ and can hence compute the shortest path between these logical group nodes. In the first iteration, i.e., $i = a_{or}$, node $o$ computes the shortest path between the ancestors of nodes $p$ and $t$ at level $a_{or}$, i.e., nodes $LGN_{a_{or}}^p$ and $LGN_{a_{or}}^t$ (see Figure 37).

23. $P_{l}^{old}$: Path of the existing connection segment through the peergroup $x_{l-1}$, i.e., the existing path between the LGNs $I_l$ and $E_l^{old}$. This information is extracted from the hierarchical source route of the existing connection.
24. $x_i$: The minimal crossover peergroup at level $l$ is computed by determining the intersection between the new path ($P_{l}^{\text{new}}$) and the old path ($P_{l}^{\text{old}}$). In other words, the new and old connections have a common path up to the peer group $x_i$. Therefore, the minimal crossover node is a node in the peer group $x_i$. In the case of level $a_{ol}$, $x_{a_{ol}}$ is determined by comparing the two paths $P_{a_{ol}}^{\text{new}}$ and $P_{a_{ol}}^{\text{old}}$ as shown in Figure 37.

After computing $x_i$, the connection is traced from node $e_i$ until it reaches the egress border node of $x_i$ ($e_{i+1}$). If value of $l$ is equal to $L$, $x_i$ is the minimal crossover node (see Figure 37:). Otherwise, the new path determination/old path-new path comparison is executed by the node $e_{i+1}$ to determine the minimal crossover peergroup $x_{i+1}$. This iterative procedure is illustrated in Figure 37:. On identification of the minimal crossover node, new segment setup is initiated from this physical crossover node to the target node. Note that, by this procedure the minimal crossover node is determined in less than $L + 1 - \text{Max} \{a_{ol}, a_{fo}, a_{fl}\}$ iterations. By this selection of the minimal crossover node, the length of the new segment to be set up is minimized, thereby reducing the associated setup delays and the signaling load. However, there is processing overhead associated with this procedure. Heuristics can be used to terminate the search process at some level ($l$) instead of repeating the procedure to compute the minimal crossover node at the egress border node of $x_{i+1}$, the ingress border node of $x_{i+1}$ ($i_{i+1}$) can be selected as the crossover node.
Example: Consider the scenario in which an existing connection between A.1.1 (far-end) and B.3.1 (old node) is to be rerouted to form a connection between A.1.1 and A.1.3 (target node) (see Figure 38:). In this case \( a_{of}, a_{ot} \) and \( P_{lo} \) are 1, 1, and A, respectively.

Since \( a_{of} = a_{ot} \), the optimal crossover node is the calling switch (A.1.1). We determine the minimal crossover node following the steps shown in Figure 33:.

At B.3.1, the values of \( E_l^{old}, E_l^{new}, I_l, \) and \( I \) are \{B\}, \{A\}, \{A\} and 1, respectively. It determines the values of \( P_l^{new}, P_l^{old}, x_l, \) and \( n_l \) to be \{A\}, \{A, B\}, \{A\} and \{\}, respectively. The connection is traced from the old node to the egress border node of peer group A, A.2.2.

At A.2.2, the values of \( E_l^{old}, E_l^{new}, I_l, \) and \( I \) are \{A.2\}, \{A.1\}, \{A.1\} and 2, respectively. It determines the values of \( P_l^{new}, P_l^{old}, x_l, \) and \( n_l \) to be \{A.1\}, \{A.1, A.2\}, \{A.1\} and \{\}, respectively. The connection is then traced till the egress border node of peer group A.1, A.1.2.

At A.1.2, the values of \( E_l^{old}, E_l^{new}, I_l, \) and \( I \) are \{A.1.2\}, \{A.1.3\}, \{A.1.1\} and 3, respectively. It determines the values of \( P_l^{new}, P_l^{old}, x_l, \) and \( n_l \) to be \{A.1.1, A.1.2, A.1.3\}, \{A.1.1, A.1.2\}, \{A.1.2\} and \{\}, respectively. Since \( I = L, x_l (A.1.2) \) is the minimal crossover node. Therefore, A.1.2 is the minimal crossover node (see Figure 38:).
5.3 One-phase crankback scheme

In this section, we present a one-phase mobile connection setup scheme in which an optimal connection path is achieved while setting up the connection to a mobile. The crankback location management scheme classified as a two-phase scheme can be modified to a one-phase scheme if the call is cranked back to the optimal or minimal crossover node and then rerouted to the current location of the mobile.

In order to reroute the connection, first, the call is cranked back till the optimal or minimal crossover node, and then routed to the visiting location of the mobile. The optimal crossover node is determined using the procedure described in Section 5.2.1 by considering \( f \) as the calling party's switch, \( o \) as the home switch of the called mobile, and \( t \) as the current switch of the called mobile. The first condition for being able to determine the minimal crossover node described in Section 5.2 (and illustrated in Figure 33:) is that the hierarchical path be available. This condition is met in this one-phase crankback procedure since hierarchical routes are available during call setup [3]. If the second condition described in Section 5.2 for minimal crossover node determination, i.e., all nodes being permitted to change DTLs is met, then the connection is cranked back to the minimal crossover node. In this case, the procedure specified by flow chart in Figure 33: is used to crankback to the minimal crossover node. At the minimal crossover node, the connection is rerouted to the current location of the mobile. This rerouting procedure results in an optimal connection, thereby achieving a one-phase mobile connection setup procedure.

In order to facilitate the easy integration of wireless ATM into the existing and planned fixed ATM networks, an approach of incremental deployment of mobile-enhanced switches needs to be taken. Therefore, it is important to consider isolating mobility-related
features to mobile-enhanced switches. Examples of mobility-related features are support for crankback to reroute calls to mobiles, support for transport of mobility related signaling messages along the existing path, support for buffering during route optimization, etc. Mobile PNNI method during connection setup is easily implementable. In the mobile PNNI method only the home switch of the mobile needs to be mobile-enhanced. In the one-phase scheme, additionally, the crossover point has to be a mobile-enhanced switch.

Next, we address the question of how best to support one phase crankback scheme with limited deployment of mobile-enhanced switches. While determining a crossover point, choosing a node on the existing path between the far-end and the optimal crossover point also results in an optimal path. In a network with limited deployment of mobile-enhanced switches, if the optimal crossover point is not a mobile-enhanced switch, a mobile-enhanced switch that occurs between the optimal crossover node and the far-end is selected. If no such mobile-enhanced node is available, a suboptimal path results by choosing a node between the optimal crossover point and the old location. Another issue to be addressed is, how does the home switch (or other switches) know about the availability of mobile-enhanced switches on the path (before executing the route optimization procedure). Possible approaches are: i) To use the feature of sending reachability updates to propagate information regarding the peer groups with mobile-enhanced nodes. Since the hierarchical route of the connection is known, information regarding the mobile-enhanced peer groups can be used to determine a list of peer groups on the path that are mobile-enhanced. In this case, the border nodes of the mobile-enhanced peer groups have to be mobile-enhanced switches. This approach is of limited use, since only the hierarchical DTLs are given in the SETUP message and the home
switch may not have enough information to recognize that the existing path is through a mobile-enhanced peer group of a lower level. ii) Maintain a list of mobile-enhanced switches on the existing path. This list can be obtained during call setup, by having a list (DTL) in the SETUP message to which the mobile-enhanced switches on the path add their address.

If a certain border node is not mobile enhanced, it is assumed that the connection is further cranked back and the first mobile enhanced element on the path performs the minimal crossover peer group determination. In the scenario in which the crossover node is not mobile enhanced the connection cranked back further than the minimal/optimal crossover node and rerouted. It may be necessary to reroute the connections from a node that lies between the minimal/optimal crossover node and the old node (home switch), if that node is the first mobile enhanced node on the path. It is also possible to recognize this node either by setting a bit in the connection setup message or by having an additional information element to have this information.

There are two alternatives to computing the shortest paths. 1) Store all the shortest paths (possibly multiple per node pair) for all node pairs, compare the \( f-o \) path with the set of shortest paths for a node pair \( f-t \), and find the \( f-t \) path with the longest match with the \( f-o \) path. This alternative requires large storage space but has the advantage of low latency. Note that in a hierarchical network paths stored are for logical node pairs of the same level and not between all node pairs. ii) Run modified Dijkstra’s algorithm on-the-fly. This approach requires minimal extra storage relative to Dijkstra’s algorithm, but has the disadvantage of larger processing delay than former alternative.
5.4 Performance analysis

In this section, a comparative performance analysis of one-phase and two-phase methods for location management is performed. Two variations are considered while analyzing the performance of the one-phase scheme. In the first variation, referred to as the one-phase scheme (optimal), the call is rerouted from the optimal crossover node (see Section 5.2.1) to the current location of the mobile. In the second variation, referred to as the one-phase scheme (minimal), the call is rerouted from the minimal crossover node (see Section 5.2.2).

As discussed earlier, there is a trade-off between call setup delay and resource utilization in one- and two-phase methods. Using a simple analytic model, we analyze this trade-off between resource utilization and connection setup delays for these methods. Later, we present details of simulation modeling and the results obtained from simulation of three different topologies. These results help in validating the trends observed from the analytical results and also give accurate estimates of performance of the one- and two-phase methods for the three topologies. In this performance analysis, the second phase (route optimization) of the two-phase schemes is not considered since it does not affect the parameters used to compare these methods (amount of resources allocated to a connection during call setup and the call setup delay) significantly. This issue is discussed further in Section 5.4.1.3.

Other considerations to compare the performance of these schemes are signaling load and processing requirements to handle the signaling procedures associated with these schemes. Discussion on qualitative considerations, such as the simplicity of signaling procedures, impact on standards, networking with non-mobile-enhanced ATM switches (backwards compatibility), etc., is also presented.
5.4.1 Analytical model

In this section, we present the details of the analytic model used to analyze the performance of the one-phase crankback scheme and two-phase schemes. We first define the notation used in this analysis, and then describe a basic property of PNNI standards based hierarchical networks that is repeatedly used in the analysis. In Section 5.4.1.3 the modeling details are presented. Finally, the numerical results obtained are presented in Section 5.4.1.4.

5.4.1.1 Notation

Table 2 lists the symbols used in this analysis section, along with their definitions.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>Number of peer group levels in the network; level 1 is the topmost level and level $L+1$ represents individual switches in the network.</td>
</tr>
<tr>
<td>$a_{ij}$</td>
<td>The ancestors-are-siblings level is the level at which the ancestors of the two nodes $i$ and $j$ belong to the same peer group (see Section 5.2).</td>
</tr>
<tr>
<td>$h, v$</td>
<td>Subscripts used to represent the home switch, visiting switch of the mobile, the calling party and the crossover node, respectively.</td>
</tr>
<tr>
<td>$c, p$</td>
<td>Subscripts used to represent the home switch, visiting switch of the mobile, the calling party and the crossover node, respectively.</td>
</tr>
<tr>
<td>$S$</td>
<td>Scope indicates a stopping distance for reachability update propagations in the mobile PNNI scheme; if scope is $S$, reachability sent by a node $i$ does not propagate to any node $j$ for which $a_{ij} &lt; S$.</td>
</tr>
<tr>
<td>$N_i$</td>
<td>For a given node $i$, all nodes $j$ such that $a_{ij} \geq S$ is defined as the neighborhood of node $i$.</td>
</tr>
<tr>
<td>$m_i + 1$</td>
<td>Number of peer nodes in a peer group of level $i$ for $i = 1, 2, ..., L$; $m_i = 0$ for $i &gt; L$.</td>
</tr>
<tr>
<td>$p_i$</td>
<td>Average length of the “shortest-path” between nodes of a peer group at level $i$.</td>
</tr>
<tr>
<td>$D_{ij}$</td>
<td>Distance, in terms of the number of nodes on the route, from node $i$ to node $j$.</td>
</tr>
<tr>
<td>$S_d$</td>
<td>Delay incurred at a switch during connection setup.</td>
</tr>
</tbody>
</table>

Table 2: Notation
5.4.1.2 Property

For three nodes $x$, $y$ and $z$, the following relations hold between their ancestors-are-sibling levels:

\[
\text{Case I: } a_{xy} < a_{xz} \Rightarrow a_{yz} = a_{xy} \quad (\text{EQ 1})
\]

\[
\text{Case II: } a_{xy} = a_{xz} \Rightarrow a_{yz} \geq a_{xz} \quad (\text{EQ 2})
\]

The proof of this property can be inferred from Figure 39. If $a_{xy} < a_{xz}$ the arrangement of nodes is shown in Case I of Figure 39, from which it is clear that $a_{yz} = a_{xy}$. A similar argument extends for Case II. Also, note that since $a_{ij} = a_{ji}$, we use these terms interchangeably.

![Figure 39: Relative positions of $x$, $y$, and $z$](image-url)
5.4.1.3 Location management schemes

In this section, we analyze the performance of one- and two-phase location management schemes. In particular, we analyze the proposed one-phase method (optimal and minimal), and the two-phase location management scheme in which the first phase consists of a quick forwarding of the connection to the mobile by the home switch [34] and the second phase is the proposed route optimization scheme. Measures of analysis include the connection setup delays incurred in these schemes and the amount of network resources required. Since the bandwidth requirement of a connection is independent of the scheme used, we use the number of hops in the connection to estimate the network resource allocation required in each scheme.

In this two-phase connection setup scheme, the connection is set up with each switch believing in its reachability information. The connection setup delay, $\text{Delay}_{\text{two}}$, in this method is given in (EQ 3). The first case, $a_{cv} \geq S$, implies that the calling party is within the neighborhood (see Table 2) of the called mobile, where $S$, defined in Table 2, corresponds to the level upto which the reachability information is propagated. Therefore, the connection setup delay can be approximated by $S_d D_{cv}$, where $S_d$ denotes the delay at each switch on the path and $D_{cv}$ is the number of nodes on route from node $c$ (calling party) to node $v$ (visiting location of the mobile).

The second case of (EQ 3) corresponds to the scenario when a mobile is in its home neighborhood. In this situation the entire network has the correct reachability information about the mobile and hence the call is routed directly to the mobile. All the nodes within the neighborhood of the mobile have correct reachability information as a result of localized reachability updates. The summarized reachability information in the nodes outside the neighborhood of the mobile (default information) indicates that the mobile is
in its home neighborhood. This information is sufficient to route a call directly to the mobile, since in PNNI networks hierarchical source routing is used to route the connection.

The third case corresponds to the scenario in which the call is first routed to the home of the called mobile \( h \) and is then forwarded by the home to the current location of the mobile \( v \). Therefore, the connection setup delay comprises of the delay in setting up the connection from the calling party switch \( c \) to the home of the called mobile and the setup delay for the connection between the home and the visiting location of the mobile.

\[
\text{Delay}_{\text{two}} = \begin{cases} 
S_d d_{cv} & a_{cv} \geq S \quad \text{independent of } a_{hv} \\
S_d d_{cv} & a_{hv} \geq S \quad \text{independent of } a_{cv} \\
S_d (D_{ch} + D_{hv}) & a_{hv} < S \quad \text{and} \quad a_{cv} < S
\end{cases}
\]  

Next, we provide a method for estimating \( D_{ij} \), and derive the expression for estimating the average connection setup delay. The distance \( D_{ij} \) between nodes \( i \) and \( j \) is approximated as

\[
D_{ij} = \prod_{k=a_{ij}}^{L} \frac{p_k}{p_k}
\]

where \( p_k \) is the average (among all node pairs) length of the "shortest-path" between nodes of a peer group at level \( k \). For the worst case performance, the maximum length (among all node pairs) of the "shortest-path" can be taken. By this definition, the distance \( D_{ij} \) between nodes \( i \) and \( j \) depends on the value of their ancestors-are-siblings level \( a_{ij} \). From the property described in (EQ 1)-(EQ 2), the relationship between the ancestors-are-siblings levels of different nodes is known. Using (EQ 1) through (EQ 4), an expression to
estimate the average connection setup delay, $A_{vgD_{two}}$, is obtained in (EQ 5).

$$
A_{vgD_{two}} = \sum_{i=S}^{L+1} P(a_{cv} = i)S_dD_{cv} + \sum_{i=1}^{S-1} P(a_{cv} = i) \left\{ \sum_{j=i+1}^{S} P(a_{hv} = j)S_d(D_{hv} + D_{ch}) \right\} + \sum_{j=i}^{L+1} \left\{ \sum_{i=1}^{S} P(a_{ch} = j)S_d(D_{cv} + D_{hv}) \right\}
$$

(EQ 5)

In the two-phase scheme being analyzed, i.e., call forwarding scheme, since the connection is not "backtracked" at any point, the number of hops in the connection can be estimated from equations (EQ 3) and (EQ 5) by setting the value of $S_d$ to be 1. The expression to estimate the average number of hops in a connection, $A_{vgR_{two}}$, is given in (EQ 6). We expect that the second phase, i.e., route optimization phase, is performed soon after the connection setup is completed. Hence, the additional network resources needed for extra hops in a connection are only needed for a short period of time. Note that in this analysis the second phase (route optimization) of the two-phase method is not considered. Second phase is not considered because of the following two reasons. i) It does not affect the amount of resources allocated to a connection during connection setup. However, it should be noted that the number of hops in a connection after the second phase is executed is equal to the number of hops in the connection set up by the one-phase scheme. ii) It does not affect the connection setup delay significantly.

---

3. Increase in signalling load due to the second phase results in a higher call setup delays due to increase is queuing delays.
\[
\text{Avg} R_{\text{tw}} = \sum_{i=S}^{L+1} P(a_{cv} = i) D_{cv} + \sum_{i=1}^{S-1} P(a_{cv} = i) \left\{ P(a_{hv} \geq S) D_{cv} \right\} 
+ \sum_{j=1}^{S} P(a_{hv} = j) (D_{cv} + D_{hv}) + \sum_{j=i+1}^{S} P(a_{hv} = j) (D_{cv} + D_{hv}) + \sum_{j=i}^{S-1} P(a_{hv} = j) \left\{ \sum_{j=i}^{S} P \left( \frac{a_{ch} = j}{a_{cv} = a_{hv} = i} \right) (D_{ch} + D_{hv}) \right\} 
\]

In the one-phase connection setup scheme (optimal), like the two-phase scheme, the connection is set up with each switch “believing” in its reachability information. But on connection arrival to the home switch of the mobile, instead of forwarding the connection, the connection is cranked back to an optimal crossover point and rerouted to the mobile. The connection setup delay, \( Delay_{\text{one}} \), in this method is given in (EQ 7). The first two cases of this equation are the same as those of (EQ 3). The third case corresponds to the scenario in which the connection proceeds to the home of the called mobile \( (S_d D_{ch}) \), then cranked back from the home to the crossover node \( (C_d D_{hp}) \), where \( C_d \) denotes the delay at each node to perform crankback), and then rerouted from the crossover node \( (p) \) to the current location of the mobile \( (S_d D_{pv}) \).

\[
Delay_{\text{one}} = \left\{ \begin{array}{l}
S_d D_{cv} \\
S_d D_{cv} \\
S_d (D_{ch} + D_{pv}) + C_d (D_{hp})
\end{array} \right\} \text{ independent of } a_{hv} \text{ and } a_{cv} \leq S
\]  

Using the approximation given in (EQ 4) (the distance \( D_{ij} \) between nodes \( i \) and \( j \) depends on the value of their \( \text{ancestors-are-siblings level } a_{ij} \)), the expression to estimate the average call setup delay (given in (EQ 8)) can be obtained from (EQ 7).
Applying the properties described in (EQ 1)-(EQ 2) (which give the relationship between the ancestors-are-siblings level of different nodes), (EQ 8), can be rewritten as below:

\[
AvgD_{one} = \sum_{i=S}^{L+1} P(a_{cv} = i) S_d D_{cv} + \sum_{i=1}^{S-1} P(a_{cv} = i) \left\{ P(a_{hv} \geq S) S_d D_{cv} \right\}_{i=S}
\]

\[
+ \sum_{j=1}^{S} P(a_{hv} = j) \left( S_d (D_{ch} + D_{pv}) + C_d D_{hp} \right)
\]

\[
+ \sum_{j=i+1}^{S} P(a_{hv} = j) \left( S_d (D_{ch} + D_{pv}) + C_d D_{hp} \right)
\]

\[
+ P(a_{hv} = i) \left\{ \sum_{j=i}^{S-1} P\left( a_{ch} = j, a_{cv} = a_{hv} = i \right) \left( S_d (D_{ch} + D_{pv}) + C_d D_{hp} \right) \right\}_{i=S}
\]
The third term of the equation corresponds to Case IIb of Figure 40: where $c$ and $v$ are equidistant from the home of the mobile ($a_{ch} = a_{hv}$). The fourth and fifth terms correspond to Cases I and IIa of Figure 40:, respectively.

Next, we consider the variation of the one-phase scheme (minimal) in which the call is cranked back to the minimal crossover point. The average delay while cranking back to the minimal crossover point is expected to be lower than while cranking back to the optimal crossover point. In order to compute the average delay while cranking back to the

![Diagram](image)

Figure 41: Estimation of number of crankback nodes

minimal crossover node and then rerouting the call, we make the assumption that the minimal crossover peergroup (refer Section 5.2.2) at each level is equidistant from the corresponding level LGN of the ingress border node of that peergroup and the egress peer node performing the computation. In other words, the minimal crossover peergroup at
level $k$ is assumed to be equidistant from LGNs $l_k$ and $E_k^{old}$ (see Figure 41:). Assuming that there are odd number of peergroups on the path between these two nodes (i.e., $p_k$ is odd), the number of nodes involved in the crankback at level $k$, $M_k$, can be approximated by (EQ 10), where $p_k$ is the average (among all node pairs) length of the "shortest-path" between nodes of a peer group at level $k$, and $D_k$ is the average number of nodes on the route between nodes $i$ and $j$ whose $a_{ij}$ is equal to $k$, i.e., $D_k$ is equal to $D_{ij}$ (see (EQ 4)). Note that the number of nodes on the common path (including the minimal crossover node) in this peergroup of level $k$ is also equal to $M_k$.

$$M_k = \begin{cases} \frac{p_k}{2} D_k + M_k + 1 & 1 \leq k \leq L \\ 1 & k = L + 1 \end{cases}$$  \hspace{1cm} (EQ 10)

The average delay while cranking back to the minimal crossover node can be approximated by:

$$AvgD_{one}^{min} = \sum_{i=1}^{L+1} \sum_{j=1}^{S-1} P(a_{cv} = i) S_d D_{cv} + \sum_{i=1}^{S-1} P(a_{cv} = i) \left\{ P(a_{hv} \geq S) S_d D_{cv} \right\}$$

$$+ \sum_{j=1}^{S} P(a_{hv} = j) \left( S_d \left( D_{hv} + M_{a_{hv}} \right) + C_d \left( D_{hv} - M_{a_{hv}} \right) \right)$$

$$+ \sum_{j=i+1}^{S} P(a_{hv} = j) \left( S_d \left( D_{cv} + M_{a_{hv}} \right) + C_d M_{a_{hv}} \right)$$

$$+ P(a_{hv} = i) \left\{ \sum_{j=i}^{S-1} P \left( \frac{a_{ch} = j}{a_{cv} = a_{hv} = i} \right) \left( S_d \left( D_{ch} + D_{cv} - M_{a_{ch}} \right) + C_d M_{a_{ch}} \right) \right\}$$

The third term of the equation corresponds to Case IIb of Figure 42:, where $c$ and $v$ are equidistant from the home of the mobile ($a_{ch} = a_{hv}$). The fourth and fifth terms correspond to Cases I and IIa of Figure 42:, respectively. Note that in the worst case, the
optimal crossover node is the minimal crossover node. Hence, in the worst case, the maximum call setup delay in the one-phase (minimal) scheme is equal to the maximum call setup delay in the one-phase (optimal) scheme.

In the one-phase crankback scheme (in both the variations), the connection is rerouted during call setup such that the resulting connection traverses an optimal path. Hence, the number of hops in the connection depend only on the optimal path between calling party and the current location of the mobile \( (D_{cv}) \). The expression to estimate the average number of hops in a connection, \( \text{Avg}R_{one} \), is given in (EQ 12).

\[
\text{Avg}R_{one} = \sum_{i=1}^{\infty} P(a_{cv} = i) D_{cv}
\]  
(EQ 12)

5.4.1.4 Numerical results

In this section, we quantitatively compare the one-phase (route optimization during call setup) and the two-phase (route optimization after call setup) schemes. The measures of comparison include the average and maximum call setup delays and the amount of network resources required (number of hops in the connection).
**Input Data:** Values of input parameters assumed for this numerical computation are shown in Table 3 (see Table 2 for definitions of these parameters). Input parameters $S_d$ and $C_s$ are estimated from measured data [53]. The probability distributions used for calling patterns and user location are presented in the Appendix. We observe that the exact numerical results are dependent on the exact values chosen for these parameters. However, the trends observed are more or less independent of these values. Sensitivity of the comparative results to these input parameters has been studied.

The variation of connection setup delays (average and maximum delays) and the resource requirement (average and maximum number of hops) with the variation of $S$ (which defines the extent of the "neighborhood") for the one- and the two-phase methods is shown in Figure 43:. From the plot showing the variation in connection setup delay, the following four observations are made.

1. The average call setup delay in the one-phase scheme (minimal) is lower than that of the two-phase scheme.
2. The average call setup delay in the one-phase scheme (optimal) is higher than that of the two-phase scheme.
3. The maximum call setup delay in the one-phase scheme is higher than that of the two-phase scheme. Since in this section we are not considering any specific topology, the maximum delay in the one-phase (minimal) scheme is assumed to be equal to the maximum delay in the one-phase (optimal) scheme. Maximum delays in the one-phase
iv) The estimates of the average delays for both the methods are comparable, but there is significant difference in the maximum delay incurred in both these methods. For example, the average delay in the one-phase scheme (optimal) is approximately 10% higher than that of the forwarding scheme ($S = 6$), but the maximum delay in the one-phase scheme is approximately 30% higher than that of the forwarding scheme.
It should be noted that the call setup delay depends on the locations of the calling party, called mobile and the home of the called mobile, and the topology and loading conditions of the network. Therefore, changing the calling scenario (fixing the locations of the calling party, called mobile and the home of the called mobile) and the topology and loading conditions of the network, the relative performance of these schemes may vary. In the one-phase scheme (optimal), the worst case scenario occurs when the calling party, the home of the called mobile and the current location of the mobile are far-away from each other. In this scenario, the call is routed to the home of the called mobile, cranked back all the way to the calling party switch and then rerouted to the current location of the mobile leading to large delays. Such a scenario can occur in the one-phase scheme (minimal) also when there is no common segment between the old and the new paths.

The variation of $S$ has a similar effect on the call setup delays in all the schemes. At lower values of $S$ (neighborhood to which reachability data is propagated is larger), the call setup delays are lower, since most of the calls are routed directly to the mobile, instead of routing to the home and then rerouting to the mobile. When $S = 1$, since all the calls are routed directly to the mobile in both the schemes, the call setup delays are minimum. The disadvantage of having a large neighborhood (i.e., low value of $S$) is that, cost of updating reachability information regarding the location of a mobile, when a mobile moves from one location to another, is very high [35].

From the plot showing the variation in the number of hops in a connection, we observe that both the average and the maximum number of hops in connections resulting from the two-phase scheme are higher than that in the one-phase scheme. This is expected trend since the one-phase scheme results in optimal connection paths. It should be noted that both the variations of the one-phase scheme (optimal and minimal) result in a connection
with the same number of hops. The plot of average number of hops in a connection show
that significant improvements in resource utilization can be achieved by the one-phase
scheme. In the worst case (the plot for maximum number of hops), the two-phase
forwarding scheme (prior to route optimization) requires twice the amount of
communication resources as the one-phase scheme (route optimization during call setup).
In the two-phase scheme, the number of hops in connection after the route optimization
phase is equal to the number of hops in a connection set up by the one-phase scheme. At
low values of $S$ (i.e., neighborhood to which reachability data is propagated is large), the
number of hops in a connection are lower in the two-phase method, since most of the calls
are routed directly to the mobile, instead of routing to the home and then rerouting to the
mobile. Since the one-phase method results in an optimally routed connection, the number
of hops in a connection does not vary with the variation in $S$.

5.4.2 Simulation model

In this section, we describe the details of the simulation model developed to study the
performance of the one- and two-phase schemes and discuss the results obtained from the
simulations. The simulation modeling was performed using the OPNET\textsuperscript{4} modeling tool
[75].

In this simulation study, the network model consists of switches and links
interconnecting them. Each switch is modeled as a combination of a queue, message
processing module, call generator, and message transmitters and receivers as shown in
Figure 44: The queue is an infinite queue with a specifiable service rate, and accounts for
the queuing delays experienced in the switch. The message processing module contains

\textsuperscript{4}OPNET is a registered trademark of MIL 3, Inc.

100
the logic to perform topology discovery and summarization, shortest path computation, call setup, call forwarding, crankback to the optimal crossover node and crank back to the minimal crossover node. During the topology discovery phase the switch sends "hello" packets on all its outgoing links (transmitters) and learns about its neighbors (switch address) from the hello packets received from them. This information is used to set up the routing tables that are used to route call setup and crankback messages. After exchanging hello packets, nodes broadcast their link information to other nodes in their lowest level peergroups (level-L). Summarized information regarding these links is broadcast to nodes of other peergroups. From the information received through these broadcast messages, nodes compute their local view of the network.

The call generator module is used to generate calls with an exponential distribution and a specified call interarrival rate. The mobile to which this call is intended (called party) is chosen randomly. The mobility pattern is also assumed to follow a random distribution, i.e., the switch at which the called mobile is currently located is chosen randomly. On call arrival at a switch the hierarchical source route is computed base on the link weights and the summarized topology information and the call setup packet is routed
to the next switch. The message processing module also implements other functions of call setup, call forwarding [35], crankback to the optimal crossover node (see Section 5.2.1), and crankback to the minimal crossover node (see Section 5.2.2).

A desired network configuration is obtained by creating as many switch modules as there are nodes in that configuration and then interconnecting the switches according to the topology of that configuration. An interconnecting link between two switches is specified in each direction by a transmitter and a receiver. A transmitter transmits the messages it receives from the message processing module to the corresponding receiver. The receiver in turn sends the received message to the queue module of the switch.

5.4.2.1 Simulation results

In this section, we present the simulation results comparing the one-phase (route optimization during call setup) and the two-phase (route optimization after call setup) schemes. The measures of comparison include the average and maximum call setup delays, the amount of network resources required (number of hops in the connection) and the number of nodes involved in the crank back procedure (in the one-phase scheme). The results obtained in the simulation study help in validating the trends observed from the analytical results. They also provide accurate estimates of the call setup delay and the number of hops in a connection in the one- and two- phase methods for the topologies being considered.
**Input Data:** In this study we simulated three different network configurations (see Table 4). The network configurations that are simulated include: i) a network consisting of a of twenty nodes organized in single peergroup. The topology of this network configuration was generated using a random connected graph generator [76]. The connectivity (fraction of the number of edges possible in a complete graph) of the random graph generated 0.15. ii) A network consisting of 200 nodes organized as a hierarchical network of two levels. The 200 nodes are arranged into 10 level-1 peergroups each consisting of 20 switches. The topology of this network configuration is also generated using a random connected graph generator with a connectivity of 0.15. iii) A network consisting of 480 nodes organized as a three-level hierarchical network. These 480 nodes are arranged into 5 level-1 peergroups each consisting of 8 level-2 peergroups. Each of the level-2 peer groups comprise of 12 switches. The topology for the level-2 peergroup was generated using a random connected graph generator with a connectivity of 0.15.

### Table 4: Network configurations

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Level</th>
<th>No. of Nodes</th>
<th>Topology(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Configuration 1</strong></td>
<td>1</td>
<td>20</td>
<td>Random</td>
</tr>
<tr>
<td>(one level hierarchy)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Configuration 2</strong></td>
<td>1</td>
<td>20</td>
<td>Random</td>
</tr>
<tr>
<td>(two level hierarchy)</td>
<td>2</td>
<td>10</td>
<td>Random</td>
</tr>
<tr>
<td><strong>Configuration 3</strong></td>
<td>1</td>
<td>5</td>
<td>Star</td>
</tr>
<tr>
<td>(three level hierarchy)</td>
<td>2</td>
<td>8</td>
<td>Random</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>12</td>
<td>Random</td>
</tr>
</tbody>
</table>

\(^a\) Random topology refers to a topology generated by the random graph generator.
network configuration was obtained by making five replicas of the level-2 peergroup and then interconnecting these level-2 peergroups in a star topology. Values of the input parameters that were used in this simulation study are shown in Table 5.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Call interarrival time at a switch</td>
<td>1.5 sec.</td>
</tr>
<tr>
<td>Queue service rate</td>
<td>1000 messages/sec</td>
</tr>
<tr>
<td>Call setup handling time at a switch</td>
<td>2.33 ms</td>
</tr>
<tr>
<td>Crankback handling time at a switch</td>
<td>2.0 ms</td>
</tr>
</tbody>
</table>

Table 5: Input parameters

The average call setup delay in the one- and two phase schemes for the three configurations is shown in Figure 45. From Figure 45 we observe that for all the configurations, the average call setup delay in the one-phase scheme (optimal) is the highest and the average call setup delay in the one-phase (minimal) is the lowest. These observations are in agreement with the results obtained in the Section 5.4.1.4.
The maximum call setup delay in the one- and two-phase schemes for the three configurations is shown in Figure 46. The maximum call setup delay in the one-phase scheme (optimal) is higher than in the two-phase and the one-phase (minimal) schemes. The maximum call setup delay in the one-phase (minimal) scheme is marginally greater than that in the two-phase scheme for first two network configurations and lower than that in the two-phase scheme for the third network configuration (see Table 4). In Section 5.4.1.4, plots of maximum call setup delay (see Figure 43) show that the maximum call setup delay in both variations of the one-phase scheme is the same and it is significantly greater than the maximum call setup delay in the two-phase scheme. This discrepancy can be explained as follows: in section 5.4.1.4, while computing the maximum call setup delay, we considered the worst case scenario in which the minimal crossover node is always the optimal crossover node. But if we consider a specific network configuration, the worst case scenario in these schemes could be different and will depend on the topology of the network. In other words, the worst case scenario of these schemes is a
function of the topology of the network. Since we are considering specific network configurations in this section, the plots of the maximum call setup delay show a different trend than the ones shown in Figure 43. It should be noted that for certain topologies, the maximum call setup delay in the one-phase (minimal) scheme is smaller than that in the two-phase scheme. An example of such a topology is the star topology (see maximum call setup delays corresponding to configuration 3 in Figure 46).

Figure 47: shows the average and maximum number of hops in a connection in the one- and two-phase schemes. Both in the average case (the plot for average number of hops) and the worst case (the plot for maximum number of hops), the two-phase forwarding scheme (prior to route optimization) requires twice the amount of communication resources as the one-phase scheme. Since the number of hops in a connection after the route optimization phase in the two-phase scheme is equal to the number of hops in the one-phase scheme, the above observation indicates that the route optimization results in a significant reduction of resources required for a connection. These results are in agreement with the trend observed in Section 5.4.1.4.
The number of nodes that participate in the crankback procedure (crankback nodes) during the one-phase (optimal) and the one-phase (minimal) schemes for the different configurations is shown in Figure 48. We observe that the number of crankback nodes (both average and maximum number of nodes) in the one-phase scheme (minimal) is lower than the number of crankback nodes in the one-phase scheme (optimal). Since the number of crankback nodes are smaller in the one-phase (minimal) scheme, a larger part of the existing segment is identified as the common segment (see Figure 36), thereby resulting in smaller call setup delays as shown in Figure 45. We can also infer that the signalling load due to the one-phase (minimal) scheme is lower than the signalling load due to the one-phase (optimal) scheme, since the average number of nodes involved in the crankback process is lower than in the one-phase scheme (minimal) and the average number of nodes in a connection is the same in both the variations.

Figure 48: Crankback nodes involved in a connection setup
In summary, the following observations are made about the average case performance of the one- and two-phase schemes. *i*) The average call setup delay in the one-phase (minimal) scheme is lower than that in the two-phase scheme, and the average call setup delay in the one-phase (optimal) scheme is higher than that in the two-phase scheme. *ii*) The number of hops in a connection in the two-phase scheme (prior to route optimization) is significantly greater than the number of hops in the one-phase scheme. Besides the extra amount of resources required prior to route optimization phase, the two-phase call setup method has an overhead of route optimization signaling and buffering of cells. In order to minimize the amount of resources required for a connection and to avoid the additional signaling overhead of the second phase in the two-phase scheme, it is preferable to execute route optimization during call setup (one-phase scheme).

However, from the worst case analysis, we observe that the one-phase scheme can lead to high worst case call setup delays. These high worst case call setup delays associated with the one-phase scheme may not be acceptable for certain applications. Therefore, in such situations a fast local rerouting of the connection is required, and in order to minimize the number of hops in a connection, route optimization should be executed subsequently. Since the connection setup delay in these schemes depends on the topology of the network, and the exact locations of the calling party, called mobile and home of the called mobile. We propose that, on call arrival at the home of the called mobile, the home switch decides which scheme to execute, the one-phase scheme or the two-phase scheme. Some of the factors on which this decision can be based include: QoS (Quality of Service) requirements of the connection, duration of the call, the associated call setup delays, location of the calling party and the called mobile, topology and loading conditions of the
network, and number of hops in the connection. From the analysis, we can also infer that the one-phase (minimal) scheme shows significant performance gains over the one-phase (optimal) scheme, both in terms of call setup delay and signaling load.

5.5 Summary

In this Chapter, we proposed an one-phase crankback scheme. A comparative performance analysis of the one-phase crankback scheme and two-phase connection setup schemes was presented. Measures of comparison are call setup delay and the amount of network resources allocated to a connection. The maximum call setup delay (worst case call setup delay) is lower in the two-phase scheme, but the average call setup delay, interestingly, is lower in the one-phase scheme. The amount of resources required for a connection in the two-phase scheme (prior to route optimization) is significantly greater than that in the one-phase scheme.
CHAPTER 6

FLAT LOCATION MANAGEMENT SCHEME

As mentioned in Chapter 3, there are two approaches to preforming location management. In the first approach (overlay approach), a location management architecture is overlaid on the backbone network to manage the location of the mobiles. The location management architecture consists of location registers (databases) interconnected by a singling network. In this approach, an explicit location query can be issued to the associated location registers, and after determining the exact location of the mobile the connection is set up. This approach is being used in the current connection-oriented networks, e.g., cellular networks. In the second approach (integrated approach), the mobility management is integrated into the fixed network. In Chapters 4 and 5, two new methods of the integrated approach are presented.

In this Chapter, we propose a new overlay approach scheme, flat location management scheme. Overlay approach location management procedures are defined as part of the North American IS-41 cellular standard [4], and the European GSM standard for mobile communications [9]. Several other schemes have also been proposed to track and locate users within these networks [7, 10, 12]. Previous studies [80][81] have shown that with predicted levels of usage, there will be significant loads on the signaling network and on
network databases. References [7, 10, 12] illustrate how their location management schemes reduce the load on the signaling network. In this Chapter, we propose a flat location management scheme with the goal of not only reducing the signaling load (communication cost) but also the number of database accesses (computation cost), and mobile location delay (and hence call setup delay).

The rest of the Chapter is organized as follows. In Section 6.2, we describe the proposed flat scheme for location management. A comparative analysis of the flat scheme with existing schemes is included in Section 6.3 Conclusions are presented in Section 6.4

6.1 Prior work

Location management schemes track mobiles with the help of a hierarchy of location registers. The number of levels in the hierarchy are different in different schemes (see Chapter 3 for details). The IS-41 standards scheme [4] for location management is based on a hierarchical arrangement of MSCs (Mobile Switching Centers), VLRs (Visitor Location Registers) and HLRs (Home Location Registers). A minimum of two levels of hierarchy and a maximum of three levels of hierarchy are possible in an IS-41 based network. To reduce signaling costs relative to the IS-41 scheme. the forwarding scheme [7], the anchor scheme [10] and the hierarchical schemes [12, 8] propose increasing the number of levels of hierarchy. The first two schemes reduce "long-distance" signaling costs incurred by mobile tracking, while the hierarchical approach reduces "long-distance" signaling costs for both mobile tracking and mobile locating. However, the computation costs of adding extra levels of location registers is not considered.
We observe that instead of increasing the number of levels of hierarchy for all mobiles, a single level of hierarchy could be used for home mobiles to lower both computation and communication costs. In small networks, where the communication costs are not high, the computation cost becomes the dominant factor. In such cases, the need to create VLRs at MSCs, consuming additional computational resources, becomes unnecessary. Even if the network is large, it can be divided into "areas", with mobiles in their home areas directly tracked by their HLRs and mobiles outside their home areas being assigned a proxy HLR while in the remote area. This creates a two-level hierarchy only on an as-needed basis. We thus propose the flat scheme to allow for networks to be implemented with a single-level tracking for home mobiles and a "two-level" tracking for "remote" mobiles.

In this Chapter, we analytically compare the proposed flat scheme with the IS-41 scheme, the anchor scheme, and the forwarding scheme (see Chapter 3 for description). The hierarchical scheme has been excluded from this comparison, but is considered in detail in the next chapter. In the rest of this section, we review details of the IS-41 scheme, the forwarding scheme, and the anchor scheme needed in the subsequent performance analysis.

6.2 Proposed flat scheme

In this section, we propose a "flat" scheme for location management in PCNs with the goal of reducing both communication and computation costs. In this approach, we use one level of tracking for mobiles in their "home area", and a two-level tracking approach only for mobiles visiting other areas, referred to as "remote" mobiles. For home mobiles, the location is tracked directly by the HLR of the mobile. For each remote mobile, a Proxy HLR is assigned in the current "area" of the mobile to track its location. The HLRs of
remote mobiles track the Proxy HLR assigned to each such mobile. The "home area" of a mobile could include the entire network of its service provider, if the service provider's network is reasonably contained. For service providers with large networks, local signaling networks are defined, each corresponding to an "area". The home area of a mobile corresponds to the coverage area of the local signaling network in which its home switch is located. Local signaling networks are interconnected by a wide area network for such large service providers.

We first describe the mobile location procedure of the flat scheme in Section 6.2.1 and then present its mobile tracking procedure in Section 6.2.2.

6.2.1 Mobile location

Figure 49 shows the mobile location procedure in the presence of such a flat tracking scheme. If the mobile is located in its home network, mobile location only requires a one-hop message exchange. Given that the HLR knows the current MSC of its home mobiles, we note that the HLR can return a TLDN (Temporary Location Directory Number) identifying the current MSC of the mobile without sending a message to the MSC as in the IS-41 scheme. This solution of HLRs assigning TLDNs is adopted for the location of home mobiles as shown in Figure 49a. For remote mobiles, a two-hop location request procedure is carried out as shown in Figure 49b. The originating MSC sends a location query to the HLR of the called mobile. Since the HLR knows the address of the Proxy HLR tracking the called remote mobile, it generates a new location query to the Proxy HLR. The Proxy HLR selects a TLDN based on its knowledge of the current MSC of the called mobile, which is returned in two hops as shown in Figure 49b. Call setup then
proceeds from the originating MSC to the current MSC using the returned TLDN as in the case of home mobiles. Thus, if most mobiles are home mobiles, mobile location delay is reduced by only requiring a one-hop location request.

Next, we consider the issue of TLDN to MIN (Mobile Identification Number) mappings, i.e., the mapping of the temporary address (TLDN) to the actual address (MIN) of the mobile. Two solutions are possible for this TLDN to MIN mapping problem. The first solution is for the HLR/Proxy HLR to assign a default TLDN (Temporary Directory Local Number) using the identifier of the current MSC at which the mobile is located (given that it knows the MSC locations of its mobiles). For example, if the MSC is identified by 908-872-xxxx, then a default TLDN 908-872-0000 is used for all mobiles.
Call setup proceeds using the TLDN from the originating MSC to the MSC of the called mobile since routing of connections to the appropriate switches are based on the first six digits (908-872). Since the TLDN only indicates that the called party is a mobile without an exact association to a specific mobile, the MIN (Mobile Identification Number) of the mobile should be passed as an additional parameter in the call setup message as shown in Figure 49:. Thus, the mobile’s actual address (MIN) is *tunneled* in call setup messages which carry the temporary address (TLDN) of the mobile as the called party address. This can be achieved using a “mobility-related service control” parameter to carry the MIN, much like a CUG (Closed User Group) indicator used for virtual private networking [83].

A drawback of the above solution is that the format of the call setup message needs to be redefined to allow for the tunnelling of MINs. Our *second* solution avoids this drawback by having the current MSC of a mobile assign a TLDN during mobile registration. This assigned TLDN is sent in the REGNOT message to the HLR. In this approach, since the MSC has assigned the TLDN, it can keep a mapping of MINs to assigned TLDNs, which can then be used to determine the specific MIN corresponding to a TLDN during call setup. Thus, MINs do not need to be tunneled in call setup messages. This in turn implies that no changes are needed to the existing signaling standards. However, this approach was considered in GSM standards as reported in [84]. The reason this scheme was not accepted in the GSM standards is that it is likely that an MSC may run out of TLDNs given that only four digits are typically reserved for endpoints at an MSC, which, in addition to mobiles may also have fixed endpoints. A solution to the problem is arrived at by the observation that the last four digits of the TLDN are BCD (Binary Coded Decimal) numbers and a nibble is used to represent each of the BCD digits. This allows for 64K numbers of which only 10,000 correspond to valid TLDNs. We refer
to the rest of the representations as Invalid TLDNs (ITLDNs). To handle the issue of MSCs running out of numbers, the MSC can assign an ITLDN, rather than a valid TLDN, during mobile registration. By maintaining a map of MINs to assigned ITLDNs, on call arrival to a particular ITLDN, the MIN of the mobile is readily determined.

6.2.2 Mobile tracking

The basic mobile tracking procedure of the flat scheme is fairly simple. As stated earlier, mobiles in their home areas are tracked directly by their HLRs. Remote mobiles are assigned Proxy HLRs. When such a mobile moves from one area to the next, a new Proxy HLR is assigned to track the mobile in its new area. The assignment of Proxy HLRs for remote mobiles localizes the registration traffic. In other words, only mobiles moving out of the current network area will incur long distance registration signaling. This hybrid scheme will incur a lower communication cost than an "all-out flat" scheme, where all mobiles, irrespective of whether they are home mobiles or remote mobiles, register directly with their HLRs. Thus, the choice of choosing a two-level hierarchy for remote mobiles will reduce mobile tracking cost.

To further reduce mobile tracking costs without increasing mobile location delay, we propose an add-on concept to the basic tracking scheme by creating a logical base station controller-HLR interface to send registration messages directly from the BSC (Base Station Controller) to HLR bypassing the MSC. This proposal results from the observation that the current reference point definitions, with different protocols defined for the base station-MSC interface (IS-634) and for the MSC-HLR interface (IS-41), has certain inefficiencies. Figure 50 shows the protocol stacks for this proposed scheme in comparison with the one used in the current standards. In our proposed scheme, an IS-41
MAP REGNOT message is created by the base station/base station controller. This message is routed through datagram routers, STPs (Signaling Transfer Points), where only the MTP (Message Transfer Part) layer is terminated. Termination of the IS-41 MAP layer is only done at the HLR. On the other hand, in the current standards approach, an MSC needs to receive and parse IS-634 messages, and then create IS-41 MAP messages to the HLR. This will clearly add to both the computation and communication costs of mobile
tracking. In the proposed scheme, base stations still retain the notion of “zones” and only receive zone-change registrations. Thus, the same number of registrations are received by the network irrespective of whether the base station is sending the REGNOT directly to the HLRs or the MSC sends the REGNOT.

In order for a base station to send the REGNOT to the appropriate location register, the base station first needs to determine whether the registering mobile is a home or remote mobile. Since cellular networks use hierarchical network-based addressing, the base station can determine this from the MIN of the mobile. Next, for home mobiles, it needs to know the identities of the HLRs of mobiles. Given the use of hierarchical addressing, each range of MINs can be associated with an HLR address. For remote mobiles, if there is only one Proxy HLR per network area, then the base stations simply send all registrations for remote mobiles to the Proxy HLR. However, if there are multiple Proxy HLRs, then a base station to which a mobile moves needs to know which Proxy HLR to send the registration for a given MIN. One approach to solve this problem is to have each remote mobile know the identity of the Proxy HLR tracking it. This identifier can be sent by the network in response to the first registration. Subsequently, as the remote mobile moves, it includes the identifier of its Proxy HLR in its registration allowing the base station to send the REGNOT to the appropriate Proxy HLR.

A major implication of the proposal shown in Figure 50a is that Base Station Controllers (BSCs) will now require IS-41 software. This will clearly increase the cost of base station controllers. Furthermore, if the registration messages do not pass through the MSC, the second solution described in Section 6.2.1 requiring the MSCs to assign ITLDNs to the mobiles during the registration procedure cannot be used. Therefore, only the first solution of having HLR/Proxy HLRs assign default TLDNs can be used, which
implies that the call setup message format needs to be changed. Due to these implications, for analysis purposes, we consider the flat scheme without this add-on proposal of sending registration messages directly to the HLR/Proxy HLR\(^1\).

6.3 Performance analysis

In this section, we analytically compare four schemes for their mobile tracking and locating costs. The four schemes are the IS-41 scheme with collocated MSC/VLRs, the forwarding scheme, the anchor scheme, and the flat scheme proposed in Section 6.2. We use a generalized cost formulation for the computation of tracking and locating costs. This generalized formulation allows for the comparison of signaling loads (communication costs), database accesses (computation cost), and call setup delays (mobile location delays) of the different schemes.

---

1. In spite of the implications, we recorded the add-on scheme in this Chapter for future consideration in designing new PCN configurations.
For analysis purposes, we use a simplified PCN architecture as shown in Figure 51.

This reference architecture is used for all the schemes being analyzed. The entire coverage area is divided into cells, and cells are grouped into zones or registration areas. Cell sites belonging to a zone are connected to an MSC/VLR (Mobile Switching Center/Visitor Location Register). MSC/VLRs within an “area” are interconnected by a Local Signaling Network (LSN). Messages are sent between LSNs via the Wide Area Network. Messages sent over the wide area network are assumed to be “long distance signaling messages”, and hence incur a larger cost than messages sent within an LSN. HLRs within an area are connected to MSC/VLRs in that area via the Local Signaling Network of that area. The coverage area of the Local Signaling Network in which a mobile’s HLR is located, is termed the “home area” as mentioned in Section 6.2 When away from the home area, the mobile is termed a “remote mobile”. The HLRs and MSC/VLRs shown in Figure 51 are used in the IS-41 scheme, the anchor scheme, and the forwarding scheme. For the flat
scheme, we introduce a new set of location registers called *Proxy HLRs* also connected to the LSNs as shown in Figure 51. In the flat scheme, VLR functionality is not needed in the MSC. The notation used in the analysis is listed in Table 6.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>Cost to transport a signaling message in a Local Signaling Network</td>
</tr>
<tr>
<td>$R$</td>
<td>Cost to transport a signaling message in a Wide Area Network</td>
</tr>
<tr>
<td>$f$</td>
<td>Fraction of the mobiles that are remote (i.e., not in their home network area)</td>
</tr>
<tr>
<td>$Q_h$</td>
<td>Cost to query/update a HLR/Proxy HLR</td>
</tr>
<tr>
<td>$Q_s$</td>
<td>Cost incurred at an MSC</td>
</tr>
<tr>
<td>$Q_{sv}$</td>
<td>Cost incurred at an MSC/VLR</td>
</tr>
<tr>
<td>$k$</td>
<td>Maximum number of MSC/VLRs in the forwarding chain allowed in the forwarding scheme</td>
</tr>
<tr>
<td>$t$</td>
<td>Fraction of calls that originate within the current Local Signaling Network of the called mobile</td>
</tr>
<tr>
<td>$q$</td>
<td>Fraction of moves of the mobile that result in movement out of the current Local Signaling Network</td>
</tr>
</tbody>
</table>

Table 6: Notation

6.3.1 Tracking/locating costs in the flat scheme

For analysis purposes, as stated in Section 6.2.2 we assume that the flat scheme is implemented without the add-on proposal shown in Figure 50a. In other words, mobile tracking is as shown in Figure 50b with MSCs generating REGNOTs and the mobile location procedure is as shown in Figure 49. The mobile tracking procedure used in the flat scheme is illustrated in Figure 52. Mobile tracking cost consists of (i) cost incurred at the new MSC ($Q_s$), (ii) cost incurred at the HLR (for home mobiles) or the Proxy HLR (for remote mobiles)($Q_h$), and (iii) cost of signaling between the HLR/Proxy HLR and the
new MSC \((2L)\). When the mobile moves across from one network area (area served by a Local Signaling Network) to another, it incurs an additional cost of \((2(L + R + L) + Q_h)\), which is the cost of an HLR access \((Q_h)\), and the associated signaling between HLR and the Proxy HLR. Cancellation of the pointers at the old locations are assumed to the timer-based\(^2\). This assumption has also been applied to the

\[\text{Figure 52: Mobile tracking procedures in the flat scheme}\]

\[\text{new MSC (2L). When the mobile moves across from one network area (area served by a Local Signaling Network) to another, it incurs an additional cost of (2(L + R + L) + Q_h), which is the cost of an HLR access (Q_h), and the associated signaling between HLR and the Proxy HLR. Cancellation of the pointers at the old locations are assumed to the timer-based\(^2\). This assumption has also been applied to the}\]

---

\(^2\) In IS-41 standards, all the location requests are sent to the HLR independent of the location of the called mobile. Therefore, timer-based cancellations of information at the old locations are justified, since the outdated information at old locations are never consulted.
anchor scheme and the forwarding scheme. The average tracking cost per move incurred by the flat scheme \( M_{fl} \) is given by (EQ 13), where \( q \) denotes the probability of movement from one network area to another.

\[
M_{fl} = Q_s + Q_h + 2L + q (2 (L + R + L) + Q_h)
\]  

(EQ 13)

The mobile location procedure used in the flat scheme is shown in Figure 49. The location cost for home mobiles only consists of the cost incurred at the HLR for handling a location query \( Q_h \). For remote mobiles, an additional cost of querying the Proxy HLR \( Q_h^\) and the cost of the associated signaling is incurred. The average locating (Find) cost per call incurred by the flat scheme \( F_{fl} \) is given by (EQ 14). The cost of signaling between the originating MSC and the HLR is not included in the equation, since this cost is incurred by the location procedures of all the schemes being analyzed.

\[
F_{fl} = Q_h + \frac{f \times (2 \times (R + 2 \times L) + Q_h)}{f}
\]  

(EQ 14)

6.3.2 Tracking/locating costs in the IS-41 scheme

In this section, we analyze the cost of the IS-41 location management scheme for the collocated MSC/VLR configuration. The mobile tracking cost consists of (i) cost of updating the new MSC/VLR \( Q_{sv} \), (ii) cost of cancelling the old MSC/VLR \( Q_{sv} \), (iii) cost of updating the HLR \( Q_h \) and (iv) cost of signaling between the MSC/VLRs and the HLR. The average tracking cost per move incurred by the IS-41 scheme \( M_{is} \) is given by (EQ 15). For a remote mobile, the signaling between the MSC/VLRs and the HLR occurs over the Wide Area Network for both intra-LSN and inter-LSN moves. On the other hand,
for a home mobile, the signaling between the new MSC/VLR and the HLR occurs over the Wide Area Network only during an inter-LSN move. The scenario in which a remote mobile moves back to its home area is ignored in this analysis.

\[
M_{is} = Q_{sv} + Q_{h} + Q_{sv} + f(4(2L + R)) \\
+ (1 - f)(1 - q)(4L) + q(2L + 2(2L + R))
\]

(EQ 15)

The mobile location cost incurred by the IS-41 scheme consists of cost of querying the HLR \(Q_h\), cost of querying the VLR/MSC \(Q_{sv}\), and the cost of the associated signaling. The average locating cost per call incurred by the IS-41 scheme \(F_{is}\) is given by (EQ 16).

\[
F_{is} = Q_h + Q_{sv} + \frac{2(L + f \times (R + L))}{2}
\]

(EQ 16)

6.3.3 Tracking/locating costs in the forwarding scheme

In this section, we analyze the cost of tracking and locating procedures used in the forwarding scheme [12]. Tracking in the forwarding scheme is done by setting a pointer to the new MSC/VLR at the old MSC/VLR. If the maximum length of the forwarding chain has \(k\) nodes, forwarding pointers are set for \(k - 1\) moves and on the \(k^{th}\) move, an IS-41 registration procedure is performed. So, the tracking cost consists of (i) cost of setting up forwarding pointers for the first \(k - 1\) moves, and (ii) cost of an IS-41 tracking in the \(k^{th}\) move. We assume that the cancellation of the pointers at the \(k - 2\) intermediate MSC/VLRs is done by a timer-based mechanism as explained in Section 6.3.1. The average tracking cost per move incurred by the forwarding scheme \(M_{fp}\) is given by (EQ 17).

\[
M_{fp} = \{ (k - 1) \times 2 \times \left( Q_{sv} + \frac{L}{q} + q \times (R + L) \right) + M_{is} \} / k
\]

(EQ 17)
In the forwarding scheme, on location query, the HLR returns information about the anchor MSC to the originating MSC. The originating MSC then sends a location query to the anchor MSC, which is sent along the forwarding chain until the current MSC of the mobile is reached. The current MSC assigns a TLDN, which is returned to the originating MSC.

The location cost in the forwarding scheme consists of (i) cost of querying the HLR \(Q_h\), (ii) cost of signaling between the originating MSC and the anchor MSC \(L + (1 - t)(R + L)\), where \(t\) denotes the fraction of calls that originate within the current Local Signaling Network of the called mobile), (iii) cost of querying the anchor VLR/MSC \(Q_{sv}\), and (iv) cost of traversing the forwarding pointers. The average number of forwarding pointers traversed is \(\frac{k - 1}{2}\). The average locating cost per call incurred by the forwarding scheme \(F_{fp}\) is given by (EQ 18).

\[
F_{fp} = Q_h + \frac{2(L + (1 - t)(R + L))}{2(L + q(R + L))} + Q_{sv} + \left(\frac{k - 1}{2}\right)\left(\frac{2(L + q(R + L))}{2(L + q(R + L))} + Q_{sv}\right)
\]

**6.3.4 Tracking/locating costs in the anchor scheme**

In this section, we analyze the cost of locating/tracking for the static case of the anchor scheme. Tracking cost consists of the cost of updating the anchor \(Q_{sv}\), the cost of updating the new MSC/VLR of the mobile \(Q_{sv}\), and the cost of signaling between the local anchor and the new MSC/VLR of the mobile. The cancellation at the old MSC/VLR...
is assumed to be timer-based as explained in Section 6.3.1. The average tracking cost per move incurred by the anchor scheme ($M_{la}$) is given by (EQ 19). The scenario in which a mobile moves back to its local anchor area is ignored in this analysis.

$$M_{la} = 2Q_{sv} + \frac{2(L + q(R + L))}{1 - P_0}$$  \text{(EQ 19)}

The location cost in the anchor scheme consists of (i) cost of querying the HLR ($Q_h$), (ii) cost of signaling between the HLR and the anchor MSC/VLR, and (iii) cost of querying the MSC/VLR ($Q_{sv}$). If the mobile has moved since the last call arrival (i.e., the mobile is not at the anchor MSC/VLR), the location cost also consists of the cost of signaling between the local anchor and the current MSC/VLR, and the cost of querying the current MSC/VLR. The equilibrium probability that a mobile is in its local anchor area is $P_0$ and is computed to be $\frac{\rho}{1 + \rho}$ in [10], where $\rho$ is the CMR (Call-to-Mobility Ratio).

The average mobile locating cost per call incurred by the anchor scheme is given by (EQ 20).

$$F_{la} = Q_h + 2(f \times (R + L) + L) + Q_{sv} + (1 - P_0) \left(2 \left(L + q(R + L)\right) + Q_{sv}\right)$$  \text{(EQ 20)}

### 6.3.5 Average total costs

The total cost of a location management scheme depends on the tracking cost and the locating cost of that scheme. In order to be able to estimate the average total cost, the rate of call arrivals at a mobile, $\lambda_c$, and the rate at which the mobile moves between registration areas, $\lambda_m$, are needed. The average tracking cost per unit time, average locating cost per unit time, and the average total cost per unit time, are given in (EQ 21), where $M$ and $F$ represent the average tracking cost per move and average locating cost per call, respectively.
\[ M = \lambda_m M \quad F = \lambda_c F \quad T = M + F \]  
(EQ 21)

Since the main objective of the forwarding scheme, anchor scheme, and the flat scheme is to improve the performance over the IS-41 standards approach, we define the normalized cost of a scheme to be the ratio of the average total cost in the scheme to that of the average total cost in the IS-41 scheme. The normalized cost of a scheme, \( T^n \), is given below.

\[ T^n = \frac{\lambda_m M + \lambda_c F}{\lambda_m M_{is} + \lambda_c F_{is}} \]  
(EQ 22)

Since we do not have exact numbers for \( \lambda_c \) and \( \lambda_m \), we use the CMR (Call-to-Mobility Ratio), denoted as \( p \) (defined as the number of call arrivals per move), to study the performance of these schemes. Using this definition of CMR, the normalized cost, \( T^n \), is given by

\[ T^n = \frac{M + \rho F}{M_{is} + \rho F_{is}} \]  
(EQ 23)

where \( \rho = \frac{\lambda_c}{\lambda_m} \).

6.3.6 Numerical results

In this section, we present the quantitative results of our comparison. The measures of comparison include computation costs (number of database accesses), communication costs (signaling load), the average total cost, and the call setup delay. The numerical comparison uses the formulas derived in Sections 6.3.1 to 6.3.5.
**Input Data:** Values of input parameters assumed for this numerical computation are shown in Table 7 (see Table 6 for definitions of these parameters). For the sake of this quantitative comparison, we assume that the values of $Q_v$, $Q_z$, and $Q_h$ are equal unless otherwise specified and the value is referred to as $Q$. The value of $q$ is estimated to be 0.13, considering each LATA (Local Access and Transport Area) to be a local signaling network area [7]. We observe that the exact numerical results are dependent on the exact values chosen for these parameters. We also note that the cost incurred at a network entity could be potentially different in different schemes. Sensitivity of the comparative results to the input parameters is also analyzed.

### 6.3.6.1 Computation costs

Computation cost consists of the cost of querying/更新 databases, namely MSC/VLRs, HLRs, and Proxy HLRs. The computation costs of the schemes are calculated by setting the values of $L$ and $R$ to 0 in equations 13 to 20. In other words, the signaling costs are set to 0. Plots in Figure 53: show the variation of the normalized computation cost with
Figure 53: Comparison of computation costs

CMR for different values of $f$ (fraction of remote mobiles) and $q$ (fraction of inter-LSN moves). Plots with varying values of $f$ and $q$ are shown only for the flat scheme, since the computation resource requirements of the other schemes are independent of the values of $f$ and $q$ (refer (EQ 15) to (EQ 20)). The plots show that at high CMRs, the flat scheme has the least computation cost for all values of $f$ and $q$, but at low CMRs, the computation cost depends on the values of $f$ and $q$. The plots also show that the computation cost of the flat scheme increases with an increase in the values of $f$ and/or $q$.

Next, we consider the sensitivity of these results to a variation in the values of $f$ and $q$. With an increase in the value of $f$, the computation cost of the location procedure in the flat scheme increases, since the number of mobiles that require location queries to Proxy HLRs increases. From (EQ 13), we observe that the computation cost of the tracking procedure is independent of $f$. Since the computation cost of only the location procedure increases with increase in $f$, the computation cost of the flat scheme increases with increase in $f$ at high CMRs. With an increase in the value of $q$, the computation cost of
the tracking procedure increases, since the tracking procedure of the flat scheme updates the HLR on every inter-LSN move. From (EQ 14), we observe that the computation cost of the locating procedure is independent of $q$. Since the computation cost of only the tracking procedure increases with increase in $q$, the computation cost of the flat scheme increases with increase in $q$ at low CMRs. It is interesting to note that the computation cost of the flat scheme is less than or equal to that of the IS-41 scheme (limiting to the IS-41 scheme, when the values of $f$ and $q$ are equal to 1). In general for low-tier PCS applications, where the CMR is high, the flat scheme incurs a lower computation cost.

6.3.6.2 Communication costs

Communication cost consists of the cost of signaling over the LSN and the Wide Area Network. The communication costs of the schemes are computed by considering only the cost of signaling (i.e., ignoring the cost incurred by database accesses). In other words, the value of $Q$ is set to 0 in evaluating equations 13 to 20. Plots in Figure 54: show the variation of the normalized communication cost with CMR. These plots show that at high CMRs, the flat scheme has the least communication cost for all values of $f$ and $q$, but at low CMRs, the communication cost depends on the values of $f$ and $q$.

In general, as the number of levels of hierarchy decreases, the location management scheme is expected to incur a higher communication cost. On the contrary, Figure 54: shows that, for a wide range of CMRs the flat scheme has a lower communication cost than the other schemes. This trend in the decrease of communication cost in the flat scheme is due to: i) the improved mobile location procedure, where signaling messages to the MSC/VLR for the assignment of the TLDN (Temporary Location Directory Number) are avoided, and ii) the use of Proxy HLRs for remote mobiles, which eliminates long-
distance signaling messages for moves within the Local Signalling Network area. For a move across LSNs, long-distance signaling cost across the Wide Area Network is incurred in all the schemes.

Figure 54: Comparison of the communication costs
Next, we consider the sensitivity of these results to a variation in the values of $f$ and $q$. Table 8 shows the terms in the location and tracking signaling (communication) cost formulations that are dependent on the values of $f$ and $q$. From Table 8 we observe that, for the flat scheme, the signaling cost of the tracking procedure increases with an increase in $q$, and the signaling cost during locating procedure increases with an increase in $f$. At low CMRs, the signaling cost of the tracking procedure dominates, while at high CMRs, the signaling cost during the locating procedure dominates. Therefore, at low CMRs, the total signaling cost in the flat scheme is affected significantly by the variation in $q$ and not by the variation in $f$. We observe that the rate of increase in signaling cost with an increase in $q$ is greater for the flat scheme than in any other scheme. Therefore at low CMRs, with increase in the value of $q$, the communication cost in the flat scheme increases relative to that of the other schemes.

### Table 8: Tracking and locating costs

<table>
<thead>
<tr>
<th>Cost</th>
<th>Flat Scheme</th>
<th>Anchor Scheme</th>
<th>Forwarding Scheme</th>
<th>IS-41 Scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tracking</td>
<td>$2q(R + 2L)$</td>
<td>$2q(R + L)$</td>
<td>$\frac{2(q(1-f)+\beta)(R+L)}{k}$</td>
<td>$2(q(1-f)+\beta)(R+L)$</td>
</tr>
<tr>
<td>Location</td>
<td>$2f(R + 2L)$</td>
<td>$2f + q(1-P_\alpha)(R + L)$</td>
<td>$q(k-1)(R+L)$</td>
<td>$2f(R + L)$</td>
</tr>
</tbody>
</table>

6.3.6.3 Average total cost

The average total cost of a scheme is computed using, the equations derived in sections 6.3.1 to 6.3.5, and the input values from Table 7. Plots of the average total cost show the combined effect of the computation and communication costs. Therefore, these plots will vary with variations in the ratio of the cost of computation to communication ($Q/L$, $Q$ being the cost incurred at a database, and $L$ being the cost of signaling over an
LSN). For high values (low values) of $Q/L$, these plots will show a similar trend as the plots of computation costs (communication costs). For example, if the computation cost (communication cost) of scheme is lower than that of another for values greater than a certain CMR, as the value of $Q/L$ increases (decreases) the crossover point of the total cost plots will approach the value of that CMR. We demonstrated in Sections 6.3.6.1 and 6.3.6.2 that, in most of the cases, the trends in the communication and computation costs are similar. Therefore, in general, the $Q/L$ ratio which determines, whether the computation cost should be emphasized or the communication cost, will effect all the curves in a similar manner. This is demonstrated in Figure 55:, which plots the normalized average total cost against CMR. An exception is seen while comparing the IS-41 and the forwarding schemes for values of $f = 0.9$, and $q = 0.13$. In Section 6.3.6.2, the plot (see Figure 54:b) shows that the forwarding scheme incurs a lower communication cost than the IS-41 scheme at all CMRs but in Section 6.3.6.1 the plot shows that the
 forwarding scheme has a crossover CMR above which it incurs a high computation cost. So while comparing the total costs of the forwarding and the IS-41 schemes, the $Q/L$ ratio matters. On the other hand, both computation and communication costs of the flat scheme show similar trends in that both costs are lower relative to the other schemes at high CMRs but could be higher at low CMRs. Hence the total cost behavior of the flat scheme relative to other schemes varies only marginally with different $Q/L$ ratios. From the plots showing the variation of the normalized average total cost with CMR (see Figure 55:), we observe that, for very low values of CMR (<0.07, i.e., greater than 14 moves per call) the anchor scheme incurs the least cost, and for higher values of CMR (> 0.07) the flat approach incurs the least cost.

6.3.6.4 Call setup delay

Incoming calls to mobiles require a mobile location procedure prior to actual connection setup. If we assume that the connection setup procedure is the same in all the schemes, by comparing the mobile location delay, we compare the call setup delays in the schemes. Normalized location delay of a scheme is computed as the ratio of location cost in that scheme to the location cost in the IS-41 scheme. Figure 56: shows the variation of mobile location cost, hence call setup delay, with CMR. Plots show that the flat scheme has lower mobile location cost than any of the other schemes. This is because for home mobiles, the flat scheme only requires one database access, while for all the other schemes at least two database accesses are required. Mobile location delay varies with CMR only for the anchor scheme, since at high CMRs the mobile is typically located at the local anchor, while at low CMRs, a pointer will most likely exist from the local anchor to the current MSC/VLR.
Variation of $q$ and $f$ affects the mobile location cost. With an increase in the value of $q$ (i.e., increase in inter-LSN moves), the cost of the mobile location procedure increases in the anchor scheme and the forwarding scheme. With an increase in the value of $f$, the cost of the locating procedure increases in the flat scheme, the anchor scheme and the IS-41 scheme. For high values of $f$, the mobile location delay of the flat scheme is similar to that of the IS-41 scheme.

6.4 Summary

This Chapter presented a “flat” mobile location management scheme. The objective of this scheme is to reduce both mobile tracking and location costs. Mobile tracking costs are reduced by (i) the use of one level of tracking for mobiles in their “home area”, and a two-level tracking approach only for mobiles visiting other areas, and (ii) creating a logical base station controller-HLR interface to send registration message directly from the BSC
to the HLR, bypassing the MSC. Mobile location costs are reduced by using an improved mobile location strategy, which requires a one-hop message exchange for “home” mobiles, and a two-hop message only for “remote” mobiles.

We analytically modeled the costs incurred during the tracking and locating procedures of the different schemes, and compared their performance. The measures of comparison used were the computation cost (database accesses), the communication cost (signaling messages), the average total cost, and the mobile location cost (call setup delay). Analysis results showed that the proposed flat scheme reduces the number of database accesses, number of signaling messages, and the call setup delay when compared to the location management strategy proposed in the IS-41 standard. Results also show that for medium and high values of CMR (> 0.07, i.e., less than 14 moves per call), the flat scheme performs better than the forwarding scheme and the anchor scheme. In most cases, the performance of the flat scheme relative to other schemes varied only marginally with different $Q/L$ ($Q$ being the cost incurred at a database, and $L$ being the cost of signaling over an LSN) ratios.
CHAPTER 7

GENERALIZED OVERLAY SCHEME: LR SCHEME

In this Chapter, we describe a generalized overlay approach scheme, LR (Location Register) Scheme. Prior work on overlay approach location management includes the cellular IS-41 MAP (Mobility Application Part) standard [4] and several improvements proposed in [7, 8, 10, 12, 55]. The cellular IS-41 scheme consists of using a two-level hierarchy of location registers called Home Location Registers (HLRs) and Visitor Location Registers (VLRs) to track mobile locations using REGistration NOTification (REGNOT) messages. An HLR is assigned to a mobile based on its permanent address, while a VLR, which is typically collocated with an Mobile Switching Center (MSC), is assigned based on the current location of the mobile. Incoming calls to mobiles are delivered after executing a mobile location phase, wherein the call originating switch generates a mobile LOCATION REQuest (LOCREQ) to the HLR of the mobile, which, in turn, generates another query to the VLR/MSC. Among the improvements proposed to this scheme are the extremes of the “flat” scheme [55] and the “hierarchical” scheme [8, 12]. The former proposes using a single-level hierarchy of location registers, while the latter proposes building a rooted tree of location registers. In the flat scheme, upon receiving a LOCREQ, the HLR assigns a temporary address based on the VLR/MSC at
which the called mobile is located rather than require an additional message exchange from the HLR to the VLR/MSC to obtain a temporary address assignment. The mobile’s permanent address is tunnelled in the call setup message while the temporary address is used to route the connection from the call originating switch to the mobile’s current switch. The hierarchical scheme uses a hierarchy of location registers to localize both mobile tracking and mobile locating messages. A registration is propagated up the hierarchy until it reaches a location register beyond which there is no change of information regarding the mobile’s location. The call setup message (or an explicit location query) is sent up the hierarchy until it reaches a location register that knows the location of the mobile from which point the hierarchy is traced in the downward direction to reach (or determine) the exact switch where the mobile is located.

The flat scheme results in lower computation costs but incurs larger communication costs than the cellular scheme, while the hierarchical scheme achieves the opposite (lower communication costs, but higher computation costs). By building a rooted tree of location registers, the need for home location registers is eliminated, thus removing the need for “long-distance” signaling messages. On the other hand, to determine the location of a mobile, the location query needs to be stopped and processed at a much larger number of location registers (on the average). This increases the computation costs of the network. In contrast, the flat scheme increases the overall signaling load (communication costs) on the network since registrations and location queries need to be sent “long-distance” to the HLRs of mobiles, but it also decreases the computation costs since processing is needed only at one node for both registrations and location queries. The flat scheme also results in a lower mobile location delay due to the one-hop location query processing, which, in
turn, leads to a lower overall call setup delay. Other improved schemes, such as the forwarding scheme of [7] and the anchor scheme of [10], are in between these two extreme schemes in terms of computation and communication costs.

<table>
<thead>
<tr>
<th>Cost</th>
<th>Flat scheme</th>
<th>Hierarchical scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication costs</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Computation costs</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Mobile location delay (and hence call setup delay)</td>
<td>Low</td>
<td>High</td>
</tr>
</tbody>
</table>

Table 9: Comparison of Flat and Hierarchical Schemes

If signaling load and mobile location delay are parameters of comparison the hierarchical scheme perform well at low CMR (Call-to-Mobility Ratio) and the Flat scheme performs well at high CMR. The Location Register (LR) scheme is proposed to support different classes of users and also to study the effect of different levels of hierarchy in the location management architecture.

The Location Registers (LR) scheme proposed is a hybrid scheme whose parameters can be set to default to one of the two improved schemes, i.e., the flat scheme or the hierarchical scheme. It essentially uses a hierarchy of location registers but limits the hierarchy by lopping off the tree at some level $S$, beyond which it resorts to the flat scheme approach of updating/consulting a home location register. It also uses the concept of tunnelling the mobile’s permanent address in the call setup message as was proposed for the flat scheme. Such a hybrid scheme allows a network provider to implement one of the two extremes or some in-between scheme by selecting $S$ according to computation and
communication costs. The novelty of the LR scheme lies in this concept of lopping off the tree at level $S$, which allows for the simultaneous minimization of both communication and computation costs.

The LR (Location Registers) scheme, isolates the effect of mobility from the PNNI routing protocol. The cellular concept of using location registers to handle mobile users is introduced into PNNI based ATM networks. Location registers (databases) are placed within the peer group structure of these ATM networks. However, instead of directly adopting the IS-41-based location management scheme, we use a combination of the two improved schemes reviewed in Chapter 3. In this approach, location registers track the location of mobiles, and respond to location queries generated prior to connection setup. Thus, unlike the mobile PNNI scheme, the LR scheme uses an explicit mobile location phase prior to connection setup.

7.1 Location Registers (LR) scheme

In this scheme, the routing protocol reachability information is disregarded for mobile endpoints. Instead, an explicit tracking and locating procedure is overlaid on a network using location registers. The LR scheme architecture, the mobile tracking procedure and the mobile locating procedure are defined in the following subsections.

7.1.1 LR scheme architecture

Figure 57 shows hierarchically-organized location registers (LRs). The switches are represented by circles. Location registers only exist from level $L$ up to some level $S$ (the tree is lopped-off at level $S$). We make an assumption that each peer group is assumed to have one LR. This assumption can be relaxed and multiple LRs may be located in each
peer group. This is effectively equivalent to creating a sublayer under the lowest layer of switches, and applying the same concept of allocating one LR per peer group of this new sublayer.

Location register A.1.L (we use the .L extension to avoid confusing this node with the A.1 logical node) is assumed to track all mobiles attached to switches within peer group A.1 (i.e., mobiles located at base stations connected to switches A.1.1, A.1.2, A.1.3 and
A.1.4). Similarly, A.2.L is assumed to track all mobiles located at base stations connected to switches A.2.1 and A.2.2. A home LR is assigned to a mobile based on its permanent address, e.g. A.1.1.L is the home LR of the mobile A.1.2.3.

The hierarchy of location registers helps localize mobile tracking and locating costs. However, if the hierarchy is carried to the topmost level \((l = 1)\) as in the hierarchical scheme of [12], the processing requirements could be high. If computation costs are more than communication costs, it is more expensive to stop and process REGNOT (Registration Notification) or LOCREQ (Location Request) at each LR in the hierarchy, than to send one such request as a connectionless message to the home. Hence, we limit the hierarchy to level \(S\) and resort to the flat scheme approach of updating and/or querying the home LR of the mobile. However, if the home LR were to track the lowest level \((l = L)\) LR currently tracking the mobile as in the flat scheme, the long-distance signaling costs of updating or querying the home LR would be high. Hence, the home LR only tracks the \(S^{th}\) level LR for each mobile, and only receives location queries if none of the LRs up to level \(S\) of the calling mobile's switch can respond to the query. The parameter \(S\) allows the LR scheme to be flexibly implemented as a flat structure, or as a rooted hierarchical tree, or as a mixed structure combining these extremes.

7.1.2 Mobile tracking

When a mobile powers on, the switch connected to its base station receives a power-on registration message. This switch sends a REGNOT (Registration Notification) to its LR at level \(L\). This, in turn, causes REGNOTs to be generated to the ancestor LRs upstream up to an LR at level \(S\). If the visiting switch is distinct from the home switch, the LR at level \(S\) sends a REGNOT to notify the home LR of the mobile that the mobile is currently
in its domain. REGNOTs are sent as connectionless packets using the ATM NSAP address of the mobile as the destination. The home LRs of all mobiles visiting at switches other than their home switch track the $S^{th}$ level LR of the mobile in its current location. Power-off registrations are handled in a similar manner as power-on, whereby LRs up to level $S$ are informed that a mobile powered off, and if the mobile was visiting (away from home), its home LR is also notified. Figure 59: shows how location registers are updated in the LR scheme.

Next consider zone-change registrations, which are generated as mobiles move from a base station connected to one switch to a base station connected to another switch. The hierarchy of LRs is exploited to limit the propagation of registration information for such movements. On receiving the registration message, the new switch sends a REGNOT message to its level $L$ LR. This, in turn propagates the REGNOT message upwards up to the LR which is a common ancestor of the LR corresponding to the old switch and the LR
corresponding to the new switch, or up to level $S$, whichever is lower in the hierarchy (higher in numerical value). A message is sent by the new switch to the old switch, informing the old switch about the movement of the mobile. The old switch then generates a REGCANC (Registration Cancellation), which is sent to its level $L$ LR. This, in turn, is propagated upwards, cancelling the old information in the LRs. If the $S^{th}$ level LR tracking the mobile changes due to the move, then the home LR of the mobile is updated.

For example, if the mobile A.1.1.4, shown in Figure 57:, moves to a base station connected to switch A.1.2, only LR A.1.L needs to be updated. One cancellation is required at the switch. On the other hand, if it moves from switch A.1.1 to a switch A.2.2, then REGNOTs are sent to LR A.2.L from switch A.2.2, and subsequently from LR A.2.L to LR A.L, since LR A.L is the common-ancestor LR of the LRs corresponding to the old and new switches. Since the LR at level $S$ (A.L) did not change, there is no REGNOT sent to the home LR of the mobile. However, a cancellation message is sent from A.2.2 to A.1.1, which in turn generates a REGCANC from A.1.1 to A.1.L. Finally, if the mobile moves from switch A.2.2 to switch B.1.1, REGNOTs propagate from switch B.1.1 to LR B.1.L, and then to LR B.L. Since there is a change in the $S^{th}$ level LR tracking the mobile, LR B.L notifies home LR A.1.L. In addition, a REGCANC is generated by B.1.1 to A.2.2, which passes upwards to LR A.2.L, and then to A.L.
7.1.3 Mobile locating

To find a mobile prior to call setup, a chain of location registers is traced. The length of the chain depends on the location of the calling party and the current location of the mobile. The called party's switch begins by checking to see if the called mobile is located at a base station in its domain. If so, it completes the call without generating any LOCREQs (Location Requests).

If the called mobile is not located at a base station within its domain, it generates a LOCREQ to its LR. Such requests are forwarded upwards in the hierarchy of LRs. If an LR at some level $k$ has information (pointer to a child LR) regarding the location of the mobile, then it sends LOCREQs downwards toward the called mobile's current location. The location query will be resolved by the level $L$ LR of the switch at which the called mobile is located, and the response will be sent directly to the calling party's switch.

If, however, none of the LRs, from the level $L$ LR of the calling party's switch to the $S^{th}$ level LR know the location of the called mobile as shown in Figure 60:, the $S^{th}$ level LR sends a LOCREQ to the home LR of the called mobile. It uses connectionless transport (see Chapter 10) with the destination address set to that of the called mobile. The called mobile's home switch will then forward this message to the home LR of the mobile. Since the home LR tracks the $S^{th}$ level LR of its mobiles, it forwards the LOCREQ to the $S^{th}$ level LR tracking the mobile in its current location. This LR generates downward LOCREQs according to the information it has about the called mobile. The LOCREQ will reach the level $L$ LR of the called mobile's switch. The response is sent directly from this LR to the calling party's switch as shown in Figure 60:. The address tunnelling concept of the flat scheme described in Chapter 3 is also used in the LR scheme.
As examples, we consider call originations from three endpoints, B.2.2.5, B.1.1.5 and A.2.2.5, all targeted at mobile A.1.2.3 (see Figure 57). In the first example, when switch B.2.2 generates a LOCREQ for mobile A.1.2.3 to its LR B.2.L, the latter can immediately respond since the called mobile A.1.2.3 is located within its region. In the second example, switch B.1.1 sends the LOCREQ (in response to the call setup request from its endpoint B.1.1.5 to mobile A.1.2.3) to its LR B.1.L. Since it has no pointer regarding this mobile, it simply generates a LOCREQ to the higher-level LR B.L. This register has a pointer indicating that B.2.L is tracking the mobile. Hence a LOCREQ is sent to this LR. Since B.2.L is the level L LR for the called mobile, it responds indicating that the mobile is located at switch B.2.1. This response is sent directly to switch B.1.1 (instead of retracing the pointers backwards) allowing it to initiate call setup to the called mobile’s switch. In the third example, where endpoint A.2.2.5 generates the call setup to mobile A.1.2.3, the LOCREQ sent by switch A.2.2 traverses the chain of LRIs, A.2.L and A.L. Since neither of these LRIs have information on the location of the called mobile and $S = 2$, A.L sends a LOCREQ to the home LR of the called mobile A.1.L. This LR forwards the LOCREQ to
LR B.L, since each home LR tracks the $S^{th}$ level LR of its mobiles. LOCREQs are then sent downwards from B.L to LR B.2.L, which responds with a temporary address for the mobile indicating that the mobile is located at switch B.2.1.

7.1.4 IP Mobility Support

In the previous sections, the LR scheme is described with in a connectionless environment. The techniques of the LR scheme can be used in a connectionless environment with modifications. For example, can be used to provide IP mobility support. As mentioned earlier, in mobile IP the datagrams are tunneled from the Home Agent to the Foreign Agent/Mobile. For scalability, accounting, and security reasons it may be desirable to use hierarchical tunnels using the LR architecture.

Most implementations of routers can support limited number of tunnels. This will limit the number of foreign agents to which a home agent can maintain a tunnel to. Which in turn limits the number of mobile/areas to which mobiles can move. If an architecture similar to the LR architecture is used to tunnel traffic, then a Home Agent can be logically connected to more number of Foreign Agents.

The use of hierarchical tunnels facilitates efficient security management. In the Flat architecture, it is necessary to configure (in the absence of any key exchange protocol as it is currently) a home agent with security associations of all the possible level L+1 Foreign Agents. In the LR architecture, the Home Agent has to maintain security associations for the level S Agents. In this Architecture, Security gateways at Firewalls could be Agents, thus facilitating efficient traversal of Firewalls. The difference in LR implementation for the connection-oriented case and the connectionless case is that in the connection-oriented case, an optimally routed is set up between the far-end node and the mobile. Whereas in
the connectionless case, the data is tunneled and follows the path of the LOCREQ as shown in Figure 60. During updating (registration) the LR architecture contains the effects of mobility to local area, and in the connectionless case will support efficient, secure and fast handoffs.

7.2 Performance analysis

In this section, we analyze the LR scheme and the mobile PNNI scheme (Chapter 4) and compare their costs. We determine the tracking and locating costs for a mobile in each of these schemes using analytical models. The analysis also allows us to provide insights into the effects of different parameters on the performance of each scheme. In the mobile PNNI scheme, we do not account for the route optimization phase in this analysis. After addressing certain preliminaries in Section 7.2.1, Sections 7.2.2, 7.2.3 and 7.2.4 describe how the average tracking cost per move, average search cost per call setup, and average total cost per move are computed, respectively. Numerical results are provided in Section 7.2.5.

7.2.1 Preliminaries

We first define the notation used in this analysis, and then describe a basic property of PNNI standards based hierarchical networks that is repeatedly used in the analysis.

Figure 61: Illustration of the hierarchy numbering notation
7.2.1.1 Notation

Table 10 lists the symbols used in this analysis section, along with their definitions.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L )</td>
<td>Number of peer group levels in the network; level 1 is the topmost level and level ( L+1 ) represents individual switches in the network (see Figure 61).</td>
</tr>
<tr>
<td>( a_{ij} )</td>
<td>The ancestors-are-siblings level of two nodes ( i ) and ( j ).</td>
</tr>
<tr>
<td>( h, v, o, n, c )</td>
<td>Subscripts used to represent the home, visiting, old and new locations of a mobile, and the calling party, respectively.</td>
</tr>
<tr>
<td>( S )</td>
<td>Scope indicating the stopping distance for reachability update propagations.</td>
</tr>
<tr>
<td>( G_i )</td>
<td>The neighborhood of node ( i ).</td>
</tr>
<tr>
<td>( U_K )</td>
<td>Cost of updating reachability data sent by a node ( i ) to all nodes ( j ) such that ( a_{ij} \geq K ).</td>
</tr>
<tr>
<td>( m_i )</td>
<td>Length of the longest MST (minimum spanning tree) among all peer groups at level ( i ) for ( i = 1, 2, \ldots L ); ( m_{L+1} = 1 ); ( m_i = 0 ) for ( i &gt; L + 1 ). There are ( m_i + 1 ) peer nodes in a peer group of level ( i ).</td>
</tr>
<tr>
<td>( p_i )</td>
<td>Average (among all node pairs) length of the “shortest-path” between nodes of a peer group at level ( i ).</td>
</tr>
<tr>
<td>( N_b )</td>
<td>Number of base stations in a registration area (i.e., a switch in this case).</td>
</tr>
<tr>
<td>( D_{ij} )</td>
<td>Distance, in terms of the number of nodes on the route, from node ( i ) to node ( j ).</td>
</tr>
<tr>
<td>( h )</td>
<td>Cost of sending a long distance message (such as the registration message).</td>
</tr>
<tr>
<td>( R_k )</td>
<td>Cost of updating or querying LRs up to level ( k ) from a switch.</td>
</tr>
<tr>
<td>( c_i )</td>
<td>Cost of updating or querying a level ( i ) location register, where ( c_i = 0, i &gt; L ).</td>
</tr>
<tr>
<td>( \rho )</td>
<td>CMR (Call-to-Mobility Ratio). It represents the number of calls per move.</td>
</tr>
</tbody>
</table>

Table 10: Notation
7.2.1.2 Property

For three nodes \(x\), \(y\) and \(z\), the following relations hold between their *ancestors-are-sibling* levels:

Case I: \(a_{xy} < a_{xz} \Rightarrow a_{yz} = a_{xy}\)  \hspace{1cm} (EQ 24)

Case II: \(a_{xy} = a_{xz} \Rightarrow a_{yz} \geq a_{xz}\)  \hspace{1cm} (EQ 25)

The proof of this property can be inferred from Figure 62. If \(a_{xy} < a_{xz}\) the arrangement of nodes is shown in *Case I* of Figure 62, from which it is clear that \(a_{yz} = a_{xy}\). A similar argument extends for *Case II*. Also, note that since \(a_{ij} = a_{ji}\), we use these terms interchangeably.

7.2.2 Mobile tracking costs

The cost of tracking a mobile includes the costs incurred during power-up, move, and power-down procedures. We first characterize these costs, and then compute the average move cost.
7.2.2.1 Tracking cost in the mobile PNNI scheme

Since the tracking procedure in this scheme uses the PNNI routing protocol for sending reachability updates, we first quantify the cost of a reachability update, $U_K$ (defined in Table 10). The cost of updating reachability data for a mobile using the PNNI routing protocol at all nodes whose ancestors-are-siblings at levels $l$ such that $l \geq K$ is given below:

$$U_K = m_K + \sum_{i = K + 1}^{L} m_i \left( \prod_{j = K + 1}^{i} (m_{j-1} + 1) \right) = \left( \prod_{i = K}^{L} (m_i + 1) \right) - 1. \quad \text{(EQ 26)}$$

(EQ 26) can be explained as follows. We assume that the PTSPs flooded in a peer group are routed on an MST (Minimum Spanning Tree). Though peer groups at the same level may have MSTs of different lengths, we make a pessimistic assumption that all peer groups at a given level have MSTs of the same length as the peer group with the longest MST. If the MST of a peer group at level $i$ is of length $m_i$, then the cost of sending a PTSP with the updated reachability information within a peer group of level $i$ is $m_i$. This explains the first term in (EQ 26) for the topmost level peer group witnessing this update ($K$). The cost of updating all other peer groups from levels $K + 1$ to the lowest level $L$ is the second term in (EQ 26).

The costs of individual tracking procedures in the mobile PNNI scheme are summarized in Table 11. The cost of tracking a mobile in the mobile PNNI scheme includes the cost of updating reachability data for a mobile, and the cost of sending messages to the home (and old) locations of the mobile to set forwarding pointers. The
cost of sending a message from node \( i \) is assumed to be 1 if the recipient is within the neighborhood \( G_i \) (as defined in Table 10). If the recipient is outside the neighborhood \( G_i \), the cost is \( h \), where \( h > 1 \).

The cost of updating reachability information depends upon the relationship between the ancestors-are-siblings levels, and the scope parameter \( S \). If a mobile powers on such that \( a_{hv} \geq S \) (where \( a_{hv} \) is the ancestors-are-siblings level as defined in Table 10 for the home and visiting nodes of the mobile), then reachability updates are sent to only to those nodes whose siblings are ancestors at the \( a_{hv} \) level or at a lower level in the hierarchy. This explains the reachability update cost \( U_{a_{hv}} \) shown in Table 11 for this case, where

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Relative locations</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power on</td>
<td>( a_{hv} \geq S )</td>
<td>( U_{a_{hv}} + 1 )</td>
</tr>
<tr>
<td></td>
<td>( a_{hv} &lt; S )</td>
<td>( U_S + h )</td>
</tr>
<tr>
<td>Move</td>
<td>( a_{on} \geq S )</td>
<td>( U_{a_{on}} + 1 )</td>
</tr>
<tr>
<td></td>
<td>( a_{on} &lt; S ) ( a_{hn} \geq S )</td>
<td>( U_S + U_{hn} + h + 1 )</td>
</tr>
<tr>
<td></td>
<td>( a_{on} &lt; S ) ( a_{hn} &lt; S ) ( a_{oh} \geq S )</td>
<td>( U_S + U_{oh} + 2h )</td>
</tr>
<tr>
<td></td>
<td>( a_{on} &lt; S ) ( a_{hn} &lt; S ) ( a_{oh} &lt; S )</td>
<td>( 2U_S + 2h )</td>
</tr>
<tr>
<td>Power off</td>
<td>( a_{hv} \geq S )</td>
<td>( U_{a_{hv}} + 1 )</td>
</tr>
<tr>
<td></td>
<td>( a_{hv} &lt; S )</td>
<td>( U_S + h )</td>
</tr>
</tbody>
</table>

Table 11: Tracking costs in the mobile PNNI scheme
$U_{as}$ is given by (EQ 26). The cost of setting a forwarding pointer at the home location is assumed to be 1 since the home is in the neighborhood of the mobile ($a_{hv} \geq S$). The cost of the second row in Table 11 can be similarly reasoned.

Move costs depend upon the relative distances between the old and new locations of the mobile, and the home and the new locations of the mobile. Four cases are possible as shown in Table 11. Depending on the case, the reachability update cost depends upon one or more of the following: $a_{on}$, $a_{hn}$, $a_{oh}$ and $S$, where the subscripts $o$, $n$, and $h$ represent the old, new, and home location of the mobile, respectively. For example, in case $a_{on} < S$ $a_{hn} < S$ $a_{oh} \geq S$, the reachability update has to be sent to the entire neighborhood $G_n$ around the new switch. Reachability cancellations around the old switch need only propagate up to level $a_{oh}$ since other nodes within the neighborhood $G_o$ will continue storing the same reachability as before (pointing toward home for this mobile). Other cases can be similarly reasoned. Before generating these reachability updates, registrations are sent to set forwarding pointers. The terms $1$ and $h$ in the move costs shown in Table 11 represents the costs of these registrations, depending upon whether the registration has to be sent inside or outside the neighborhood of the new location. In cases when a mobile powers on/off at the home, it is not necessary to send an explicit registration message to home. We ignore this minor optimization in the analysis. The power-off costs are the same as power-on costs as shown in Table 11.

The average tracking cost per move in the mobile PNNI scheme ($\bar{M}_{PNNI}$) is given by (EQ 27).
The first two terms account for the first two rows of the move cost in Table 11. The third row of the move cost from Table 11 accounts for the last term of the equation. It is nonzero only if \( a_{nh} = a_{on} \). This is because if \( a_{hn} \neq a_{on} \), (EQ 24) shows that \( a_{oh} \) is the smaller of the two values. But since both values are smaller than \( S \), \( a_{oh} \geq S \) cannot occur.

For the case when \( a_{nh} = a_{oh} \), a multiplicative factor showing the conditional probability \( P\left( a_{oh} = j \mid a_{nh} = a_{on} = i \right) \) is needed, since \( a_{oh} \) is dependent on the values of \( a_{nh} \) and \( a_{on} \). The fourth row of the move cost shown in Table 11 corresponds to the third and fourth additive terms in (EQ 27). The third term shows the case when \( a_{nh} \neq a_{oh} \), and the fourth term shows the case when \( a_{nh} = a_{oh} \). The latter requires a conditional probability since it involves all three terms \( a_{on} \), \( a_{nh} \) and \( a_{oh} \). Expressions for the probabilities in (EQ 27) are derived in Section 7.5.
7.2.2.2 Tracking cost in the LR scheme

Mobile tracking costs in the LR scheme are shown in Table 12. Tracking in the LR scheme essentially requires updating pointers at LRs (see Figure 59). The cost of updating pointers at LRs up to and including a level $k$ LR is given by $R_k$ as defined in Table 10.

The cost of updating/querying LRs in the path from the level $L$ LR to the level $k$ LR is given by:

$$R_k = \sum_{i = k}^{L} c_i, \text{ where } c_i \text{ is defined in Table 10.} \quad \text{(EQ 28)}$$

The entries in Table 12 are explained as follows. When a mobile powers on/off, all LRs in the path up to level $S$ are updated, and a message is sent to the home LR. In cases when a mobile powers on/off at the home, it is not necessary to send an explicit
registration message to the home LR. We ignore this minor optimization in the analysis. The message to the home is treated as a long distance message of cost $h$ if the home LR is not in the neighborhood of the mobile.

Tracking cost during a move consists of the cost incurred to set up new pointers, delete old pointers, and send a message to the old switch, and/or to the LR of the home switch. When $a_{on} \geq S$, pointers are updated at LRs up to level $a_{on}$ relative to the new location, and deleted at LRs up to level $a_{on} + 1$ relative to the old location. When $a_{on} < S$, pointers are updated at LRs up to level $S$ relative to the new location, and deleted at LRs up to level $S$ relative to the old location. The cost of updating the home LR is given by $h$ or 1, depending on $a_{hn}$. The average tracking cost per move in the LR scheme, $\overline{M_{LR}}$, is given below:

\[
\overline{M_{LR}} = \sum_{i = 1}^{S-1} P(a_{on} = i) \left( R_{a_{on}} + R_{a_{on} + 1} + 1 \right) + \sum_{i = S}^{L} P(a_{on} = i) \left\{ \sum_{j = S}^{L} P(a_{hn} = j) \left( 2R_S + h + 1 \right) + \sum_{j = 1}^{S-1} P(a_{hn} = j) \left( 2R_S + 2h \right) \right\}
\]  
\text{(EQ 29)}

The above equation can be readily explained from the move costs shown in Table 12. We also refer to the cost $\overline{M_{LR}}$ as the "average move cost".

7.2.3 Mobile search costs

In this section, we define the average "search" cost incurred during call setup to a mobile. In the LR scheme, this is the cost of determining the mobile location, since this scheme has an explicit location phase. This cost may be mobile location delay or the signaling load (in Mb/s) incurred to send location queries and receive responses. On the
other hand, the search cost for the mobile PNNI scheme is more difficult to define. If bandwidth is of concern, the overhead of the scheme can be characterized by the average extra bandwidth required for the connections that are routed inefficiently. We first define the search costs for these schemes, and then formulate the average search costs in these schemes.

7.2.3.1 Search cost in the mobile PNNI scheme

Given that there is no explicit mobile location phase in the mobile PNNI scheme, we define the "search" cost in this scheme as the number of extra hops needed to route forwarded connections. In other words, for connections that are misrouted, the search cost in this scheme is

\[ S_{\text{misroute}} = D_{ch} + D_{hv} - D_{cv}, \]  

(EQ 30)

where \( D_{ij} \) is defined in Table 10 as the number of nodes on the route from node \( i \) to node \( j \), and the subscripts \( c, h, \) and \( v \) represent the calling party, home location of the called mobile, and visiting location of the called mobile, respectively. Some of the calls originating at nodes outside the neighborhood of the current location of a called mobile will be routed inefficiently. In addition, some calls originating within this neighborhood may also be routed inefficiently if the call setup request arrives soon after the called mobile moved, and the reachability update has not propagated to all the relevant nodes. For purposes of this analysis, we ignore this cost for two reasons. First, if the reachability update limiting level \( S \) is high (numerical value is large), reachability updates will presumably propagate fast, allowing us to ignore this cost. On the other hand, if \( S \) is low.

157
the mobile tracking costs become significant, thus allowing us to ignore search costs while comparing the total costs of the two schemes. The search cost in the mobile PNNI scheme is listed below:

\[
S_{\text{PNNI}} = \begin{cases} 
0 & a_{hv} \geq S \quad \text{independent of } a_{cv} \\
0 & a_{cv} \geq S \quad \text{independent of } a_{hv} \\
D_{ch} & a_{hv} < S \quad a_{cv} < S \quad a_{hv} < a_{ch} \\
D_{hv} & a_{hv} < S \quad a_{cv} < S \quad a_{ch} < a_{hv} \\
2D_{ch} - D_{cv} & a_{hv} < S \quad a_{cv} < S \quad a_{ch} < S \quad a_{ch} = a_{hv}
\end{cases} \quad (\text{EQ 31})
\]

The first case of (EQ 31) is justified because when a mobile is in its home neighborhood, the entire network has the correct reachability information about the mobile and hence no calls are misrouted. All the nodes within the neighborhood of the mobile have correct reachability information due to the reachability updates propagated as part of the mobile PNNI scheme. The summarized reachability information in the nodes outside the neighborhood of the mobile (default information) indicates that the mobile is in its home neighborhood. In the second case of (EQ 31), since the calling party is located in the neighborhood of the called mobile, the call is not misrouted.

The last three cases shown in (EQ 31) represent the cases where the mobile PNNI scheme may incur a search cost by setting up a connection that later requires route optimization. Pessimistically, we treat all such calls as misrouted calls. First, we provide a method for estimating \( D_{ij} \). The distance \( D_{ij} \) between nodes \( i \) and \( j \) is approximated as

\[
\gamma_{ij} = \prod_{k=a}^{b} p_{k} \quad (\text{EQ 32})
\]
where $p_k$ is the average (among all node pairs) length of the "shortest-path" between nodes of a peer group at level $k$. For the worst case performance, the maximum length of the "shortest-path" can be taken. By this definition, the distance $D_{ij}$ between nodes $i$ and $j$ depends on the value of their ancestors-are-siblings level $a_{ij}$.

From the property described in (EQ 24) and (EQ 25), we know the relationship between the ancestors-are-siblings level of different nodes. When $a_{hv} < a_{ch}$, the cost associated with misrouting, given by (EQ 30), $D_{ch} + D_{hv} - D_{cv}$ is equal to $D_{ch}$ since $a_{cv} = a_{hv}$. Similarly, when $a_{ch} < a_{hv}$, the cost associated with misrouting, given by (EQ 30), $D_{ch} + D_{hv} - D_{cv}$ is equal to $D_{hv}$. This case is illustrated in Figure 63:. It also demonstrates that this estimate of the cost is approximate because the border node in the peer group PG1 which receives the call setup may be closer to the home node than to the visiting node, making the exact number of extra hops in the misrouted connection different from $D_{hv}$. Finally, if $a_{hv} = a_{ch}$, then the search cost is $2D_{ch} - D_{cv}$ and $a_{cv} \geq a_{ch}$ (as per (EQ 25)).

The average search cost per call in the mobile PNNI scheme, $s_{PNNI}$ is given by (EQ 33). Since the search cost is 0 for $a_{hv} \geq S$, the first ($i^{th}$ index) summation is from 1 to $S - 1$. The first and the second additive terms are due to the third and fourth cases of (EQ 31). In the third case, $a_{hv} < a_{ch}$. Hence the $j^{th}$-index summation varies $a_{ch}$ from $i + 1$ to

![Figure 63: Distances between calling party, and home and visiting locations](image-url)
The condition that \( a_{cv} < S \) is automatically satisfied, since \( a_{cv} = a_{hv} \) as per (EQ 24), and \( a_{hv} < S \). Similarly, the fourth case of (EQ 31) requires only the probability distributions of \( a_{hv} \) and \( a_{ch} \). In this case, \( a_{ch} \) should be varied from 1 to \( i - 1 \) since \( a_{ch} < a_{hv} \). The last term of (EQ 33) originates from the second and fifth cases of (EQ 31). In the previous two terms, since the value of \( a_{cv} \) is determined from the values of \( a_{hv} \) and \( a_{ch} \), and these latter values were less than \( S \), the condition \( a_{cv} > S \) is always false. However, if \( a_{hv} = a_{ch} \), \( a_{cv} \geq a_{hv} \) allowing \( a_{cv} \) to vary from \( i \), where \( a_{hv} = a_{ch} = i \), to \( L \). The second case of (EQ 31) shows that if \( a_{cv} \geq S \), the search cost is 0. This implies that \( a_{cv} \) should only be varied from \( i \) to \( S - 1 \) as indicated in the last term in (EQ 33).

Expressions for the probabilities in (EQ 33) are derived in Section 7.5.

\[
\overline{S}_{PNNI} = \sum_{i=1}^{S-1} P(a_{hv} = i) \left( \sum_{j=i+1}^{L} P(a_{ch} = j) D_{ch} + \right)
\]

\[
+ P(a_{ch} = i) \left\{ \sum_{j=i}^{S-1} P(a_{cv} = j \mid a_{ch} = a_{hv} = i) \left( 2D_{ch} - D_{cv} \right) \right\}
\]

7.2.3.2 Search cost in the LR scheme

In the LR scheme, since there is an explicit mobile location phase, the search cost is the cost of sending LOCREQs (Location Requests) up and down the chain of LRs while locating a mobile. This search cost is given below:
\[ S_{LR} = \begin{cases} 
R_{a_{cv}} + R_{a_{cv}+1} + 1 & a_{cv} \geq S \quad \text{independent of } a_{hv} \\
2R_{s} + 2h + 1 & a_{cv} < S \\
2R_{s} + 1 + 2h & a_{cv} < S \\
2R_{s} + 3h & a_{cv} < S \\
\end{cases} \]  
(EQ 34)

where \( R_{i} \) is defined in (EQ 28).

When the calling party is located in the neighborhood of the called party (case 1 of (EQ 34)), search cost consists of the cost of querying LRs from the level \( L \) up to level \( a_{cv} \), and then down from the LR at level \( a_{cv} + 1 \) to the LR at the level \( L \) LR of the called mobile’s switch. These two costs correspond to the terms, \( R_{a_{cv}} \) and \( R_{a_{cv}+1} \), respectively. The cost of sending the final response directly from the level \( L \) LR to the calling party’s switch, as shown in Figure 60:, is 1, since the two ends of this response message are within the same neighborhood.

When the calling party is outside the neighborhood of the called party’s current location (cases 2, 3, and 4 of (EQ 34)), the relative location of the home address of the mobile and its current location becomes relevant. If it is in its home neighborhood (case 2), then the search propagates up to the level \( S \) LR from the calling party (at a cost \( R_{s} \)), which then sends a message to the home LR of the mobile (at a cost \( h \) since the home LR of the called mobile is outside the neighborhood of the calling party). The home LR sends a location request to the \( S^{th} \) level LR currently tracking the mobile at a cost of 1 unit (since both these LRs are in the same neighborhood). This is followed by a set of location requests which follow the trace of the pointers from the LR at level \( S \) to level \( L \) (at a cost \( R_{s} \)). The final reply costs \( h \) units since the LR at the visiting location of the mobile is outside the neighborhood of the calling party. Costs shown in cases 3 and 4 of (EQ 34) can be similarly reasoned.
The average search cost per call in the LR scheme ($S_{LR}$) is given by (EQ 35). The first two terms correspond to the first two cases, and the fifth term corresponds to the third case of (EQ 34). The third and the fourth terms represent the fourth case of (EQ 34).

\[
S_{LR} = \sum_{i=S}^{L} P(a_{cv} = i) \left( R_{a_{cv}} + R_{a_{cv}} + 1 + 1 \right) + \sum_{i=1}^{S-1} P(a_{cv} = i) \tag{EQ 35}
\]

\[
\left\{ \sum_{j=S}^{S-1} P(a_{hv} = j) \left( 2R_S + 2h + 1 \right) + \sum_{j=1, j \neq i}^{S-1} P(a_{hv} = j) \left( 2R_S + 3h \right) + P(a_{hv} = i) \left\{ \sum_{j=i}^{S-1} P(a_{ch} = j | a_{hv} = a_{cv} = i) \left( 2R_S + 3h \right) \right. \\
+ \sum_{j=S}^{L} P(a_{ch} = j | a_{cv} = a_{hv} = i) \left( 2R_S + 1 + 2h \right) \right\} \right\}
\]

7.2.4 Average total costs

The total cost of a location management scheme depends on the move cost and the search cost of that scheme. In order to be able to estimate the average total cost, the rate of call arrival at a mobile, $\lambda_c$, and the rate at which the mobile moves between base stations, $\lambda_m$, are needed. The average move cost per unit time, average search cost per unit time, and the average total cost per unit time, are given in (EQ 36), where $\bar{M}$ and $\bar{S}$ represent the average move cost and average search cost per call, respectively.

\[
\bar{M} = \lambda_m \bar{M} \quad \bar{S} = \lambda_c \bar{S} \quad T = \bar{M} + \bar{S} \tag{EQ 36}
\]

162
Since we do not have exact numbers for $\lambda_c$ and $\lambda_m$, but are interested in the impact of these parameters, we use the CMR (Call-to-Mobility Ratio), denoted $\rho$ (defined as the number of calls arrivals per move), and quantify the dependence of the average total costs on the CMR. The average total cost per move, $T^m$, for the two schemes is given by

$$T^m_{PNNI} = \overline{M_{PNNI}} + \rho \overline{S_{PNNI}} \quad T^m_{LR} = \overline{M_{LR}} + \rho \overline{S_{LR}} \quad (EQ \ 37)$$

where $\rho = \lambda_c / \lambda_m$, and $\overline{M_{PNNI}}$, $\overline{M_{LR}}$, $\overline{S_{PNNI}}$ and $\overline{S_{LR}}$ are given in (EQ 27), (EQ 29), (EQ 33), and (EQ 35), respectively.

7.2.5 Numerical results

In this section, we quantitatively compare the mobile PNNI scheme, and the LR scheme described in Section 7.1. The measures of comparison include the average move cost per mobile (derived in Section 7.2.2), the average search cost per call (derived in Section 7.2.3), and the average total cost per move (derived in Section 7.2.4). This analysis also provides insights into the effect of key parameters of these algorithms, such as $S$, the reachability update limiting scope, which is also the highest-level of the hierarchy of location registers in the LR scheme, $h$, the cost of “long-distance” signaling, and CMR (Call-to-Mobility Ratio).
Input Data: Values of input parameters assumed for this numerical computation are shown in Table 13 (see Table 10 for definitions of these parameters). We observe that the exact numerical results are dependent on the exact values chosen for these parameters. However, the trends observed are more or less independent of these values. Sensitivity of the comparative results to these input parameters has been studied.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_i$, $1 \leq i \leq L$</td>
<td>4</td>
</tr>
<tr>
<td>$p_i$, $1 \leq i \leq L$</td>
<td>2</td>
</tr>
<tr>
<td>$c_i$, $1 \leq i \leq L$</td>
<td>1</td>
</tr>
<tr>
<td>$N_b$</td>
<td>37</td>
</tr>
<tr>
<td>$L$</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 13: Input values

7.2.5.1 Comparison of the two schemes

Plots of (EQ 37), showing the variation of average total cost in the two schemes with CMR are given in Figure 64: This figure shows that the mobile PNNI scheme incurs a lower average total cost at high CMRs, while, at low CMRs, the LR scheme performs better. These plots depend on the value of $S$, which can potentially be different in the two schemes. For example, if the CMR is 0.02, the LR scheme when operated with $S = L - 1$ gives the lowest average total cost. But if CMR is 0.03 the mobile PNNI scheme should be chosen and operated with $S = L + 1$. We observe that the CMR at which the mobile PNNI scheme does better than the LR scheme is at 0.025 (we designate this the “break point CMR”).

164
Figure 64: Comparison of the average total costs of the two schemes

A second level of comparison is to understand the behavior of the two schemes relative to increasing CMR and $S$. Since $S$ is a parameter of the two algorithms, these results provide significant insight for the selection of this parameter. To study the effect of $S$ on the average total costs of the two schemes, consider the plots shown in Figure 65: The
Figure 65: Effect of $S$ on the average total costs in the two schemes 

Mobile PNNI plots demonstrate that at higher CMRs, a low value of $S$ should be selected, while at low CMRs, high values of $S$ should be chosen. This is illustrated in Figure 66: 

The opposite behavior is seen in the plots for the LR scheme.
In the mobile PNNI scheme, the move costs increase with a decrease in $S$, since reachability updates have to propagate to a wider area. Thus, if the CMR is low, a high value of $S$ should be chosen to limit the move costs. On the other hand, the search cost increases with an increase in $S$, since more calls are likely to be misrouted as the neighborhood containing the reachability information is small. For high CMRs, a low $S$ should be chosen to limit the search costs. Figure 65:(a) for the mobile PNNI scheme shows that the $S = L + 1$ plot offers the lowest average total cost at very low CMRs (0 to 0.49), the $S = L$ plot becomes the best (minimum average total cost) solution for the next range of CMRs (0.49 to 2.55), the $S = L - 1$ plot becomes the best for CMRs ranging from 2.55 to 6.13, and the trend continues.

A similar behavior is observed in the plots for the LR scheme, shown in Figure 65:(b) with the exception that the trends are in the opposite direction. In other words, the LR scheme executed at lower values of $S$ tend to perform better at lower CMRs, and at higher values of $S$, it tends to perform better at higher CMRs. The LR scheme plots in Figure 65:(b) show that for CMRs below 0.053, the $S = L - 1$ choice is a better one, while for CMRs above this value, the $S = L$ choice is better (this point has been chosen to illustrate the trend, though it is higher than the break point CMR). Unlike the mobile PNNI scheme, as CMR increases, larger values of $S$ should be chosen in the LR scheme.

<table>
<thead>
<tr>
<th>LR Scheme</th>
<th>Mobile PNNI Scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$</td>
<td>CMR</td>
</tr>
</tbody>
</table>

Figure 66: Contrasting behavior of the two schemes with varying $S$ and CMR
Interestingly, in the LR scheme plots, for some values of $S$ there does not exist a range of values of CMR where operating with that value of $S$ gives the minimum average total cost. For example in Figure 65:(b) plots $S = L - 2$ and the $S = L + 1$ plots incur higher costs at all values of CMR. This is explained by the effect of the parameter $h$, an important parameter excluded from the above discussion. Next we address the effect of this parameter.

7.2.5.2 Effect of key parameters $h$ and $S$

Variation of average costs with changing $h$ for different values of $S$: The value of $h$ affects the average move cost in both schemes (see (EQ 27) and (EQ 29)), and the average search cost in the LR scheme (see EQ 35), as seen in the plots shown in Figure 67: It does

![Figure 67: Effect of $h$ on the average costs in the two schemes](image)

168
not effect the search cost in the mobile PNNI scheme (see EQ 33). For large values of $S$, the move cost in both schemes is dominated by the cost of setting forwarding pointers (in the mobile PNNI scheme) or updating the home/old LR (in the LR scheme), and hence, $h$ becomes a more significant parameter (the average move plots corresponding to $S = L + 1$ in both schemes have a steep slope). If $S = 1$, there are no updates to the home LR of a mobile as a user moves, since the $S^{th}$ level LR never changes (there being only one $S^{th}$ level LR). Searches also do not require location queries to distant location registers, since every location query will be resolved during the upward propagation of queries. Hence, in Figure 67: the plots corresponding to $S = 1$, for the LR scheme, are flat.

Variation of average costs with changing $S$ for different values of $h$: From the plots in Figure 67:, we can make observations about the effect of $S$ on the average costs in the two schemes at different values of $h$ (note CMR is not involved in these plots since they show the average move and search costs and not the average total costs). Instead of selecting specific values of $h$ and showing the variation of the average costs with respect to $S$, we show that at different ranges of $h$, the average move and search costs in the two schemes follow certain trends. These trends are shown for the two schemes in Figure 68: and Figure 69: (arrows pointing upward indicate an increase in cost). These results provide important information in helping us select values of $S$ for a given value of $h$.

For the mobile PNNI scheme, Figure 68: shows that at "low and medium values" of $h$, the average move cost decreases with increasing $S$. This is observed in Figure 67:., which shows that up to $h = 7.5$, the average move cost incurred is consistently higher for lower values of $S$. However, at high values of $h$, Figure 68: shows that the average move cost decreases up to a value, and then increases as $S$ increases. This is seen in the mobile PNNI
Figure 68: Effect of change in $S$ on the average costs in the mobile PNNI scheme

average move plot of Figure 67:, where, for example, if $h = 12$, changing $S$ from $L - 1$ to $L$ to $L + 1$, causes the cost to first drop and then increase. The mobile PNNI search cost simply increases with $S$ irrespective of $h$ (seen from Figure 67: and shown in Figure 68:).

For the mobile PNNI scheme at high values of $h$, the move cost decreases with increase in $S$ up to a value, $S_{\text{max}}$, as shown in Figure 68:, beyond which it increases. Since the mobile PNNI scheme search cost increases monotonically with increasing $S$, to minimize the average total cost, for the mobile PNNI scheme, the values of $S$ should be chosen such that $S \leq S_{\text{max}}$. Such a statement cannot be made if $h$ is in the low or medium ranges, since the average move cost decreases with increasing $S$, while the average search cost increases with increasing $S$. The optimal value for $S$ is then determined by the value of CMR. For example, Figure 65: showed the variation of the average total costs with varying CMR for different values of $S$ at an operating point where $h$ is in the medium range ($h = 6$).

The LR costs show a slightly different trend. At “low” values of $h$, the LR scheme experiences decreasing average move and search costs with increasing $S$, at “high” values of $h$, both costs increase with increasing $S$, and at “medium” values of $h$, both costs first decrease and then increase, as shown in Figure 69:. This is seen in the LR scheme plots of
Figure 69: Effect of change in value of $S$ on average costs in the LR scheme

Figure 67.: The reason for such behavior is that if $h$ is small, the LR scheme should be operated more like the "flat" scheme, by choosing a large $S$. In other words, all nodes know the home LRs of mobiles and directly send registrations and location queries to the home LRs. At large values of $h$, the LR scheme should be operated more like the "hierarchical" scheme, by making $S$ equal to 1, since the best result (lowest average total cost) is obtained at the smallest value of $S$. For the medium range of $h$ (such as $h = 6$, for which we provided detailed plots in Figure 65:), the optimal value of $S$ depends upon the CMR. In this range, the average move cost decreases with increasing $S$ up to $S_{min}$, and the search cost decreases up to a value of $S = S_{max}$. The minimum value of average total cost is obtained for some value of $S$ that lies between $S_{min}$ and $S_{max}$. For example, in Figure 65:(b), $S_{min} = L - 1$ and $S_{max} = L$.

In summary, there are three important parameters, $S$, the reachability update limiting scope, which is also the highest-level of hierarchy in the LR scheme, $h$, the cost of "long-distance" signaling, and CMR (Call-to-Mobility ratio). As to which location management scheme incurs a lower average total cost, the mobile PNNI scheme or the LR scheme, depends on these three parameters. Typically, at low values of $h$, operating the LR scheme...
at high values of $S$, leads to a lower average total cost than the PNNI scheme, because at high values of $S$, while the mobile PNNI scheme does not incur a move cost, it does incur a search cost, while in the LR scheme, by virtue of $h$ being low, both move and search costs are small. On the other hand, if the cost of long-distance signaling $h$ is high, either LR scheme or the mobile PNNI scheme could lead to minimal average total cost, provided the correct value of $S$ is selected. For the mobile PNNI scheme, this depends on the CMR expected, but in the LR scheme, one needs to select a low $S$ (preferably $S = 1$). Finally, if $h$ is in the medium range (which we expect will be the range of operation), there will be a break-point CMR below which the LR scheme will perform better, and above which the PNNI mobile scheme will incur lower costs. A significant point to note is that for a number of cases, for example, when mobiles are located close to their home locations (which we expect will be a high-percentage), or if the calling party is close to the visiting location of the called mobile, the mobile PNNI scheme incurs a zero search cost. This leads to the mobile PNNI scheme performing better in most regions of operation expected in low-tier (i.e., slowly moving users) PCS applications.

7.3 Qualitative comparisons

In this section qualitative comparison of the integrated scheme (mobile PNNI scheme) and the LR scheme is presented. The LR scheme requires the partitioning of the address space between fixed and mobile terminals, where as in the mobile PNNI scheme such a partitioning is not required. It is desirable to use a scheme in which there is no address partitioning. The mobile PNNI scheme could result in a connection path routed suboptimally, where as in the LR scheme the connection path is routed via the best path (at the expense of an additional phase of querying). Connection setup latency tends to be
lower in the mobile PNNI than the LR scheme. Implementing the mobile PNNI scheme or other integrated approach schemes requires considerable modifications to the existing signaling/routing protocols. Whereas LR scheme can be implemented without considerable modifications to the signalling and routing protocols but requires additional overlay network.

7.4 Summary

This Chapter described a generalized overlay scheme called the LR scheme. The LR (Location Registers) scheme introduces location registers (such as the cellular home and visitor location registers) into the PNNI standards-based hierarchical networks. This scheme uses a hierarchical arrangement of location registers with the hierarchy limited to a certain level $S$. It also requires the update of home and old location registers at a cost $h$.

Analytical models were set up to compare the average move, search, and total costs per move, of these LR scheme and the mobile PNNI scheme for different values of the CMR (Call-to-Mobility Ratio), and to provide guidelines for selecting the critical parameters of the algorithms. Results showed that at low CMRs (CMR < 0.025), the LR scheme performs better than the mobile PNNI scheme. We also observed that the two schemes show a contrasting behavior in terms of the value to be used for the parameter $S$ to achieve the least average total cost. For the mobile PNNI scheme, the parameter $S$ should be high at low CMRs (within the range in which the mobile PNNI scheme should be used), and low at high CMRs. However, in the LR scheme, the parameter $S$ should be low at low CMRs (within the range in which the LR scheme should be used), and high for high CMRs. These observations are made for a region of operation in which $h$, the cost of
setting forwarding pointers and updating distant LRs, is of medium value. If this cost is low, the LR scheme always outperforms the mobile PNNI scheme. For other ranges of \( h \), the scheme selected, and the \( S \) at which the scheme is operated, depends on the CMR.

### 7.5 Probability Distributions

In this section, we describe the probability distributions, \( P(a_{cv}) \), \( P(a_{ch}) \), \( P(a_{hv}) \), \( P(a_{hn}) \), \( P(a_{ho}) \), \( P(a_{on}) \) and the conditional probabilities used to compute the average costs in (EQ 27), (EQ 29), (EQ 33), and (EQ 35).

We assume that the distribution of calls to a mobile follows a uniform distribution, i.e., it is equally probable for a mobile to receive a call originating at any switch. So the probability of a call originating from a switch for which \( a_{cv} = x \) depends on the number of nodes, for which \( a_{cv} = x \) and the total number of nodes in the network. Thus, the probability \( P(a_{cv} = x) \) is given by:

\[
P(a_{cv} = x) = \frac{\left( \prod_{i=x+1}^{L} (m_i + 1) \right) m_x}{\prod_{i=1}^{L} (m_i + 1)}
\]  

(EQ 38)

This distribution makes long distance calls very highly probable. For e.g., \( (a_{cv} = 1) \approx 0 \) for \( L = 10, m_i = 4 \). This, in effect makes our cost estimates pessimistic. Since a call can originate from anywhere, the distribution of \( a_{ch} \) is the same as the distribution of \( a_{cv} \).

Next, we consider the distribution of \( a_{hv} \). The choice of this distribution is based on the premise that a majority of the mobiles roam in and around their respective homes. We choose a distribution such that the probability of being located close to home is high. The
distribution is such that the probability of being located at the home switch is \( f \), in the \( L^{th} \) level peer group of the home switch as \( f^2 \), in the \( (L-1)^{th} \) peer group as \( f^3 \), etc. Given that a mobile is located somewhere in the network,

\[
f + f^2 + f^3 + \ldots + f^{L+1} = 1 \quad \text{or} \quad \frac{1 - f^{L+1}}{1 - f} = 2 \quad \text{(EQ 39)}
\]

For \( L = 8, f = 0.5 \). This can be interpreted as 50\% of users being in their home registration area. This is also somewhat pessimistic, since majority of the users are typically in and around their home location. Under this assumption, the distribution of \( a_{hv} \) is given by (EQ 40).

\[
P (a_{hv} = x) = f^{L-x+2} \quad \text{(EQ 40)}
\]

The probability distributions of \( a_{oh} \) and \( a_{hn} \) are assumed to be same as the distribution of \( a_{hv} \).

In order to determine the probability distribution of \( a_{on} \), we assume that the base stations under a peer node of a peer group are arranged in a hexagonal fashion. We model each peer node as a macrocell in which all the base stations under the peer node are arranged in a hexagonal fashion. The results from [20] are used to approximate \( P (a_{on}) \).

Base stations in a macrocell are arranged in layers, with the \( j^{th} \) layer base stations arranged around the \( j-1^{th} \) layer base stations. A macrocell of \( i \) layers with all its base stations arranged in a hexagonal fashion has \( 3i^2 - 3i + 1 \) base stations in it. The layers are numbered 0 through \( i-1 \) with layer \( j \) \((j > 0)\) having \( 6j \) base stations in it [20]. For example, a 3 layer macrocell consists of 19 cells, with layer 0, layer 1 and layer 2 having 1, 6, and 12 base stations respectively. The probability that a mobile in an \( i-1^{th} \) layer base station moves to an \( i^{th} \) layer base station, \( P (i-1 \rightarrow i) \) \( (i.e., \) out of the macrocell, since an layer-\( i \) macrocell has layers numbered 0 to \( i-1 \)) is given by (EQ 41) [20].
Let \( N(k) \) be the number of base stations in a peer node of level \( k \). \( N(k) \) is given by (EQ 42), where \( N_b \) is the number of base stations under a switch (peer group of level \( L+1 \)).

\[
N(k) = \left( \prod_{i=K+1}^{L} (m_i + 1) \right) N_b
\]  

(EQ 42)

The total number of base stations in a \( k \)-level macrocell is also given by \( 3 \times (I(k))^2 - (3 \times I(k)) + 1 \), if we arrange the \( N(k) \) base stations in this macrocell as a hexagon of \( I(k) \) layers. Thus

\[
3 \times (I(k))^2 - (3 \times I(k)) + 1 = N(k)
\]  

(EQ 43)

Solving (EQ 43) for its positive root, we obtain \( I(k) \).

\[
I(k) = \frac{3 + \sqrt{9 + 12(N(k) - 1)}}{6}
\]  

(EQ 44)

"Border" base stations are base stations that lie on the boundary of the macrocell (i.e. base stations from which a user can move to other registration areas). Let \( B(k) \) denote the number of border base stations in a level \( k \) peer group. Since there are \( I(k) \) layers in a level-\( k \) peer node and the layers are numbered 0 to \( I(k) - 1 \), the border base stations of a level-\( k \) peer node are the base stations in level \( I(k) - 1 \). Thus, there are \( 6(I(k) - 1) \) border base stations for level-\( k \) peer node as shown in (EQ 45).

\[
B(k) = 6(I(k) - 1)
\]  

(EQ 45)

Probability of moving out of a level-\( k \) peer group, \( P(k) \), can be approximated as
\[
P(k) = \frac{B(k)}{N(k)} (P(I(k) - 1) \rightarrow I(k)) \quad (\text{EQ} \ 46)
\]

Assuming that each level-\(k\) peer node contributes equally to the border base stations of level-(\(k - 1\)) peer node, the probability distribution of \(a_{on}\) is obtained from \(P(k)\) and \(P(k - 1)\) and is given in (EQ 47).

\[
P(a_{on} = k) = P(k) - P(k - 1) \quad (\text{EQ} \ 47)
\]

The conditional probability \(P\left( a_{ch} = j \mid a_{cv} = a_{hv} = k \right)\) is given by (EQ 48).

\[
P\left( a_{ch} = j \mid a_{cv} = a_{hv} = k \right) = \frac{\prod_{i=j+1}^{L} (m_i + 1) m_j}{\prod_{i=k}^{L} (m_i + 1)} \quad (\text{EQ} \ 48)
\]

Conditional probability \(P\left( a_{oh} = j \mid a_{nh} = a_{on} = i \right)\) is given by (EQ 49).

\[
P\left( a_{oh} = j \mid a_{nh} = a_{on} = i \right) = \frac{P\left( a_{oh} = j \right)}{\sum_{k=i}^{L} P\left( a_{nh} = k \right)} \quad (\text{EQ} \ 49)
\]
In mobile communication networks, handoff procedures and location management procedures are needed to support user mobility. Handoff procedures are needed to reroute connections on which the mobile user is communicating while moving. Location management consists of tracking mobiles and locating them for incoming call deliveries. In both these sets of procedures, paths taken by connections could become “suboptimal\(^1\).” For example, most handoff schemes propose performing a local connection reroute rather than an end-to-end connection reroute to keep handoff latencies low. Such reroute operations could result in making the connection path suboptimal. Similarly, location management schemes that propose setting up the connection to the home location of a mobile and then rerouting the connection to the mobile’s current location based on location data provided by the home node could result in suboptimal connection paths. In other words, suboptimality that is introduced during connection setup is primarily due to the lack of information about the exact location of the mobile at the call originating.

\(^1\)The path of a connection is classified as being “suboptimal” if it is not the shortest path between the two endpoints of the connection.
switch, and suboptimality occurs after connection setup because of user movements by a communicating mobile. It is important to optimize routes of connections since suboptimal paths imply an inefficient usage of network resources.

In Chapter 5, we proposed an algorithm for optimizing the route of a connection and applied it to the location management problem. This Chapter presents a significantly modified version of that algorithm and applies it to perform route optimization after a suboptimal connection is setup or after a fast handoff.

Prior work on handoff management and route optimization is presented in Section 8.1. For prior work on location management refer to Chapter 3. In Section 8.2, we state the "base" rerouting problem, and provide a solution to this problem in Section 8.3. Section 8.4 describes how this solution can be used to perform route optimization after single and multiple handoffs and after a two-phase connection setup. Finally, Section 8.5 presents results of a comparative performance analysis, and Section 8.6 provides a summary of the Chapter.

8.1 Background

In this section, we summarize prior work on handoff management procedures that cause suboptimal routes and prior work on route optimization.

8.1.1 Prior work on handoff management

In this section, we briefly describe handoff management schemes that result in connections with suboptimal routes, thus motivating the need for route optimization.

Handoff management consists of procedures to reroute connections on which a mobile user is communicating while moving. Various handoff schemes have been proposed in [86-91]. [24] proposed a generalized handoff scheme, which allows for the
implementation of most of the handoff schemes proposed in [86-91]. It is based on the observation that by changing the node serving as the *CrossOver Switch* (COS), from which point connections are rerouted to the mobile at its new location, different handoff schemes can be realized. Figure 70 shows a mobile moving from base station I to base station II. A COS selection procedure is used to select a crossover switch. In Figure 70, this is determined to be ATM Switch III. A new segment is set up from the COS to the new base station. Figure 71 shows that if the COS is chosen to be the far end switch on the connection, handoff can be accomplished in *one-phase*, i.e., an optimal route is achieved.
requiring no further route optimization after the completion of handoff. Clearly, handoff latency becomes an issue with this scheme. At the other extreme, if the COS is statically assigned to be the old base station, low handoff latencies can be achieved if adjacent base stations are interconnected with direct physical or logical links. But the path taken by the rerouted connection could become suboptimal. The path extension scheme, anchor switch scheme, and dynamic COS search schemes shown in Figure 71 lie in-between these extremes. In the path extension scheme, the COS is assigned to be the switch attached to the old base station. In the anchor switch scheme, the COS is statically assigned to be the switch in whose domain the mobile was located at the time of call delivery. In the dynamic COS search scheme, the COS is selected dynamically with the search typically being limited to a local area for fast handoffs. Figure 71 classifies the dynamic COS search and the static COS assignment schemes as two-phase schemes since they require route optimization after the completion of the handoff. We believe that two-phase handoff schemes will be needed due to considerations of handoff latency. Hence, in this Chapter we propose a scheme to optimize the route of a connection that becomes suboptimal due to handoffs. From the spectrum of schemes shown in Figure 71, we observe that if the dynamic COS search scheme is applied to find the optimal COS, such a scheme can be classified as a one-phase dynamic COS search scheme. The proposed one-phase dynamic COS search scheme will be described in Chapter 9.

8.1.2 Prior work on route optimization

A route optimization procedure has been proposed for IP (Internet Protocol) networks to work in conjunction with mobile IP [13, 14]. Mobile IP is an extension to the Internet Protocol (IP), which enables hosts to change their point of attachment to the Internet.
without changing their IP addresses. In this protocol, a packet destined to a mobile host is routed to the home network of the mobile as identified by its permanent IP address. The home network tracks the current location of the mobile and tunnels the packet to the current network of the mobile. As an improvement to this "triangular" routing, route optimization extensions have been proposed [14]. These extensions provide a means for communicating nodes to maintain a binding between the mobile and its current location, and use this binding to tunnel datagrams directly to the mobile. Extensions are also provided to allow datagrams in flight when a mobile node moves, and datagrams sent based on an outdated binding information, to be forwarded directly to the mobile. The primary difference between the route optimization extensions proposed in [13] and the route optimization procedure being proposed in this Chapter is that the former extensions have been proposed for the connectionless Internet Protocol, while the latter is being proposed for connection-oriented networks (PNNI-based ATM networks). In ATM networks, in order to reroute a connection, a new connection segment needs to be set up between a "crossover" node and one of the ends of the connection, and the user data needs to be switched from the old path to the new path without loss of ATM cell sequence.

8.2 Problem statement

In this section, we state the base rerouting problem route optimization problem (also referred to as the base route optimization problem) and demonstrate how the route optimization problems created in the handoff management and location management schemes can be mapped to the base rerouting problem. The solution to this problem is presented in Section 8.3. The base rerouting problem is stated below.
Consider a connection between two end points $f$ and $o$ in Figure 72 (where $f$ is used to denote the "far end" and $o$ is used to denote the "old point"). The problem is to reroute the connection $f - o$ to create a connection from node $f$ to node $t$ (target node). This problem is referred to as the base rerouting problem or the base route optimization problem. Next, we demonstrate how the route optimization problems created in the handoff management schemes can be mapped to this base problem.

The suboptimality resulting from the two-phase handoff schemes shown in Figure 71: is illustrated in Figure 72, where node $c$ is the constant end\(^2\), node $o$ is the COS, and node $n$ is the switch attached to the new base station (or the new base station). In the case of path extension scheme (see Figure 71), the handoff procedure is performed by extending the path from the old switch to the new switch. Therefore, node $o$ corresponds to the switch attached to the old base station, also referred to as the old switch. In this case, the

---

\(^2\)In order to make the description of the schemes easier to understand, we explain the schemes with the assumption that the mobile is communicating with a fixed terminal. It should be noted that these schemes are general and can be applied to the case where both communicating terminals are mobiles.
suboptimal path from c to o (old switch) to n (new switch) needs to be rerouted to form an optimal path from c to n by either rerouting the c-o segment toward the target node n (case A in Figure 72), or by rerouting the o-n segment toward the target node c (case B in Figure 72). In either case, the problem reduces to that illustrated in Figure 72, where a connection from node f (far end) to node o (old node) needs to be rerouted to create a connection from node f to node t (target node).

In the two-phase anchor switch scheme shown in Figure 71, the COS, i.e., node o in Figure 72, is the anchor switch (the switch in whose domain the mobile was located at the time of call delivery). In the base-station-based handoff scheme, node o is the old base station, and in the two-phase dynamic COS scheme, node o is a local switch through which the connection was rerouted during the fast handoff. In all the above schemes, the route optimization problem reduces to the problem of Figure 72 after identifying whether the segment c-o or o-n needs to be rerouted to the corresponding target node (n or c, respectively).

Table 14 summarizes how the rerouting problem created in one- and two-phase handoff management schemes can be mapped to the base rerouting problem shown in Figure 72. In the two-phase schemes, after the completion of the first phase (handoff procedure) rerouting is needed on one of the two segments. Definition of the "closeness" of two nodes needed to select between case A and case B is provided in Section 8.4.1.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Case</th>
<th>f</th>
<th>o</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-phase schemes</td>
<td>o is &quot;closer&quot; to c than n</td>
<td>n</td>
<td>o</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>o is &quot;closer&quot; to n than c</td>
<td>c</td>
<td>o</td>
<td>n</td>
</tr>
<tr>
<td>One-phase dynamic COS scheme</td>
<td>for all cases</td>
<td>c</td>
<td>o</td>
<td>n</td>
</tr>
</tbody>
</table>

Table 14: Mapping to the common rerouting problem
The suboptimality resulting from the call forwarding location management scheme is illustrated in Figure 73. The suboptimal path, from \( c \) (calling party's switch) to \( h \) (home switch) to \( v \) (visiting location of the mobile) needs to be rerouted to form an optimal path from \( c \) to \( v \) by either rerouting the \( c-h \) segment toward the target node \( v \) (case A in Figure 73), or by rerouting the \( h-v \) segment toward the target node \( c \) (case B in Figure 73). In either case, the problem reduces to that illustrated in Figure 72, where a connection from node \( f \) (far-end) to node \( o \) (old node) needs to be rerouted to create a connection from node \( f \) to node \( t \) (target node).

In the two-phase crankback scheme, the resulting connection is similar to that shown in Figure 73, except that the connection is routed through some local node \( l \) instead of the home node \( h \), where the local node \( l \) represents the point up to which the call was cranked back during call setup (first phase). This problem again reduces to the problem of Figure 72 after identifying whether the segment \( c-l \) or \( l-v \) needs to be rerouted to the corresponding target node (\( v \) or \( c \), respectively).
Table 15 summarizes how the rerouting problem created in the three location
management schemes can be mapped to the base rerouting problem shown in Figure 72.

Next, we address the issue of selecting the node that initiates the route optimization procedure. To determine the crossover node (see Figure 72), the network needs to compute the shortest path between nodes \( f \) and \( t \), and then determine the overlap between the new path \( f - t \) and the old path \( f - o \). In the rerouting required for the suboptimal paths caused by the two-phase schemes, any of nodes \( f \), \( o \), or \( t \) could initiate the route optimization procedures. As shown in Table 14, the assignment of nodes (\( c \), \( o \), and \( n \)) in the two-phase schemes to nodes \( f \), \( o \) and \( t \) is such that \( o \) and \( t \) are close while \( f \) is the distant node (hence the term "far-end"). This implies that selecting either nodes \( o \) or \( t \) as the initiating node will lead to smaller processing delays. In this scenario, the target node can be selected as the route optimization initiating node only if the detailed path of the existing connection is available. As mentioned earlier, it is possible to use the base rerouting algorithm to reroute during a handoff (one-phase dynamic COS search scheme). In the one-phase dynamic COS scheme, since it is desirable to keep the handoff latency low, it

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Case</th>
<th>( f )</th>
<th>( o )</th>
<th>( t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Call forwarding</td>
<td>( A ) (( h ) is &quot;closer&quot; to ( v ) than ( c ))</td>
<td>( c )</td>
<td>( h )</td>
<td>( v )</td>
</tr>
<tr>
<td></td>
<td>( B ) (( h ) is &quot;closer&quot; to ( c ) than ( v ))</td>
<td>( v )</td>
<td>( h )</td>
<td>( c )</td>
</tr>
<tr>
<td>Two-phase crankback</td>
<td>( l ) is &quot;closer&quot; to ( c ) than ( v )</td>
<td>( v )</td>
<td>( l )</td>
<td>( c )</td>
</tr>
<tr>
<td></td>
<td>( l ) is &quot;closer&quot; to ( v ) than ( c )</td>
<td>( c )</td>
<td>( l )</td>
<td>( v )</td>
</tr>
<tr>
<td>One-phase crank back</td>
<td>for all cases ( (o, c, h, n) )</td>
<td>( c )</td>
<td>( h )</td>
<td>( v )</td>
</tr>
</tbody>
</table>

Table 15: Mapping to the common rerouting problem during connection setup
becomes necessary for node o or r to initiate the route optimization procedure. The choice of the nodes between o and r depends on several factors such as type of handoff (backward or forward), and availability of existing path information. This procedure is described in Chapter 9.

8.3 Route optimization algorithm

In this section, we present a solution to the base rerouting problem. There are four steps involved in solving the base rerouting problem shown in Figure 72 (see Figure 74),

(i) determining a crossover node p, (ii) establishing a new segment between p and t, (iii) switching user data from the old segment to the new segment, and (iv) releasing the old segment from p to o. Procedure for switching of user data from the old segment to the new segment depends on the scenario in which the route optimization is applied and is therefore discussed in Section 8.4. The fourth step of releasing the old segment can be performed using the standard ATM signaling procedures. Next, we present a solution to
the problem of determining a crossover node $p$ for the base rerouting problem illustrated in Figure 72, and discuss how the new segment can be set up. Ideally, the crossover node $p$ should be such that the path $f - p - t$ is the shortest path between $f$ and $t$, while at the same time, there should be a maximal overlap in the paths of the old connection $f - o$ and the new connection $f - t$. We refer to such a crossover node as the *minimal crossover node* for the two paths $f - o$ and $f - t$.

While finding the minimal crossover node has both the advantages of minimizing the resources required by the new connection and minimizing the new segment setup/old segment release overhead, there are certain constraints in the PNNI standard (in its current form) that do not allow for the determination of this node. In the PNNI standard, there is currently

1. no requirement mandating that all nodes retain the hierarchical path of a connection after it is established, and
2. a restriction that only nodes that created a DTL (Designated Transit List) can change that DTL.

The minimal crossover node determined under these constraints is defined as the *optimal crossover node* since it is a node $p$ such that $f - p - t$ is an optimal path, and as much overlap as is possible between the old and new paths is achieved within the constraints of the PNNI standard.

Thus, we have defined two types of crossover nodes, the optimal crossover node and the minimal crossover node. We presented our solution for determining the optimal crossover node in Chapter 5. In this Section, we present a modified procedure for determining the minimal crossover node under the assumption that it is possible to make minor modifications to the PNNI standard to allow for the relaxation of the above-listed
constraints. The relationship between the optimal and minimal crossover node determination procedures is illustrated in Figure 75. It shows that if either the hierarchical path information is not available or only the nodes that created the DTL can change the DTL, the procedure terminates after determining the optimal crossover node. Otherwise, the procedure to determine the minimal crossover node is also executed. Figure 75: also shows the details of the two procedures. The notation used in the description of these procedures is listed in Table 16.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{ij}$</td>
<td>Ancestors-are-siblings peer group of two nodes $i$ and $j$ is the lowest peer group in the hierarchy at which ancestors of both nodes belong to the same peer group. For example, in Figure 9, ancestors-are-siblings peer group of nodes A.1.1 and A.2.1 is A.</td>
</tr>
<tr>
<td>$a_{ij}$</td>
<td>The ancestors-are-siblings level of two nodes $i$ and $j$, is the level at which the ancestors of the two nodes $i$ and $j$ belong to the same peer group, i.e., the level of their ancestors-are-siblings peer group. For example, in Figure 9, the ancestors-are-siblings level of A.1.1 and A.2.1 is 2, since the level of peer group A is 2.</td>
</tr>
<tr>
<td>$LGN^\text{phy}_l$</td>
<td>Logical group node ancestor of the physical switch $\text{phy}$ at level $l$. For example, in Figure 9, the $LGN^A_{1.1}$ is A.</td>
</tr>
<tr>
<td>$x_l$</td>
<td>Minimal crossover peergroup of level $l$. $x_L$ (value of $l$ is $L$) is the minimal crossover node.</td>
</tr>
<tr>
<td>$e^o_l$</td>
<td>Egress border node of the peergroup $x_{l-1}$ on the old path. The value of $e^o_{a_{o_{a_{o}}}}$ is $o$ (old node).</td>
</tr>
<tr>
<td>$e^n_l$</td>
<td>Ingress border node of the peergroup $x_{l-1}$ on the new segment. The value of $e^n_{a_{n_{a_{n}}}}$ is $t$ (target node).</td>
</tr>
<tr>
<td>$E^\text{old}_l$</td>
<td>Logical group node ancestor of $e$ at level $l$, i.e., $LGN^e_l$.</td>
</tr>
</tbody>
</table>

Table 16: Notation
Optimal crossover node determination

Yes

\[ a_{ol} < a_{ol} \]

\[ p = \text{ingress border node of } P_{ol} \]

No

\[ p = f \]

Is hierarchical path available?

Yes

Can any node modify the DTLs?

No

Node \( P \) is the optimal crossover node.
Set up the new segment

\[ l = a_{ol}, e^n = t, e^o = o, i = p \]

Minimal crossover node determination

Compute at \( e^n \)

\[ I_l = LGN_l, E_{l}^{old} = LGN_l, E_{l}^{new} = LGN_l \]

Compute \( P_{l}^{new} \), the shortest path between \( I_l \) and \( E_{l}^{new} \)

Extract \( P_{l}^{old} \), the existing path between \( I_l \) and \( E_{l}^{old} \)

Determine \( x_l \), the intersection between \( P_{l}^{new} \) and \( P_{l}^{old} \)

False

\[ l = L? \]

True

\[ l = l + 1 \]

\[ i = \text{ingress border node of } x_{l-1} \] on the old path.
\[ e^o_h = \text{egress border node of } x_{l-1} \] on old path.
\[ e^n = \text{ingress border node of } x_{l-1} \] on new path.
Set up connection till \( e^n \)

Node \( x_l \) is the minimal crossover node.
Set up the connection between \( e^n \) to \( x_l \)

Node \( x_l \) is the minimal crossover node.
Set up the segment between and

\[ E_{l}^{old} = LGN_l, x^o, I_l = LGN_l \]

Compute \( P_{l}^{new} \), the shortest path between \( I_l \) and \( E_{l}^{new} \)

Extract \( P_{l}^{old} \), the existing path between \( I_l \) and \( E_{l}^{old} \)

Determine \( x_l \), the intersection between \( P_{l}^{new} \) and \( P_{l}^{old} \)

Determine \( n_l \), the node following the LGN \( x_l \) on the new segment. If \( x_l = E_{l}^{new}, n_l = \{ \} \).

True

\[ l = L? \]

False

\[ l = l + 1 \]

\[ i = \text{ingress border node of } x_{l-1} \] on the old path.
\[ e^o_h = \text{egress border node of } x_{l-1} \] on old path.
\[ e^n = \text{ingress border node of } x_{l-1} \] on new path.
Trace path till \( e^n \)

Figure 75: Determination of crossover nodes and new segment setup
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_l$</td>
<td>Node following the LGN $x_l$ in the new path $P_{l}^{new}$.</td>
</tr>
<tr>
<td>$i_l$</td>
<td>Ingress border node of $x_{l-1}$ on the old path. If $l$ is equal to $a_{ol}$, $i_l$ is the optimal crossover node $p$.</td>
</tr>
<tr>
<td>$I_l$</td>
<td>Logical group node ancestor of $i_l$ at level $l$, i.e., $LGN_{l}^{i_l}$.</td>
</tr>
<tr>
<td>$E_{l}^{new}$</td>
<td>Logical group node of ingress border node $e_l^n$ or a node adjacent to $x_{l-1}$ on the new path $P_{l-1}^{new}$.</td>
</tr>
<tr>
<td>$P_{l}^{old}$</td>
<td>Hierarchical path of the existing connection segment between the logical group nodes $I_l$ and $E_{l}^{old}$. This path is specified in terms of the logical groups nodes of level $l$.</td>
</tr>
<tr>
<td>$P_{l}^{new}$</td>
<td>Shortest path between the logical group nodes $I_l$ and $E_{l}^{new}$. This path is specified in terms of the logical groups nodes of level $l$.</td>
</tr>
</tbody>
</table>

Table 16: Notation

a. The numbering of levels in the hierarchy is such that lower level peer groups have higher level values assigned to them, i.e., the level number increases from top to bottom in the hierarchy.

The optimal crossover node determination algorithm shown in Figure 75 is explained in Chapter 5. Figure 75 presents a modified algorithm to determine the minimal crossover node. Procedures explained in Chapter 5 assume that the node $o$ initiates the crossover node determination procedure. As explained earlier node $o$ or node $t$ can initiate the crossover node determination procedure. Node $t$ can initiate the minimal crossover node determination only if information regarding the detailed path of the existing connection is available. If node $t$ initiates the minimal crossover node determination procedure, i.e., the detailed path of the existing path is available, minimal crossover node determination procedure can be executed while setting up the new segment. In other words, the new
segment setup is initiated by specifying the hierarchical source route to the minimal crossover peergroup. The ingress border node of this minimal crossover peergroup computes the next lower-level minimal crossover peer group (as explained above) and specifies the route internal to the minimal crossover peer group. This procedure is applied recursively until the minimal crossover node determined is a physical ATM switch. The details of the minimal crossover node determination algorithm are presented in the flowchart in Figure 75.

8.4 Application to handoff management

In this section, we integrate the route optimization algorithm described in Section 8.3 into the handoff management schemes described in Section 8.1.1. In order to make the description of the schemes easier to understand, we explain the schemes with the assumption that the mobile is communicating with a fixed terminal. It should be noted that these schemes are general and can be applied to the case where both communicating terminals are mobiles.

8.4.1 Second phase of two-phase handoff schemes

In this section, we present a route optimization procedure that can be used to reroute a connection after a fast handoff procedure is completed, as the second phase of two-phase handoff schemes. Instead of optimizing the route after every handoff, this route optimization procedure could be carried out after multiple handoffs. In the first phase, a fast local rerouting of the connection to the new location is performed. This could result in a connection that is routed suboptimally. In order to optimize the connection route that
resulted from the fast rerouting, a route optimization procedure is executed. The need for route optimization is dictated by several factors, such as optimality of the current path, duration of the call, QoS (Quality of Service) requirements of the connection, etc.

The route optimization procedure after handoff consists of four steps. First, the connection segment to reroute and the corresponding target node are selected. Second, a crossover node between the old and the new paths is determined, and a new segment is set up between that crossover node and the target node. Third, using “Tail” signals and buffering, user data is switched over from the old path to the new path. Fourth, the old segment is released using standard ATM signaling procedures.

**Step1: Connection segment and target node determination:** As illustrated in the two cases of Figure 72, the first task in optimizing the route of a connection locally rerouted through node o (see Figure 72:) is to identify which of the two segments, c - o or o - n needs to be rerouted. The corresponding target nodes for the rerouting operation become n and c, respectively.

Table 14 presents a method of choosing the connection segment and the target node based on the relative “closeness” of the nodes to each other. Ideally, the choice of the connection segment and the target node should be such that the length of the new connection segment is minimized, thereby, reducing the signaling load, route optimization delay, and the number of cells that need to be buffered (see Step 3). A definition of closeness (based on the assumption that the distance between two nodes is inversely proportional to their ancestors-are-siblings level) with the objective of minimizing the length of the new segment is as follows: For three nodes x, y and z, node x is closer to node y than to node z if the ancestors-are-sibling level of nodes x and z is less than or equal to the ancestors-are-sibling level of nodes x and y, i.e., \( a_{xz} \leq a_{xy} \). If \( a_{oc} < a_{on} \)
(Case I of Figure 76), the old switch is closer to the new switch of the mobile. Therefore, the new switch is selected as the target node and the corresponding connection segment to be rerouted is $c-o$. If $a_{oc} > a_{on}$ (Case IIa of Figure 76), the constant end switch is selected as the target node (the old switch is closer to $c$) and the corresponding connection segment to be rerouted is $o-n$. If $a_{oc} = a_{on}$ (Case IIb of Figure 76), either the new switch or node $c$ can be selected as the target node, with the corresponding connection segment to be rerouted being determined therein. It should be noted that Case I of Figure 76 is expected to occur frequently while rerouting during handoffs as majority of the movements occur from one base station to another that are typically “near” each other. Using this definition of closeness and the information given in Table 14, the connection segment $f-o$ and the target node $t$ can be determined. It should be noted that the optimality of the resulting connection path is independent of the choice of the target node.

### Figure 76: Connection after the handoff is completed

![Figure 76: Connection after the handoff is completed]

**Step 2: Determination of the crossover node and setting up of the new segment:** The crossover node is determined using the procedures described in Section 8.3. Using the target node (and the corresponding connection segment to be rerouted) determined in the first step described above, the mappings listed in Table 14 is used to determine the identities of nodes $f$, $o$ and $t$ needed for execution of the crossover node determination procedures.
If either hierarchical path information is not available or only the nodes that created the DTL can change the DTL, the procedure (see Figure 75) determines the optimal crossover node. In order to set up a new segment between the target node and the crossover node, the identity of the crossover switch is required. Therefore, the existing connection is traced (i.e., a COS_SELECT message is sent along the existing path) till the optimal crossover node and the new segment is then set up. If the hierarchical path information is available and any node can change the DTLs, the connection is traced till the minimal crossover node and the new segment is set up. Furthermore, if the detailed path of the existing connection is available, the target node can initiate the COS determination and new segment setup procedure (see Figure 75). Otherwise, the old switch (node o in Figure 72) initiates the COS determination procedure and the new segment is set up.

**Step 3: Switching of user data from the old to the new path:** Procedures to switch user cells from the old to the new path depend on the acceptable amount of cell loss and out-of-sequence cells, and the support for switching and buffering available in various network elements. For applications that do not require lossless handoffs, once the new segment is formed, user data is switched to the new path as soon as the new segment is set up. In other words, the switching fabric in the target node (or COS) is configured to transmit/receive on the new path without coordinating with the COS (or target node).

For applications requiring lossless and in-sequence cells, “Tail” signals and buffering are used to switch user data from the old path to the new path. Tail signals are special cells sent on the same virtual circuit as the user cells (in-band signals). Tail signals should be easily distinguishable from user cells. For instance, they could be special RM (Resource Management) cells [6]. These Tail signals are sent to signal the switching of user data to
the new segment. For bidirectional connections, user data is to be switched in both the
directions, downstream (from COS to the Target node) and upstream (from the target node
to the COS). For each direction, we can buffer either at the COS or at the target node.
Therefore, there are four possible alternatives for switching bidirectional connections. We
will illustrate the procedures for the following two alternatives: 1) upstream cells are
buffered in the COS and the downstream cells are buffered in the target node, and 2) both
upstream and downstream cells are buffered in the target node. Procedures for the other
two alternatives can be obtained similarly.

Buffering of upstream cells at the COS and downstream cells at the target node:

Figure 77: illustrates the procedure for buffering cells and switching data flow while
maintaining cell sequence. After the new segment is set up, Tail signals are sent in the
downstream direction (towards the target node) by the crossover switch, and upstream
direction (towards the crossover switch) by the target node. After sending the Tail signals,
the crossover switch configures the switching fabric in the downstream direction, and the
target node configures the switching fabric in the upstream direction. Since cells sent on
the new path may arrive before cells are depleted on the old path (given the old path could
be longer), the target node and the COS need to buffer cells received on the new segment
until they each receive the Tail signal sent by the other end. After receiving the Tail signals from the other end, the buffered cells are first sent and then, the switching fabric is configured for the opposite direction in both the target node and the COS. All the transit nodes on the old path can release the connection in the appropriate direction as and when they see the corresponding Tail signal.

**Buffering of both upstream and downstream cells at the target node:** Figure 78:

![Figure 78: Switching of user cells by buffering at the target node](image)

illustrates the procedure for switching of user cells by buffering only at the target node. After the new segment is set up, the upstream cells and the downstream cells received on the new path are buffered and a Tail signal (shown as Tail-1 in Figure 78:) is sent in the upstream direction (towards the COS) by the target node. After receiving this Tail signal, the switch fabric (to receive/transmit cells on the new path) is configured at the COS, and a Tail signal (shown as Tail-2 in Figure 78:) is sent in the downstream direction. On receipt of the Tail signal from the COS, cells in the upstream and downstream buffers are sent first, and then the switch fabric at the target node is configured. All the transit nodes on the old path can release the connection in the appropriate direction as and when they see the corresponding Tail signals. Analysis of these buffering alternatives is beyond the scope of this Chapter.
8.4.2 Multiple handoffs

The route optimization procedure can be used to perform route optimization after multiple fast-handoffs. Figure 72 shows suboptimal paths after $m$ handoffs. The procedure to reroute is similar to the rerouting after single handoff with the exception that the procedure to choose the to be rerouted will vary slightly. In this scenario, if $n$ is closer to $o_m$, the segment $n - o_m$ will be rerouted towards $c$ as shown in Case B of Figure 72. Otherwise, the segment $c - o_1$ will be rerouted towards $n$ as shown in Case A of Figure 72.
8.4.3 Second phase of two-phase location management schemes

In this section, we present a route optimization procedure that can be used to reroute a connection after a connection is set up to a mobile, thus proposing the second phase of two-phase mobile location/connection setup schemes. In the first phase, when a call setup arrives at the home switch of the mobile, a fast local rerouting of the connection to the mobile is performed. This results in a connection suboptimally routed through the home switch (or some other local node). In order to optimize the connection route that resulted from the fast rerouting, a route optimization procedure is executed. The need for route optimization is dictated by several factors, such as optimality of the current path, duration of the call, QoS (Quality of Service) requirements of the connection, etc. A simple criterion is the optimality of the current path, which can be determined at the home switch by comparing the “local view” of existing and the optimal paths between the calling switch and the current switch of the mobile. Once the need for optimizing the existing route is detected, this route optimization procedure is executed.

The route optimization procedure after call setup consists of Four steps. First, the connection segment and the corresponding target node are selected. Second, a crossover node between the old and the new paths is determined, and a new segment is set up between that crossover node and the target node. Third, using “Tail” signals and buffering, user data is switched over from the old path to the new path. Fourth, the old segment is released using standard ATM signaling procedures.

Step 1: Connection segment and target node determination: As illustrated in the two cases of Figure 73, the first task in optimizing the route of a connection set up through the home node (local node l instead of the home node h in the case of two-phase crankback
scheme) is to identify which of the two segments, \( c-h \) or \( h-v \) (\( c-l \) or \( l-v \)) needs to be rerouted. The corresponding target nodes for the rerouting operation become \( v \) and \( c \), respectively.

Table 15 presents a method of choosing the connection segment and the target node based on the relative "closeness" of the nodes to each other. Ideally, the choice of the connection segment and the target node should be such that the length of the new connection segment is minimized. Thereby, reducing the signaling load, route optimization delay, and the number of cells that need to be buffered (see Step 3). A definition of closeness (based on the assumption that the distance between two nodes is inversely proportional to their ancestors-are-siblings level) with the objective of minimizing the length of the new segment is as follows: For three nodes \( x, y \) and \( z \), node \( x \) is closer to node \( y \) than to node \( z \) if the ancestors-are-sibling level of nodes \( x \) and \( z \) is less than the ancestors-are-sibling level of nodes \( x \) and \( y \), i.e., \( a_{xz} \leq a_{xy} \). Using this definition of closeness and the information given in Table 15, the connection segment \( f-o \) and the target node \( t \) can be determined. It should be noted that the optimality of the resulting connection path is independent of the choice of the target node.

**Step 2: Determination of the crossover node (COS) and setting up of the new segment:**

The crossover node is determined using the procedures described in Section 8.3. If either hierarchical path information is not available or only the nodes that created the DTL can change the DTL, the procedure (see Figure 75) determines the peergroup whose ingress border node is the optimal crossover node. In order to set up a new segment between the target node and the crossover node, the identity of the crossover switch is required. Therefore, the existing connection is traced till the optimal crossover node and the new
segment is then set up. If the hierarchical path information is available and any node can change the DTLs, the connection is traced till the minimal crossover node and the new segment is set up.

Step 3: Switching of user data from the old to the new path: Procedures to switch user cells from the old to the new path depend on the acceptable amount of cell loss and out-of-sequence cells, and the support for switching and buffering available in various network elements. For applications that do not require lossless handoffs, once the new segment is formed, user data is switched to the new path as soon as the new segment is set up. In other words, the switching fabric in the target node (or COS) is configured to transmit/receive on the new path without coordinating with the COS (or target node).

For applications requiring lossless and in-sequence cells, “Tail” signals and buffering are used to switch user data from the old path to the new path. Tail signals are special cells sent on the same virtual circuit as the user cells (in-band signals). Tail signals should be easily distinguishable from user cells. For instance, they could be special RM (Resource Management) cells [6]. These Tail signals are sent to signal the switching of user data to the new segment. For bidirectional connections, user data is to be switched in both the directions, downstream (from COS to the Target node) and upstream (from the target node to the COS). For each direction, we can buffer either at the COS or at the target node. Therefore, there are four possible alternatives for switching bidirectional connections. We will illustrate the procedures for the following two alternatives: 1) upstream cells are buffered in the COS and the downstream cells are buffered in the target node, and 2) both upstream and downstream cells are buffered in the target node. Procedures for the other two alternatives can be obtained similarly.
Buffering of upstream cells at the COS and downstream cells at the target node:

Figure 77: illustrates the procedure for buffering cells and switching data flow while maintaining cell sequence. After the new segment is set up, Tail signals are sent in the downstream direction (towards the target node) by the crossover switch, and upstream direction (towards the crossover switch) by the target node. After sending the Tail signals, the crossover switch configures the switching fabric in the downstream direction, and the target node configures the switching fabric in the upstream direction. Since cells sent on the new path may arrive before cells are depleted on the old path (given the old path could be longer), the target node and the COS need to buffer cells received on the new segment until they each receive the Tail signal sent by the other end. After receiving the Tail signals from the other end, the buffered cells are first sent and then, the switching fabric is configured for the opposite direction in both the target node and the COS. All the transit nodes on the old path can release the connection in the appropriate direction as and when they see the corresponding Tail signal.

Buffering of both upstream and downstream cells at the target node: Figure 78: illustrates the procedure for switching of user cells by buffering only at the target node. After the new segment is set up, the upstream cells and the downstream cells received on
the new path are buffered and a Tail signal (shown as Tail-1 in Figure 78) is sent in the upstream direction (towards the COS) by the target node. After receiving this Tail signal, the switch fabric (to receive/transmit cells on the new path) is configured at the COS, and a Tail signal (shown as Tail-2 in Figure 78) is sent in the downstream direction. On receipt of the Tail signal from the COS, cells in the upstream and downstream buffers are sent first, and then the switch fabric at the target node is configured. All the transit nodes on the old path can release the connection in the appropriate direction as and when they see the corresponding Tail signals.

8.5 Results

Figure 82 shows the average number of hops in a connection after a connection is set up by the two-phase call forwarding scheme (mobile PNNI) scheme and the number of hops in the connection after the route optimization is performed on the original connection (see Chapter 5). From the plots we can observe that the amount of resources required for a connection in the two-phase scheme can be significantly reduced by performing the route
optimization. Since the number of hops in a connection after the route optimization phase in the two-phase scheme is equal to the number of hops in the one-phase crankback scheme we use (EQ 6) and (EQ 12) to obtain the plots of Figure 82 (Chapter 5).

![Figure 82: Number of hops in a connection after connection setup](image)

Figure 83 shows the average number of hops in a connection after a fast-handoff is performed by a two-phase path extension scheme and the number of hops in the connection after the route optimization is performed on the connection (see Chapter 9). From the plots we can observe that the amount of resources required for a connection after...
a two-phase handoff scheme can be reduced by performing the route optimization after the handoff. Instead of performing route optimization after every handoff, it can be performed after multiple handoffs.

![Diagram showing number of hops in a connection after a handoff before and after route optimization.](image)

Figure 83: Number of hops in a connection after a handoff

8.6 Summary

Two-phase connection setup and handoff schemes result in suboptimally routed connections. In this Chapter, a route optimization procedure that can be used to reroute suboptimal connections that result due to connection setup to a mobile or single and multiple handoffs. Results show that the amount of resources required for a suboptimally routed connection can be significantly reduced by performing the route optimization.
CHAPTER 9

HANDOFF REROUTING

Handoff procedures are needed to reroute connections on which the mobile user is communicating while moving. As mentioned in the previous Chapter, most handoff schemes propose performing a local connection reroute rather than an end-to-end connection reroute to keep handoff latencies low. Such reroute operations could result in making the connection path suboptimal. In Chapter 8, we proposed an algorithm for optimizing the route of a connection and applied it to the location management and handoff management problems. Based on this algorithm, a new handoff scheme called the one-phase dynamic COS search scheme has been developed. This procedure is described in this Chapter. In this method, the procedure to determine the optimal/minimal crossover is integrated into the handoff procedure. In this method, as a user moves from one base station to another, an optimal path to the new base station is determined and the connection is rerouted.
A brief summary of prior work on handoff management is presented in Chapter 9. In Section 9.1, we show how the state the "base" route optimization problem. Section 9.2 describes how this solution can be integrated with the handoff management procedures. Finally, Section 9.3 presents results of a comparative performance analysis of the one- and two-phase schemes for handoff management, and Section 9.4 concludes this Chapter.

9.1 Problem statement

In the previous Chapter, the base rerouting problem shown in Figure 84 was defined. In this Chapter, we demonstrate how the rerouting to achieve one-phase handoff scheme can be mapped to the base rerouting problem.

The dynamic COS search handoff scheme can be converted from a two-phase scheme to a one-phase scheme if the old connection is rerouted from an optimal crossover node to the new location. Such a scheme can be viewed as one in which optimal rerouting is performed during the handoff. In this case, the same problem illustrated in Figure 84: needs to be solved wherein node $f$ corresponds to the constant end $c$, node $o$ corresponds to the switch attached to the old base station, and node $t$ corresponds to the switch attached to the new base station. The new segment is setup between the optimal crossover node $p$ and the new switch.
Next, we address the issue of selecting the node that initiates the COS determination procedure. To determine the crossover node (see Figure 84), the network needs to compute the shortest path between nodes $f$ and $t$, and then determine the overlap between the new path $f-t$ and the old path $f-o$. In the one-phase dynamic COS scheme, since it is desirable to keep the handoff latency low, it becomes necessary for node $o$ or $t$ to initiate the route optimization procedure. The choice of the nodes between $o$ and $t$ depends on several factors such as type of handoff (backward or forward), and availability of existing path information.

9.2 One-phase dynamic COS search handoff scheme

As mentioned earlier, the dynamic COS search scheme classified as a two-phase scheme can be modified to a one-phase scheme if the rerouting of the existing connection during the handoff results in an optimally routed connection. In this section, we present a one-phase dynamic COS (CrossOver Switch) search handoff scheme in which an optimal connection path is achieved while rerouting the connection from the old base station to the new base station.

Handoff procedures, typically involve functions such as identification of the new base station, rerouting of the wireline connection between the far end and the old base station to the new base station, radio link establishment between the mobile and the new base station, etc. In this Chapter, we are only concerned with the connection rerouting aspect of the handoff. This rerouting of the connection during a handoff involves four steps, 1) determination of the crossover node (optimal or minimal), 2) setting up of the new segment, 3) switching of user data onto the new segment, and 4) releasing the old segment.
Handoff signaling depends on factors such as the type of handoff initiation (mobile initiated, mobile assisted, and base station initiated), type of handoff detection (forward handoff: handoff indication detected by new base station; backward handoff: handoff indication detected by old base station), and type of data transmission during handoff (soft handoff: data can be sent to/received from mobile through the old and new base stations; hard handoff: data transfer is switched over from the old base station to the new base station). In this Chapter, we will not attempt to address all the above issues. We will illustrate the procedures for a subset of the scenarios, and procedures for other scenarios can be developed similarly.

Steps 1 and 2: Determination of the crossover node and setting up of the new segment:
The crossover node is determined using the procedure described in Chapter 8 by considering $f$ as the constant end switch, $o$ as the switch attached to the old base station, and $t$ as the switch attached to the new base station. As mentioned earlier, the crossover determination procedure can be initiated by the node $t$ (new switch) or node $o$, and node $t$ can initiate the procedure only if the detailed path of the connection (list of switches on the path) is available.

*If the detailed path of the existing connection is not available*, i.e., node $o$ initiates the route optimization procedure, the crossover node is first determined and then the new segment is set up. The identity of the crossover switch (optimal or minimal crossover switch) is determined by tracing the existing connection (i.e., a COS_SELECT message is sent along the existing path). Once the identity of the crossover node is determined, the new segment is set up.
If the detailed path of the existing connection is available, i.e., node $t$ can initiate the route optimization procedure, the new segment can be set up directly by the target node, i.e., the procedures of determining the crossover node and setting up the new segment are combined. In this case, since the connection that is to be rerouted is not traced and several connections could exist through the crossover switch, it is necessary to uniquely identify the connection that is being rerouted. In order to uniquely identify this connection, we assume that during connection setup every connection is assigned a global identifier, and this global identifier is sent along in the new segment setup message. In Section 9.3, we analyze the performance of the one-phase dynamic COS search handoff schemes with and without the detailed path information.

Steps 3 and 4: Switching of user data from the old to the new path and releasing of the old connection: Procedures to switch user cells from the old to the new path depend on type of data transmission during handoff, type of handoff detection, acceptable amount of cell loss and out-of-sequence cells, and the support for switching and buffering available in various network elements. For applications that do not require lossless handoffs, once the new segment is formed, user data is switched to the new path as soon as the new segment is set up, and the old segment is released. In other words, the switching fabric in the new switch (or COS) is configured to transmit/receive on the new path without coordinating with the COS (or new switch).

For applications requiring lossless and in-sequence cell delivery, “Tail” signals described in the previous section and buffering are used to switch user data from the old path to the new path. Unlike in the two-phase scheme, the new segment is between the COS and the new switch (switch attached to the new base station) and the old is between the COS and the old switch (switch attached to the old base station).
During a soft handoff, data can be sent to and received from both the old and new base stations. In other words, there are two active paths, the old path and the new path between the mobile and the COS. This scenario resembles the switching of user data in the two-phase scheme (see Chapter 8), except that in this case the mobile (see Figure 85) is the endnode that participates in the switch over instead of the Target node. Therefore, procedures similar to the ones explained in the previous section can be used.

During a hard handoff, data transfer is switched over from the old base station to the new base station, i.e., data cannot be sent to and received from both the old and new base stations. Typically, the radio link is disconnected from the old base station and a new radio link is established to the new base station. In order to switch user data in the upstream direction (from the mobile), upstream cells are buffered by the mobile when radio link to the old base station is disconnected until the radio link with the new base station is established. Once the radio link is established, the buffered upstream cells are sent before sending the new upstream cells. If the new radio link is established before the new connection segment is established, as in a forward handoff, the new base station or the switch attached to the new base station can buffer the upstream cells and send them after the new segment is formed.

Figure 85: Old and new path during soft handoff
For switching user data in the downstream direction, two buffering alternatives are possible: buffering at the new base station (or the switch attached to the new base station) and buffering at the COS. Figure 86 illustrates the procedure for switching of user downstream cells by buffering at the new switch. After the new segment is set up, the switch fabric is configured at the COS and a Tail (Tail-1) signal is sent by the COS to the old switch indicating that no more cells will be sent on the old path. On receipt of the Tail-1 signal, the old switch sends the buffered cells followed by a Tail (Tail-2) signal to the new switch. The new switch first sends the cells received from the old switch and on receipt of the Tail signal, sends the cells received on the new path to the mobile. This procedure requires a temporary connection between the new switch (or base station) and the old switch (or base station).

Figure 87 illustrates the procedure for switching of user downstream cells by buffering at the COS. After the new segment is set up, the COS starts buffering the downstream cells received from the constant end, and sends a Tail signal (Tail-1) to the old switch (or old base station). On receipt of the Tail signal, the old switch sends the buffered cells followed by a Tail (Tail-2) signal to the COS. The COS first sends the cells received from the old switch and on receipt of the Tail signal, sends the cells in its buffer and then
configures the switch fabric. A signaling alternative for backward handoff and one for forward handoff are shown in Figures 88 and 89, respectively. Thus, using these buffering and forwarding techniques, lossless one-phase handoff is achieved.

9.3 Analysis and discussion

In this section, we present a comparative performance analysis of one-phase (route optimization during handoff) and two-phase (route optimization after handoff) handoff management schemes. As discussed earlier, there is a trade-off between handoff latency and resource utilization in one- and two-phase methods. Using a simple analytic model,
we analyze this trade-off between resource utilization and handoff latencies for these methods. Other metrics for comparison are signaling load, buffering requirements, communication disruption period, and processing requirements to handle the signaling procedures associated with these schemes. These metrics are not considered in this Chapter.

9.3.1 Analytical model

In this section, we present the details of the analytic model used to analyze the performance of the one- and two-phase schemes. We first define the notation used in this analysis in Section 9.3.1.1, and then describe a basic property of PNNI standards based hierarchical networks that is repeatedly used in the analysis (Section 9.3.1.2). In Section 9.3.1.3, the modeling details are presented. Finally, the numerical results are presented in Section 9.3.2.

9.3.1.1 Notations

Table 17 lists the symbols used in this analysis section, along with their definitions.
9.3.1.2 Property

For three nodes $x$, $y$ and $z$, the following relations hold between their ancestors-are-sibling levels:

\[
\text{Case I: } a_{xy} < a_{xz} \Rightarrow a_{yz} = a_{xy} \quad \text{(EQ 50)}
\]

\[
\text{Case II: } a_{xy} = a_{xz} \Rightarrow a_{yz} \geq a_{xz} \quad \text{(EQ 51)}
\]

The proof of this property can be inferred from Figure 90:. If $a_{xy} < a_{xz}$ the arrangement of nodes is shown in Case I of Figure 90:, from which it is clear that $a_{yz} = a_{xy}$. A similar argument extends for Case II. Also, note that since $a_{ij} = a_{ji}$, we use these terms interchangeably.

Table 17: Analysis notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>Number of peer group levels in the network; level 1 is the topmost level and level $L+1$ represents individual switches in the network.</td>
</tr>
<tr>
<td>$a_{ij}$</td>
<td>The ancestors-are-siblings level is the level at which the ancestors of the two nodes $i$ and $j$ belong to the same peer group.</td>
</tr>
<tr>
<td>$o$, $n$, $c$, $p$</td>
<td>Subscripts used to represent the old location of the mobile, new location of a mobile, the constant-end and the crossover node, respectively.</td>
</tr>
<tr>
<td>$m_i + 1$</td>
<td>Number of peer nodes in a peer group of level $i$ for $i = 1, 2, \ldots L$; $m_i = 0$ for $i &gt; L$.</td>
</tr>
<tr>
<td>$p_i$</td>
<td>Average length of the &quot;shortest-path&quot; between nodes of a peer group at level $i$, $p_{L+1} = 1$.</td>
</tr>
<tr>
<td>$D_{ij}$</td>
<td>Distance, in terms of the number of nodes on the route, from node $i$ to node $j$.</td>
</tr>
<tr>
<td>$S_d$</td>
<td>Delay incurred at a switch during connection setup.</td>
</tr>
<tr>
<td>$C_m$</td>
<td>Delay incurred at a transit switch to transport a COS_SELECT message.</td>
</tr>
</tbody>
</table>

215
In this section, we analyze the performance of one- and two-phase handoff schemes. In particular, we analyze the proposed one-phase dynamic COS search handoff scheme (optimal and minimal), and the two-phase handoff scheme in which the first phase consists of a quick path extension from the old switch to the new switch and the second phase consists of the application of the proposed route optimization scheme. Measures of analysis include handoff latencies and the amount of network resources. Since the bandwidth requirement of a connection is independent of the scheme used, we use the number of hops in the connection to estimate the network resource allocation required in each scheme. Next, we determine expressions to estimate handoff latencies in various schemes.

In this Chapter, we are primarily concerned with the connection rerouting aspect of handoff. Therefore, in this analysis, handoff latency is estimated by the time required to determine the crossover node and to set up the new segment. Time required to determine the crossover node depends on the choice of the crossover node, i.e., optimal crossover node (one-phase optimal scheme), minimal crossover node (one-phase minimal scheme), or the old switch (two-phase scheme being analyzed). In the one-phase schemes (optimal and minimal), it is possible to combine the operations of determining the crossover node.
and setting up of the new segment, if detailed information regarding the existing path is available. This alternative, one-phase scheme with detailed path information, is analyzed later.

In the one-phase dynamic COS search scheme, handoff latency includes the delay in sending a signaling message along the existing path to determine the COS \( (p) \) and the delay in setting up the new connection segment as shown in (EQ 52), where \( C_m \) and \( S_d \) are the cost of sending a signaling message and the cost of connection setup through a switch respectively, and \( D_{op} \) and \( D_{pn} \) are the number of hops on the path between COS \( (p) \) and old switch \( (o) \), and between the COS and the new switch \( (n) \), respectively.

\[
Delay_{one} = C_m D_{op} + S_d D_{pn} \tag{EQ 52}
\]

In a PNNI network with \( L \) levels of hierarchy, the distance \( D_{ij} \) between nodes \( i \) and \( j \) is approximated as:

\[
D_{ij} = \prod_{k = a_{ij}}^{L} p_k \tag{EQ 53}
\]

where \( p_k \) is the average (among all node pairs) length of the “shortest-path” between nodes of a peer group at level \( k \). For the worst case performance, the maximum length of the “shortest-path” can be taken. By this definition, the distance \( D_{ij} \) between nodes \( i \) and \( j \) depends on the value of their ancestors-are-siblings level \( a_{ij} \).

In the one-phase optimal handoff scheme, the COS is the optimal crossover point. The expression for the average handoff latency in the one-phase optimal scheme when the ancestor-are-siblings level of the old and the new locations is \( i \), can be derived from (EQ
50)-(EQ 53). This expression is given in (EQ 54), where $\overline{D}_k$ is the average number of nodes on the route between nodes $i$ and $j$ whose $a_{ij}$ is equal to $k$, i.e., $\overline{D}_{a_{ij}}$ is equal to $D_{ij}$ (see (EQ 53)).

$$\text{Delay}^{\text{opt}}_{\text{one}}(a_{on} = i) = \sum_{j=1}^{L+1} P(a_{co} = j) (P_{m} + S_d) \overline{D}_i + P(a_{co} = i) \sum_{j=i}^{L+1} P(a_{cn} = j) \left( P_{m} \overline{D}_i + S_d \overline{D}_j \right) + \sum_{j=i+1}^{L+1} P(a_{co} = j) \left( P_{m} \overline{D}_j + S_d \overline{D}_i \right)$$

(EQ 54)

In (EQ 54), the first term corresponds to the scenario in which the constant end is equidistant from both old and new switches, i.e., the old and new switches are closer to each other than the constant end ($c$). The second term corresponds to the scenario in which the constant end is closer to new switch and the third term corresponds to the scenario in which the constant end is closer to the old switch. The worst case handoff latency in the one-phase optimal scheme when the ancestor-are-siblings level of the old and the new switches is $i$ is obtained by finding the relative positions of the nodes $o$, $n$, and $c$ that lead to the maximum delay, and is estimated as shown in (EQ 55). Average delay is estimated as shown in (EQ 56).

$$\text{MaxDelay}^{\text{opt}}_{\text{one}}(a_{on} = i) = \text{Max} \{ \text{Max}_{j=1, i+1, i-1} \left( P_{m} + S_d \right) \overline{D}_i \}$$

(EQ 55)

$$\text{Max}_{j=1, i+1, i-1} \left( P_{m} \overline{D}_i + S_d \overline{D}_j \right),$$

$$\text{Max}_{j=1, i+1, i-1} \left( P_{m} \overline{D}_j + S_d \overline{D}_i \right)$$

218
As mentioned earlier, it is possible to combine the operations of determining the crossover node and setting up of the new segment if detailed information regarding the existing path is available. In this scenario, sending of signaling messages along the existing path to determine the COS are not required. The minimal crossover peergroup is determined while computing the route during the new segment setup. Assuming that the additional time required to compute the minimal peergroup during connection setup is negligible, handoff latency for the one-phase optimal scheme with detailed path information can be obtained by setting the value of $C_m$ to 0 in (EQ 54).

Next, we consider the variation of the one-phase scheme in which the minimal crossover point is selected as the COS (one-phase minimal scheme). The average delay in the one-phase minimal scheme is expected to be lower than in the one-phase optimal scheme. In order to compute the average delay to determine the minimal crossover node and to set up the new segment, we make the assumption that the minimal crossover peergroup at each level is equidistant from the corresponding level LGN of the ingress border node of that peergroup and the egress peer node performing the computation. In other words, the minimal crossover peergroup at level $k$ is assumed to be equidistant from

\[
\text{Avg} \text{Delay}^\text{opt}_{\text{one}} = \sum_{i=1}^{L} P(a_{on} = i) \text{Delay}^\text{opt}_{\text{one}}(a_{on} = i) \tag{EQ 56}
\]
LGNs $I_k$ and $E_k^{old}$ (see Figure 91). Assuming that there are odd number of peergroups on the path between these two nodes (i.e., $p_k$ is odd), the number of nodes involved in tracking the connection at level $k$, $M_k$, can be approximated by (EQ 57), where $p_k$ is the average (among all node pairs) length of the "shortest-path" between nodes of a peer group at level $k$, and $\bar{D}_k$ is the average number of nodes on the route between nodes $i$ and $j$ whose $a_{ij}$ is equal to $k$, i.e., $\bar{D}_{a_{ij}}$ is equal to $D_{ij}$ (see (EQ 53)). Note that the number of nodes on the common path (including the minimal crossover node) in this peergroup of level $k$ is also equal to $M_k$.

$$M_k = \begin{cases} \frac{p_k}{2} \bar{D}_{k+1} + M_{k+1} & 1 \leq k \leq L \\ 1 & k = L + 1 \end{cases} \quad \text{(EQ 57)}$$

The average delay to determine the minimal crossover node and then to set up the new segment can be approximated by:

$$Delay_{\text{one}}^{\text{min}}(a_{on} = i) = \sum_{j=1}^{i-1} P(a_{co} = j) \left( C_m + S_d \right) M_{a_{on}} + \frac{1}{2} \sum_{j=i}^{L+1} P(a_{co} = j) \left( C_m \left( D_{on} - M_{a_{on}} \right) + S_d M_{a_{on}} \right)$$

$$+ \sum_{j=i+1}^{L+1} P(a_{co} = j) \left( C_m M_{a_{co}} + S_d \left( D_{cn} - M_{a_{co}} \right) \right) \quad \text{(EQ 58)}$$

The second term of the equation corresponds to Case IIb of Figure 92, where $c$ and $n$ are equidistant from the old switch of the mobile ($a_{co} = a_{on}$). The first and third terms correspond to Cases I and IIa of Figure 92, respectively. Note that in the worst case, the
optimal crossover node is the minimal crossover node. Hence, in the worst case, the maximum handoff latency in the one-phase minimal scheme is equal to the maximum handoff latency in the one-phase optimal scheme.

In the two-phase handoff method being analyzed (path extension followed by route optimization), the COS is assigned to be the switch attached to the old base station, i.e., the new segment is formed between the old and the new switch. Since, the identity of the old and the new switch are known, the new segment can be set up directly. The time required to reroute the connection in this two-phase scheme can be estimated by the delay in setting up the connection segment between the old and the new switches (given in (EQ 59)). It should be noted that the route optimization procedure is executed after the fast handoff is completed. Therefore, the time required to perform the route optimization is not included in the estimate of the handoff delay.

\[ Delay_{two} = S_d D_{on} \]  

(EQ 59)

\( D_{on} \) can be estimated using (EQ 53). But this estimate could be very pessimistic, since with careful network design, the number of hops between old and the new switches (\( D_{on} \)) can be kept to a minimum. Therefore, we also present results for the scenario in which there is a direct link between the old and the new switch (i.e., the number of nodes involved in the new segment connection setup is 2).
Next, we derive expressions to estimate the amount of communication resources allocated to the resulting connection after a handoff in each of the schemes. As explained earlier, the usage of communication resources is estimated by the number of hops in the resulting connection. The one-phase (minimal and optimal) scheme results in an optimally routed path. Therefore, number of hops in a connection after a one-phase dynamic COS handoff (optimal and minimal) is equal to the number of hops in an optimally routed connection between the constant end and the new switch. Expression to estimate the number of hops in a connection resulting from a one-phase handoff is given in (EQ 60). The first, second and third terms of (EQ 60) correspond to Case I, Case IIa and Case IIb of the optimal crossover determination procedure.

\[
\text{Route}_{\text{one}} (a_{on} = i) \quad \text{(EQ 60)}
\]

\[
= P (a_{co} < i) D_{co} + \sum_{j = i + 1}^{L + 1} P (a_{co} = j) \bar{D}_i + P (a_{co} = i) \left( \sum_{j = i}^{L + 1} P \left( \frac{a_{cn} = j}{a_{co} = a_{on} = i} \right) \bar{D}_j \right)
\]

In the two-phase method, the number of hops in the resulting connection prior to the route optimization phase consists of the number of hops in the original connection \(D_{co}\) and the number of hops in the new segment \(D_{on}\). The average number of hops in a connection resulting from the two-phase handoff method is given in (EQ 61). In the case when direct path is assumed between old and new location, the extra communication resources will be the communication resources allocated on that direct connection.
should be noted that in a two-phase method, the resulting connection after the route optimization procedure is executed is an optimally routed connection between the constant end and the new switch.

\[
Route_{two}(a_{on} = i) = \left( \sum_{j=1}^{L+1} P(a_{co} = j) D_{co} \right) + D_{on}
\]  

(61)

9.3.2 Numerical results

In this section, we quantitatively compare the one-phase and two-phase schemes. In order to compare these schemes, a 10-level large hierarchical network consisting of 100,000 switches is considered. Values of input parameters assumed for this numerical computation are shown in Table 18 (see Table 17 for definitions of these parameters). We observe that the exact numerical results are dependent on the exact values chosen for these parameters. However, the trends observed are more or less independent of these values. Sensitivity of the comparative results to these input parameters has been studied. The probability distributions used for calling patterns and user mobility are derived in Section 9.5. Input parameters \(S_d\) and \(C_m\) are estimated from measured data [53].

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(m_i), (1 \leq i \leq L)</td>
<td>10</td>
</tr>
<tr>
<td>(p_i), (1 \leq i \leq L)</td>
<td>3</td>
</tr>
<tr>
<td>(L)</td>
<td>5</td>
</tr>
<tr>
<td>(S_d)</td>
<td>2.33 ms</td>
</tr>
<tr>
<td>(C_m)</td>
<td>1 ms</td>
</tr>
</tbody>
</table>

Table 18: Input values
The variation of handoff delays (average and maximum delays) and the resource utilization (number of hops) for different values of $a_{on}$ (ancestors-are-siblings level of the old and the new switches) for the one- and the two-phase methods is shown in Figure 93.

![Figure 93: Delay and resource comparisons of one- and two-phase handoff methods](image)

From the plot showing the variation in *inter-switch handoff delay*, three observation are made. *i)* One-phase optimal scheme (EQ 54) incurs a higher average delay than that of the two-phase path extension scheme (EQ 59). *ii)* One-phase minimal scheme (EQ 58) incurs a lower delay than that of the two-phase path extension scheme (with no direct link). As
mentioned earlier, the delay estimate in the two-phase scheme with no direct links is pessimistic. The actual value of this delay will depend on the way the network is engineered, and will lie in between the estimates for the case in which there is a direct link and the case in which there is no direct link. Simulation studies are being done to accurately quantify the performance of these schemes in networks of varying topologies.

iii) Maximum delay in the one-phase scheme (EQ 55) is significantly higher than in the two-phase scheme. Note that, the maximum delay in the one-phase minimal scheme is assumed to be equal to the maximum delay in the one-phase (optimal) scheme. It is important to keep the maximum (worst case) handoff latency low, since high latency leads to a degradation of the quality of service.

Next, we discuss the variation in handoff delay with a variation in the value of $a_{on}$ (ancestor-siblings level of the old and the new switches). The handoff delay in all the schemes increases with a decrease in the value of $a_{on}$ (the lower the value of $a_{on}$, the greater the distance between the old and new locations). The handoff latency for the one-phase optimal scheme is similar to that of the two-phase scheme for high values of $a_{on}$ (handoff across lowest level peer groups), but significantly higher than in the two-phase scheme for low values of $a_{on}$. It is interesting to note that majority of the inter-switch handoffs (estimated to be 90% of the inter-switch handoffs) occur at high $a_{on}$ values (see Figure 93). Therefore for most of the handoffs, the handoff latency will be low in the one-phase handoff scheme. Estimated average handoff delays in the one-phase optimal

---

1. Maximum delays in the one-phase scheme for the two variations (optimal and minimal) can be different for a specific topology (e.g., star topology).
2. Majority of the handoffs are intra-switch handoffs.
scheme, the one-phase minimal scheme, and the two-phase scheme are 28ms, 14ms, and 19ms, respectively. Average delays were computed using the probability distributions derived in Section 9.5 and (EQ 56).

If the detailed path of the existing connection is available, as described in Section 9.2, the new segment can be set up directly by the new switch, instead of first determining the COS by tracing the old connection and then setting up the new segment. Figure 94 show

![Figure 94: Delay in one-phase handoff with and without detailed path information](image)

the variation in inter-switch handoff delay in the one-phase handoff scheme with and without detailed path information. From the plots we observe that the handoff delay can be significantly reduced (about 27% in the one-phase optimal scheme) if detailed path information is available (results obtained by setting $c_m$ to 0 in (EQ 54) and (EQ 58)).

The variation in the number of hops in the resulting connection after rerouting by the one- and the two-phase methods is given in Figure 93. The one-phase method shows a slight increase in efficiency of connection routes compared to the two-phase method when there is a direct link\(^3\) between the old and the new switches, and a significant increase in

---

3. Even if there are direct links between adjacent locations, multiple handoffs can result in inefficiently routed connections.
efficiency of connection routes when there is no direct link between the old and the new switches. In the two-phase handoff scheme, the connection path after route optimization is performed will be the same as the resulting connection after the one-phase handoff. Therefore, the route optimization phase of the two-phase handoff method will result in a significant improvement in the utilization of communication resources. However, there is an overhead associated with the route optimization procedure. This overhead can be reduced by performing route optimization after multiple handoffs instead of performing the route optimization procedure after every handoff.

Though efficient resource utilization is desirable, for some applications, a large handoff latency may not be acceptable. Since an increase in handoff latency leads to degradation of the quality of service, handoff latency should be minimized. Minimizing the handoff latency also reduces cell loss, buffering requirements, and the number of cells to be forwarded from the old switch to the new switch in the case of lossless handoff protocols.

9.4 Summary

In the previous Chapter, we presented an algorithm for optimizing the route of a connection that becomes suboptimal due to operations such as handoffs. In this Chapter, the COS determination algorithm has been integrated into the handoff rerouting procedure, thus developing a one-phase handoff scheme. A comparative performance analysis of the one-phase and two-phase handoff schemes was presented. The one-phase optimal scheme incurs an average handoff latency that is twice as much as the one-phase minimal scheme, while the two-phase scheme incurs an average handoff latency that is about 30% more than the one-phase minimal scheme. The variation in which the COS
determination is initiated at the target node is better than the one in which the COS
determination is initiated at the old node (by about 27% using the optimal variation).
Finally, the amount of network resources saved by selecting the optimal crossover node in
the one-phase handoff scheme or by performing the route optimization in the second phase
in two-phase schemes is shown to be significant.

9.5 Probability Distributions

In this section, we derive the probability distributions, $P(a_{on})$ and $P(a_{co})$, and the
conditional probabilities used to compute the average handoff latency and the number of
hops in a connection.

In order to determine the probability distribution of $a_{on}$, we assume that the base
stations under a peer node of a peer group are arranged in a hexagonal fashion. We model
each peer node as a macrocell in which all the base stations under the peer node are
arranged in a hexagonal fashion. The results from [19] are used to approximate $P(a_{on})$.
Base stations in a macrocell are arranged in layers, with the $j^{th}$ layer base stations
arranged around the $j - 1^{th}$ layer base stations. A macrocell of $i$ layers with all its base
stations arranged in a hexagonal fashion has $3i^2 - 3i + 1$ base stations in it. The layers are
numbered 0 through $i - 1$ with layer $j$ ($j > 0$) having $6j$ base stations in it [19]. For
example, a 3 layer macrocell consists of 19 cells, with layer 0, layer 1 and layer 2 having
1, 6, and 12 base stations respectively. The probability that a mobile in an $i - 1^{th}$ layer
base station moves to an $i^{th}$ layer base station, $P(i - 1 \rightarrow i)$ (i.e., out of the macrocell,
since an layer-$i$ macrocell has layers numbered 0 to $i - 1$) is given by (EQ 62) [19].

$$P(i - 1 \rightarrow i) = \frac{2((i - 1) + 1)}{6(i - 1)}$$  \hfill (EQ 62)
Let $N(k)$ be the number of base stations in a peer node of level $k$. $N(k)$ is given by (EQ 63), where $N_b$ is the number of base stations under a switch (peer group of level $L + 1$).

$$N(k) = \prod_{i = k+1}^{L} (m_i + 1) N_b$$  \hspace{1cm} (EQ 63)

The total number of base stations in a $k$-level macrocell is also given by $3 \times (I(k))^2 - (3 \times I(k)) + 1$, if we arrange the $N(k)$ base stations in this macrocell as a hexagon of $I(k)$ layers. Thus

$$3 \times (I(k))^2 - (3 \times I(k)) + 1 = N(k)$$  \hspace{1cm} (EQ 64)

Solving (EQ 64) for its positive root, we obtain $I(k)$.

$$I(k) = \frac{3 + \sqrt{9 + 12(N(k) - 1)}}{6}$$  \hspace{1cm} (EQ 65)

"Border" base stations are base stations that lie on the boundary of the macrocell (i.e. base stations from which a user can move to other registration areas). Let $B(k)$ denote the number of border base stations in a level $k$ peer group. Since there are $I(k)$ layers in a level-$k$ peer node and the layers are numbered 0 to $I(k) - 1$, the border base stations of a level-$k$ peer node are the base stations in level $I(k) - 1$. Thus, there are $6(I(k) - 1)$ border base stations for level-$k$ peer node as shown in (EQ 66).

$$B(k) = 6(I(k) - 1)$$  \hspace{1cm} (EQ 66)

Probability of moving out of a level-$k$ peer group, $P(k)$, can be approximated as

$$P(k) = \frac{B(k)}{N(k)} (P(I(k) - 1) \to I(k))$$  \hspace{1cm} (EQ 67)
Assuming that each level-\( k \) peer node contributes equally to the border base stations of level-\( (k-1) \) peer node, the probability distribution of \( a_{on} \) is obtained from \( P(k) \) and \( P(k-1) \) and is given in (EQ 68).

\[
P(a_{on} = k) = P(k) - P(k-1) \tag{EQ 68}
\]

Next, we consider the distribution of \( a_{co} \). We assume that the distribution of calls to a mobile follows a uniform distribution, i.e., it is equally probable for a mobile to receive a call originating at any switch. So the probability of an existing connection from a switch for which \( a_{co} = x \) depends on the number of nodes, for which \( a_{co} = x \) and the total number of nodes in the network. Thus, the probability \( P(a_{co} = x) \) is given by:

\[
P(a_{co} = x) = \frac{\prod_{i=x+1}^{L} (m_i + 1) m_x}{\prod_{i=1}^{m} (m_i + 1)} \quad \tag{EQ 69}
\]

Conditional probability \( P\left(\frac{a_{cn} = j}{a_{co} = a_{on} = i}\right) \) is given by (EQ 70).

\[
P\left(\frac{a_{cn} = j}{a_{co} = a_{on} = i}\right) = \frac{\prod_{k=j+1}^{L} (m_k + 1) m_j}{\prod_{k=i+1}^{m} (m_k + 1) m_i} \quad \tag{EQ 70}
\]
CHAPTER 10

TRANSPORT OF SIGNALING MESSAGES

In order to support mobility in connection-oriented networks, exchange of messages such as registration messages, location query messages, authentication messages, and handoff related messages. The current methods that exist for transporting datagrams suffer from the disadvantages that they either involve significant latencies in situations when only short messages are to be exchanged or they increase complexity of network management. This chapter describes a simple mechanism to transport datagrams in connection-oriented networks that does not suffer from the large latencies associated with connection setups. The desired low latencies for messages are achieved by an efficient mechanism for connectionless transport in ATM networks by using the routing data present in switches. For example, the information distributed by the Private Network-Network Interface (PNNI) routing protocol can be used to route datagram packets in this technique. In this Chapter, “CL-ATM” (Connectionless ATM) network-layer protocol, and a proposed system called an “ATM router,” are described.
10.1 Mobility applications

in order to support mobility short message exchanges are required between network entities. Specifically, such a mechanism is required to provide efficient support in ATM networks. Examples of scenarios in which such support is useful include:

- Mobile location in ATM networks, and
- Handoff in wireless ATM networks.

10.1.1 Locating mobile endpoints

One approach to locating a mobile before connection setup is to use the cellular-standards-based approach of consulting the Home Location Register (HLR) of the mobile (we refer to the approach as the overlay approach). A simplified version of the cellular standards approach to location management is depicted in Figure 95. Consider an endpoint A generating a connection setup request to mobile B. In this example, when endpoint A located in network X requests a connection setup to mobile B, switch 1, the switch to which endpoint A is connected, must first determine the location of mobile B before connection setup can be initiated. To enable this mobile location operation, each mobile is assigned a home switch with an attached HLR which tracks the current location of the mobile. When a mobile moves between switches it informs the HLR of its new
location which is recorded in the HLR database. Therefore, to locate mobile B, switch 1 first sends a location request message, LOCREQ, to the HLR of mobile B and obtains a response which indicates switch 3 as the current location of mobile B. Switch 1 can then send a SETUP message for mobile B towards switch 3 which then completes the call to the destination mobile. Setting up a connection between switch 1 and the HLR in order to transport these two short messages, LOCREQ and its response, is obviously wasteful and will lead to large latencies in call setup. Hence, an efficient datagram transport technique is required.

10.1.2 Handoff in wireless ATM networks

In wireless ATM networks, as a mobile moves from one base station to the next, ATM connections need to be rerouted if the mobile is actively involved in one or more connections. Figure 96 shows a case where mobile B moves from base station IV to base station III. Assuming that handoffs are mobile-initiated, and that the mobile sends the
Handoff request to the new base station, the new base station has to somehow send a Handoff message to the old base station (that has been identified in the original Handoff request received from the mobile). Once the Handoff request is conveyed to the old base station, the old path of the connection can be traced until a crossover switch is identified (Switch III in Figure 96) and the new segment setup from this switch to the new base station. Data can then be switched from the old path to the new path, thus completing the handoff procedure.

This example illustrates the need for a method to route the Handoff message from the new base station to the old base station which are two nodes that may or may not be connected via a direct physical link. In this application also, the need for an efficient method to route the Handoff message is required, since any delay in this process will increase the handoff latency, which, in turn, increases cell loss, or the buffer requirement if lossless handoffs are to be ensured.

10.2 Potential solutions

To send messages between two nodes, one of five methods can potentially be used:

- Set up a SVC (Switched Virtual Circuit) between the two nodes;
- Use a PVC (Provisioned Virtual Circuit) that has been preestablished between the two nodes;
- Send in IP packets and use one of the standard IP-over-ATM approaches: Classical IP over ATM [95], Routing Over Large Clouds (ROLC) [97], or Multi-Protocol Over ATM (MPOA) [98];
- Send in IP packets and use PVCs established between ATM switches and IP routers;
Use an SS7 overlay network as in connection-oriented telephony networks.

Consider, for example, the Location Request message, LOCREQ. To set up an incoming call to a mobile, the calling party’s switch needs to first locate the mobile. For this purpose, it generates a LOCREQ to the called mobile’s Home Location Register in the overlay approach (see Chapters 6 and 7). If an SVC *is to be set up* between these two nodes in order to send this message, the total call setup delay for the call incoming to the mobile includes:

- SVC setup delay for the connection from the calling party’s switch to the called mobile’s HLR,
- processing delay of the LOCREQ message and its response, with link emission and propagation delays, and
- the SVC setup delay for the actual connection setup from the calling party to the called mobile.

For this single message exchange, any technique which removes or shortens the first component of delay will be helpful in reducing set up latencies.

The *second* option of using PVCs removes the overhead associated with setting up an SVC between the calling party’s switch and the called mobile’s HLR. However, it could be very expensive to engineer PVCs between all pairs of nodes that could potentially need to exchange signaling messages or other connectionless data. Resource utilization of PVCs often tends to be poor unless a management tool for moving bandwidth and buffer resources as needed is introduced but such a tool adds complexity.
The third approach of using one of the standard IP-over-ATM models is worse than the first solution of using SVCs because, besides requiring the setup of SVCs, it requires address translations from the IP address format to the ATM NSAP address format.

The fourth approach of using an IP overlay network with PVCs is quite feasible. This scheme is shown in Figure 97. All network nodes, switches, location registers, servers, and endpoints are connected to IP routers using PVCs. Assuming that the datagram-sending node knows the IP address of the datagram-receiving node, the datagram is sent in an IP packet on the PVC from the sending node to an IP router. This IP packet is routed until it reaches the router which has a PVC to the receiving node allowing it to convey the datagram to the receiving node. Given the predominant use of IP in the worldwide Internet, this approach is a feasible means for sending signaling and other datagram packets from one node to any other node. The drawbacks for this approach include ATM nodes having to know IP addresses of other nodes, and the need for accurately sizing the PVCs, i.e., estimating the resource allocations required for the various PVCs to and from IP routers.
The fifth approach is used in current telephony networks. Since these networks are also connection-oriented, a similar need for routing short signaling messages exists in such networks. For this purpose, a datagram-based network, called the SS7 (Signaling System No. 7) network is overlaid on PSTNs (Public Switched Telephony Networks). SS7 networks use STPs (Signaling Transfer Points) to route messages as datagrams between any two PSTN nodes. The disadvantage of using such an SS7 overlay network in ATM networks is that the SS7 datagrams are routed using "Point Codes" as network addresses. This address format is distinct from both the ATM NSAP address format and the IP address format and therefore makes it more complex for ATM nodes to track these addresses.

10.3 Proposed solution

Unlike the schemes presented in Section 10.2, our proposed solution is a datagram transport mechanism that uses ATM NSAP addresses and is based on using the routing data in ATM switches. For example, in the case of private ATM networks, the PNNI routing protocol standard has been specified so that switches can communicate with each other in order to construct a topology database. For public networks, there is no comparable routing protocol. We assume that public ATM networks will adopt the "optimal-routing" technique [94] in which a centralized network management station computes optimal routes from a network-wide perspective and downloads this data to switches. Our proposed solution works in ATM networks using either of these modes of operation for updating routing data.
In this section, a method for transporting datagrams in ATM networks is described. As indicated earlier, this method consists of a "CL-ATM" (Connectionless-ATM) network layer protocol, and a system called an "ATM router." These two concepts are described in Section 10.3.1. This solution is developed primarily to transport mobility related signaling messages. However, it also offers an efficient method for transporting "datagram" IP packets in an ATM network.

10.3.1 CL-ATM network layer protocol and ATM router

CL-ATM network-layer protocol is a connection-less protocol that uses ATM addresses in the packet header. This creates the need for "ATM routers" to terminate the CL-ATM network-layer protocol for packet forwarding. The term "ATM router" is coined from the ability of this entity to function as a standard ATM switch combined with the ability to route connectionless traffic using our CL-ATM technique.

The CL-ATM technique for the transport of datagrams is illustrated in Figure 98. Figure 98 shows two types of network nodes, endpoints, which generate the CL-ATM packets, and ATM routers, which route these packets through the network. The format of a CL-ATM network-layer packet is shown in Figure 99. Note that, as will be seen in Section 10.3.2, an endpoint is not necessarily an end user host; it could be a switch, server, or location register based on the application using this CL-ATM service. Each ATM router consists of an ATM switch, with a switch fabric and a topology/routing database, and a CL-ATM server. The topology/routing database is typically used by the ATM switches to set up ATM connections. However, in this proposal, we use the topology/routing database to route CL-ATM packets. The CL-ATM server software is the new software module added to the ATM switch to convert it to an "ATM router". It terminates the CL-ATM
network layer and performs datagram routing of the CL-ATM packets. The *PNNI* routing *protocol software module* is shown in dotted lines to indicate that an ATM switch may or may not have this software. In private ATM switches, this PNNI routing software is used to update the topology/routing database. Finally, other software, such as UNI (User Network Interface) [16] and PNNI signaling software, may be present in the ATM switch if it supports SVC capability. This is not shown in Figure 98 as it is not relevant to CL-ATM packet routing.
The method of performing CL-ATM transport in both private and public ATM networks is described next. The protocol stack at the endpoint in Figure 98 shows that the transport layer may or may not be used by the application. Nevertheless, assuming that the application has the ATM NSAP address of the destination, which it passes to the CL-ATM network layer, a packet of the format shown in Figure 99 is generated by the endpoint. The CL-ATM packet is then segmented, using ATM Adaptation Layer 5 (AAL5) functionality, into ATM cells and sent to the ATM router on a known connectionless-ATM VCI, say VCI 16. At the first ATM router connected to this endpoint, the CL-ATM packet is reassembled (from the ATM cells) and routed according to the information in the CL-ATM packet header. The CL-ATM server examines the destination NSAP address of the packet in order to determine the route for the CL-ATM packet. The computation of the route is performed using the information stored in the topology/routing database.

In private networks, the topology/routing database is maintained by the PNNI routing protocol software as it receives PTSPs. An appropriate shortest-path selection algorithm, such as Dijkstra's algorithm, can be used to determine routes, either a priori, or at the time of arrival of a CL-ATM packet. Thus, when a CL-ATM packet arrives at the first ATM router, it computes a path for the datagram to reach the destination endpoint, using a process similar to the DTL computation for connections executed upon receiving a SETUP message. The computed set of DTLs is inserted into the source-route field of the CL-ATM packet, which is then routed (after AAL5 segmentation into ATM cells) to the next ATM router indicated by the DTL. As described in Chapter 2, a set of DTLs is used to specify a "hierarchical" route, in that it includes the exact physical ATM router addresses for only the first peer group, and identifies logical group nodes for subsequent peer groups.
This implies that at each subsequent ATM router in the source route, the actions performed may differ based on whether the router is a border node of a peer group or a non-border node.

- If the router is a border node of a peer group, the detailed list of transit ATM routers through the peer group needs to be computed by the node. The CL-ATM server updates the source-route field with this recomputed DTL, containing the route from the current switch to the destination endpoint.

- If the router is not a border node of a peer group, the ATM router simply forwards the packet to the next node indicated in the source-route field.

Thus, in private ATM networks, using a combination of source routing and hop-by-hop routing, as is done by the PNNI signaling software for SETUP messages in connection-oriented ATM transport, the CL-ATM servers in ATM routers cooperate to route CL-ATM datagrams through the networks. It is also possible to perform hop-by-hop routing instead of hierarchical source routing.

In public ATM networks (presumably without PNNI routing), a topology/routing database must still exist in ATM switches in order to set up ATM connections between them. As stated earlier, one approach may be to use an "optimal routing" scheme, whereby the topology/routing database is updated by some network management system. This topology/routing database provides switches the information needed to determine the route of an ATM connection to a given destination NSAP address. Either source routing or hop-by-hop routing may be used in such public ATM switches. If source routing is supported by these switches, the mode of operation for the CL-ATM servers is fairly similar to that described above for private ATM switches. If, however, the topology/routing data only indicates the next switch by which to reach a destination NSAP address,
then the CL-ATM servers at each ATM router will need to perform hop-by-hop routing. In this technique, the servers at each node do not perform a full path computation to the endpoint but instead forward the CL-ATM packet to the next node based on their topology/routing information. The source-route field option is not required in this usage (which is the reason for indicating it as “optional” in Figure 99). Thus, the CL-ATM network-layer protocol and ATM routers can be used for routing datagrams in both private and public ATM networks.

Another option which may be of use for streams of datagrams is the “flow label” field in the CL-ATM packet header shown in Figure 99. This field is similar to the IPv6 flow label and will allow an ATM router to create a cut through connection across the switch fabric for long-lived datagram flows. This usage of flow labels is comparable to the use of fast-select transport technique in X.25 networks [101].

10.3.2 Applications

As mentioned in the introduction to this section, the primary motivation for developing this form of datagram transport in ATM networks is to send mobility management messages. However, this scheme also provides a solution to the IP-over-ATM problem. The need to route IP packets over ATM networks has arisen because most desktop applications continue to be IP-based, while for fast networking, ATM switching technology appears to be the solution of choice. In this section, we demonstrate how our proposed CL-ATM solution can be used.
10.3.2.1 Mobility management messages

Two example applications that require efficient datagram service in ATM networks are described in Section 10.1. In all these applications, the source "endpoint" of the CL-ATM packet knows the ATM NSAP address of the receiving node. In the location management example described in Section 10.1.1, the calling party's switch is assumed to know the ATM NSAP address of the HLR of the called mobile. In current cellular networks, this is achieved easily since telephony networks use geographical addressing. In private ATM networks, the ATM NSAP address of the HLR is not known. Hence, we propose a method in which the mobile's "home" NSAP address is used to route the LOCREQ to a node which can issue a response with the current location of the mobile. In PNNI-based networks, reachability data is propagated through PNNI PTSPs. A mobile's "home" reachability is propagated through this mechanism of sending summarized reachability data. The LOCREQ is generated as a CL-ATM packet with the Destination NSAP address set to that of the mobile's home address. This implies that it will be routed to the home switch of the mobile. A separate indicator is set in the message to indicate that the message is for the Home Location Register and not to the addresses mobile. This will prevent the ATM switch from forwarding the datagram to the mobile. When the LOCREQ CL-ATM packet reaches the home switch of the mobile endpoint, the datagram is delivered to the mobility management process. The latter generates a response indicating the current location of the mobile. Thus, the CL-ATM solution can be used to transport the LOCREQ to the appropriate node and obtain a response.
In the handoff example of Section 10.1, we simply assume that the mobile sends the ATM NSAP address of the old base station as a parameter in the Handoff message which it sends to the new base station. This allows the new base station to send the Handoff message to the old base station (see Figure 96:) as a datagram using the CL-ATM network-layer protocol.

10.3.2.2 IP-over-ATM usage

Different approaches have been proposed for IP-over-ATM transport since, increasingly, IP data is viewed as the main initial application for ATM networks. These include

- the standards-based IP-over-ATM models: Classical IP over ATM [95], LAN Emulation [96], ROLC (Routing Over Large Clouds) [97], and MPOA (Multi-Protocol Over ATM) [98],
- the IP overlay network using PVCs,
- the tag switching approach of Cisco [99], and
- Ipsilon's IP Switch [100] approach.

The first two approaches are described in Section 10.2. The Cisco and Ipsilon approaches focus on improving the IP routing throughput using ATM hardware switching rather than the software processing of packets that IP routers typically utilize. This is done by identifying "long-lived flows" and setting up an ATM SVC for these IP flows. For short-lived flows, IP routing is done at each node.

The CL-ATM network-layer protocol and ATM router solution proposed in this Chapter for datagram transport in ATM networks can be used for IP packets using the approach shown in Figure 100. Assuming the end host applications continue to generate
IP data on non-ATM LANs such as Ethernet, Figure 100 shows that the IP packets are routed as CL-ATM packets beyond the IP router. If the destination NSAP address has been previously discovered and stored in a cache (for example, address information from a CL-ATM packet incoming to an endpoint can provide the necessary mapping between IP and NSAP addresses) then it may be obtained from this cache and inserted in the packet header. If this information does not exist locally, the host generates an ATM-ARP (ATM-Address Resolution Protocol) request with the destination IP address to an IP-ATM ARP server which returns the ATM NSAP address to be inserted in the CL-ATM packet. This first address resolution phase is the same as that used in any of the standard IP-over-ATM models. However after this resolution, an SVC is not set up. Instead the packets are routed as CL-ATM datagrams through ATM routers. Thus, compared to any of the standard IP-over-ATM solutions, this solution is better in the sense that a low overhead is experienced by the "datagram" IP packets when routed through ATM networks. The use of ATM
routers for this purpose also allows for graceful infrastructure upgrades from IP routers to ATM routers, and, later, with the addition of ATM signaling software, from ATM routers to ATM switches.

This technique is primarily a solution for "short-lived IP flows." If a long exchange of IP packets is expected, it is better to incur the SVC setup overhead up front and move IP packets rapidly through switch fabrics rather than incur a CL-ATM routing overhead for every packet. The flow label field option described in Section 10.3.1 can be used in order to determine if an SVC should be setup. Since an ATM router consists of an ATM switching fabric, this cut-through switching function is an inherent component of an ATM router. It is also possible to use IP addresses instead of ATM addresses in this protocol. This modification will eliminate the need to perform an ATM-ARP to determine the ATM address.

10.4 Summary

In this Chapter, we presented a new connectionless ATM (CL-ATM) network layer protocol for transporting datagrams in ATM networks. Essentially a CL-ATM packet is similar to an IP packet with the destination and source addresses being ATM NSAP (Network Service Access Point) addresses. Software in ATM switches which enables them to function as "ATM routers" is then used to route CL-ATM packets within ATM networks. Routing data, for example, the data collected by the PNNI (Private Network-Network Interface) routing protocol in private ATM switches, is used to route packets from one ATM router to the next. Such a scheme is necessary for transporting signaling messages in order to efficiently support mobility in ATM networks. Using this motivation, we defined CL-ATM to be a generic "datagram" service, allowing it to carry
connectionless packets from several applications. Since the "IP application" is a predominant datagram generator, the CL-ATM solution is also a suitable solution for IP-over-ATM transport.
CHAPTER 11

A PROTOTYPE IMPLEMENTATION

This chapter describes the implementation details of a wireless network prototype. This prototype was built by the Broadband Systems Research Department at Bell Laboratories, Holmdel, NJ. In this prototype network the integrated location management scheme (see Chapter 4), handoff mechanism described in Chapter 9, CL-ATM protocol, and interworking of IP and ATM (see Chapter 10) were implemented. The motivation behind this prototype is to demonstrate the effective mobility management using the techniques developed in this thesis. Data, voice, video, and interactive applications were shown to work effectively without degradation in performance in this prototype.

11.1 Network architecture

The wireless network prototype is shown in Figure 101. The backbone network consists of five ATM switch routers (ASRs). Each ASR consists of an ATM switch (FORE ATM Switch), a Single Board Computer (SBC) implementing the CL-ATM forwarding engine (Connectionless ATM router), and a Sun Sparc host running PNNI routing protocol and a prototype signalling protocol. Single Board Computer (SBC) runs pSOS operating system and implements routing functionality to routes CL-ATM packets. It
Figure 101: Wireless Network Prototype

supports both hierarchical source routing and hop-by-hop routing based on destination address. The routing tables are computed from the routing data obtained by the PNNI routing protocol. Routing table is computed by the Routing Update process in the Sparc Station and downloaded to the SBC. SBC is connected to the ATM Switch via an OC-3 interface. Architecture of SBC is shown in Figure 102.
The LANs (Local Area Networks) are connected to the ASR network via CL-ATM IP Gateways. Architecture of the CL-ATM IP gateway is shown in Figure 103.

Some of the CL-ATM IP gateways also implement mobility management functions such as sending advertisements at regular intervals, and performing registrations and handoffs. Two WaveLAN access points provide wireless access to mobiles and these Access Points are connected to the CL-ATM IP gateways via an Ethernet as shown in Figure 101. Laptop computers equipped with WaveLAN PCMCIA cards are used as mobile terminals. Architecture of the mobile terminals are shown in Figure 104. The mobility management process performs the registrations, deregistrations, and initiates a handoff.
11.2 Mobility management

In this network, IP addresses are used for all the network elements and hosts. ATM encapsulated IP addresses are used in the ASR network. All applications that were used to demonstrate the effectiveness are end-to-end IP applications.

Mobiles tracking is done with the help of registrations and deregistrations. The two Access Points shown in Figure 101 are on two different IP subnets. The mobility management process in the Laptop computer listens to advertisement beacons sent by the Mobility management process on the Gateway. This advertisement is used to determine the current location of the mobile and send the registration message. It is also used to determine the current default gateway. Registration is sent in a UDP packet to the home
gateway of the mobile (mobile node is configured with this information). On receiving a registration the home gateway and the foreign gateway update their routing tables and location information of the mobile.

When an IP datagram destined to a node arrives at a gateway, the gateway determines if there is an existing connection (SVC) corresponding to the destination host. If a connection corresponding to the host exists, the datagram is sent via that SVC. Otherwise it is encapsulated in an CL-ATM packet and sent to the host. The encapsulated datagram is routed to corresponding gateway. It also checks for a pending connection setup corresponding to the destination host. If no such connection setup is pending it initiates a connection setup to the corresponding gateway, to be used to route future datagrams destined to the same host (with the assumption of the flow is a longlived flow). This connection is released using a timer based mechanism. During connection setup the entire path of the connection is stored in the originating and the terminating gateways. When a gateway receives a CL-ATM packet, it decapsulates the CL-ATM packet and routes the IP datagram based on the IP header information. If the IP packet is destined to mobile and the mobile is not at its home network. The IP packet is encapsulated in a CL-ATM header (with the destination address as the visiting gateway of the mobile) and routed to the corresponding gateway. When setting up a connection to a mobile host, the connection setup proceeds to the home gateway of the mobile. If the mobile is not at its home network, the connection is forwarded from the switch connected to the home gateway to the visiting gateway of the mobile. Figure 105 shows the path taken by data destined to a mobile in the prototype. When a mobile moves detects that it has moved into a new
network area, it sends a registration message to the gateway. Using this registration information, the routing data is updated in the home gateway and the connections corresponding to the mobile are rerouted.

This prototype demonstrated the effective transport of data from different applications while the mobile terminal moves from one Access Point to another without degradation in quality. Effective data transfer of large files was demonstrated using FTP (File Transfer Protocol). Interactive applications such as Telnet and accessing the World Wide Web using Netscape Navigator operated without degradation in quality while the mobile moved between Access Points. While demonstrating transport of plasticize voice using Catlike and streaming Video using Optivision, there was a drop in voice packets and Video frames for about a second.
11.3 Summary

This chapter described the implementation details of a wireless network prototype built at the Broadband Systems Research Department, Bell Laboratories, Holmdel, NJ. This prototype demonstrated the working and effectiveness of the integrated location management scheme (see Chapter 4), handoff mechanism described in Chapter 9, CL-ATM protocol, and interworking of IP and ATM (see Chapter 10). It also demonstrated the effective working of data transfer, voice, video, and interactive applications without degradation in performance.
CHAPTER 12

CONCLUSION AND DIRECTIONS FOR FUTURE WORK

This chapter concludes this dissertation with a summary of research contributions and with a brief examination of future research directions.

12.1 Research contributions

This dissertation examined the network layer aspects of mobility support in hierarchically routed networks such as IP-based networks or ATM networks. It addressed the issues of Location Management, Handoff Management, Route Optimization, and Transport of Signaling Messages. The objective is to provide ubiquitous communication to a wide range of mobile terminals supporting wide range of services. The research contributions of this dissertation is the development, performance evaluation, and implementation of location management schemes, handoff management schemes, a route optimization algorithm, and an efficient techniques to transport of signaling messages. These schemes are applied to support mobility in the IP-based networks, ATM-based networks, and cellular networks.
Location management schemes are required to route connections or datagrams to mobile hosts. We developed the integrated scheme or the mobile PNNI (Private Network-to-Network Interface) scheme, the one-phase crankback scheme, the LR (Location Register) scheme, and the flat scheme. The integrated scheme uses features of dynamic routing and signaling protocols with minimal modifications to handle mobile users in networks with hierarchical routing protocols. The concepts of the integrated scheme are used to support mobility in connection-oriented ATM networks and in connectionless IP networks. Integrated scheme provides efficient IP mobility support with hierarchical dynamic protocols such as OSPF and BGP. In the Integrated scheme, connection or data could follow a suboptimally routed path. The one-phase crankback scheme uses topology information collected by link state routing protocols and the technique of crankback to set up optimally routed connections. While the Integrated and the one-phase crankback schemes perform location management by integrating (Integrated approach) it with routing and signaling protocols, the LR scheme and the flat scheme use an overlay (Overlay approach) of external location registers (such as the cellular home and visitor location registers) to perform location management. LR scheme is a generalized overlay-approach scheme. This work also provides fundamental understanding of the effect of user and network characteristics on the design of optimal location management architectures. Results show that at low (CMR < 0.025) CMRs (Call-to-Mobility Ratio), the LR scheme performs better than the Integrated scheme.
Handoff management procedures are required to reroute connections or datagrams as a mobile host moves from one location to the other. We developed a one-phase dynamic handoff scheme that results in an optimal path and incurs low latencies. As part of this work, we have also developed and analyzed signaling and buffering alternatives to perform lossless and in-sequence delivery of data during handoffs.

Route optimization algorithms are required to reroute suboptimal connections that result during connection setup to a mobile host or during handoffs. A route optimization procedure that can be used to reroute suboptimal connections that result due to connection setup to a mobile or single and multiple handoffs has been developed. Results show that the amount of resources required for a suboptimally routed connection can be significantly reduced by performing the route optimization.

In managing mobile hosts, there is a need to exchange short mobility management messages. We developed a connectionless ATM protocol (CL-ATM) to transport such messages in connection-oriented networks such as ATM Networks. Salient feature of this protocol is the realization that connectionless service can be provided by utilizing the routing information existing in a connection-oriented network. The developed protocol significantly enhances the effectiveness of the ATM networks by providing an efficient technique for transporting short-lived data flows with minimal enhancements to the existing infrastructure. Using this protocol we have demonstrated an efficient method to transport IP traffic on a ATM backbone.

The techniques developed are independent of wireless access technologies, and therefore can be used with any access technology. These techniques can also be applied to provide efficient IP-mobility support and can be deployed to enhance performance in existing wireless networks such as PCS. This research involved implementation of a
wireless network prototype in which, mobility in an ATM network, working of CL-ATM concept, and interworking of IP and ATM were demonstrated. It also demonstrated the effective working of data transfer, voice, video, and interactive applications without degradation in performance in this prototype.

12.2 Future work

This dissertation addressed some of the network layer issues to support mobility. Issues such as security and authentication, multicast and broadcast support, QoS (Quality of Service) support, and redundancy and fault tolerance were briefly discussed. These are important network layer issues and need to be investigated further. There are several link layer, transport layer, and application layer issues that arise because of mobility and the nature of wireless connections. A brief examination of future research directions is given below:

12.2.1 Security and authentication

Wireless environment has inherent limitations in terms of security due to the broadcast nature of the communications. In this dissertation, issues of security have not been addressed. For example, before setting the forwarding pointer in the home switch or sending reachability information in the integrated scheme, the authenticity of the registration message needs to be verified. Issues such as denial of service attacks, unauthorized access, unauthorized redirection of traffic, and firewall traversal need to be investigated further.
12.2.2 Broadcast and multicast support

Ideally, the mobile user should be able to access services as if the mobile user was stationary. Broadcast support is needed to provide virtual presence to a mobile as if it was on its home network. Multicast support is required to support group communications. Currently, only unicast data delivery is supported to/from mobile terminals. Multicast and broadcast support for mobile hosts can be achieved by extending the basic unicast support mechanisms. For example, perform the multicast and broadcast operations through the home switch/router (by using bidirectional connections in the case of connection-oriented protocols or by bidirectional tunnels in the case of connectionless protocols). Though this simple extension will provide a method of providing broadcast and multicast support, it is an inefficient method. Efficient techniques and extensions need to be developed.

12.2.3 Fault tolerance

Failures in network components such as routes/switches is unavoidable. The failure of nodes such as Home Agent in mobile IP could disrupt the service to the mobile users. It is desirable that the failure of one or more network elements does not disrupt the connectivity of the mobile terminals to the network. In order to be able to tolerate certain network failures, protocols need to be designed to tolerate faults. In Chapter 4, techniques of redundancy and spreading of reachability information were discussed as methods of achieving fault tolerance. These techniques along with other new techniques need to be investigated to incorporate fault tolerance so that mobile users are provided service in spite network node failures.
12.2.4 Quality of Service (QoS) support

Quality of Service (QoS) guarantees are required for certain applications, such as voice applications, real-time applications, etc. There is substantial work being done to provide QoS guarantees in IP and ATM networks for stationary hosts and network elements. In a mobile environment, the techniques developed for stationary hosts may fail or be inefficient. For example, when a mobile user moves from one cell to another the connection need to be rerouted and the new connection should provide the QoS guarantees assured by the old path or renegotiate a set of new QoS guarantees. Availability of resources also becomes an issue. In order to provide support for applications that require QoS guarantees in a mobile environment, it is necessary to either extend the mechanisms used for stationary hosts to operate in mobile environment or develop new techniques to provide QoS support. In this dissertation connection setup delay during connection setup or data loss and rerouting delay during handoff procedure were discussed briefly. Support for these QoS parameters along with other parameters need to be investigated in a mobile environment and new techniques to provide guarantees need to be developed.

12.2.5 Ad-Hoc networks and mobile networks

Traditional wireless and mobile networks are based on cellular network infrastructure. In contrast, ad-hoc networks are not dependent on any preexisting infrastructure. Recently, due to proliferation of small communication capable devices, ad hoc networking technology has emerged as an exciting area for research activity. The methods developed in this dissertation need to be adapted to work in an ad-hoc environment and several new techniques are needed to be developed in the areas of mobility management, routing
protocols, organizing topologies, and network management. Another important mobility scenario that has been briefly discussed earlier is that of mobile networks. In a mobile network the entire network along with the router/switch moves.

12.2.6 Physical, link, transport, and application layer issues

This dissertation addresses the network layer problems associated with mobility. Mobility of terminals also impact the other layers such as the MAC (Media Access Control) Layer protocols, Transport Layer, Application Layer, etc. In order to provide seamless mobility and support wide range of services, it is necessary to examine the impact of mobility on these protocol layers and develope techniques to be incorporated at different layers. Some of the Transport Layer issues are effect of latency, data loss, and out of sequence packets during handoff, and effect of high bit error rate and low bandwidth wireless link on the Transport Layer Protocols. Another important issue is support of QoS (Quality of Service) on the wireless link.
BIBLIOGRAPHY


262


264


