Mobility Management Techniques
in Wireless Networks

by

YI-LI TRACY TUNG
M.Eng.Sci., University of New South Wales, 1999
B.E., University of Auckland, 1998

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

In
GRADUATE SCHOOL OF ENGINEERING

THE UNIVERSITY OF SYDNEY

APRIL 2003
Copyright in relation to this Thesis

Under the Copyright Act 1968 (several provisions of which are referred to below), this material must be used only under the normal conditions of scholarly fair dealing for the purposes of research, criticism or review. In particular, no results or conclusions should be extracted from it, nor should it be copied or closely paraphrased in whole or in part without the written consent of the author. Proper written acknowledgement should be made for any assistance obtained from this material.

Under Section 35 (2) of the Copyright Act 1968 'the author of a literary, dramatic, musical or artistic work is the owner of any copyright subsisting in the work'. By virtue of Section 32 (1) copyright 'subsists in an original literary, dramatic, musical or artistic work that is unpublished' land of which the author was an Australian citizen, an Australian protected person or a person resident in Australia.

The Act, by Section 36 (1) provides: 'Subject to this Act, the copyright in a literary, dramatic, musical or artistic work is infringed by a person who, not being the owner of the copyright and without the licence of the owner of the copyright, does in Australia, or authorises the doing in Australia of, any act comprised in the copyright'.

Section 31 (1) (a) (i) provides that copyright includes the exclusive right to 'reproduce the work in a material form'. Thus, copyright is infringed by a person who, not being the owner of the copyright, reproduces or authorises the reproduction of a work, or of more than a reasonable part of the work, in a material form, unless the reproduction is a 'fair dealing' with the work 'for the purpose of research or study' as further defined in Sections 40 and 41 of the Act.

Section 51 (2) provides that "Where a manuscript, or a copy, of material of other similar literary work that has not been published is kept in a library of a university or other similar institution or in an archives, the copyright in the material or other work is not infringed by the making of a copy of the material or other work by or on behalf of the officer in charge of the library or archives if the copy is supplied to a person who satisfies an authorized officer of the library or archives that he requires the copy for the purpose of research or study'.

* Thesis includes 'treatise', 'dissertation' and other similar productions.
Statement of Originality

The novel research results reported in this thesis represent an original work by the author, at the School of Electrical and Information Engineering, Sydney University, Australia, under the supervision of Associate Professor Abbas Jamalipour. None of the content of this thesis has been previously submitted for the consideration for a degree or any qualification. Some of the results have been published or submitted for publication in technical journals or conference proceedings.

Yi-Li Tracy Tung
April 2003
Abstract

With wireless networks advancing so fast, mobility is no longer an exception but a basic criterion to be incorporated in the 3rd generation mobile communications. This thesis was dedicated to enhancing the existing operations that were tackling the challenge of location tracking. The research work was divided into two parts; the first part emphasized applications in the cellular network (i.e. homogeneous mobility), the second part focused on the task of managing mobility between heterogeneous networks.

Regarding the applications of cellular networks, in order to better utilize the distinct characteristics of individual users on traveling directions, two tracking schemes, the directivity-aware location updating scheme and the Kalman-filter based update scheme, as well as a corresponding sectional paging scheme were developed to optimize performance. Other advances include the development of a new mobility model that simplifies the existing presentation of the Markovian movement model.

As an extension to the conventional distance-based update scheme, the directivity-aware location updating scheme made the selection of the optimal threshold adaptive not only to the mobility rate, but also to the traveling patterns. The Kalman-filter based update scheme improved prediction accuracy with the incorporation of a Kalman filter. The Gauss Markov distribution was used to model the level of correlation between traveling directions. Accordingly, based on the Kalman filter theory, the shape of an update area that best reflects the distinct features of the mobile concerned was able to be determined. A corresponding sectional paging scheme was developed aiming to further reduce the operational cost. By identifying the section of most likely residence, paging areas were assigned according to the location area.

For managing mobility between heterogeneous networks, new tracking strategies incorporating cellular techniques, the pointer forwarding scheme (PFS) and user profile replications (UPR) were designed. Preliminary analysis of the proposed solutions showed encouraging results. Subsequently, to enable more precise quantification of the improvement, a set of design criteria was proposed in addition to a matching set of quantifying characteristics. It is worth further exploring the underlying concept of mobility management techniques that we have presented in this thesis in order to meet the challenge of more efficient location tracking.
Acknowledgements

My highest gratitude goes to my supervisor, Associate Professor Abbas Jamalipour. He has not only been energetic in offering insightful comments and constructive suggestions about the research but has also been willing to give his fullest support at all times. Despite his many responsibilities involving academic research/teaching and IEEE chapters, he has always made himself available for individual weekly consultations. Often, additional last minute appointments were required and he filled them in to his seemingly impossible schedule. The endless patience and support he has unconditionally offered have made this research experience unique. I can’t begin to thank him enough, but to continue to strive for bigger goals in order to thank him for the many great lessons he taught me during my candidature.

I am very grateful to Professor Branka Vucetic, the director of Telecommunications research laboratory and the head of Electrical and Information Engineering School. Her initial approval of my application to graduate studies allowed me to be part of a stimulating research team, which later lead to the exploration of new research topics and confirmed my great interest in this field.

I would like to acknowledge the financial support from the University of Sydney through the University Postgraduate Award and Norman I. Price Scholarship.

I would also like to acknowledge my indebtedness to Dr. Seungcheon Kim. During his time at the Telecommunications Lab, he unreservedly shared his views and opinions about various research trends. Being a newbie to the OpNet package, I appreciated his initial guidance and it is still very much appreciated even at this date.

To those, who were at one stage lab-mates and now are themselves achieving researchers in various points around the world, Rubaiyat, Vishaken, Slavica and Louis, I am grateful for their extended friendship. The many “off-site” dinner gatherings gave us an opportunity to chat in a relaxing atmosphere about various aspects of life, ranging from current affairs to the discovery of new recipes.

Not to forget other friends I collected from the telecoms lab during this candidature, Wenching, Kegen, Chakree, Jade, Elaine, Joseph, Matthew, Robert and Fei. It is a blessing to have the opportunity of working with such a group of outstanding postgraduates.

Appreciation also goes to my dear friends Rev. Kai, Joey, Li-Wei, Amy, GiGi, Carol, Sophie for their continuing support. I realized how important it is to have a bunch of non EE friends. They open up your living spectrum, and make you experience life a little differently when seeing things from their perspective.
Thanks to Yang for taking me through a couple of the darkest valleys, as well as up some steep hills. I thank him not only for sharing technical insights, but for the courage to take up the responsibility of cheering me up when things get a little out of hand. His company late at night in uni makes me feel safe to stay behind during hectic times in order to get work done. His encouragements make sure that I always give my best.

Special thanks to my parents, Kai-Chung and Hua-Mei. Their endless and unconditional love reassures me it is ok to go for my dreams. From many past experiences, I am assured that no matter what happens in life I can always turn back and be showered with their most generous love. It is their pampering and nurturing support that gives me enough strength for another challenge. Appreciation also goes to my only and beloved sibling, Loraine, who recently received her Ph.D. degree from Cambridge University, UK. The many long-distance “sisterly” phone calls have helped tremendously in calming a sometimes anxious heart.

Finally, to my grandparents and the wonderful ladies I encountered during my visits to the nursing homes, it is the spirit of living, wisdom, and caring they shared that inspired me to be strong. I had always been very close to my grandmother, and there was always a very strong bonding between us. It is only through my experiences at the nursing homes, that I was astonished to learn how when you extend love and care to those lonely hearts of strangers, you get even more love in return. I am extremely thankful to the Buddhist society for giving me the opportunity to have the strength to make myself a better, stronger and more loving person each day. My deepest thoughts are to my grandparents, all three of them, who passed away during the four years of my Ph.D. candidature. Their spirits will always remain in my heart. To them, I dedicate this thesis.
# Table of Contents

Abstract ii

Acknowledgements iii

List of Publications x

List of Figures xii

List of Tables xvi

Chapter 1 Introduction 1

1.1 Problem Definition and Contributions 2
1.2 Thesis Outline 5

Chapter 2 Mobility Management in Wireless Networks 8

2.1 Introduction 9
2.2 3G Cellular Solution 10
   2.2.1 Observed deficiencies 13
2.3 Mobility in IP Networks 14
   2.3.1 Registrations 15
   2.3.2 Tunneling and routing 15
   2.3.3 Problems with Mobile IP
      2.3.3.1 Location management 16
      2.3.3.2 Routing management 17
2.4 Mobility Management in Satellite Systems 19
2.5 Conclusions 21

Chapter 3 Location Management in Cellular Networks 23

3.1 Introduction 24
3.2 Update Strategies 24
**Table of Contents**

3.2.1 Zone-based update scheme 26
3.2.2 Three-location area (TrLA) update scheme 27
3.2.3 Profile-based update scheme 28
3.2.4 Probabilistic location update scheme 28
3.2.5 Compression-based update scheme 29
3.2.6 Space-based update scheme 30
3.2.7 Threshold assignment 30
3.2.8 State-based update scheme 31
3.2.9 Load adaptive threshold system (LATS) 32
3.2.10 Direction-based location update scheme 33
3.2.11 Adaptive distance-based update scheme 33
3.2.12 Dynamic location area (LA) assignment update scheme 34
3.2.13 Predictive distance-based update scheme 34
3.2.14 Activity-based update scheme 35

3.3 Analytical Framework 36
3.3.1 Influential factors 36
3.3.2 Overall cost function 38
3.3.3 Update cost analysis based on markovian model 39

3.4 Conclusions 46

Chapter 4 Directional-Aware Location Updating Strategies 47

4.1 Introduction 48
4.2 Summary of Related Works 49
4.3 Adaptive Scheme Theoretical Rationale 50
  4.3.1 Mobility model 51
  4.3.2 Concept of directivity 54
4.4 Analytical Framework 57
4.5 Simulation Framework 60
4.6 Numerical Results and Discussions 61
  4.6.1 Effects of CMR 61
  4.6.2 Effects of $\omega$ 62
  4.6.3 Effects of $U/V$ 65
4.7 Applications and Implementations 68
4.8 Future Work 69
4.9 Conclusions 70
# Table of Contents

## Chapter 5 Kalman-Filter Based Update Schemes

5.1 Introduction 71
5.2 Related Works 72
5.3 Kalman-Filter Based Location Tracking 74
   5.3.1 Mobility model 75
   5.3.2 Concept of kalman filter 76
   5.3.3 Update boundary assignment 79
   5.3.4 Performance tradeoff 83
5.4 Simulation Framework 86
5.5 Results Discussions 87
5.6 Future Work 92
5.7 Conclusions 92

## Chapter 6 Sectional Paging

6.1 Introduction 94
6.2 Related Work 95
   6.2.1 Blanket polling 96
   6.2.2 Sequential paging 97
   6.2.3 Shortest distance first paging 97
   6.2.4 Time-elapsed paging 98
   6.2.5 Velocity paging 98
   6.2.6 Highest probability first paging 98
   6.2.7 Selective paging 98
6.3 Sectional Paging 99
   6.3.1 Design details 101
6.4 Analytical Model 104
6.5 Results Discussions 105
6.6 Conclusions 107

## Chapter 7 Mobility Management in IP Networks

7.1 Mobility Between Heterogeneous Networks 109
7.2 Macro-Mobility versus Micro-mobility 113
7.3 Macro-Mobility Solutions 114
   7.3.1 An overview 114
Table of Contents

7.3.2 Proxy agent architecture 115
7.3.3 More examples 117
    7.3.3.1 Cellular IP 117
    7.3.3.2 HAWAII 119
7.4 Extensions From Cellular Concepts 123
7.5 Conclusions 124

Chapter 8 More Studies on Cellular Networks 125

8.1 Introduction 125
8.2 Study of the Cellular Systems 126
    8.2.1 User profile replication scheme (UPR) 127
    8.2.2 Local anchoring scheme (LAS) 128
    8.2.3 Pointer forwarding scheme (PFS) 128
8.3 Simulation Framework 129
8.4 Programming Philosophy 136
8.5 Results and Discussions 139
8.6 Application Summary 145
    8.6.1 UPR 146
    8.6.2 PFS 146
8.7 Future Work 147
8.8 Conclusions 148

Chapter 9 Applications of Cellular Strategies onto IP Networks 150

9.1 Introduction 151
9.2 Concept of Network-Initiated Handoff 151
9.3 Location Tracking Schemes for IP Networks 154
    9.3.1 Design criteria 155
    9.3.2 Design procedure 155
9.4 Inspiration From Cellular Technologies 159
9.5 The Proposed Design 162
    9.5.1 An overview 162
    9.5.2 Specific design details 163
9.6 A Preliminary Feasibility Study 168
9.7 Performance Quantification Matrix 169
9.8 Conclusions 172
### Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 10</td>
<td>Summary of Contributions</td>
<td>173</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>178</td>
</tr>
<tr>
<td>Appendix A</td>
<td>An Overview of the Simulation Philosophy used in Chapter 8</td>
<td>185</td>
</tr>
<tr>
<td>Appendix B</td>
<td>The Generation of Various Call Arrival Rate for LAS</td>
<td>186</td>
</tr>
<tr>
<td>Appendix C</td>
<td>The Implementations of LAS</td>
<td>187</td>
</tr>
<tr>
<td>Appendix D</td>
<td>The Calculation of Database Querying load for LAS</td>
<td>188</td>
</tr>
<tr>
<td>Appendix E</td>
<td>The Calculation of Database Querying Load for UPR</td>
<td>189</td>
</tr>
<tr>
<td>Appendix F</td>
<td>An Illustration of Reductions in Querying Delay for UPR</td>
<td>190</td>
</tr>
<tr>
<td></td>
<td>Glossary of Acronyms</td>
<td>191</td>
</tr>
</tbody>
</table>
List of Publications

A. Book Chapter


B. Journals


C. International Conferences


D. Technical Reports


List of Figures

Figure 1.1  The structure of the thesis
Figure 2.1  The GPRS network
Figure 2.2  The GPRS 3-state model
Figure 2.3  Detailed signaling transfer for performing location updates
Figure 2.4  Packets flow in Mobile IP
Figure 2.5  Mobile IP tunneling (with IP-within-IP encapsulation)
Figure 2.6  Situations where registrations with HA is necessary
Figure 2.7  Route optimization mechanisms
Figure 2.8  Satellite application in global communications network
Figure 2.9  Interconnection of different terrestrial and satellite networks through interworking units
Figure 3.1  An overview of update schemes proposals optimized for individual mobile
Figure 3.2  An overview to the implications of various location updating strategies
Figure 3.3  An example of the Markovian-motion state machine
Figure 3.4  Markov chain based motion mobility model
Figure 3.5  A graphical representation of the transitional information required to evaluate the steady state probability $Q_{0,R}$
Figure 3.6  A modified Markovian-motion state machine
Figure 3.7  A modified Markov chain to present the motion model
Figure 3.8  An outline of the discussion on the analytical framework
Figure 4.1  The state Markov model for the 4-subarea cluster
Figure 4.2  Inter-relations between cells and state indices
Figure 4.3  State transitioning diagram for a ring cluster of threshold 6
Figure 4.4  Averaged state transitioning diagram for using an update threshold of 6
Figure 4.5  The state diagram illustrating the inter-ring transition probabilities in a hexagonal cell
List of Figures

Figure 4.6  A brief illustration of the overall operation

Figure 4.7  A brief illustration of the simulation framework

Figure 4.8  Impact of CMR on the selection of $d$

Figure 4.9  Impacts of $\omega$ on the selection of $d$ (a) $\omega=0.5$ (b) $\omega=1$ (c) $\omega=1.5$ (d) CMR $=0.05$

Figure 4.10  Impacts of $U/V$ on the selection of $d$ (a) $U/V=1$ (b) $U/V = 5$ (c) $U/V = 10$ (d) CMR $=10$

Figure 5.1  An example of cellular network

Figure 5.2  (a) Example of an actual mobile movement on a two-dimensional system

(b) The corresponding Gauss-Markov distributions for modeling mobile’s travel directions at boundary crossing

(c) A zoomed in picture of mobile’s movement demonstrating straight line approximations

(d) A zoomed in picture of mobile’s movement in the vicinity of the last update cell (0, 0)

Figure 5.3  Summarized operations for the Kalman filter

Figure 5.4  Fourier transforms for the analysis of mobile's movement characteristic (directional consistency level)

Figure 5.5  Inter-relations between the location updates and call arrivals in the time domain

Figure 5.6  Procedures for the assignment of update boundaries

Figure 5.7  Symmetrical versus asymmetrical update boundary

Figure 5.8  Steps for the evaluation of the updates number

Figure 5.9  Scenario 1($f_D = 0.015$) — a relatively directive movement pattern shown on a two-dimensional framework

Figure 5.10  Scenario 1($f_D = 0.015$) — the actual travelling directions at each boundary crossing (a sampling period of 15 steps)

Figure 5.11  Scenario 2($f_D = 0.04$) — a sampling period of 8 steps

Figure 5.12  Scenario 2($f_D = 0.04$) — a sampling period of 15 steps

Figure 5.13  Performance comparisons between the proposed scheme and the conventional distance-based scheme
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>An overview of paging schemes classified according to the nature of system definitions</td>
</tr>
<tr>
<td>6.2</td>
<td>Cell division method for (a) ring-structured paging and (b) sectional paging</td>
</tr>
<tr>
<td>6.3</td>
<td>The model of directivity quantification</td>
</tr>
<tr>
<td>6.4</td>
<td>The anticipated probability distribution functions for (a) $P_d = 0.33$, $P_f = 0.5$ (b) $P_d = 0.6$, $P_f = 0.8$ (c) $P_d = 0.8$, $P_f = 0.8$</td>
</tr>
<tr>
<td>6.5</td>
<td>An overview of the sectional paging scheme</td>
</tr>
<tr>
<td>6.6</td>
<td>Concept of sectional paging — An illustration of the bounded size selection</td>
</tr>
<tr>
<td>6.7</td>
<td>Paging cost evaluations for different directivity measures under varying bound size definitions</td>
</tr>
<tr>
<td>6.8</td>
<td>Performance comparisons for varying update thresholds</td>
</tr>
<tr>
<td>7.1</td>
<td>Inter-technology mobility scenarios</td>
</tr>
<tr>
<td>7.2</td>
<td>A possible scenario of network overlapping</td>
</tr>
<tr>
<td>7.3</td>
<td>A brief illustration of the relations between different levels of mobility</td>
</tr>
<tr>
<td>7.4</td>
<td>A list of design concepts frequently used in designing location management techniques</td>
</tr>
<tr>
<td>7.5</td>
<td>An illustration of the correlation between different location management techniques</td>
</tr>
<tr>
<td>7.6</td>
<td>A graphical presentation of hierarchical management and mobile network concepts</td>
</tr>
<tr>
<td>7.7</td>
<td>Cellular IP Operations</td>
</tr>
<tr>
<td>7.8</td>
<td>Parameter selection criteria in Cellular IP</td>
</tr>
<tr>
<td>7.9</td>
<td>Routing path for mobiles located in the same network</td>
</tr>
<tr>
<td>7.10</td>
<td>Handoff management in HAWAII</td>
</tr>
<tr>
<td>7.11</td>
<td>Movement within home domain (with no HA involvement)</td>
</tr>
<tr>
<td>7.12</td>
<td>Movement within a foreign domain</td>
</tr>
<tr>
<td>7.13</td>
<td>Handoff management in HAWAII</td>
</tr>
<tr>
<td>8.1</td>
<td>A graphical representation of the signaling transfers</td>
</tr>
</tbody>
</table>
List of Figures

Figure 8.2 The simulation framework
Figure 8.3 A brief illustration of the simulated network architecture
Figure 8.4 Illustration of the reduced updates from BSs to the corresponding VLR
Figure 8.5 The macro-cell mobility scenario ($n = 5$)
Figure 8.6 The micro-cell mobility scenario ($n = 2$)
Figure 8.7 Flow chart showing the generation of the trajectories in a micro-cell mobility scenarios
Figure 8.8 An overview of the simulated system
Figure 8.9 Impact of mobility on HLR updates (macro-cell scenario)
Figure 8.10 Impact of mobility on HLR updates (micro-cell scenario)
Figure 8.11 Impact of mobility on total updates (macro-cell scenario)
Figure 8.12 Impact of mobility on total updates (micro-cell scenario)
Figure 8.13 Impact of mobility on database querying loads (macro-cell scenario)
Figure 8.14 Impact of mobility on database querying loads (micro-cell scenario)
Figure 8.15 Impact of mobility on the total cost in managing mobility (macro-cell scenario)
Figure 8.16 Impact of mobility on the total cost in managing mobility (micro-cell scenario)
Figure 8.17 Combined operations of different location tracking strategies
Figure 9.1 A simple illustration of the network overlapping scenario
Figure 9.2 Illustration of the overall operational philosophy
Figure 9.3 A simplified network structure
Figure 9.4 The design philosophy of a multi-mode location management technique
Figure 9.5 The concepts of location distributions
Figure 9.6 A simple network scenario
Figure 9.7 The implementation of modulo ‘AND’ to determine the replication variables
List of Tables

Table 4.1  Impact of $\omega$ on the transitional probability
Table 5.1  Exchange of control messages
Table 5.2  Cellular IP versus HAWAII
Table 5.3  A comparison of design criteria between IP and Cellular networks
Table 7.1  Exchange of control messages
Table 7.2  Cellular IP versus HAWAII
Table 7.3  A comparison of design criteria between IP and Cellular networks
Table 8.1  Brief summary of the simulated location management techniques
Table 8.2  20 generated trajectory paths (for the macro-cell mobility scenario)
Table 8.3  Generation of trajectory (course 0) with random directional changes
Table 8.4  Main factors that would influence the efficiency of the system operations
Table 8.5  A comparison of the specific savings seen for various location management techniques
Table 9.1  A list of design criteria to be complied with in the protocol developments
Table 9.2  A summary of the complete design problem
Table 9.3  A comparison of the signaling transfers: assuming MN moves from FA1 to FA2 and then to FA3
Table 9.4  Possible criteria for the initiation of mode changes
Table 9.5  A comparison of the registration load
Table 9.6  A comparison of the call delivery load
Table 9.7  A matching between QoS parameters and corresponding measurements
Table 9.8  A list of the ultimate system requirements
Courage is the thing. All goes if courage goes.

Sir J.M. Barrie (1860-1937)

Chapter 1

Introduction

Technically, 3G mobile communication systems evolve by integrating three essential domains — broadband, mobile and Internet. In such a milieu, the increasing feasibility of virtual connections allows mobile users not only to roam freely between heterogeneous networks, but also to remain engaged in various forms of multimedia communications. Depending on whether it is geographical coverage, bandwidth or delay that is required, it would then be up to the users to decide when and how to switch from one access network to another depending upon the availability and appropriate cost/performance considerations.

This would represent a considerably advance towards an era of all-IP based communications. To cope with this, it will be necessary to implement the 3G system as a universal solution that prompts transparent user roaming among different wireless networks while delivering the widest possible range of cost effective services [1,2]. Already, mobility is no longer an exception but a norm. Mobile users demand continuous access to the networks during the lifetime of a connection without any unexpected interruptions.

Mobility management is the key to successfully enable seamless mobile services. Its process can be divided into three steps, namely, location management, routing management and handoff management. Very simply, it presents the problem of finding the moving mobile node’s location, and delivering data packets (destined for it) along the best route. As the
mobile node roams between networks, the question then is when, and how to perform the necessary actions in order to preserve the connection [3-6].

Among all challenges, developing an efficient location management technique is possibly the most important step in working towards the determination of an optimal solution to the problem of managing mobility. Given that this is a crucial component of the location tracking operations, the thesis is dedicated to presenting a detailed study of the problem. The importance of location management can be largely divided into two categories namely, the efficient utilization of the wired and wireless links [7].

In the case of wired links utilization, information is sent by the mobile users to the backbone network whenever the users move from cell to cell. This action involves not only information storing from the location database, but also information retrieving from those directories whenever it is necessary to locate a roaming user. In the case of wireless links utilization, the main problem is how to find an adequate trade-off between the registration load and the paging load. In general, the biggest challenge in framing location management is to find the most favorable tradeoff between the combined effects of the updating (or information) load and the searching load — the two parameters are frequently referred to as location registration cost and call delivery (or paging) cost. Intuitively, one would assume load increases when searching has to take place to locate the current address of the destination mobile terminal. Conversely, information load would increase due to frequent updates if a more complete location information database were to be maintained.

1.1 PROBLEM DEFINITION AND CONTRIBUTIONS

Along with a growing interest in other wireless technologies, the need for efficient operation in personal communications services (PCS) networks is increasingly present given that the number of subscribers has continued to rise in recent years. Given the irregular nature of the cell sizes in a cellular network, the behavior of mobile movement changes from cell to cell and from user to user. Thus, the need for designing an adaptive algorithm for tracking a roaming mobile has become imperative. As a result, the number of relevant publications found in the literature is extensive.

However, most of such location tracking algorithms have been proposed with the underlying assumption of random movement patterns. Given that mobile communications are now used so universally, it becomes unrealistic to assume random movements as mobiles frequently inherit traceable patterns. Despite the importance of this, there seems to be little research focusing on the potential of utilizing such directivity information, directivity being a measure of the correlation level in traveling directions observed between adjacent cell boundary crossings.
Thus, in the first part of the thesis, the focus was on making the proposed location tracking schemes adaptive to mobility patterns; more specifically, to the distinct features of transition directions. The aim was to demonstrate how variations in the movement pattern affect the selection of the updating threshold, which would in turn, influence the optimal performance achievable. In this thesis, the development of two such predictive update schemes is detailed.

In the first design, a directivity-aware update scheme is proposed. The anticipation of movement directions is realized by the introduction of a transitional directivity index $\omega$. The selection of optimal threshold is made as an adaptive process with regards to $\omega$. It is assumed that when $\omega > 1$, the mobile has a greater tendency of roaming “outwards” and hence a higher probability of transition to the next ring during each movement. As a result, the defined threshold will be reached sooner incurring more update costs. The idea is to explore the possibility of making the selection of optimal threshold an adaptive process to $\omega$. It aims to demonstrate how variations in the roaming characteristics would affect the selection of an optimal updating threshold (i.e. incur a minimal operational cost), and consequently, to study whether the information revealed by the transitional directivity index alone is sufficient to give reliable predictions for the mobility patterns and subsequently the location probability distributions.

In the second design, a novel Kalman-filter based update scheme is proposed. The idea is to explore the possibility of observing a mobile’s past movement history, and to anticipate each individual’s future roaming characteristic, be it at a certain time of the day or in a specific geographical area.; more specifically, by appropriately adjusting the feedback periods of a Kalman filter, the essence of distinct movement patterns with varying degrees of consistency in traveling direction is considered in selecting the optimal update trend.

As expected, simulation results confirmed the original hypothesis that the more “focused” the traveling pattern is (in terms of directivity), the higher the optimal threshold can be. In other words, the higher the directivity, the more precisely the location can be predicted and consequently, the update threshold can be increased to minimize the required registration cost. This is particularly advantageous in reducing the update portion of the operational cost especially when the mobile is traveling fast. However, acknowledging the tradeoff between registration and paging cost evident in most existing solutions [7-11], it is certain that as the update threshold increases, the paging cost will grow in direct proportion.

This observation signifies the need for designing a paging scheme that corresponds to the update operations. This scheme would minimize the necessary paging cost without having to sacrifice the merits of allowing for a higher update threshold. The sectional paging scheme was thus developed to fulfill such a requirement. As a result, it will be now possible to assign
each mobile user an optimal threshold, which, while being large enough to promote savings from registration, is also small enough to avoid having to incur too much paging expense.

Given identical network conditions and mobile characteristics, simulation results reveal that sectional paging outperforms the ring-structured paging scheme as long as the mobile has a granularly traceable movement pattern. Considering ring-structured paging is the most commonly applied paging scheme in the literature, such a research outcome is remarkable/significant. With a notable reduction in total paging costs, it is concluded that the usage of the newly proposed scheme is most suitable when the roaming pattern is either traceable or can be predicted with reasonable precision. As an additional advantage, it is noted that the proposed technique could be used for all basic update schemes. Provided that the prediction of location probability is reliable, a movement pattern with increasing directivity will justify having increased update thresholds without incurring additional paging costs.

Key design criteria for such location tracking mechanisms include simplicity, scalability, reliability, efficiency/cost measures and power consumption at mobile terminals. Therefore, apart from being easy-to-implement, the proposed techniques allow the network to cope with large variations in traffic loading without too much compromise on the quality of service (QoS) provisioning and cost measures. A constrained dependence on specific network entities is also necessary to ensure that the scheme operates with some guaranteed reliability independent of certain unpreventable infrastructure failures.

The developed schemes in the first part of the thesis are mainly for applications in cellular networks, however in the second part of the thesis, the focus is on the development of mobility management techniques within the IP framework. Technically, instead of restricting movements to one single network, the mobile will have been freed to roam between heterogeneous networks.

Mobile IP has gained significant attention especially in the Internet IETF community for its capability of offering a solution to this inter-domain roaming problem. Although it provides a first step toward movements between access networks, it does not give sufficient consideration to frequent movements at the subnet level; this in turn, results in non-optimal local mobility operations. Foreseeing such a deficiency, the fundamental idea is to develop a mobility management technique that would enable mobility at the IP level to improve terminal mobility. A brief study of the existing schemes indicates that one of the fundamental problems in managing mobility between heterogeneous networks is the selection of the location to perform the registrations. Specifically, the question is where would be the optimal place to have the new location information registered for mobiles that are roaming, yet remain to be idle. Could registrations other than the home network provide an improved performance in terms of connection delay and operational cost? The task is not only to provide seamless
roaming between fixed and mobile networks, but also to allow developments in new IP applications.

1.2 THESIS OUTLINE

Chapter 2 reviews the state-of-the-art in the developments of mobility management techniques currently available in wireless networks. Not only are the fundamental problems in the challenge of location tracking outlined, but reasons behind the importance of developing efficient mobility managing techniques are described. Through brief descriptions of the existing solutions, the remaining challenges awaiting further research are identified and examined. This chapter also describes the origin, formation and justification of this thesis.

Chapter 3 provides an overview of the existing location tracking mechanisms. Despite a large number of proposals available in the literature, the number of publications that have actually provided an overview is small, let alone attempts to identify the inter-relations between developments. The intention of the chapter is not only to give a complete overview of the existing developments for cellular networks, but also to provide a critical analysis of the existing technologies. In the discussion, the main merits and downfalls of the existing techniques will be summarized to indicate areas of possible improvements. This chapter is structured such that it not only gives a brief description of individual location management strategy, but also highlights the interrelations between various proposals.

Chapter 4 details the development of a new adaptive tracking strategy — a directivity-aware location updating scheme. In this scheme, an optimal distance-based update threshold is selected not only as a function of the call-to-mobility ratio, but also as a transitional directivity index $o$; a new parameter introduced to give some measures of the mobile's traveling patterns. More specifically, the emphasis is on making the proposed scheme adaptive in response to the periodic change of transition directions. With different transition probabilities assigned for movements in varying directions, it is possible to distinguish individual mobile traveling patterns. This chapter aims to demonstrate how variations in the inter-ring transitional probability would affect the selection of an optimal update threshold such that a minimum cost incurs.

The developmental details of the proposed Kalman-filter based update scheme are summarized in Chapter 5. The purpose of the chapter is two-fold: (1) to suggest a distribution model that is capable of describing and thus simulating a wide range of movement patterns with varying correlations between traveling directions and (2) to explore the possibilities of applying Kalman filter for the assignment of update boundaries. Hence, instead of simply assuming a random movement model, this particular research work focuses
on methods of "detecting" directivity correlation between movements. Specifically, it is intended to demonstrate how a measurement can be obtained from the traveling characteristics; and secondly, how feedback values can be used to decide on the shape of the registration areas. Assuming the defined area can reflect the mobile's roaming characteristics with reasonable precision, the challenge of bandwidth conservation would be well addressed through reductions in the number of location registrations.

Chapter 6 outlines the development of sectional paging, a new paging scheme designed to lessen the dependence between update and paging load. Thus, although complying with the required delay constraints, QoS measures will not need to be sacrificed as a result of increasing the update threshold. Without having to install much additional intelligence, the proposed scheme predicts the likelihood of residence and assigns the optimal paging boundaries. Consequently, it is expected that paging cost will reduce significantly for roaming mobiles with traceable patterns. Provided that the prediction of location probability is reliable, a movement pattern with increasing directivity will justify having increased update thresholds without incurring additional paging costs.

Having identified the problems that have made the existing solutions with Mobile IP unsatisfactory, analyses are made in Chapter 7 to show firstly, how other proposals differ from the current approach, and secondly, how effective are the present solutions in solving the inefficiencies evident in Mobile IP.

While chapters 4 and 5 focused on the determination of an optimal time in performing location registrations, chapter 8 goes one step further and looks at the issues involved in searching for an optimal location for update. The question is whether updates at the home network will in fact give an optimal performance, and if not, where are the alternative centers for registration.

With the merits of various individual schemes becoming clearer, in chapter 9, it is appropriate to consider similar implementations of cellular-based tracking techniques within the IP framework. While a more detailed operational design will be necessary for further development, this chapter gives an overview of the design criteria and the incorporation of special characteristics. In addition, a brief discussion of the anticipated performance is included to provide a preliminary feasibility study of the proposed schemes.

Even with such an extensive study in the sphere of location management, an all-embracing solution to the challenge of mobility management is still far from completion. Concerns such as increased security risks and the need for fairer billing mechanisms all require more comprehensive treatment to optimize the overall operation. In recognizing such a requirement, Chapter 10 summarizes the main contribution of the research carried out for this Ph.D. candidature.
Figure 1.1 summarized the structure of the thesis.

![Diagram showing the structure of the thesis]

**Figure 1.1** The structure of the thesis
In all things, success depends upon previous preparation, and without such preparation there is sure to be failure.

Confucius (c.550-c.478 BC)

Chapter 2
Mobility Management in Wireless Networks

Despite GSM having shown some satisfactory operations in the 2nd generation mobility systems, with increasing usage of data transmissions, several inefficiency problems associated with traditional GSM location updating operations have gradually become apparent. Thus, there seems to be a need for better mechanisms in managing frequent mobility. In particular, the control signaling is necessary to be maintained at a minimum such that a better management can be achieved over the scarce radio resources. To better justify the intention of this research work, brief descriptions of the location management technique used in the GSM system, and its 3G extension, GPRS systems are included. The level of signaling details provided in the following sections should be sufficient to highlight some of the obvious deficiencies associated with this standardized technique.
2.1 INTRODUCTION

With the increasing usage of portable computers, wireless networks and satellite, mobile computing has generated great interest in the last few years. It is envisaged that mobile-related traffic will soon be comparable in volume with that related to fixed networks. Thus, instead of restricting connections to be maintained always at a fixed position in the network, users in this environment are now free to roam. Indeed, the research community is at a stage to search for a common solution that would be capable of accommodating the gradual convergence of the two technologies: internet access and mobile communications. Consequently, the main stream of data traffic will no longer be limited to simple voice transmissions. High bandwidth applications, such as multimedia and video, will eventually take over most of the available media resources [12].

There is no doubt that the trend is toward the networking world’s ultimate goal: a continuously connected computing environment, anywhere anytime [13-18]. As a result, the need is obvious to have a computing environment capable of automatically “switching” to the best available processing and access mechanisms at any given time. Ideally, such a capability of system realization will not only enable cost management to be put in place, but also allow mobile services to remain “transparent” to users. In short, any on-going communications will be maintained regardless of any changes made in user locations, the communication devices in use (e.g. modem, Ethernet card), or the communication environment (i.e. noisy wireless vs. wired ATM).

Obviously, there are two possible approaches in tackling the problem: cellular system based and IP based solutions [4]. Intuitively, while a cellular based solution enhances the current mobile communications by extending capacity for data and multimedia applications, IP based solutions allow user mobility by maintaining on-going connections even for frequent changes in network point of attachments.

On the one hand, IMT-2000 (International mobile Telecommunications) is a unified 3G mobile system that supports both packet-switched and circuit-switched data transmissions with high spectrum efficiency, making the vision of anywhere, anytime communications a reality. Basically, it is a collection of standards that provides direct mobile access to a range of fixed and wireless networks. Among all, the three most significant developments are UMTS, cdma2000 and UWC-136 that represent 3G successors to the main 2G technologies of GSM (Global System for Mobile communications), IS-95 and AMPS, respectively [19]. The general idea was to make the development of 3G wireless technologies a gradual process from circuit-switched to packet-switched. Take GSM for example, in order to have the system enhanced with improved services (by means of increased capacity, coverage, quality and data
rates), the evolution to 3G was made possible through the incorporation of an intermediate stage called GPRS, the general packet radio services.

On the other hand, the most important barrier in developing mobile internetworking is the way Internet Protocol (IP) operates. It supports the interconnection of multiple networking technologies into a single logical internetwork and is the most widely used internetworking protocol. An IP address is used to identify a host and contains information used to route the packets.

Internet Protocol depends on the IP addresses that are assigned to the fixed source and destination. Any solution for supporting mobility in the Internet is constrained by the requirement of existing Internet foundation and networking applications. As the mobile user is roaming between foreign networks, it will acquire a new IP address causing the established connection of the node to be lost. Hence, it contradicts the main idea of a mobile node!

This underlying concept results in the necessity to develop Mobile IP, a modification that allows IP to accommodate mobile nodes. Mobile IP provides a means of delivering packets addressed to the mobile node. By defining special entities, Home Agents (HAs) and Foreign Agents (FAs), a Mobile Node (MN) is able to roam freely without changing its IP address [20]. This in turn, provides a means for Mobile IP to deliver packets addressed to a specific MN.

This chapter summarizes the operations of the existing technologies and provides the necessary background for understanding the state-of-the art of current technologies including GPRS (general packet radio services), Mobile IP and mobile satellite systems. Section 2.2 summarizes the evolution of the 3G cellular wireless networks. The main operations of GPRS, a packet-switched version of the 2G GSM system, are briefly discussed. Following this in section 2.3, a general overview is presented to show how Mobile IP, an enhancement of the existing IP networks, provides the basis for Internet roaming. The way in which mobility is handled in Mobile IP will also be addressed in greater detail. Section 2.4 briefly inspects the role satellites play in 3G networks. For all cases, specific problems identified in the operation of mobility management are summarized to justify the research work outlined in this thesis.

2.2 3G CELLULAR SOLUTION

GPRS completes today’s GSM cellular networks with a packet-switched network that increases data transmission rate to a maximum of 115 kbps. In fact, it is noted that the offered bandwidth is expected to increase further with the introduction of an access extension EDGE (Enhanced Data rate for GSM Evolution) reaching a tripled rate of 384 kbps. Moreover, as the
available bandwidth is now used adaptively according to the service requirements, it is envisaged to achieve higher bandwidth efficiency [21-24].

In general, the development of GPRS is significant; it has not only provided a seamless and direct connection to the external data networks (e.g., the Internet), but also created a bridge allowing further evolution to 3G with only minimal changes to the existing infrastructures. Technically, such additional network capability is achieved by the incorporation of two main components SGSN (serving GPRS support node) and GGSN (gateway GPRS support node). On one end of the connection, SGSN sends and receives data from and to mobile users through connections to a base station subsystem (BSS). Residing at the same hierarchy level as mobile switching center (MSC) in the network, it also keeps track of GPRS users, performing a similar function as MSC to the conventional GSM subscribers.

On the other end of the connection, a gateway node GGSN is used to maintain connection with external data networks (e.g., Internet, X.25, etc). To illustrate the interconnections between various network entities, Figure 2.1 highlights the corresponding signaling transfers involved in setting up a connection.

![Figure 2.1 The GPRS network](image)

Two encapsulation schemes are used for routing purposes. On the one hand, the interconnection between the GGSN and multiple SGSNs are made available through interface Gn with packets transmitted between the nodes being encapsulated by a special GPRS tunneling protocol (GTP) [25]. On the other hand, when packets are encapsulated simply to decouple logical link management from network-layer packet data protocols (PDPs), there is no particular restriction on the actual encapsulating mechanisms.
Generally, the mobility management technique deployed in GPRS allows both an independent and in parallel use of conventional GSM services. An initial setup phase is required to connect the roaming mobile station (MS) to the GPRS system. This is to simply establish a logical link between the serving GPRS support node (SGSN) and the mobile station, which (i.e. the MS), is then assigned a temporary logical link identity (TLLI).

As far as location management is concerned, the connected MS negotiates with SGSN for routing contexts in one or more packet data protocols. Having verified the accessing status of MS with each of the requesting PDPs, SGSN notifies its corresponding gateway GPRS support node, GGSN, to update the stored routing context with relevant tunneling information and the serving SGSN address.

GPRS uses a three-state model to track the location of the MS as shown in Figure 2.2. This in turn, varies the frequency at which location updates are generated. Briefly, during the ready state, MS informs SGSN about every cell change. In the case where there are changes involved in routing areas (RAs), a routing update request, with information outlining RAs of both new and old cells will be sent during the standby state. Updating notifications to GGSN and GR/HLR is only required when a change in RAs also results in changing SGSNs. In this case, as the routing context is now different, MS' PDP context will have to be updated in the new SGSN before corresponding information is removed from the old SGSN.

![Figure 2.2 The GPRS 3-state model](image)

Note that as this simulation work focuses mainly on the interactions between various VLRs and the HLR, the presence of all the immediate stages, for example SGSN, justified to be ignored.

Furthermore, it is noted that based on the enhanced core network of GPRS, UMTS (universal mobile telecommunications system) is designed to be the back-compatible 3G standard for GSM. UMTS is the European proposal for a 3G mobile system aiming to support multimedia services with extended intelligent network features and functions. As a first step
of the "integration," UTRAN (the UMTS terrestrial access radio network) will coexist with GSM access networks. The idea was to develop the UMTS core network by gradually incorporating the desired UMTS features to the GSM/GPRS core network. At this stage, UTRA supports time division duplex (WCDMA-DS TDD) and frequency division duplex (WCDMA-DS FDD) modes with the combined operation offers an optimized solution to coverage areas of all sizes. A further multi-carrier (MC-FDD) mode is to be established at a later date intending mainly for the use of cdmaOne/cdma2000 evolutions [26].

### 2.2.1 OBSERVED DEFICIENCIES

As far as the procedure of location management is concerned, GPRS and EGPRS inherit a similar location management technique of GSM — both use a two-level architecture for managing users' mobility. In general, each location area is normally assigned with a visiting location register (VLR), and thus, whenever the mobile moves out of the current location boundary, an update is due. In extreme cases where a separate VLR is assigned at the cell levels, registration operations are performed upon cell boundary crossings. While the detail of signaling transfers is illustrated in Figure 2.3, the basic concept is that all location updates have to register firstly at the local visiting location register, and then at the home location register (HLR) regardless of the geographical location of roaming. Thus, every arriving call will need to send a query directly to the called mobile's HLR requesting a call setup.

![Diagram](image)

**Figure 2.3** Detailed signaling transfer for performing location updates

In comparison to the GSM operations, the only distinguishable difference is observed handling active mobiles during the location tracking operation. Specifically, instead of
updated the HLR for every cell crossing as is the case for GSM, only RA crossings need notifications in GPRS. However, to compensate for the reduced VLR to HLR updates, additional paging is necessary before the location of a roaming mobile is identified. Providentially, location updates would normally create a lot more load on the network compared to that contributed by paging signaling. Thus, despite the additional paging, greater system capacity can still be expected from GPRS/EGPRS. Nevertheless, there still exists some fundamental inefficiency with the basic update operations.

As cell sizes are continuously reduced to comply with the ever-increasing capacity and coverage, the increasing signaling traffic can become a somewhat costly process. This is especially a concern as the act of global roaming where the mobile is far from its assigned HLR has gradually become the trend of usage.

2.3 MOBILITY IN IP NETWORKS

Having considered the cellular-based network solutions, it is now time to consider the other alternative, where mobility is incorporated to the conventional data services framework of IP networks. Generally, the work currently performed involving Mobile IP aims to extend the IP protocol in order to deal with mobile nodes.

One of the main barriers in developing mobile internetworking is the way Internet Protocol (IP) operates. IP supports the interconnection of multiple networking technologies into a single, logical inter-network and is the most widely used internetworking protocol. An IP address is used to identify a host and contains information used to route the packets.

Internet Protocol depends on the IP addresses that are assigned to the fixed source and destination. Any solution for supporting mobility in the Internet is constrained by the requirement of existing Internet foundation and networking applications. As the mobile user is roaming between foreign networks, it will acquire a new IP address causing the established connection of the node to be lost. Hence, it contradicts the main idea of a mobile node!

This underlying concept results in the necessity to develop Mobile IP, a modification that allows IP to accommodate mobile nodes. Mobile IP provides a means of delivering the packets addressed to the mobile node. By defining special entities, Home Agents (HAs) and Foreign Agents (FAs), a Mobile Node (MN) is able to cooperate in moving without changing its IP address [36]. This in turn, provides a means for Mobile IP to deliver packets addressed to a specific MN.

In Mobile IP, each mobile node is given a virtual home network. This remains unchanged, and is used to assign the mobile node a constant IP address in the same manner that a standard IP address is given to a stationary host. On the home network, a location information database is maintained for each of its subscribed mobile nodes (MNs) which are currently visiting
other networks. The accuracy of this information becomes vital when routers are to deliver any MN addressed packets.

The core operations involved in Mobile IP include agent discovery, registration, and packets tunneling. This is exactly what mobility management is defined to be: to detect MN’s change of location, register the new location with HA (either directly or via FA, the foreign agent), and finally to perform handover as MN moves to the new network.

The following sections detail the main actions required from both the nodes and agents to ensure a successful implementation of mobility management [27-30].

2.3.1 REGISTRATION

Upon detecting a change in location, the roaming MN acquires a new IP address, Care-of-Address (CoA), from either the received foreign agent advertisement (FACoA), or dynamic host configuration protocol, a collocated CoA (CCoA). MH then notifies HA of the new location through the process of registration. Figure 2.4 gives a brief illustration of how Mobile IP works.

![Figure 2.4 Packets flow in Mobile IP](image)

2.3.2 TUNNELING AND ROUTING

Data packets (from the correspondent nodes CNs) are generally routed by default to the MN’s home address. HA attracts packets destined for those nodes that are away from their home network, and re-delivers them according to the corresponding CoAs being registered by each roaming node.

After the registration with the HA is complete, the mobility management protocols should secure a way for packets to be routed to the current point of attachment, a process referred to as tunneling. The method used to forward data to roaming MN is known as encapsulation. Though Mobile IP assumes an IP-within-IP encapsulation methodology, other encapsulation
mechanisms are applicable upon agreement made between relevant network entities. An example of such other alternatives is *minimal encapsulation*.

Figure 2.5 briefly illustrates the operations of the IP-within-IP encapsulation methodology. This operation involves the encapsulation of one datagram into another datagram as its payload. Once the encapsulated datagram arrives at its intermediate destination (i.e. FA), it is decapsulated, yielding the original IP datagram. Subsequently, the IP datagram is delivered to its intended destination based on the information specified in the *destination address field*.

![Diagram of Mobile IP tunneling (with IP-within-IP encapsulation)](image)

**Figure 2.5** Mobile IP tunneling (with IP-within-IP encapsulation)

### 2.3.3 PROBLEMS WITH MOBILE IP

In spite of the fact that Mobile IP has been recognized as "the" solution in go about combining operations of mobile computing and wireless communications, there are still quite a few problems that are awaiting to be solved. Basically, the inefficiencies can be classified into three main categories according to each step of the mobility management process: location management, routing management, and handoff management.

#### 2.3.3.1 LOCATION MANAGEMENT

Refer back to Mobile IP operations, a serious inefficiency is widely evident in that a registration with HA, is required at every handoff. This includes not only changes in networks (Figure 2.6 c), but also any change of links, a common scenario in a campus network where one faculty is assigned a different “link” from the other (Figure 2.6 b). Such a registration
requirement applies also to an even smaller scale of mobility, say from one point of attachment to another on the same link (Figure 2.6 a).

As a result, it is not hard to imagine the much wasted resources that are associated by the frequent location updates arising from just one single MN’s frequent movement — not to mention the thousands, if not millions of MNs that are simultaneously attached to the network at any instant time. Although the need of notifying the home network of mobile users’ current locations is always present, it is questionable whether accurate location information is essential for MNs that are not active (i.e. no data transmission). It is thus logical to separate the location tracking operations for idle and active mobile nodes. Consequently, the update frequency at which MNs’ movement notifications are sent can be appropriately adjusted according to the actual roaming characteristics.

![Diagram](image)

**Figure 2.6** Situations where registrations with HA is necessary

The main problem then, is how to find an adequate tradeoff between the searching load and the information load. Load increases when searching has to take place to locate the current address of the receipt MN. Conversely, more information load would result (due to frequent updates) if a more complete location information database is to be maintained.

### 2.3.3.2 ROUTING MANAGEMENT

One of the biggest concerns in Mobile IP is the inefficiency associated with the way packets
are delivered to roaming MNs, namely the triangle routing, an asymmetric routing with respect to topology.

From the users’ perspective on the one hand, this operation will not only result in a high latency (due to a possibly distant HA), but will also cause mobility related disruptions; for example, packet losses during handoffs. From the network operators’ perspective on the other hand, frequent movements of MN will incur high control overheads which then reduces the efficiency of the link (or network) resource usage during tunneling. Moreover, the support of QoS will become complex.

If the corresponding node were to know, somehow, MN’s CoA, then it can tunnel the packets straight to the mobile node without having to bypass HA. This is what’s known as Route Optimization [31,32]. The main idea is to allow potential corresponding nodes to keep an updated mobility binding for the mobile node of interest. A ‘binding’ associates each MN’s home address with its current location, CoA, and is maintained by CNs in a binding cache. When a CN is to initiate communication with a roaming mobile, the location information is retrieved, and packets can be sent accordingly. Note that for security reasons, it is the HA’s responsibility to provide binding updates (BU) to any concerned nodes upon the completion of an authentication process.

Even with route optimization, the packet-loss problem during handoff still remains unsolved. As a packet is routed to the receipt MN, there is a possibility that the packet arrive at the FA just after the mobile node leaves. Ideally, these packets should not be dropped, but to be further redirected to the new FA (possibly by the previous FA). One solution is to let previous FAs also maintain a binding cache for their formerly visited MNs. Thus, by encapsulation, misdirected packets can be re-delivered, and hence to promote a smooth handoff. Figure 2.7 provides a graphical illustration of the route optimization mechanism.

![Route Optimization Mechanism](image_url)
Though with route optimization, Mobile IP does seem to have an enhanced operation, there are however other complications that should be taken notice of. For example, data transfer may be disrupted while CN is obtaining new bindings. Moreover, additional traffic load can be imposed on the network even for roaming mobiles that are idle.

### 2.4 MOBILITY MANAGEMENT IN SATELLITE SYSTEMS

A satellite-based solution allows users mobility by maintaining all on-going Internet connections even in the presence of frequent handoffs or changes in the network point of attachments. Satellite UMTS (S-UMTS) for example is considered as a part in 3G networks [33]. The satellite segment of the network connects through appropriate interworking units (IWU) to the ground segment. An illustration of this incorporation of satellites in providing mobile Internet connectivity is shown in Figure 2.8. IWU for the satellite has similar functionality as the gateways used to interconnect 2G and 3G networks for interoperation of those networks during the transition period from 2G to 3G as well as the gateways used for interconnection of different operator networks of the same kind (e.g. GSM). Such a concept is shown in Figure 2.9.

![Figure 2.8 Satellite application in global communications network](image_url)

Furthermore, mobile satellite systems using non-geostationary orbits such as the lower-earth-orbit (LEO) and the medium-earth-orbit (MEO) have very similar characteristics with cellular networks. The main similarity comes from the resulting handoff issues due to the arrangement of a cellular-like coverage area used to increase the total system capacity. However, there is also one essential concept that differentiates the specific operations of
mobility management in the two systems. Specifically, the “entity” being considered as the moving object is different; the former encounters mobile movements within fixed network architecture and the latter incurs satellite movements in reference to fixed mobile nodes. Consequently, while the prediction of the mobile entity’s deterministic location is relatively easy to obtain in satellite systems; such certainty is not guaranteed in cellular systems. In other words, the difficulties encountered in cellular networks do not seem applicable to its satellite-mobile network counterparts. As a result, this suggests that the usage of the mobility management technique should be different. Essentially, appropriate modification of the existing terrestrial-based solutions will be necessary before similar operations are suitable for applications on satellite-incorporated third generation systems.

![Diagram](image)

**Figure 2.9** Interconnection of different terrestrial and satellite networks through interworking units

The use of LEO satellites on the one hand, is most favorable for its high traffic capacity and reduced user power requirements both in satellites and in ground terminal. However, as individual LEO satellites rotate relatively fast along the earth surface, handoff becomes a particular concern in coping with the non-stationary nature of the coverage area. Depending on the relations of the two satellites involved in the handoff operations, three types of handoff are classified. Basically, intra-satellite handoffs are used to describe changes between spot beams under the management of the same satellite. With the increasing involvement of network management, inter-satellite handoffs indicate handoffs between satellites and link
handoffs incur due to changes in the connection pattern of satellite footprints. As an example of the latter scenario, such handoffs occur when links to adjacent orbits are turned off when the concerning satellite moves near to the polar region. Thus, the task is not only to efficiently utilize the available frequency spectrum, but to also minimize unnecessary forced termination of connections due to handoff failures. In other words, it would be essential to at least attempt to anticipate user motions and to reserve the resources accordingly for the predicted residential time. A brief literature survey indicates that there are two major prioritization techniques that were designed specifically for such purposes; (1) use of guard channels and (2) queuing of handoff requests when the resources were not currently available. Consequently, the operation of call admission control also becomes important as it decides whether sufficient resources would be available to accommodate the newly arrived transmission requests [34].

Cellular networks on the other hand, focused more on the efficient operations of location management specifically for idle mobile users. Thus, although the basic problem of managing mobility is the same, the actual emphasis on the system developments is different for satellite and cellular systems. In fact, due to the movement of LEO satellites, definitions of location area cannot be fixed even for the duration of a connection life.

Despite of the differences, there are, however, certain aspects where the two systems show similar operational purposes. For example, in terms of routing, [35] has developed a protocol aiming to reduce the frequency rerouting attempts during a link handoff. Basically, target probability is defined to quantify the estimated duration of residency of a mobile terminal on one particular inter-satellite link (ISL). During the route establishment of a new call, only links that can demonstrate a lifetime of greater than the target probability will be considered to form segments of the route. Though it might not seem obvious, this idea is similar to the predicting method used in location management operations in cellular networks; specifically, the optimal sequence of paging is selected based on the information obtained about the probability of residence in individual cells. Thus, while acknowledging the fact that the prediction method would have been easier in satellite systems (due to the fact that its motion is deterministic and predictable), the goal of determining efficient operations in both systems is the same. Consequently, although potential re-applications of the designed cellular techniques on satellite applications are not considered in the thesis, the possibility of performing similar tasks should be acknowledged.

2.5 CONCLUSIONS

In the sphere of terrestrial networks, there are at present, two possible approaches for
establishing the task of mobile computing: cellular-based and IP-based solutions. Intuitively, while a cellular-based solution enhances current mobile communications by extending capacity for data and multimedia transmissions, an IP-based solution allows user mobility by maintaining all on-going Internet connections even in the presence of frequent handoffs or changes in the network point of attachments.

In view of the problems in the existing solutions, clearly, some dynamic schemes will be required to handle user mobility. The remainder of the thesis focuses on developing new alternatives to enhance the current operations. To start with, chapter 3 summarizes the available alternatives for applications in cellular systems. The intention of that chapter is to provide an overview of the available techniques, and to determine areas for further improvements.
Study the past, if you would
Divine the future.

Confucius (c.550-c.478 BC)

Chapter 3
Location Management in Cellular Networks

Despite certain fundamental differences between various technologies, the need for an efficient solution to the problem of location tracking is present in all systems. The task of searching for an optimal tradeoff between the update and paging operations remains the biggest challenge in location management. With the tremendous growth in usage started in the last decade, research interests oriented around cellular networks continue to increase even to this date. We feel the time is now appropriate to further open up the spectrum for developments in cellular networks. Thus in this chapter, a comprehensive literature review of existing and specifically designed location management techniques for cellular networks is presented. The examination aims to be both informative and critical. Thus, while giving a comprehensive overview of existing developments, the study also provides a critical analysis of the significance of the various developments.
3.1 INTRODUCTION

Mobility is no longer an exception, but a criterion that is to be complied in today's communication era. An efficient location tracking mechanism is a crucial step in complying with the desired QoS measures. With the irregular nature of the cell sizes in a cellular network, the behavior of mobile movement changes from cell to cell and from user to user. Thus, the need for designing an adaptive algorithm for tracking a roaming mobile becomes more imperative than ever. As a result, the number of relevant publications found in the literature is extensive. However, as some have succeeded in being referenced more frequently than others, it is easy to lose track of the complete set of the available solutions. Despite a need for it, the number of publications that have actually provided an overview is small, let alone attempts to identify the inter-relations between developments. This observation provides genesis of this chapter. It intends not only to give a complete overview of the existing developments for cellular networks, but also to provide a critical analysis of the existing technologies. In the discussions, the main merits and downfalls of the existing techniques will be summarized to indicate areas of possible improvements. This chapter is structured such that it not only gives a brief description of individual location management strategy, but also highlights the interrelations between various proposals. The findings are significant not only to aid the basic operations of cellular networks including more recent developments of GPRS, but also to enhance various aspects of mobile communications.

In addition to a brief description of the operations, the merits and downfalls of each model are critically examined. A list of specific location techniques is compiled in sections 3.2 to highlight differences and similarities in the key design philosophy of various proposals. Through a critical evaluation of the designed operations, the significance of distinct features observed from each protocol is accessed and compared. Section 3.3 then compares the characteristics of the commonly applied analytical framework. Although there are other measures which are also important to evaluate the effectiveness of a location tracking solution, a general comparison guideline is that the fewer the combined numbers of updates and paging, the better the location tracking technique.

3.2 UPDATE STRATEGIES

In studies that do perform similar tasks [7,34,37,39], individual strategies are classified principally into static and dynamic schemes. The term static is used to describe specific techniques where the assignation of location area is fixed independently of the individual mobile characteristics. Conversely, the term dynamic refers to update strategies where the formation of location areas may vary according to the actual mobile characteristics.
Such categorization is general and almost in some ways simple as little additional information is revealed to reflect the more specific nature of individual update technique. To avoid similar deficiencies, in this chapter, update schemes are more carefully categorized to ensure their individualities are appropriately highlighted; specifically, the individual scheme is classified according to the mechanism that is used to decide the need for initiating an update operation.

The central aim of an update scheme is to define the time or location at which a new registration is required such that the network will at all times have a granular view of the mobile's location. Thus, upon the detection of arrival at a new cell, a decision is made to decide whether a new registration is required depending on the actual location. Although the final effect will be the same, the actual operation details might vary.

For most proposed update operations, the need for update is determined by a simple table lookup. By and large, the mobile terminals keep in their memory a list of cell identities (ID) retrieved from the network normally during the last registration operation. Upon the detection of roaming to a new cell, the mobile compares the current cell ID with that stored in the database. In most conventional schemes where the location area is fixed, the mobile is required to update the network when exiting the currently registered area. In other words, the mobile has moved out of the expected area, where the current location has not been previously included in the record.

The more recent proposals however are designed to operate based on the opposite philosophy. In those schemes, an update is due when mobiles arrive at those locations. As the location areas are dynamically assigned according to the actual mobile movements, a new location area for the mobile will become available only upon the registration of a new update. Since a new location update area is dynamically defined upon registration, the mobile is required to maintain a set of require to update cell identifications. Thus, an update is required when the mobile is detected to have reached one of those cells specified in the cache providing details to the update boundaries.

This distinction forms the basis in our classification of various location management schemes. Figure 3.1 classifies update algorithms based on the time when an update is due.

![Figure 3.1](image)

An overview of update schemes proposals optimized for individual mobile
At the top level, the classification of update schemes can be made according to the nature of decision. In other words, while in some cases where an update is performed when mobiles arriving at certain database centers, others register the new location upon exiting the defined area. Although the actual mechanism that was used to derive the set of cell IDs varies significantly between different schemes, any update that demonstrates this type of table lookup is classified under the category of location area oriented.

There are, however, other schemes that demand for their update decision more operations than a table lookup. In these update strategies, additional computations are generally involved to evaluate the need for a location update in the mobile terminals. The concept of location area does not apply and the decision of update depends greatly on the actual circumstances at the time a cell boundary is crossed. The term non location area oriented seems to be the obvious choice for describing an update scheme under this category. Clearly, depending on the complexity of the operations, the required computation load and thus, power consumption at mobile terminals, will differ. The tradeoff between the infrastructure expense and the achievable performance gain determines the suitability of incorporating this strategy into a given network scenario.

3.2.1 ZONE-BASED UPDATE SCHEME

Static update algorithms refer to the type of zone-based schemes where the locations of the update centers are fixed for each network. Independent of individual mobile's roaming characteristics, a registration is required whenever a mobile arrives at such cells [38,39]. While such schemes require little computational load to implement, the achievable performance gain is often limited. Clearly, without distinguishing the roaming characteristics of individual mobiles, it is difficult, if not impossible, to assign update cells such that the overall performance is optimized for all mobiles. To illustrate such deficiencies, consider the two situations where the application of the static update algorithms is particularly infeasible. On the one extreme, when a mobile frequently passes through the selected reporting centers, high update rate will result and thus incurring unnecessary operational expenses. On the other extreme, when a mobile does not roam into any of the selected reporting centers, no update will be registered and consequently a high paging load is likely to result upon call arrivals.

Recognizing such an inadequacy, update strategies such as three-location area (TrLA) are proposed to minimize the number of updates initiated by the repetitive motion characteristics observed from the movement patterns of a mobile. The actual implementation methods can be quite different. As an example, a list of the most probable visiting locations can be defined for each mobile according to its movement history. Thus, an update at the reporting centers will
be necessary only if the specific update cell has not been previously compiled in the database [42].

3.2.2 THREE-LOCATION AREA (TrLA) UPDATE SCHEME

In the three-location area (TrLA) update scheme, the mobile will have three neighboring LAs stored in the memory instead of only one which is the currently residing LA as for the general GSM/GPRS standard. Thus, for every occurrence of a boundary transitioning, the mobile checks whether a registration is required by comparing the obtained LA identification with the three location areas cached at MN. Classified also under the category of update when exiting, an update is due only when the mobile moves out from the big-location area (BLA) comprised of the three location areas. The information cached at MN thus gets refreshed during the update operation. Effectively, the proposed scheme suggests reducing the update cost by increasing the size of the update area.

The actual mechanism used to select the additional LAs for caching is, however, not specified. It was not possible to determine whether some sort of “prediction” had been implemented based on the available documentation. To a large extent, the scheme appears to be only a slightly varied version of the conventional static location update scheme.

Not only is the network assumed to have a complete knowledge of the LA layout, with additional information about the identification of neighboring location areas cached at the terminals and the network, the size of the database increases in direct proportion with the number of connecting subscribers. With mobile subscribers continuing to increase, it is likely that scalability issues will restrict the applications [46].

Discussions up to this point consider only static assignments of the location area. Although the actual positions of the registration centers might be different for individual mobile users, the general shapes and sizes of the location areas remain the same. Hence, no specific distinction or consideration is given to individual mobiles that inherit different mobility characteristics. As a result, although performance gain over the conventional cellular standards might be evident, the extent of improvement is limited on average.

To enhance operations, more user-specific location areas are assigned to reflect actual movement patterns. The ultimate goal is to define an update boundary that maximizes the number of transitions between cells before the mobile finally exits the location area. A large number of specific update strategies proposed under this category are based on the assignation of thresholds. The defined parameter will then be used to determine whether the need of update is necessary upon transitioning to a different cell. To enable a comprehensive comparison between different techniques, the operations for all schemes under the category of dynamic update will be discussed within the same framework. Specifically, discussions will
be structured to answer the particular questions of *how*, *what*, and *why*; How does the scheme operate? What are the novelties introduced? And why is the proposal, or in some cases modification, significant.

### 3.2.3 PROFILE-BASED UPDATE SCHEME

As the name suggests, with the profile-based update scheme, a profile is maintained for each user in different time periods. Upon registration, the network assigns the mobile a list of regions in which a high probability of residency is expected based on what was learned from previous movements. A region can be a cell, or where appropriate, groups of cells. Mobiles are not expected to perform updates unless the region entered was not already included in the list. Thus, in extreme cases, where a region is used to represent a regular-sized location area, the operation turns into a modified version of the two location area (TLA) and the three location area (TrLA) update schemes. This is possible however only if regular mobility patterns have been observed from past movement history [46].

For the systems to establish a reliable set of database describing the mobile's movements for different occasions, the subscriber is assumed to demonstrate a certain degree of consistency in movements. It was concluded in [47] that when a medium to high predictability is observed from mobile movements, a lower location management can be obtained in comparison to other update schemes where the boundaries are more generally defined. Other more specific analyses for the profile-based scheme can be acquired from [48].

### 3.2.4 PROBABILISTIC LOCATION UPDATE SCHEME

Classified under static update schemes, [49] proposed a probabilistic location update scheme (PLU); in which, the mobile registers with a probability of $P$ that varies according to the call and mobility characteristics of individual users upon entering a new cell. Basically, the key conclusion obtained from the chapter illustrates that as mobility increases (i.e. with decreasing CMR), the probability of needing a location update at the cell decreases. This somehow is complementary with our justification which states that the more direct the mobility pattern is, the less update load is required since the probability of predicting the correct location of a roaming mobile would be higher. Thus, in terms of categorizing the mobility characteristic, while one is using the mobility rate information with respect to call arrivals, the other focuses on mobility directions. Both however are aiming at reducing the necessary updating loads by increasing the probability of accurate prediction.
3.2.5 COMPRESSION-BASED UPDATE SCHEME

Similar to the profile-based scheme, the compression-based update algorithm utilizes location probabilities profiles. The actual mechanisms that are used to establish the probability profiles are, however, different. While the former only requires a counter mechanism at the terminal end, the latter requires a lot more computation and analysis from both the network and mobiles, however, at the benefit of a more reliable profile database. Specifically, in compression-based update, there is no concept of a location area definition. Thus, different from its profile-based counterpart in which a set of location areas is predefined according to the mobile’s previous movement history, each cell is the equivalent of a location area on its own in the compression-based update.

To save on unnecessary memory requirement at the mobile end, the LeZi-update algorithm is proposed to be implemented on top of one or more threshold-based update schemes. Thus, based on an always update strategy, instead of triggering an update whenever the threshold is reached, the algorithm delays the update operation and attempts to process the information in chunks. At the time of registrations, rather than sending information about the mobile’s current residency, the past movement history having been encoded in a compressed form is registered at the network. A trie diagram is used based on the Ziv and Lempel compression algorithms. Thus, an update is registered only when the path is not yet recorded in the database. Clearly, the proposed update schemes will have optimal performance gain only when regular mobility patterns are evident, and the network is confident that the mobile’s future movements will continue to show such predictability [43]. Effectively, the mobile and the system form an encoder/decoder pair. Considering that it is actually the movement history being reported during updates, this location tracking strategy is considered more of a path-oriented than a zone-based solution. It should be noted however that the application of the scheme is based on the assumption that a mobile’s movement pattern is generally repetitive and can be learned over time.

The fact that additional computation is required before a decision is made about the need for an update makes the compression-based update scheme appear to be similar to [49]. In which case, the previous information based on residence time will be used to evaluate the probability of having call arrivals before the mobile exits the current cell. Clearly, with the call arrival rate fixed, the longer the mobile resides in the area, the higher the possibility that the mobile can be located in the cell. Thus, in comparison to the compression-based scheme, [49] will incur a lot more intelligence at the mobile terminals. This is no longer an issue of database conservation, but also a concern because of increased computation load.
3.2.6 SPACE-BASED UPDATE SCHEME

Essentially as the distance-based update scheme, with the actual traveling velocity considered in the assignment of optimal update boundary, space-based scheme defines the update boundary in terms of the actual coordinates instead of in cell numbers. The general concept of fixed-boundary location area remains relevant [48].

3.2.7 THRESHOLD ASSIGNMENT

A series of threshold-based update schemes represents some important and fundamental developments for basic user-specific update strategies. A specific parameter is assigned to the mobile by the network during registrations, the quantity is then monitored continuously such that an update is performed when what was observed from the actual movement exceeds the quantity that is predefined by the network. The most commonly used parameters are time-, movement-, and distance-based update schemes.

A time-based strategy defines the frequency at which mobiles are required to register their new locations. At defined interval \( T \), the mobile compares its current location to the previously registered cell identification (cell ID). An update is performed when found both values differ. The implementation prerequisite for such a scheme is simple and requires only a timer at the mobile terminals. There are however certain challenges that need to be overcome before actual applications are realizable. It is true that the number of registrations can be controlled by the update duration \( T \), but there is no guarantee over the effectiveness of such registration information. Clearly, when a mobile continues to engage in a repetitive movement pattern, a multiple number of registrations will incur without attempts to reflect the actual mobility patterns. Those registrations are redundant and have minimal values to aid the paging operations. With paging unbounded, neither the cost of operation nor the quality of services can be controlled upon call arrivals [7,44,45].

Classified also under the category of threshold assignment update strategies, the movement-based scheme keeps track of a mobile’s movements by counting the actual number of cell boundary transitions. Specifically, for a defined threshold of \( M \), an update is due whenever the number of cell transitions exceeds the defined constant. In comparison to its counterpart of time-based update strategy, the maintained information about movement numbers minimizes the incurring of unnecessary updates that could possibly result from low mobility users when the threshold is quantified by time. In addition, the maximum paging area can be defined as it is certain that the furthermost point where the mobile can possibly reside upon call arrival will have a separation of \( M \) cells from its last registration. There are however other concerns that need to be addressed before the actual implementation is feasible.
in practical systems. Primarily, as the update scheme focuses more on the rate of traveling rather than the way of traveling, actual movement patterns of individual users are not well tracked by the network. Consequently, as the mobiles are asked only to trace the number of movements but not the actual cell residencies, any repetitive movements are unable to be detected by the system. Not only will the unnecessary updates create additional operational cost, the bounded paging area may be needlessly oversized and thus incur further paging load. A potential solution is to invest additional database capacity in the terminals, thereafter, a counter is incremented only when the currently visited cell has not yet been recorded in the movement history [7,38,39].

This solution however would increase the necessary intelligence at the mobile node, a consequence that is undesirable. An obvious solution is thus to assign a distance threshold in terms of cell numbers and to have an update registered only when the actual distance of traveling exceeds a predefined threshold \( D \); hence the formation of a distance-based update strategy [8,10].

Among all these, [8,53-54] show that the distance-based scheme gives the best performances in terms of reduced update numbers over the conventional GSM/GPRS update schemes. One factor that contributes to such performance gains is that the effects of "ping-ponging" in the movement patterns have actually been accounted for in the design of the system operations.

### 3.2.8 STATE-BASED UPDATE SCHEME

The state-based update strategy is one of the few classified under the category of *hybrid* techniques. The actual interpretations of state are considered to be a variable, and can be uniquely defined for individual users. It is possible to combine operations of individual techniques to optimize the performance gain. For example, an update strategy combining the essential elements of time-based and movement-based schemes takes advantage of the movement-based scheme of defining a paging boundary while utilizing the merit of timing information to adjust the optimal threshold size. Subsequently, a further reduction in the update load can be expected with a minimal requirement of additional infrastructure.

[41,55] analyzed a state-based update scheme where the state is considered to be a combination of the current location and the time elapsed since the last update. On the one hand, the fact that the state can take one or more parameters of the threshold-based update schemes makes it part of the location area oriented solutions. On the other hand, the need for additional computations to be performed at the mobiles after each cell crossing causes its linkage to the non location area oriented category.
\[55\] considers a greedy algorithm where the decision of update is made at each cell boundary not necessarily to optimize the overall performance, but to minimize the operational cost for the current roaming interval only. To extend the application of the location tracking strategy, the level of greediness is adjustable by the introduction of an additional parameter, \(\alpha\). With an allowable range of values defined between 0 and 1, \(\alpha\) determines how far back previous roaming intervals should be taken into consideration for the minimization of the cumulative operational cost. An update is due only when the resultant paging cost exceeds the expected optimal.

Assuming a registration to be due at time \(s + t\), given the current position, the elapsed time since the last update, and the past history, a state-based policy \(\theta_s\) is generated to determine an optimal \(t\) that minimizes the combined cost of paging and update \(\eta\). The extent of involvement of the past history in the optimization is controlled by \(\alpha\). The algorithm is said to be completely greedy when the optimal \(t\) is assigned to minimize operational cost for the current roaming interval only. Effectively, an optimal registration area will vary as a function of time and will be assigned by the mobile each time a boundary crossing is detected.

To some extent, there is a great similarity between \[55\] and \[10,11\] in terms of the method being used to predict the likelihood of residence. The biggest difference observed between the two however is the fact that while most computations are done at the mobiles in the former, a similar workload is imposed on the network in the latter. Consequently, with the expectation that \[55\] will assume a mobile to have certain knowledge about its specific motion trajectories, equivalent assumptions of such information disclosure are not evident in \[10\].

### 3.2.9 LOAD ADAPTIVE THRESHOLD SYSTEM (LATS)

Also categorized as a hybrid update technique is the load adaptive threshold scheme (LATS). In fact, this scheme is similar to that outlined in the state-based update strategy, in that, the decision of performing an update is dependent on some additional parameters other than the predefined threshold. In the load adaptive threshold system, the extra consideration used to determine the need for update is the network loading. Upon entering a cell, the mobile compares its predefined threshold to that of the loading threshold generated by the cell. An update is thus due only when the need of the update defined by a priority level has exceeded that constrained by the loading conditions at the cell \[57\].

The application of LATS is most promising when the network is heavily loaded. With the maximum number of allowable updates constrained by the cell's loading threshold, the problem of location tracking can be managed without having to sacrifice other QoS measures for example, throughput.
3.2.10 DIRECTION-BASED LOCATION UPDATE SCHEME

A recent publication [58] proposed a direction-based location update scheme with a line-paging strategy for PCS networks. Basically, the scheme works as follows: whenever the mobile enters a new cell, it determines whether a change of roaming direction has occurred since the last movement. An update will be registered only if the answer is affirmative, and hence the name “direction-based” location update scheme. Upon a call arrival, the last registered cell will be the first place the system looks at, the search then extends in both directions until the mobile is found. To set some upper bounds of paging loads, the search discontinues after reaching a time threshold $T$.

This estimation is based on the assumption that cells are regularly arranged so that each has unique identification in terms of cell coordinates. Mobiles are expected to roam mostly in a straight line with only occasional changes of direction that can be detected when the coordinate difference between adjacent movements is not same as the previous differences. Because it is unlikely that mobiles always travel in an “absolute” straight line, there will be a fair amount of updates. As a result, this seems to contradict the goal of reducing the amount of location updates in designing an efficient location management technique. The other problem that would be encountered in the scheme proposed in [58] is its complexity (i.e. heavy signaling and high computational load) that results from the necessary computation after each movement. As the scheme depends on the prompt detection of change of roaming direction by the mobile in order to make the corresponding paging strategy feasible, additional loading (and hence the necessity of power consumption) will be imposed on the mobile terminals. Consequently, the schemes appear to have optimal performances in a system whose call arrival rate is a lot more frequent than its mobility rate. The application is thus restricted not only by the range of relevant mobile characteristics, but also the requirement of the system layout.

3.2.11 ADAPTIVE DISTANCE-BASED UPDATE SCHEME

The proposal is intended to uplift some of the unrealistic assumptions commonly imposed in the assigning of an update boundary. In this algorithm, an optimal update boundary is thus defined based on the assumptions of arbitrary cell topologies and general cell residence time distributions. Consequently, the implementation is no longer restricted by the structured cell configurations or the independent and identically distributed cell residence times. The concept of trip is also incorporated such that there might be a particular path for each defined destination [59].
To restrict the necessary computation load which results from the determination of appropriate threshold for all subscribers, it is possible to define a set of groups with significantly different mobile characteristics. Upon each registration, an appropriate group index is assigned to individual mobiles as a reference for future update operations. With the range of groups effectively identified, the averaged performance gain for all subscribers can be optimized with scalability concerns considerably catered for.

3.2.12 DYNAMIC LOCATION AREA (LA) ASSIGNMENT UPDATE SCHEME

A location area is dynamically adjusted in shape at each registration to ensure that the number of cells the mobile traverses in the area is maximized before an update incurs. Based on the shortest distance model, an optimized location area is determined using a heuristic greedy algorithm. Generally, the iterative method starts by including only the last updated cell in the location area. Estimation is made to all cells on the perimeter of the current location to determine mobile’s next most likely residency. An extra cell is included in the location area at the end of each iteration. The algorithm continues until reaching the maximum location area that is predefined. Furthermore, to minimize the possible overheads that are likely to result from the advertising and storing of the irregular location area (LA) shapes, a separate heuristic algorithm is applied to determine the optimal dimensions of a corresponding rectangular LA. Simulation results reveal that such a rectangular LA approximation will give a comparable performance gain to what would otherwise achieved from its irregular-shaped counterpart.

The evaluation of performance gain is currently based on the assumption of a Manhattan grid topology. Although it was suggested in this research work that the extension of the operations to an arbitrary cell topology will be equally straightforward, the amount of necessary computations seems to have suggested the opposite [60].

3.2.13 PREDICTIVE DISTANCE-BASED UPDATE SCHEME

The prediction of mobile’s future location is based on its location and velocity information registered during the update. Part of the challenge is to determine an optimal location area. As a general rule, the shape of the assigned area should reflect the actual mobility patterns of the individual subscriber with the area size varying as a function of the rate of incoming calls. With the shape of the location area defined to optimize the number of cell transitions, the size should adjust to keep the paging cost per unit time constant. In other words, for a specific roaming characteristic, the more frequent the rate at which call arrives, the smaller the
location area should be. The actual implementation will depend on the observed traffic characteristics and network conditions mainly the loading status [11].

### 3.2.14 ACTIVITY-BASED UPDATE SCHEME

Based almost on the same framework is the proposal of an activity-based update scheme in which a personalized location area is defined for each subscriber based on its previous mobility history. For movements within the defined location area, the frequency with which each cell is accessed from its adjacent cells is measured along with the actual time of residency. An update is due when the mobile exits the current location area. A new location area is thus defined according to the previously gathered information, cells are included in the registration area sequentially in the order of decreasing likelihood of residency until reaching the predefined maximum location area size. The procedure is recursive and stops when the location area reaches its maximum size.

The scheme is optimized for mobile characteristics where regular mobility patterns are evident. Generally, the past movement history has demonstrated a “consistency” in the roaming area and the network is certain that the trend will continue for future movements. Utilizing the information observed from the consistency over movement patterns, the adaptive scheme can effectively reduce the necessary update loads [61-63].

Figure 3.2 summarizes the implications of various location updating strategies. It should be clear by this point that depending on the underlying assumptions and the simulating natures of individual mobility models, the application of the technique is not always suitable for all systems. Clearly, activity-based location management can only be available when there is sufficient information to have the relevant databases established. The strategy is very user-specific and is most apposite to represent daily activities that are highly consistent over a long duration of time with a periodicity measured in days or weeks.

![Figure 3.2](image-url) An overview to the implications of various location updating strategies
3.3 ANALYTICAL FRAMEWORK

The effectiveness of individual update and paging techniques can be evaluated based on the operational cost incurred within a fixed time period. In effect, for a predefined mobile characteristic and call arrival rate, the operation costs resulting from the applications of different location management techniques can be assessed and compared. Given that a good location management should be able to efficiently locate a roaming mobile when a call request arrives, it is equally viable to consider the statistical measures of the updates and paging load against rate of call arrival when the sampling data involves a large enough range of mobile characteristics. This is possible as the rate of call arrival determines the number of calls within a defined time period.

Having studied the main location management techniques proposed in the literature, the chapter will not be complete without highlighting some of the general analytical frameworks that have been used in the evaluation of the various location management schemes. It should become apparent to the readers now that the key parameters used to quantify the performance of specific location management techniques include the update load, paging load and where applicable, the paging delay.

While the ultimate goal is the same in research work involving scheme evaluations, the actual analytical framework can have various representations depending on the approaches that were taken by the authors. This section is devoted to comparing, and specifically to highlighting the great similarities among the available analytical systems. The main intention is to widen up the analyzing scope and hopefully to inspire further techniques. In this section, we first describe the general expression used to estimate the total expected cost. Specific evaluation methods of the update and paging are then discussed separately in a more depth.

3.3.1 INFLUENTIAL FACTORS

The effectiveness of specific solutions to the problems of location tracking is mainly quantified by the costs that were incurred during operation. While in some, the quantification is based on costs during call arrivals, in others, the relative cost between update and paging is evaluated.

Other than the general acknowledgement of call arrival and mobility rate, specific network and mobile characteristics also have direct impact on the actual performance measures. In order to optimize the overall performances, it is important to identify those parameters and to consider them when assigning system parameters to optimize performance gains. Obviously, some of the observed parameters will have a more crucial impact on the designed techniques
than others. In this section, a list of commonly identified parameters has been compiled to extend the horizons in perceiving such location tracking problems.

Given that many of the update strategies are, in some form or another, extensions to the basic distance-based threshold, the discussion of influential performance factors is oriented on the concepts of such a scheme.

A. Cost Definitions

The definition of unit update cost has a significant impact on the overall operational characteristic, be it the selection of the optimal distance-threshold or the switching between different modes of operations. In most network scenarios, a unit update is often more costly than a unit paging due to the amount of resources both wireless and wired required to properly register the new location. In certain cases where the expense of an update operation becomes extremely high, unnecessary updates should be avoided at all times. Such additional caution taken during the implementation phase will affect the selection of optimal threshold and thus, affect the overall performance gain.

B. Call and Mobility Characteristics

The rate at which the requests of connection calls arrive plays an important part in the problem of location tracking. Clearly, in network scenarios where a mobile does not expect to receive any call-connections during its time of roaming, the need for a location registration is redundant. From the implementation point of view, the ideal threshold definition will be infinity, and thus the operation is equivalent to the never update strategy [34,39]. Although the occurrence is rare, in extreme cases where calls are expected to arrive for the roaming mobile all the time, it will be necessary to set the update threshold to a small value to avoid excessive paging load due to the frequent need of determining the mobile's whereabouts. As a general rule, the optimal threshold should decrease as call arrival rate increases.

In a busy urban environment with a pico-cell topology, cell boundaries are crossed at a higher rate than in a suburban network configured with a macro-cell topology. To ensure that the assigned threshold adequately reflects the characteristics of the specific network of interest, the rate at which a mobile traverses between boundaries should be considered to optimize the update operation.

The effects of mobility and call arrival characteristics can be considered either separately or jointly.
C. Paging Delay Constraint

Although the impact of a paging delay constraint is generally seen in the paging operations, to optimize the overall performance, appropriate consideration of such a parameter taken at the assignment of update threshold will also ensure the achievement of a better performance gain.

3.3.2 OVERALL COST FUNCTION

The total operational cost consists of two parts: the update cost and the paging cost. Although there are a few possible ways to model mathematically the cost function, the basic framework is the same. Intuitively, the challenge is to determine each part of the total cost from the product of number and the unit cost, giving

\[ C_{\text{total}} = U \times N_u + P \times N_p \]  

(3.1)

Given a set of cost parameters, \( U \) and \( P \), respectively represent the unit update and paging cost, the real challenge is to estimate the numbers of each quantity in the process of cost evaluation, \( N_u \) and \( N_p \). There are a few alternatives to quantify such parameters. For example, [60,65] express the cost relation by

\[ C(k, \mu_a) = U \cdot u_k + P \cdot k\mu_a \]  

(3.2)

For a location area containing \( k \) cells, \( u_k \) represents the update rate and \( k\mu_a \) denotes the number of paging signals required for a call arrival rate of \( \mu_a \).

The update rate is usually a function of the mobility pattern, the traveling speed, and the definition of the location update area. In this research work, the update rate is a measure of the time a mobile resides in the specific location area.

During a call, the location is reported to the network at each boundary crossing. The operation of mobility tracking is thus applicable between the inter-arrival of adjacent calls. In the comparison of various update schemes, it is thus common to evaluate the total cost incurred within this time interval. Based on the nature of the assessment, the study is categorized into various mobility models.

A. Sensitivity Analysis

With predictions incorporated in many schemes, it becomes important to evaluate the impact that prediction accuracy has on overall performance: the more sensitive the scheme, the higher the operational cost.
Apart from comparing the absolute operational cost, the range of variability in the resultant cost subject to other system parameters also gives a good indication of the stability of the performance gains. As part of the evaluation study, [55] quantifies the performance gain of different location management techniques by examining the reduction in the variability of paging and registration cost under varying mobile characteristics. Generally, the smaller the cost standard deviation, the less sensitive the performance gain is to small changes in mobility indexes, and thus, the more reliable the location tracking strategy.

It should be clear that the purpose of location management is to determine an optimal tradeoff between the update and the paging operations. Undoubtedly, while there are alternative methods available to quantify the performance gains, the general goal is the same — to determine the best location tracking strategy that minimizes the operational cost. Having discussed the main characteristics of some commonly used mobility models and various update and paging strategies, it is now appropriate to examine the range of analytical models that are available for performance evaluations. In the following subsection, discussions are included along with basic descriptions of the analytical model utilized to evaluate the numbers. The purpose is to clarify the present state of the art, comparing the similarities and differences of specific models.

3.3.3 UPDATE COST ANALYSIS BASED ON MARKOVIAN MODEL

Depending on the actual mobility characteristics, the quantifying parameters may be different, however the common theme is the same — to determine the frequency at which updates are performed. Intuitively, the general goal is to minimize the number of updates such that the cost from update operations is kept as low as possible. [9, 59, 66-67] have quantified the efficiencies of update schemes by keeping a record of the number of times update registrations are required during the inter-arrival time of incoming calls. The update strategy that results in the least updates has the optimal performances for specific mobile behaviors. Alternatively, [60, 65] obtained similar evaluations by measuring the number of movements incurred for each update that was registered. Generally, the greater the number of boundary crossings before an update is due, the better and more efficient the location tracking strategy. To have the basic concepts better understood, short discussions on the determination of update cost of each approach are included.

Markovian movement patterns are generated based on a slotted time model. At the beginning of each discrete time interval, the mobile transits to an adjacent cell with a predefined probability. Advancing from an independent and identical distribution model where the transitioning probability to all adjacent cells is the same, a Markovian model allows
assignations of varying probabilities in order to imitate better the characteristics of different movement patterns.

Earlier analysis assumes a one-dimensional system framework \([10,11,69-70]\) where each cell has only two adjacent cells. Thus, whenever a boundary crossing is initiated at the cell level, the mobile can move either to the right or to the left cell during each slotted time interval. With the transitional probability distribution also defining the likelihood that the mobile stays in the same cell, the simplest state machine needs only three states to describe the complete motion behavior. Figure 3.3 gives a general example of the mobile characteristics. In the illustration, the circles represent the states specifying the actions to take in the next time slot with the arcs indicating the probabilities of transiting from one state to the other.

\[
\begin{array}{c}
\text{R: } \text{The mobile moves rightward in the next time slot} \\
\text{L: } \text{The mobile moves leftward in the next time slot} \\
\text{S: } \text{The mobile stays put in the next time slot}
\end{array}
\]

**Figure 3.3** An example of the Markovian-motion state machine

At each instant of time, the mobile can be in any one of the states along with the transition probabilities specified in the state machine. As an example, if at time slot \(t\), a mobile is in state \(L\), a leftward movement will thus occur in the next time slot \(t+1\) from its current location. In addition, the mobile will stay at state \(L\) with a probability of \(q\), change to state \(R\) with a probability of \(v\), or stays put as state \(S\) with a probability of \(1-q-v\).

Let \(Q_d\), denote the steady-state probability of having the mobile reside \(d\) cells away from the last updated cell with state \(s\). For a distance-based update strategy of threshold \(D\), a registration area will consist of \(2D-1\) cells which allow mobiles to move freely in the range \([-D+1, D-1]\) without having to register. Figure 3.4 shows a part of the corresponding Markov chain used to describe the complete motion model.

Since each state will actually have two values; \(d\) denotes the distance from the last updated cell \(d \in \{0,1,2,3\ldots\}\) and \(s\) denotes the next action to be taken \(x \in \{R, L, S\}\). The complete Markov chain should show in two dimensions indicating the transitional probabilities between the two values. To illustrate the basic concepts, Figure 3.5 shows an extract of the Markov chain which indicates the transitional behavior that contributes to the formation of the steady-state of \(Q_d\).
Mathematically, the motion model can be described by the set of balance equations shown in (3.3)-(3.10) [7].

\[ Q_{0,R} = pQ_{0,S} + qQ_{D-1,R} + vQ_{(D-1),L} + qQ_{-1,R} + vQ_{1,L} \]  
(3.3)

\[ 2pQ_{0,S} = (1 - q - v)(Q_{D-1,R} + Q_{(D-1),L} + vQ_{-1,R} + Q_{-1,L}) \]  
(3.4)

\[ 2pQ_{d,S} = (1 - q - v)(Q_{d-1,R} + Q_{d+1,L}) \quad 1 \leq d \leq D - 2 \]  
(3.5)

\[ Q_{d,R} = pQ_{d,S} + qQ_{d-1,R} + vQ_{d+1,L} \quad 1 \leq d \leq D - 2 \]  
(3.6)

\[ Q_{d,L} = pQ_{d,S} + vQ_{d-1,R} + qQ_{d+1,L} \quad 1 \leq d \leq D - 2 \]  
(3.7)

\[ 2pQ_{D-1,S} = (1 - q - v)Q_{D-2,R} \]  
(3.8)

\[ Q_{D-1,R} = pQ_{D-1,S} + qQ_{D-2,R} \]  
(3.9)

\[ Q_{D-1,L} = pQ_{D-1,S} + vQ_{D-2,R} \]  
(3.10)

With the steady-state probabilities summed up to 1 (3.11), the steady-state probabilities can be obtained by solving the equations, giving (3.12-3.15).

\[ \sum_{d=(D-1)}^{D-1} (Q_{d,S} + Q_{d,R} + Q_{d,L}) = 1 \]  
(3.11)

\[ Q_{d,R} = \frac{p[D-d](1-q+v) + 2(q-v)}{D(1+2p-q-v)[D(1-q+v) + 2(q-v)]} \quad 0 \leq d \leq D - 1 \]  
(3.12)

\[ Q_{d,L} = \frac{p(D-d)(1-q+v)}{D(1+2p-q-v)[D(1-q+v) + 2(q-v)]} \quad 0 \leq d \leq D - 1 \]  
(3.13)
\[ Q_{4,d} = \frac{(1 - q + v)(D - d)(1 - q + v) + (q - v)}{D(1 + 2p - q - v)[D(1 - q + v) + 2(q - v)]} \quad 0 \leq d \leq D - 1 \]  

(3.14)

\[ Q_{6,3} = \frac{(1 - q + v)}{D(1 + 2p - q - v)} \]  

(3.15)

An update is due when a mobile moves to the boundary cells with \( Q_{D-1,R} \) or \( Q_{(D-1)L} \). Whether it is in the former situation where the mobile is expecting a rightward move from its current location \( D-1 \) or it is in the latter situation where the mobile is expecting a leftward move from its current location \(-D-1\), with the symmetrical threshold defined, the stationary probabilities for the two states are equal.

The expected number of updates transmitted in a time slot is thus equal to the probabilities of being in the two states \( Q_{D-1,R} \) and \( Q_{(D-1)L} \) and can be approximated by

\[ U_D = Q_{D-1,R} + Q_{(D-1)L} = 2Q_{D-1,R} = \frac{2p(1 + q - v)}{D(1 + 2p - q - v)[D(1 - q + v) + 2(q - v)]} \]  

(3.16)

Clearly, the update rate is not only a function of the predefined threshold, but also of the actual transitional probability distributions. The probability of incurring an update decreases with increasing threshold and in general a close-to-unity approximation of the ratio \( q/(q+v) \).

For example, extending from the analysis of [7], [10] examines the effects of directional consistency/persistency across stops. Thus, instead of assuming that a mobile loses its sense of directions after stopping, two additional states are introduced to record the most recent move. Figure 3.6 illustrates the modified Markovian-motion state machine.

![Figure 3.6](image)

An modified Markovian-motion state machine

Consequently, it is necessary to include a few more arcs describing the movement patterns. The definitions mainly include \( p_i \), the probability of resuming motion in the same direction as that prior to stopping and \( p_2 \), the probability of resuming motion in the opposite direction from that prior to stopping. There are also possibilities for the mobile to stay puts in the following time slot with a probability of \( 1 - p_1 - p_2 \). The graphical representation of the Markov chain also becomes more complex because of the additional states. Figure 3.7 shows an extract of the transitioning behavior around the last update cell. As the update boundary is no
longer symmetrical about the last update cell, the location of the center cell is now denoted by \( Q_{C*} \) instead of \( Q_{o*} \), as evident in the previous case.

![Diagram](image)

**Figure 3.7** A modified Markov chain to present the motion model

While the size of a registration area remains the same as in [7,53], the fact that the persistence across stops is now considered in the motion characteristics, means the resulting steady-state probability equations need to reflect the corresponding differences, giving

\[
Q_{C,R} = p_1 Q_{C,SR} + p_2 Q_{C,SL} + q Q_{D-1,R} + q Q_{-(D-1),L} + q Q_{C-1,R} + v Q_{C+1,L}
\]

(3.17)

\[
Q_{C,SR} = (1 - q - v)(Q_{D-1,R} + Q_{-(D-1),L} + Q_{C-1,R}) + (1 - p_1 - p_2)Q_{C,SR}
\]

(3.18)

\[
Q_{d,R} = p_1 Q_{d,SR} + p_2 Q_{d,SL} + q Q_{d-1,R} + v Q_{d+1,L}
\]

\[
d \neq -(D-1), C, D-1
\]

(3.19)

\[
Q_{d,SR} = (1 - q - v)Q_{d-1,R} + (1 - p_1 - p_2)Q_{d,SR}
\]

\[
d \neq -(D-1), C
\]

(3.20)

\[
Q_{d,SL} = (1 - q - v)Q_{d+1,L} + (1 - p_1 - p_2)Q_{d,SL}
\]

\[
d \neq D-1
\]

(3.21)

\[
Q_{D-1,R} = q Q_{D-2,R} + p_1 Q_{D-1,SR}
\]

(3.22)

\[
Q_{D-1,LL} = v Q_{D-2,R} + p_2 Q_{D-1,SR}
\]

(3.23)

Note that the states with arrows indicate a modification from the original balance equations obtained from [7].

Other than modifications to some of the existing equations, note also the inclusion of extra balance equations in the determination of stationary state probabilities. In addition, as it is assumed that the forward direction after each update is by default a rightward movement, the state of \( Q_{C,SL} \) does not exist in the modified motion model.
\[ Q_{C,L} = p_2 Q_{C,SR} + p_1 Q_{C,SL} + v Q_{D-1,R} + v Q_{(D-1)_L} + v Q_{C-1,R} + q Q_{C+1,L} \]  
\[ Q_{(D-1)_R} = v Q_{(D-2)_L} + p_1 Q_{(D-1)_SL} \]  
\[ Q_{(D-1)_L} = q Q_{(D-2)_L} + p_1 Q_{(D-1)_SL} \]

Furthermore, instead of evaluating \(2Q_{D-1,R}\), the boundary conditions should be separately considered for the determination of the expected number of update messages, giving

\[ U_D = Q_{(D-1)_L} + Q_{D-1,R} = \frac{r}{(D-C)[(D+C)(1-r)+2r-1]^F} \quad C \neq (D-1), D-1 \]  

Where

\[ T = 1 + (1-q-v) \frac{1}{p_1 + p_2} \]

and

\[ r = q + (1-q-v) \frac{p_1}{p_1 + p_2} \quad 0 < r \leq 1 \]

The additional term of \(1/T\) comes from the explicit update that incurs when the predefined time period expires. The advantage of adding the concept of a time-based update to the distance-based update operations is evident when mobiles remain in the same cell for a significantly long period of time. In that case, it is necessary to consider the mean holding time among cells.

Clearly, when \(p_1 = p_2\), in which it is assumed that the mobile loses its sense of direction after its stopping, \(r\) and \(T\) are simplified forming

\[ r = q + \frac{(1-q-v)}{2} \]  
\[ T = 1 + \frac{(1-q-v)}{2p} \]

If replacing \(C\) with 0, (3.28) as expected becomes (3.16) when (3.31-3.32) are substituted.

\[ U_D = \frac{r}{(D-C)[(D+C)(1-r)+2r-1]^F} = \frac{2p(1+q-v)}{D[(D(1-q+v)+2(q-v))(1+2p-q-v)]} \]

Hence, the proof shows that additional parameters can be added to the basic Markovian model with only appropriate modifications to the corresponding balance equations. Specifically, the inclusion of additional states in the motion model allows individual movement patterns to be more accurately modeled however, at the expense of increasing computational complexities.

With \(q\) fixed, \(r\) grows when the persistency \(p_1\) increases and consequently, the update rate goes up. In order to minimize the resulting expected number of updates, it is important to
locate C such that an asymmetric distance-based position-update criterion is generated. As a rightward movement is assumed to be the forward direction, the idea is to move the update center as far as possible from the right end boundary cell. In the extreme case where a mobile shows a high persistence in motion characteristics, the optimal location for situating the center point C will be at \(-D-1\). Consequently, the update rate turns into

\[
U_{opt} = \frac{p_1 + p_2}{(2D-1)(1 + p_1 + p_2 - q - v)} = \frac{1}{(2D-1)r} \tag{3.34}
\]

In a more general form however, the optimal C can be determined from the expression

\[
C_{opt} = \max \left\{ 1 - \frac{1}{2((1-r)^2(D-1))} \right\} \tag{3.35}
\]

Simulation results included in [10] have shown a reduction in the update rate with the implementation of optimal C. The success demonstrated in this work has since attracted huge research interests to explore the merits of implementing asymmetrical update boundaries. A brief section is included in later sections highlighting specific methods for the determination of an optimal asymmetrical update boundary.

While the one-dimensional analytical model provides a good framework for the evaluation of system performances, it reveals limited information about mobile characteristics in a more realistic roaming environment. Based on the underlying structure, [59-60,61,64-66] extend the analysis for a two-dimensional system with mesh and hexagonal network topologies. These analytical frameworks are based on the developments of the existing tracking mechanisms. Figure 3.8 outlines the general framework of the developments.

The purpose of location tracking is to determine an optimal update boundary. The frequency at which an update is triggered depends not only on the movement patterns of a mobile but also on the appropriate assignations of an update boundary. Generally, a large threshold incurs fewer updates at the expense of increasing paging load when allowable delay is restricted at unity. A better attempt is made to assign the update boundary such that while keeping the maximum paging load constraint, the number of updates is reduced. When the selected thresholds result in a minimal operational cost, the specific set of boundary cells is referred to as optimal. If the purposes of an update strategy are to be summarized in one sentence, it is about the determination of an optimal update boundary.
3.4 CONCLUSIONS

In this chapter, a literature survey is included to compare the operational differences between various location tracking strategies. The discussions on each scheme are thus oriented around the questions of \textit{when} to update in order to minimize the overall location tracking cost. Generally, better performances are achievable by tracking schemes that utilize the directional information of individual mobiles. Preference in certain directions might simply be due to each user's behaviors or might have resulted from specific geographical conditions.

To evaluate the capabilities of individual schemes, several analytical frameworks are also described in this chapter. Among a wide range of QoS parameters, we focus on the modeling of cost quantification. With the ultimate goal of determining the best tradeoff between the information load and the searching load, the derivation of a common comparison framework will enable a performance comparison to be made between various strategies. In general, the Markovian model provides a good underlying framework both in one and two dimensions. While being capable of modeling the consistency between traveling directions, the analysis is straightforward and can be easily extended to include specific mobility features.
We often discover what will do, by finding out what will not do; and probably he who never made a mistake never made a discovery.

Samuel Smiles (1812-1904)

Chapter 4
Directional-Aware Location Updating Strategies

In this chapter, a new adaptive scheme is developed in which an optimal distance-based update threshold is selected not only as a function of the call-to-mobility ratio, but also as a transitional directivity index $\omega$, a new parameter introduced to give some measures of the mobile's traveling patterns. It is assumed that when $\omega > 1$, the mobile has a greater tendency of roaming “outwards” and hence a higher probability of transition to the next ring during each movement. As a result, the defined threshold will be reached sooner incurring more update costs. As far as the mobility characteristics are concerned, the introduction of the directivity index has successfully demonstrated its ability to make optimum decisions on a distance-based update threshold. Its advantage becomes even more significant when the theoretically determined “ideal” optimal threshold is not obtainable due to certain restrictions imposed by the network during times of high system loading. Simulation results reveal that the additional information made available about a roaming mobile's transitional directivity is critical to ensure that the best available sub-optimal threshold is realizable.
4.1 INTRODUCTION

One general problem with most existing updating schemes is the requirement of reliable call-to-mobility ratio (CMR) information for performance optimization of the adaptive schemes. While it might be possible to relatively easily anticipate the frequency of movements, it is difficult to obtain realistic statistical anticipation about the call arrival characteristics. This observation motivates the formation of the research work. The idea is to explore the possibility of making the selection of an optimal threshold process adaptive to some new parameters other than CMR.

In this section, the research work focuses on making the proposed scheme adaptive in response to mobility patterns, more specifically to the periodic change of transition directions. The extension will be based on a distance-based strategy with the anticipation of movement directions being realized with the introduction of a transitional directivity index $\omega$. The idea is to explore the possibility of making the selection of optimal threshold an adaptive process to $\omega$. It aims to demonstrate how variations in the inter-ring transitional probability would affect the selection of an optimal updating threshold (i.e. incurs a minimal operational cost), and consequently, to study whether the information revealed by the transitional directivity index alone is sufficient to give reliable predictions for the mobility patterns and subsequently the location probability distributions.

Although it might be worth further extending the overall framework by allowing the coexistence of different threshold-based techniques, in the mean time, the focus is on anticipating the mobile’s roaming characteristic using the estimated transitional directivity measures $\omega$. The key concept in the proposed scheme is to adjust adaptively the distance-based threshold $D$ when it is detected that the roaming mobile seems to have demonstrated a change in the motion characteristics. Thus, if $\omega$ is found to be different from its previous values taken at the sampling time, $D$ will be changed to ensure that the specific mobility patterns are taken into consideration. Effectively, this sets the threshold to be adaptive to the mobile’s roaming direction in addition to the usual cell residence time (CRT) and call arrival distributions. The aim is to draw conclusions on the impact $\omega$ has on the selection of the optimal threshold. The key question to be answered was to what extent the traveling characteristic affected the determination of an optimal threshold set up for cost minimization.

Section 4.2 begins by highlighting the relevance of this work to the operations of location management in the context of cellular networks. A more thorough explanation is given to justify the mobility model used for the development of the designed technique. Followed by a brief summary of the related works, Section 4.3 gives a brief description of the simulation
philosophy in addition to defining various details for the setting of parameters. Based on the simulation results, discussions were included in section 4.4 to examine the significance and effectiveness of the introduced transitional directivity $\omega$ in selecting the optimal threshold; specifically, we wish to determine whether the information provided by $\omega$ alone is sufficient to give predictions for mobility patterns in a distance-based scheme. Following this in section 4.5, possible areas for the application of the designed technique are suggested. Section 4.6 previews some of the extension work that is under investigation to further improve performances. This chapter concludes in section 4.9 with a summary recapping the main uses of the proposed scheme.

4.2 SUMMARY OF RELATED WORKS

Despite its importance, there seems little research on the potential of utilizing such directivity information. In terms of involving directions in the design of location tracking mechanisms, a recent publication proposed a direction-based location update scheme with a line-paging strategy for cellular networks [58]. Basically, whenever the mobile enters a new cell, it determines if a change of roaming direction has occurred since the last movement. An update is registered if the answer is affirmative. Upon a call arrival, the last registered cell will be the first place the system looks at, the search then extends in both directions until the mobile is found. To set some upper bounds of paging loads, the search discontinues after reaching a time threshold. Because it is unlikely that mobiles always travel in an absolute straight line, there will be a fair amount of updates. As the scheme depends on the prompt detection of change of roaming direction by the mobile in order to make the corresponding paging strategy feasible, additional loading and hence the necessity of power consumption will be imposed on mobile terminals.

Classified under static update schemes, [49] proposed a probabilistic location update scheme (PLU); in which, the mobile registers with a probability of $P$ that varies according to the call and mobility characteristics of individual users upon entering a new cell. Basically, the key conclusion obtained from the chapter illustrates that as mobility increases (i.e. with decreasing CMR), the probability of needing a location update at the cell decreases. This is somehow complementary with our justification which states that the more direct the mobility pattern is, the less paging load is required since the probability of predicting the correct location of a roaming mobile would be higher. Thus, in terms of categorizing the mobility characteristic, while one is using the mobility rate information with respect to call arrivals, the other focuses on mobility directions. Both however are aiming at reducing the necessary updating loads by increasing the probability of accurate prediction.
Our scheme functions similarly in the sense that directivity aspects of the traveling characteristics are taken into consideration. However, we allow at the same time a certain degree of freedom in the traveling directions, directivity being in the sense of a granular tracking rather than a specific traveling direction. Basically, the aim is to find the significance of the introduced parameter, specifically to determine not only the limitations it creates but also the various factors that could influence its performance.

A few earlier references [10,65] also utilize the collected information about a mobile's direction of traveling obtained at the previous update time to form a distance-based update boundary. However, while the works might provide a good framework of the distance-based scheme, those proposed strategies are a lot more complex and require much more system intelligence. This research intends to identify the fundamental operational issues, and to propose an update alternative with minimal underlying assumptions about network's capabilities. In brief, the key idea of the proposed scheme is to adjust the distance-based threshold \( d \) adaptively only when it is detected that the roaming mobile has demonstrated a change in a certain traveling direction or a movement pattern. Though it might ultimately be worth further extending the overall framework by allowing the coexistence of different threshold-based techniques, in the mean time, the focus is on anticipating the mobile's roaming characteristic using the estimated transitional directivity measures \( \alpha \). In addition, it should be noted that the thresholds chosen for scenarios where the probability of call arrivals can be anticipated would be different from those where such expectation is not realizable. Thus, the resultant costs are always maintained within a reasonable range where applicable.

Furthermore, given that the traveling pattern of a mobile can now be predicted with the introduction of a new mobility characterizing parameter, it seems unnecessary to have the roaming mobile paged on a per-ring basis as it is almost certain that some of the paging signaling will be redundant. Thus, a corresponding cell-grouping methodology alternative is proposed with the intention of improving the performance of sequential paging.

In summary, the main justifications for the introduction of the directivity parameter are two fold: (1) to reduce the location updates, (2) to ensure that the mobile can be paged more efficiently within the constraints set in time and in signaling load. Hence, to reduce updating load without having to sacrifice paging performance.

### 4.3 ADAPTIVE SCHEME THEORETICAL RATIONALE

Generally, the number of updates is a function of mobility rate and call arrival rate. For a given CMR definition, assuming there are \( \alpha(m) \) cells traversed between inter-call arrivals, the update rate is quantified by the probability that boundary cells are reached. In most cases, the
higher the frequency of call arrival, the greater the need for precise tracking of the mobile's movements. The same general relation however does not necessarily apply to the measure of the mobility rate. In fact, the number of updates required varies according to the actual mobility characteristics. Compared to what would otherwise be obtained from a movement-based update scheme of threshold $M$ where the update rate is simply $M/C(m)$, the evaluation for the same parameter in the distance-based scheme becomes more complicated. The performance improvement depends, to a great extent, on the actual roaming characteristics of individual mobiles. Generally, with more accurate predictions, less updates will be necessary to optimize the performance gains in terms of bandwidth consumption and delay compliance.

4.3.1 MOBILITY MODEL

It has been generally assumed (where Markovian movements are considered) that the transition probabilities to all neighboring sides are the same. Consequently, one of the concerns that has raised over the operation of the basic analytical framework is the number of states that is necessary to represent the complete state information. In a $D$-layer cluster of cells, [72] elucidates that a total of $3D^2+3D-5$ number of states is essential to capture the complete transitioning characteristics of a roaming mobile modeled by a Markov movement. Clearly, for a distance-based update strategy, the number of states required increases in direct proportion to the size of the update threshold.

Given that a random movement pattern is assumed, some states are actually the replicates of other states in a structured cell configuration. Recognizing this potential waste of computational resources, [73] and [74] attempt to simplify the representations by using the symmetrical features of a hexagonal network topology. [73] has proposed a new motion model intending to simplify the number of states required capturing the motion characteristics.

Generally, the complete cluster is divided into 6 equal-pieces. Any two cells that have the same relative position on different pieces will be assigned with the same type number and are assumed to show the same roaming characteristics to their respective adjacent cells. Each cell is denoted by state $(i,j)$, with $i$ represents the mobile's current location in terms of cell separations from the last update and $j$ indicates the type number the current cell is assigned to. As an example, Figure 4.1 illustrates the state diagram for a 4-subarea cluster based on the new approach of type classification.
Thus, equivalent to a threshold of $D$ in a distance-based update scheme, for a $D$-subarea cluster with the new random walk model, the number of states required to illustrate the two-dimensional random walk will be reduced to $D(D+1)/2$. By further exploiting the characteristic of intra-sectoral symmetry, [74] shows that the number of necessary states can be reduced to $\lceil(D+1)/2\rceil \lceil D/2 \rceil + i + 1$.

Based on the same underlying assumption, it is discovered in this research work that the number of states could be further reduced to $2(D-1)$. In fact, given the mobile's current ring of residence, it is possible to use at the most two states to describe the complete transitioning patterns between adjacent cells. Thus, in addition to the actual ring of residence, each state of the Markov chain also takes a second parameter indicating the specific class of cell the mobile is currently residing in. Generally, upon entering a new cell, the mobile could possibly propagate into six directions. A cell is referred to as a class-0 cell when it has at least three neighboring cells that belong to the outward ring, or when in all other cases, assigned as a class-1 cell. Figure 4.2 illustrates the allocation of cell-classes in a hexagonal network topology.

Assume that $Q_s$ denotes the steady-state probability of having the mobile reside $s$ rings away from the last updated center in a cell of class $t$. For a distance-based update strategy of threshold $D$, the mobile is allowed to move freely in an area of radius $D-1$ without having to register. As an example, Figure 4.3 demonstrates the state-transitioning diagram for a ring cluster of update threshold 6.
Alternatively, the random movement can be represented by a transition matrix $P$, giving

$$
P = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 & 0 \\
0 & \frac{1}{6} & \frac{1}{6} & 0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 \\
0 & 0 & \frac{1}{6} & 0 & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{6} \\
0 & 0 & 0 & 0 & \frac{1}{6} & \frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\
\frac{1}{3} & 0 & 0 & 0 & 0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\
\frac{1}{3} & 0 & 0 & 0 & 0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6}
\end{bmatrix}
$$

According to the frequency of occurrence, each state is also assigned a scalar quantity of weight $G$. Generally, with more cells in a ring that belongs to the same class, a greater weight should be assigned to the corresponding state.

For the ring cluster of threshold 6 shown in Figure 4.3, the weight of individual states can be represented by
\[ G = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} & \frac{3}{3} & \frac{2}{4} & \frac{3}{4} & \frac{4}{5} \end{bmatrix} \] (4.2)

In a distance-based update scheme, the main task is to determine the frequency at which the boundary cells are reached. It is important, therefore, to be able to explicitly describe the transitional probabilities between inter-ring movements. With the weight of each state taken into consideration, the average transitioning probabilities between rings can be evaluated by combining the behavior of individual states. Figure 4.4 shows the resulting state transitioning diagram.

**Figure 4.4** Averaged state transitioning diagram for using an update threshold of 6

The corresponding transition matrix can be then obtained through (4.1) and (4.2), giving

\[
P' = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 \\
\frac{1}{6} & \frac{1}{3} & \frac{1}{2} & 0 & 0 & 0 \\
0 & \frac{1}{4} & \frac{1}{3} & \frac{5}{12} & 0 & 0 \\
0 & 0 & \frac{5}{18} & \frac{1}{3} & \frac{7}{18} & 0 \\
0 & 0 & 0 & \frac{7}{24} & \frac{1}{3} & \frac{3}{8} \\
1/30 & 0 & 0 & 0 & \frac{3}{10} & \frac{1}{3}
\end{bmatrix} \] (4.3)

For example, consider the transitional probability from ring 3 and ring 4. The summation of elements (5,7), (5,8), (6,7), (6,8) from matrix (4.1) appropriately blended with elements of (1,5) and (1,6) from vector (4.2) will give the entry for element (4,5) in matrix (4.3).

### 4.3.2 Concept of Directivity

Based on a ring-structured cellular system, Figure 4.5 gives a brief illustration of the state diagram used to model the transitional probability of movements.

**Figure 4.5** The state diagram illustrating the inter-ring transition probabilities in a hexagonal cell
While each state represents the actual “ring” the mobile is currently residing in, the arrows indicate the probability of transitioning to its adjacent states. Thus, for an update threshold of $D$, if the mobile is currently residing in the state (or ring) $s$, the transitioning probabilities of going forward, sideways and backward are denoted by $q_{s,D-1,D}$, $q_{s,s,D}$, and $q_{s,D-1,D}$, respectively. With the assumption of random movements (i.e. the same probabilities to all sides), the general expressions become:

$$q_{s,s+1,D}^{(1)} = \frac{2s+1}{6}$$  \hspace{1cm} (4.4)

$$q_{s,s,D}^{(1)} = \frac{4s-1}{3[4s-1]}$$ \hspace{1cm} (4.5)

$$q_{s,s-1,D}^{(1)} = \frac{4s^2 - 3s + \sqrt{2}}{3s[4s-1]}$$ \hspace{1cm} (4.6)

Note that the probability of moving to an adjacent inter-ring cell is always equal to 1/3. In other words, for each cell in the network, there are always two cells of the 6 neighbors that belong to the same ring.

With the widespread usage of mobile terminals, it seems however non-realistic to assume that the probability to all sides will be the same when many of the movements have in fact demonstrated a traceable purpose. Therefore, it is important to consider the transition probabilities to be different (forwards/sideways/backwards) in order to more realistically model the movement characteristics and hence the introduction of $\alpha$. Intuitively, instead of assuming that the mobile is moving to all neighboring cells with an equal probability, it is now assumed in this framework that the transition probability to a cell on the outward ring will have an additional factor $\alpha$. For example, the transition probability from state 1 to state 2 in Figure 4.4 will thus be changed from 1/4 to $\alpha/4$. Consequently, to maintain equilibrium, the transition probabilities of moving sideways and backwards will also vary accordingly.

It is therefore necessary to re-compute the transitional probabilities between cells. Effectively, for an update threshold of $D$, the general expressions derived for calculating the probability of entering adjacent rings from ring $s$ are shown for the transitioning probabilities of going forward, sideways and backward in (4.7)-(4.9), respectively.

$$q_{s,s+1,D}^{(1)} = \frac{2as + \alpha}{6s}$$ \hspace{1cm} (4.7)

$$q_{s,s,D}^{(1)} = \frac{2(3 - \alpha)s - \alpha}{3[4s-1]}$$ \hspace{1cm} (4.8)

$$q_{s,s-1,D}^{(1)} = \frac{2(3 - \alpha)s^2 - 3s + \alpha/2}{3s[4s-1]}$$ \hspace{1cm} (4.9)
Thus, the transitional probability will not only be as functions of the current ring of residence $s$, but also responsive to the directivity measures $\omega$ associated with individual mobiles. In addition, when $\omega = 1$, the two sets of equations (i.e. (4.4)-(4.6) and (4.7)-(4.9)) will become the same.

To illustrate the impact of $\omega$ on the resultant movement patterns, the transitional probabilities to the neighboring cells for different values of $\omega$ are given in Table 4.1.

<table>
<thead>
<tr>
<th>Transitional probability</th>
<th>$\omega = 0.5$</th>
<th>$\omega = 1$</th>
<th>$\omega = 1.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_{s,s+1,D}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{3}{4}$</td>
</tr>
<tr>
<td>$q_{s,s,D}$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{3}$</td>
<td>$\frac{1}{6}$</td>
</tr>
<tr>
<td>$q_{s,s-1,D}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{6}$</td>
<td>$\frac{1}{12}$</td>
</tr>
</tbody>
</table>

Generally, the greater the value of $\omega$, the more outwards (and hence directive) the traveling pattern is. This research work focuses on determining the impact of variations in the inter-ring transitional probability on the updating strategy by means of the optimal threshold selection.

Finally, in assigning a directivity index to various mobility characteristics, the basic idea was to set up two thresholds, one for distance, and the other for time. Once the distance threshold is reached, the actual time taken is compared with the time threshold to see whether it approximates to what was first anticipated. An affirmative answer would suggest that there is no change to the directivity assigned during the last registration update. On the other hand, if the difference is significant, that suggests a great possibility that the mobile’s traveling pattern was altered and the path is now either more “focused” (if the threshold was reached in a less time), or more “diverse” (if the threshold was reached in longer time). The directivity index being assigned to the mobile will then require to be adjusted to reflect such changes. Figure 4.6 gives a condensed illustration of the concept.

Furthermore, it is intended to incorporate $\omega$ for a dynamic grouping of the rings such that the operation requires only minimum paging before the mobile is located. For example, if $\omega$ is less than 1, then it implies that the mobile has a tendency of incurring intra-ring movements, thus more paging subareas should be incorporated with each subarea containing less cell rings. Similarly, if $\omega$ is estimated to be large, fewer subareas should be formed with more cell-rings covered in each subarea. The main advantage of this assignation scheme is that while paging
delay is appropriately controlled, the number of paging signals is somewhat controlled to be at the possible minimal level.

![Flowchart](image)

**Figure 4.6** A brief illustration of the overall operation

### 4.4 ANALYTICAL FRAMEWORK

With the distance-based update strategy, clearly, update cost increases with the number of times the distance threshold is reached. Given there are \( m \) movements between call arrivals, the number of times the distance threshold is reached will determine the update cost. Based on the assumption of random movement patterns, [66] proposed analytical models to study the cost of update and paging. The analytical models are significant as they provide an upper bound to the performance gains that are achievable by the dynamic update strategies over the conventional GSM standards. For the sake of comparison, the same model forms the basic framework for this analytical study. The total operational cost can thus be evaluated by using the general function

\[
C_{\text{total}} = C_u(\mu, \sigma) + C_v(\alpha(m), D_T)
\]  

(4.10)

While the update cost \( C_u \) depends on the joint effects of mobile's traveling rate (\( \mu \)) and roaming pattern (\( \sigma \)), the paging cost \( C_v \) is a function of call arrival rate \( \alpha(m) \) and the size of update area \( D_T \). To evaluate the update cost during call arrival, there are two separate parameters to be quantified for a given threshold; namely the probability of having \( m \) cell boundary crossings during call arrival \( \alpha(m) \), and the expected number of updates to be
registered $E_s(m)$. Given then that the unit update cost is $U$, the update cost can be evaluated from:

$$C_s = U \sum_{m=0}^{\infty} \alpha(m)E_s(m)$$

(4.11)

where $D$ is the predefined update threshold.

Clearly, the definition of CMR will have a significant impact on the resulting total cost. Assuming call requests arrive following a poisson distribution, [48] has provided a detailed framework to evaluate $\alpha(m)$ for a given cell residence time distribution CRT. In other words, for each specific CRT distribution defined, there is a corresponding $\alpha(m)$ derived to reflect the characteristics of the network concerned. As an example, an exponential distribution will give

$$\alpha(m) = \begin{cases} 1 - \frac{1}{\rho} \left[ 1 - \left( \frac{1}{\rho+1} \right)^m \right] & , \quad m = 0 \\ \frac{1}{\rho} \left[ 1 - \left( \frac{1}{\rho+1} \right)^m \right] \left( \frac{1}{\rho+1} \right)^{m-1} & , \quad m > 0 \end{cases}$$

(4.12)

as opposed to a gamma distribution, giving

$$\alpha(m) = \begin{cases} 1 - \frac{1}{\rho} \left[ 1 - \left( \frac{1}{\rho+1} \right)^m \right] & , \quad m = 0 \\ \frac{1}{\rho} \left[ 1 - \left( \frac{1}{\rho+1} \right)^m \right] ^{m-1} \left( \frac{1}{\rho+1} \right) & , \quad m > 0 \end{cases}$$

(4.13)

where $\rho$ is the call to mobility ratio and is approximated by $\lambda_c/\lambda_m$.

Although it was not considered in this work, it is noted that the other alternative is to make $\alpha(m)$ a function of “rings” (instead of cell boundaries), and to make $E_s(m)$ the expected number of updates for the $m$ ring boundary crossings. That way, the operation can be effectively considered as a movement-based scheme with the expressions of probability functions $\alpha(m)$ derived not only as a function of CRT, but also with weight $\omega$.

As our proposed scheme selects the optimal threshold based on the distance traveled (in terms of number of rings), another way of determining the expected number of updates is to look at the probability of the mobile moving from state 0 to state $D$ during the call’s inter-arrival time. Consequently, according to the transition probabilities derived in the preceding section, $E_s(m)$ can be evaluated as:

$$E_s(m) = \sum_{k=D}^{m} q_{0,0,D}^{(k)}$$

(4.14)

where $q_{0,0,D}^{(k)}$ defines the probability of moving from state 0 to state $D-1$ in $k$ steps. For the proposed distance-based scheme, the number of updates does not depend only on the number of movements, but also on the direction of movement; specifically, the probability of having the mobile roam toward the next ring-cell. For each update to occur, a minimum of $d$
movements (or cell boundary crossings), or a max of $m$ movements would be required depending on how many transitions were actually made to the neighboring cells. Thus, if $k=m-1$, there will be only one update during a call arrival (i.e., $E_u(m)=1$) as it takes the whole $m$ movements for the mobile to move from state 0 to state $D$. On the other hand, if $k=D-1$, the probability of transition to the next ring is found to be unity for every movement. In other words, it is guaranteed that every time the mobile makes $D$ number of transitions, it will be reaching the predefined threshold ring for update. Correspondingly, for $k$ ranging from $D-1$ to $m-1$, the mobile is wandering around between rings dependent on the transitional probability.

Substituting the general expressions derived for the transitioning probabilities of (4.7)-(4.9), the number of updates can be thus evaluated from:

$$E_u(m) = \left\{ \prod_{i=1}^{A} \left( \frac{2\omega_i + \omega}{6i} \right) \right\} \left\{ \prod_{i=0}^{A} \left( \frac{2\left[3 - \omega_i\right] j_i^2 - 3j_i + \omega_i/2}{12j_i^2 - 3j_i} \right)^A \right\} \left[ \frac{2\left[3 - \omega_i\right] - \omega_i}{12i - 3} \right]^{m-2A-d} \right.$$  \hspace{1cm} (4.15)

The first term in (4.15) takes into account the probability that the mobile transits to the next ring of cells after each movement. The actual probability varies according to the cell of residency. Clearly, a minimum of $D$ movements will be all it takes to incur a new update in some cases. Although such a movement pattern is possible, the mobile is expected to inherit a strong correlation between traveling directions and hence a restrictive movement pattern. It is thus more likely that the reaching of the update boundary requires more than $D$ transitions. Clearly, apart from the basic $D$ movements, there are $m-D$ movements where the directions of transitioning are significantly different from their corresponding predecessors. In the extreme case where a transition is made in the opposite direction to its most recent movement, the term backward is used to describe the moving characteristics. Assuming there are $A$ units of such backward movements, the second term in (4.15) takes into account the combined effects of these $A$ transitions to the final pattern. In this expression, $j_k$ refers to the ring of residence before the backward movement incurs. Apart from moving forwards and backwards, according to the definition of the network topology, it is possible that the mobile transits to an adjacent cell belonging to the same ring (or subarea) as the current cell. In this case, a sideways movement is said to have occurred. For a hexagonal cell topology, the transitioning probability is always equal to a third. Obviously, for every one backward movement incurred, there will be at least one other movement in the opposite direction required before the mobile is back on the main direction of traveling. Hence for a total of $m-D-A$ units of sideways movements, the overall effects are considered by the last term in (4.15). Consequently, the resultant operational cost will vary for different mobility patterns.

This is where the introduction of $\omega$ becomes significant. Higher values of $\omega$ mean more directive pattern in traveling, and thus, fewer boundary crossings are required before reaching
the threshold state for the same mobility rate. Consequently, we ended up having a scenario where the total update cost is a function of not only CMR and $D$, but also $\omega$.

For the computation of paging cost, the same analysis framework [9] as for an ordinary distance-based scheme is found applicable. Generally, given a per unit paging cost of $V$, the calculation comprises two parts—to estimate the probability of having the roaming mobile reside in subarea $j$ (i.e., $A_j$) and to predict the number of cells ($\varphi_j$) to be paged before the user is located. Specifically, the evaluation for such a cost can be defined by (4.16).

$$C_v = V \sum_{j=0}^{\eta-1} \rho_j \varphi_j$$  \hspace{1cm} (4.16)

where $\eta$ is a measure of the paging delay.

If based on an underlying assumption of random movement patterns, clearly, the bigger the area enclosed within the update boundary, the longer duration the mobile will remain in the current update area. Thus, the occurrence of a new update becomes less probable. However, the savings achieved in the update operations are likely to be abridged by the increasing paging load upon call arrival — the greater the update area, the higher the number of cells required to be polled during the paging phase. Given the unavoidable tradeoffs between the two terms, the ultimate solution is thus to bound one cost and minimize the other.

### 4.5 SIMULATION FRAMEWORK

Figure 4.7 provides a brief illustration of the operation of the simulation.

![Simulation Diagram](image-url)  
**Figure 4.7**  A brief illustration of the simulation framework
For each $\omega$, a set of total costs consisting of updating and paging costs is generated for the 10 possible simulated threshold values. It was indicated by [9] that a movement number of 300 would be sufficient to illustrate the maximum number of boundary crossings between call arrivals. In addition, to smooth out the variations causing the probabilistic nature of the distribution functions, it is assumed that there are a total of 100,000 users roaming in the system simultaneously, and thus running 100,000 iterations.

At the end of the iterations, we should have obtained two matrices, one for the expected number of updates $E(u(m))$ and the other for residency characteristics of the mobile $\chi(i,m)$. Effectively, the obtained statistics can be used to calculate the cost for firstly, different values of CMR and secondly, different values of $\omega$. From these, the optimal threshold $D$ can be obtained from the plotted chart.

The issues we wish to address are two fold: (1) the precise impacts of the mobility rate and the transitional directivity index on the selection of optimal threshold, (2) the inter-dependency between the two parameters. Specifically, could the information provided by $\omega$ be sufficient to compensate for the absence of reliable data about call arrival characteristics?

### 4.6 NUMERICAL RESULTS AND DISCUSSIONS

As a general rule, the selection of an optimal threshold depends on two factors — mobility rate and roaming directivity. These two parameters form an AND function for optimal system performance. The aim of this chapter is to determine how closely these two “approaches” correlate with each other; specifically, to examine whether one factor weighs more than the other and if so, how a compromise can actually be formed between the two variables.

### 4.6.1 EFFECTS OF CMR

With $\omega$ fixed, Figure 4.8 shows the impact of CMR on the selection of optimum $D$. Generally, the smaller the CMR, the higher the mobility rate (respective to the call arrival rate) and thus, the greater the update cost for a fixed threshold. Thus, for fixed $D$ and $\omega$, it is evident that the greater the CMR, the lower the expected update cost (per call arrival). Intuitively, an increasing CMR suggests a lower number of boundary crossings between call arrivals and thus, less updates will be required to maintain the reachability of the roaming mobile. The impact of CMR however seemed to have diminished with increasing $D$. In fact, it is observed that when CMR is high (say >5), the update cost would be almost negligible for all values of $D$. However, provided that the network conditions (i.e. the loading constraints) permit, it
would still be an optimal practice to perform updates at the completion of every movement. Since the call arrival rate is relatively frequent, it is necessary to minimize the required paging load where possible.

![Graph showing impact of CMR on the selection of threshold D.](image)

**Figure 4.8** Impact of CMR on the selection of D

It seems reasonable to state that efficient location management is particularly important for handling mobile users with low CMRs. In such scenarios, since the mobility rate is high, the issue is to reduce update without having to suffer the resultant paging cost. Correspondingly, the proposed scheme should be designed such that the selection of optimal threshold is an operation sensitive to even small variations in CMRs.

### 4.6.2 EFFECTS OF $\omega$

Figure 4.9 shows the impacts of $\omega$ on the selection of optimum $D$. 

Figure 4.9  Impacts of $\omega$ on the selection of $D$ (a) $\omega=0.5$ (b) $\omega=1$
(c) $\omega=1.5$ (d) CMR = 0.05
Figure 4.9  Impacts of $\omega$ on the selection of $D$ (a) $\omega=0.5$ (b) $\omega=1$
(c) $\omega=1.5$ (d) CMR = 0.05

Generally speaking, the effect of $\omega$ has a greater impact than mobility rate. A low CMR would, however, indicate, almost by default, the incursion of high update cost. Obviously, for a fixed threshold, the higher the directivity, the more wasteful the system resources are going to be. Thus, it would be wise to appropriately increase the threshold to minimize such impacts.
What needs to be done then is to quantify more precisely the variations in optimal threshold caused by changes in directivity.

For roaming mobiles with the same CMR statistical characteristics, the cost of update becomes a function of $\omega$ (Figure 4.9(d)). In other words, for the same number of mobile transitions, those that were estimated to have a more “directive” traveling pattern (i.e. higher $\omega$), would reach the same predefined threshold more quickly thus resulting in a greater cost due to the more frequent incursion of updates. Hence, the threshold will be reached faster and consequently, incur a higher updating cost to the network.

It is also encouraging to see that in terms of their proven ability in selecting the optimal threshold, both parameters CMR and $\omega$ seem to be able to reach the same decisions. In addition, the actual cost of update would vary in accordance with the traveling direction of the roaming mobile $\omega$. In other words, it is observed that while the total cost $C_{\text{total}}$ is a function of $D$, $D$ is a function of CMR and $\omega$. Thus, even with the same threshold, the actual update cost would vary depending on the estimated value of $\omega$.

However, when moving away from the optimal threshold, it is observed that the performance differences become significant with varying CMR and $\omega$. In fact, the simulation results show that the selection of sub-optimal thresholds afterwards will be different. This behavior somehow makes the introduced adaptive scheme particularly promising for applications. In reality, due to the large number of subscribers that are simultaneously connected to the network at any instant of time, the network might have to set some constraints on the maximum allowable update rate (i.e. a minimum executable threshold) to maintain the total signaling loads within a reasonable range. As a result, the roaming mobiles might not always be given permission to perform updates at their ideal frequencies. Thus, as the differences increase with varying CMR and $\omega$, the additional information made available about $\omega$ becomes significant in making such sub-optimal decisions.

For a mobile that is experiencing high mobility rate, ideally the threshold should be increased to decrease the update cost. The overall performance however, might not always improve. In fact, the actual improvement depends greatly on the mobile's traveling patterns. For example, in a scenario where the mobile moves fast in a circular motion, given that paging delay is constrained, more cells are likely to be paged than is necessary when the threshold is set high.

### 4.6.3 EFFECTS OF $U/V$

Generally, the dependence of $\omega$ in the selection of optimal threshold becomes more obvious
when the ratio between the unit update cost $U$ and the unit paging cost $V$ changes. Figure 4.10 illustrates the impact varying $U/V$ definitions have on the sensitivity of the directivity index.

\[ U/V = 1, \text{CMR} = 0.05 \]

(a)

\[ U/V = 5, \text{CMR} = 0.05 \]

(b)

**Figure 4.10** Impacts of $U/V$ on the selection of $D$ (a) $U/V = 1$ (b) $U/V = 5$
(c) $U/V = 10$ (d) CMR = 10
Figure 4.10  Impacts of $U/V$ on the selection of $D$ (a) $U/V = 1$  (b) $U/V = 5$  
(c) $U/V = 10$  (d) CMR = 10
Generally, for smaller thresholds, the performance showed the predominant contribution was from the definitions of unit update cost. Simulation results show that as $U/V$ increases, it is not necessary to have the optimal threshold settled at some small values. In fact, the optimal threshold increases in value with increasing $U/V$ ratios.

The sensitivity to $U/V$ definitions however becomes insignificant for high CMRs (Figure 4.10(d)). One justification for such system behavior is that when the threshold is high, the main cost would have come from paging. Thus, the variations in unit update would not have significant influence on the system performance. Note also that the update cost is highest when the threshold equals to 1. The need to have an efficient location management technique is reinforced as a significant reduction in the update cost is evident even when the threshold is merely increased by one unit. The improvement in performance however stabilizes eventually with increasing thresholds.

In the case where $U$ and $Y$ are both set to unity, the total cost is really only a measure of the number of required updates between call arrivals. Thus, in a real system where there are millions of subscribers simultaneously connected to the network at any instant of time, an appropriate assignment of the unit cost of updating and paging will be required to ensure that a reasonable performance is guaranteed.

In summary, it is found that as far as mobility characteristics are concerned, the actual transitional direction of roaming mobiles plays a significant role in selecting the optimal threshold in addition to the usual perception about mobility rate. As an additional selection criterion, the threshold will increase if the movement characteristics have been found to be relatively stable. On the other hand, if a traceable movement pattern cannot be clearly identified due to a greater tendency to directional changes, the selection of the optimal threshold would be more conservative. In other words, the range for acceptable thresholds would be wider. Effectively, the more directives the mobile’s roaming characteristics inherit, the more precise the prediction of the possible locations upon call arrival, and hence the lesser need for frequent registrations during movements.

4.7 APPLICATIONS AND IMPLEMENTATIONS

The most profound advantage of the proposed technique should probably be the fact of its simplicity, from both the viewpoints of operations and implementations. For a start, the adaptive technique will be relatively easy to implement in the sense that the resource and intelligence required for the determination of an optimal threshold update are minimal. In addition, as the network adaptively learns about the mobile’s roaming characteristics (e.g. the transitional directivity), no prior registration of such statistical information from mobile to network is required.
Upon the first registration of a mobile's update, each individual user is assigned a "grouping" directivity index according to the specific movement pattern in terms of directions demonstrated by the mobile's roaming history. Associated with the index is an optimal update threshold to be used by the mobile as a guide to decide whether an update is required after each cell boundary transition. The "matching up" of the two parameters is obtained from one of those "Figure 4.9"-like cost plots evaluated for varying thresholds for the specific $U/V$ definitions given by the system infrastructure. Thereafter, the appropriateness of the assigned directivity index will be checked against the actual movement every time an update is registered. On the network side, once a change occurs in the matching between $\omega$ and $D$, a beacon message will be sent to all mobile users announcing such change of optimal $D$ since their last registrations. At times when changes in $\omega$ were detected, a notification of the corresponding new update threshold could be sent as part of the acknowledgment through piggybacking such that the additional controlling signaling is minimized.

4.8 FUTURE WORK

No special concern has been directed as for to the occurrence of ping-pong effects. One of the possible solutions that we are currently evaluating is to combine the operational concepts of a movement-based updating strategy with the scheme. The issue can be tackled by a couple of different approaches; one is to make the system adaptive to individual schemes and the other (preferable) is to combine the merits of each to form a more complete solution to the problem of location management as a whole. For the latter, what could be done is to give each cell two identities. Firstly, to assign cells an absolute cell ID to maintain a record of past movement history. Thus, whenever the mobile moves back to a previously updated cell, a duplicated update procedure will not be required. Secondly, to assign cells a temporary ring ID for the minimization of the combined updating/paging cost. Therefore, while minimizing the impact of possible ping-pong effects, the original scheme can be used to optimize the selection of threshold. For example, [10] presents a predictive distance-based scheme taking full advantage of the correlation between a mobile's current velocity and location and its future velocity and location. The reliability of the correlation model seems to be something that is worth further exploration. Any such gathered information should help to increase the precision of the prediction made about the probability distribution of user locations. Thus, while keeping the number of cells to be paged in each polling cycle the same, the probability of finding the location is increased.

As far as paging is concerned, for an inconsistent or untraceable roaming characteristic (i.e. $1 \leq \omega$), the optimal strategy would be to page the roaming mobile on a per-ring basis, such that
it ensures the required paging load is always restrained at a minimum. However, this could result in a high paging delay that is undesirable in a real system. In order to better manage the operations of sequential paging, a switching mechanism between ring- and other paging schemes should be incorporated such that an optimal performance is guaranteed at all times.

4.9 CONCLUSIONS

This chapter outlines the development of a new location management strategy utilizing the directional information of mobile’s traveling patterns. In the proposed scheme, a transitional directivity index has been introduced. By using this index, it is possible to update the optimum threshold for the distance-based location tracking strategy not only adaptively, but also more accurately. From the operation and implementation points of view, simplicity is the most profound advantage of the proposed technique. Firstly, the adaptive technique will be relatively easy to implement in the sense that the resource and intelligence required for the determination of an optimal threshold update is confined to a minimum. Secondly, the network adaptively learns of the mobile’s roaming characteristics, no prior registration of such statistical information from mobile to network will be required. All in all, the simulation results were satisfactory. They not only gave some significant insights into the mobility problem, but also provided new directions for designing better strategies for location management.
Chapter 5
Kalman-Filter Based Update Schemes

In this chapter, we propose a new predictive location management strategy that reduces the update cost while restricting the paging load optimized for mobiles roaming with traceable patterns. Enhanced with directional predictive capabilities offered by Kalman filtering, new update boundaries are assigned to better reflect the movement patterns of individual mobiles upon location registration. Thus, while complying with the required delay constraints, QoS measures such as throughput will not need to be sacrificed as a result of increasing update threshold. The contribution of this chapter is two-fold: (1) to propose a distribution model that is capable of describing a wide range of movement patterns with varying correlation between traveling directions and (2) to show the capabilities (in terms of reliable performances) of the Kalman filter in predicting future movement patterns. Simulation results have successfully demonstrated the ability of the Kalman filter in assigning update boundaries capable of reflecting a mobile's roaming characteristics. The performance gains achieved mainly through a significant reduction in the number of updates indicate its potential for promoting better bandwidth conservation.
5.1 INTRODUCTION

Developing an efficient location management technique is an important step in working towards the determination of an optimal solution to the problem of managing mobility. With the irregular nature of the cell sizes in a cellular network, the behavior of mobile movement changes from cell to cell and from user to user. Thus, the need for designing an adaptive algorithm for tracking a roaming mobile becomes imperative. The biggest challenge in framing location management is to find the most favorable tradeoff between the location updates load and the searching load—the two parameters that are frequently referred to as location registration cost and the call delivery cost. Intuitively, one would assume load increases when searching has to take place to locate the current address of the destination mobile host. Conversely, more information load will result due to frequent updates if a more complete location information database were to be maintained. For location update, there are two classifications based on static and dynamic technologies. With the static algorithms, the locations of the reporting centers are fixed. Thus, all mobile users are required to transmit their update messages when roaming into the same set of cells predefined either centrally or distributively. In contrast with the dynamic algorithms, location updates are performed based on each individual mobile user’s call and mobility patterns. From the many alternatives, depending how the threshold is defined for the updates, time-based, movement-based, and distanced-based techniques are the three most commonly used methods [39]. Among all, it is shown in [7] that the distance-based scheme tends to outperform other alternatives (in terms of minimal total cost per call arrival) under the assumption of a random movement model.

Given that mobile communications are now used so universally, it becomes realistic to consider some degree of correlation between the traveling directions to adjacent registration updates in time. Despite its importance, there seems little research on the potential of utilizing such directivity information, directivity being a measure of correlation level in traveling directions observed between adjacent cell boundary crossings.

Location management is all about finding an efficient solution to the problem of mobile tracking. The main purpose of update is to restrict the area of paging upon call arrivals, it is expected that the update boundary is defined such that it reflects the network's confidence in locating the roaming mobile. It becomes apparent that the assignation of location areas should not only be specific to each mobile's traveling rates, but also be sensitive to its actual movement patterns. This observation motivates the current research work. The idea is to explore the possibility of observing a mobile's past movement history, and to anticipate each individual's future roaming characteristic, be it at a certain time of the day or in a specific geographical area. The purpose of this research work is to design an update scheme that
Chapter 5 Kalman-Filter Based Update Schemes

...dynamically assigns the optimal update boundary in order to minimize the combined operational cost of update and paging.

While similar goals have been the aim of previous research works [10,58,59,65], the novelty of this chapter lies in its inclusion of a Kalman filter during the prediction stage. Specifically, with more granularly tractable traveling directions, it is possible to predict network residency more precisely upon call arrivals. Hence, the assignment of a more mobile-specific shaped update boundary is expected to minimize the necessary registrations from what would otherwise be incurred from the conventional distance based update strategy.

The proposed new update scheme allows an asymmetrical update boundary to be assigned for individual mobile users with different mobility and call arrival characteristics. Effectively, while constraining the paging signaling loads required upon call arrivals, it is possible to reduce the necessary registration numbers by allowing asymmetrical definitions of the update threshold in various directions. In fact, recognizing a highly possible presence of correlation between traveling directions for individual mobiles, the scheme is designed not only to identify such consistency in patterns, but also to utilize such information to predict future movements. Technically, as long as some distinct mobile characteristics can be granularly identified, an asymmetrical update boundary is defined such that it captures the movement characteristics of the roaming mobile more adequately. With a fixed registration area, the mobile is likely to stay in the registration area for a longer duration before exiting compared to its symmetrical update boundary equivalent, thereby achieving reductions in the number of update registrations. Note that the asymmetrical update boundary can be of any shape, an optimal rectangular approximation is used in this chapter for a simple illustration of the concept.

In summary, the complete operation is comprised of two main procedures. In the first part, Kalman filtering is used to predict the future movement of an idle mobile from its past mobility characteristics observed from previous registrations. The level of correlation (mathematically quantified by a directivity index $f_d$) will then be used to assign an appropriate optimal update boundary that reflects the granularly traceable patterns predicted.

To evaluate the achievable performance improvement of the proposed predictive location management scheme, movement patterns with various levels of randomness were generated according to Gauss-Markov distributions. The performance of the Kalman-filter based update scheme proposed in this chapter is compared to the original distance-based scheme. This chapter demonstrates that with predictions incorporated in the operation, the mobile inherits a more directive pattern which will incur a lower update cost as a result of a better location area assignment. The simulation results not only verify the applicability of the proposed update/paging scheme, but also identify specific system parameters that have direct impacts on the overall performances.
Section 5.2 begins by highlighting briefly the development of other related work oriented about the concepts of direction and prediction. Key design philosophy and distinct features that were incorporated in the proposed Kalman filter based predictive update/paging scheme are indicated in Section 5.3. Operational details as well as various underlying network assumptions are also discussed in this section. Section 5.4 then gives a brief description of the simulation philosophy in addition to defining various details for setting parameters. Discussions are included in Section 5.5 to examine the significance and effectiveness of the proposed asymmetrical (non-uniform) update/paging scheme. Future extensions to this work are discussed in section 5.6. Section 5.7 concludes the chapter by summarizing the main features of the proposed scheme.

5.2 RELATED WORKS

Despite the existence of many specific update strategies in the literature, the number of proposals that have actually considered either directional information or predictions is small, let alone attempting a combination of both. To justify the significance of the predictive scheme proposed in this chapter, a brief review of previous developments is included in this section. The main merits and downfalls of the existing techniques will also be summarized to indicate areas of possible improvements.

In the sphere of managing mobility, [10] appears to be one of the first papers to introduce the idea of predictive location management technique. Briefly, the Gauss-Markov model is utilized to simulate the correlations between mobile's traveling speeds. Thus, as part of the update registration process, the network takes measures of the mobile's traveling speed. Based on the collected information, a prediction is made about the roaming mobile's whereabouts in advance in the form of a probability density function. Thereafter, an update is due either when the mobile moves out of or when the current residency exceeds a certain threshold from the predicted location. It is noted that the actual direction of traveling was not considered in this reference. Effectively, random movements are simulated and paging is based on the conventional ring-structure scheme. On the contrary, the Gauss-Markov model is applied in our proposed scheme to simulate mobility patterns with varying correlations in the traveling directions. The usage of such a model allows movement patterns incorporating different levels of randomness to be simulated and studied.

In [59], the concept of trip is incorporated in the generation of mobile movements. More specifically, by first defining a specific destination, the mobile is assumed to travel following a particular path. Thus, at each cell boundary crossing, the mobile evaluates both the cost of performing an update and the cost of paging in case an update was not due. An update is registered only if the estimation indicates that the network would incur a lower cost with an
update than without. However, in order to perform the necessary calculations, it should be recognized that individual mobiles are not only required to be equipped with more computational power, but are expected to have a complete knowledge of the network structure. Consequently, the implementation of such a scheme will result in a higher cost with an increased complexity.

Instead of assuming that the trajectories of mobiles are independent in different inter-update intervals, [10] introduces a linear Markovian movement model that carries direction information over from one inter-update interval to the next. Not only is an asymmetric distance-based update criterion derived, a search strategy that exploits the knowledge of the elapsed time since the last update is considered. The scheme however, is also based on a one dimension linear topology. Such assumption would in fact make the paging problem a lot simpler since it is almost guaranteed that the user can be found by paging symmetrically in two directions.

In this chapter, instead of simply assuming a random movement model, we focus on methods of detecting directivity correlations between movements. Specifically, we demonstrate how a measurement can be obtained from the traveling characteristics; and secondly, how feedback values can be used to decide the shape of the registration areas. Assuming the defined area can reflect the mobile's roaming characteristics with reasonable precision, the challenge of bandwidth conservation is well addressed through reductions in the number of location registrations.

5.3 KALMAN-FILTER BASED LOCATION TRACKING

The idea is to predict the traveling patterns leading to the next registration and to assign an appropriate update boundary such that the combined operations of information and search incur a minimum cost. Generally, a unit update is more costly than a unit paging due to higher bandwidth consumption for secured operations. The ultimate goal is thus to minimize the necessary update signaling through maximization of the registration area while imposing some limits on the required paging load. In comparison to the symmetrical update boundary definitions [8], the intention is to reduce the number of required updates while maintaining the current registration area. This in effect would justify the attempt to assign asymmetrical update boundaries in accordance to the specific movement patterns identified during each registration timeframe. Consequently, the design minimizes the dependency between the operations of update and paging.

Different from previous studies [10,48,65] where Markovian movement model is used, this chapter illustrates the impact of variations in the transitional directions on the updating strategy in terms of the optimal threshold selection. Thus, when it is found that the correlation
between traveling directions at the current registration has been changed since its last registration characteristics, the update boundary will be adjusted to ensure that the specific mobility patterns were considered. In summary, the update boundaries are dynamically assigned according to the level of correlation in mobility patterns between registrations in time.

5.3.1 MOBILITY MODEL

Consider a network with a hexagonal cell topology shown in Figure 5.1.

![Hexagonal Cell Topology](image)

**Figure 5.1** An example of cellular

Upon entering a new cell, there are 6 directions into which the mobile could possibly roam. Given the level of usage penetration in today’s mobile communications, it is unrealistic to assume that the probabilities of moving to all sides are the same. Recognizing a highly possible presence of correlation level between traveling directions in time and in certain geographical locations for individual mobiles, the need for a mobility model that could simulate such characteristics of reduced randomness becomes obvious. Specifically, the distribution model should be capable of adjusting the degree of correlation between the traveling directions for various movement patterns.

In this chapter, we propose to model the degree of randomness in movement patterns by appropriately defining the distribution parameters in a Gauss-Markov model. In fact, through adjustments in the cutoff frequency, the level of correlation between traveling directions can be simulated. The model is proved capable of representing a wide range of movement patterns; from straight lines to random patterns. As an example, Figure 5.2 illustrates a possible movement pattern simulated by the Gauss-Markov distribution. Depending on the level of correlation, each movement will incur a change of direction in the range of $[\pi, -\pi]$ in reference to the last updated cell.
Figure 5.2(a) shows a mobile's possible traveling pattern on a two-dimensional network modeled in Figure 5.1. Thus, starting from the last updated cell (0,0), the mobile moves through cells (0,-1), (0,-2), (0,-3), ..., (0,10), ..., (15,30) and finally arrives at cell (38,30) in this particular example.

Clearly, depending on the parameters defined in the Gauss-Markov model, the correlation in traveling direction changes over time. To more closely examine the impact of the mobility pattern, Figure 5.2(b) shows the corresponding direction of traveling (measured in the radians) at each of the 500 sampling times.

**Figure 5.2** (a) Example of an actual mobile movement on a two-dimensional system, and (b) the corresponding Gauss-Markov distributions for modeling mobile's travel directions at boundary crossing.

On the one hand, when the traveling directions are highly correlated in time, the points in Figure 5.2(b) are more concentrated. On the other hand, when the traveling directions become
less correlated, the corresponding points in Figure 5.2(b) are shown to be more sparsely distributed. For example, a straight line approximation between cells (21,30) and (23,4) in Figure 5.2(a), graphs as a highly concentrated point in Figure 5.2(b). Conversely, when many “turnings” are evident in the movement patterns (for example in the vicinity of cell (0,0) during the first 25 sampling times), a more diffused distribution of the sampling points is observed as shown in Figure 5.2(b). Refer to Figures 5.2(c) and 5.2(d) for the zoomed-in versions of the corresponding figures.

**Figure 5.2 (c)** A zoomed in picture of mobile’s movement demonstrating straight line approximations

**Figure 5.2 (d)** A zoomed in picture of mobile’s movement in the vicinity of the last update cell (0, 0)

Note that an abrupt change in direction (i.e. from +3.14 radians to -3.14 radians) between the 24th and 25th sampling times is only due to the definition of the traveling directions.
Assuming that an upward movement from the last update cell represents the direction of 0 radians, graphically $\pi$ and -$\pi$ will both imply the direction of moving downwards. The generated movement patterns successfully demonstrate the capability of a Gauss-Markov model in simulating different levels of correlations in traveling directions. The usage of such a movement model is thus justified for further analysis.

To more closely examine the effects of varying correlation levels, consider firstly examples where specific movement patterns were defined to be highly correlated. In such cases, it is reasonable to make the update boundary almost infinitely large (equivalent to the never update scheme of [49]) as the location of the mobile can be determined almost with the zero paging cost guaranteed. As it was summarized earlier, straight line movements are a classic example of such a highly predictive category. Another less obvious example however is the one where the movements appear to repeat themselves in certain time frames. Although adjacent movements do not correlate in such a way that they all result in traveling in the same direction, the pattern variations are periodic. It is thus justifiable to assign a theoretically infinitely large update boundary since the mobile could be located relatively easily without having to incur much searching signaling.

In reality however, it is more frequently observed that while mobiles do seem to demonstrate certain preferences in specific directions, the actual pattern is unlikely to follow such extremes. Consequently, the precision of prediction will vary in accordance to the actual mobile characteristics. The Gauss-Markov model serves well for such purposes and its application seems appropriate for this research where the mobility tracking scheme is designed to be adaptively adjusted to the actual movements.

### 5.3.2 CONCEPT OF KALMAN FILTER

The Kalman filter [51] was first introduced as a computational efficient recursive solution to the discrete-data linear filtering problem. It is a set of mathematical equations that provides estimations to the past, present and future states. Such predictions are achievable even when the precise nature of the modeled system is not known. The purpose of including Kalman filter in this scheme is mainly to predict the next residency or the granularly traceable traveling directions based on the past movement information obtained during update registrations. Thus, given that a mobile's traveling directions can be modeled as a discrete time controlled process governed by

\[ X_k = AX_{k-1} + W_k \]  \hspace{1cm} (5.1)

and the actual measured output to be $Z_k$ defined by
The Kalman filter can estimate the future state (i.e. traveling direction) using a feedback control. Note that in (5.1) and (5.2), parameter $A$ relates the state at the previous time step to the state at the current step and parameter $H$ relates the modeled state to its measured state, respectively. In addition, $W_k$ and $V_k$ represent the corresponding process noise and measurement noise, respectively. Within the particular system framework defined, while the former gives a measure of the expected modeling errors, the latter evaluates the actual system uncertainties experienced.

The goal of the Kalman filter is to find an equation that computes an a posteriori state estimate as a linear combination of a priori estimate and a weighted difference between the actual measurement and a measurement prediction. The complete operation consists of two parts: (1) to obtain the a priori estimate for the next time step and (2) to obtain an a posteriori estimate with the incorporation of the new measurement. In general, the evaluations can be made through the applications of discrete Kalman filter predictor equations and measurement update equations, respectively.

Based on the prior knowledge of state $k$, $\hat{X}_k^{-}$ defines a priori state estimate and is evaluated by the time update equations as

$$\hat{X}_k^{-} = A\hat{X}_{k-1}$$

(5.3)

The corresponding a posteriori estimate can then be obtained based on the measurement update equations. The concept of feedback control is briefly illustrated in Figure 5.3.

![Figure 5.3](image)

**Figure 5.3** Summarized operations for the Kalman filter

It is easily evident that the improved estimation will be a linear combination of the a priori estimation and the product of Kalman gain and the estimation error. Hence,

$$\hat{X}_k = \hat{X}_k^{-} + K_k (Z_k - H\hat{X}_k^{-})$$

(5.4)

where $K_k$ represents the Kalman gain and is evaluated by
\[ K_K = P_K^{-1} H^T \left( HP_K^{-1} H^T + R \right)^{-1} \]  

(5.5)

with \( P_K^{-} \) represents the corresponding \textit{a priori} estimate error covariance and \( R \) being the covariance of \( V_k \). \( P_K^{-} \) is a function of the estimate error \( e_k \), and can be evaluated by

\[ P_K^{-} = E[e_k e_k^T] \quad \text{where} \quad e_k = X_k - \hat{X}_k \]  

(5.6)

Given that the Kalman gain is a function of the noise and the signal covariance, it should be obvious that by dynamically adjusting its value, the effects of estimation errors can be minimized.

One crucial operation of the Kalman filter is to determine the range of possible directional changes in the movements that lead to the next registration. Such prediction is based on the movement patterns, or more specifically, the correlation level, observed in traveling directions from past registration information. To study the mathematical description of such deterministic signals, Fourier Transform provides the means for determining the linkage between the time-domain and frequency-domain representations. Figure 5.4 gives the \( F(\theta) \) equivalents of the Gauss-Markov distribution shown in Figure 5.2(b) in frequency domain.

The correlation of movements is modeled as a function of \( \theta \). Thus, once information with regard to the cutoff frequency becomes available, the sampling rate required to fully reconstruct the signal can be determined based on the Nyquist criterion.

*Figure 5.4* Fourier transforms for the analysis of mobile's movement characteristic (directional consistency level)

To consider the effects of correlation between movement directions in time, consider the previous two cases again where the mobile characteristic is highly predictive. Clearly, when taking the Fourier transform functions of the two movement patterns (i.e. \( \theta \) variations in the Gauss-Markov model), impulse functions will result. For the periodic function, the center point for the impulse function will depend on the frequency of periodicity. For all other general movement patterns where \( \theta \) undergoes a different level of directional changes, it is possible to further classify the patterns into two categories: \( \theta \) changes gradually (high correlation observed) and \( \theta \) changes abrasively (movement patterns approach random).
Observing the gradual changes when the impulse function is represented as a limiting form of the Gaussian pulse, it is clear that the smoother the function in the time-domain, the narrower the corresponding spectrum will result in the frequency-domain [52]. Effectively, it is logical to conclude that when $\theta$ changes gradually in time, the resultant Fourier transform will be constrained within a limited spectrum. As the frequency in the variations of $\theta$ increases, the correlation between adjacent points in time decreases, and results in an increasing cutoff band in the frequency-domain of the signaling representation. When the correlation becomes infinitely small, the Fourier transform becomes unity. In such extreme cases, the mobile is roaming in an absolute random fashion with no correlation in traveling directions between boundary crossings, as shown in (7). It is therefore no longer possible to predict future movements.

$$E[\theta(t_i)\theta(t_j)] = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

(5.7)

Kalman filter operates satisfactorily when there is a reasonable correlation between adjacent values in the time domain. This is because, if a strong correlation is observed from a mobile's past movements, there is a greater probability that the mobile will continue with similar movement patterns in the time that leads to the next registration. It is, however, observed that when a sudden change is incurred in the movement patterns, the accuracy of the prediction degrades. In other words, the limitation of Kalman filter comes in handling huge directional changes. More specifically, the precision of prediction reduces as the possible variation ranges become more diverse. Thus, in order to guarantee a minimum performance gain, the maximum acceptable range of the variations in traveling directions needs to be precisely identified. Let $\Delta \theta$ represent the change in direction of the mobile upon entering and exiting a cell and let $T$ represent the time spent within each cell. Taking the mean value of such measures of all transiting cells over the duration between two updates, the level of correlation $f_\theta$ can be evaluated by:

$$f_\theta = E\left[\frac{\Delta \theta}{T}\right]$$

(5.8)

Clearly, the larger the cutoff frequency, the smaller the correlation between a mobile's traveling directions and thus, the less significant the application of the Kalman filter will be. Thus, in order to have the minimum performance gain guaranteed, it is decided that whenever $f_\theta$ is found to be greater than a given threshold $F_\theta$, the observed correlation will be insufficient to provide reliable predictions of future movement patterns. Therefore, it is a better alternative to assign update boundary according to the conventional distance-based for such mobiles of
interest. The selection process of the specific mode of operation can be briefly summarized by

\[
\begin{align*}
|f_0 - f_i| > F_0 & \quad \text{symmetrical update scheme} \\
|f_0 - f_i| < F_0 & \quad \text{asymmetric update scheme}
\end{align*}
\]  
(5.9)

A distinct value of \( F_0 \) is determined by the network according to the specific characteristics and requirements of the system. It is, however, important to periodically re-evaluate the suitability of the selection in order to guarantee optimal performance gains. Re-adjustments of \( F_0 \) might be necessary to take into account changes in the network behavior, for example variations in the network loading. The actual selection criteria of \( F_0 \) could differ from system to system, from network to network and is beyond the scope of this chapter.

### 5.3.3 UPDATE BOUNDARY ASSIGNMENT

Upon call arrivals, the precise location of the roaming mobile should be determined within constrained delay times. Figure 5.5 illustrates the relations between the occurrence of a call arrival and update events in time.

![Figure 5.5 Inter-relations between the location updates and call arrivals in the time domain](image)

Clearly, the smaller the CMR, the greater the time elapsed since the last update (labeled as \( t_1 \) in the graph) upon a call arrival. Given that the mobile is roaming at a speed known to the network, a greater time elapse leads to an increased uncertainty about mobile’s precise location.

To optimize the applications of the designed location tracking scheme, a low call-mobility-rate mobile characteristic is assumed. In other words, the scenario where an efficient location management technique becomes most desirable is when the mobility rate is much greater than the call arrival rate (i.e. the mean cell residence time is much shorter than the call inter-arrival times). The solution can thus be independent of the actual residence time distributions.

To minimize paging cost, an optimal design would be to define the update boundary such that a call would arrive right after an implicit update is performed. The operational cost would
therefore equate to that of a unit update cost. There are however, certain difficulties that need to be overcome before one could guarantee that such method works. Given that it is desirable to require as little additional intelligence as possible while achieving significant performance gains, the requirement of precision prediction will need to be compressed as far as it is possible.

Once the movement patterns of the immediate future are predicted by the Kalman filter with appropriate numerical values assigned for various directivity indices, a mobile-specific update boundary can be defined.

Ideally, the optimal solution would be to adjust the size of the update boundary according to the predicted level of traveling directivity. Thus, the greater the certainty of the mobile's residency distribution, the less frequent the update registrations are necessary. However, as the update area increases, the number of maximum paging loads also increases. To minimize such dependency, an alternative solution is to fix the number of cells in one registration area, and to vary the dimensions of each (width and length) to reflect the traveling patterns. Specifically, while the size of the registration area is the same for mobiles inheriting similar call to mobility (CMR) characteristics, the actual shapes vary depending on their different mobility patterns. The procedure of assigning the next update boundary upon registration is briefly summarized in Figure 5.6.

![Diagram](image)

**Figure 5.6** Procedures for the assignation of update boundaries

As shown in the graphical presentation, the actual shape and size will depend on the actual movement patterns and the call arrival characteristics respectively for individual mobile users. Clearly, the greater the correlation between transiting directions, the narrower the width of the update area can be due to a higher certainty about mobile's movement patterns. The ultimate goal is to maximize the number of movements in the location area before the mobile exits.
Figure 5.7 illustrates the differences between the assignment of a symmetrical and an asymmetrical update boundary.

With conventional distance-update schemes on the one hand, upon registrations, a symmetrical boundary of threshold $D$ (illustrated by a circular approximation) centered at the last update cell is defined for the roaming mobile. In other words, an update is due whenever a mobile moves more than $D$ cells away from its last update cell. With our proposed new tracking scheme on the other hand, there are two steps involved in assigning the next update boundary upon registrations. Firstly, the Kalman-filter is used to predict the mobile’s future roaming direction (in a granularly traceable manner) based on its previous movements leading to the last update. Subsequently, while keeping the total number of cells constant, an asymmetrical update boundary that best reflects the certainty the network has about the prediction is assigned. Though the asymmetrical update boundary can be of any shape, an optimal rectangular approximation is used in this chapter for a simple illustration of the concept. Thus, for an assumed hexagonal network topology, the area enclosed by the width and length of the rectangular update boundary will contain an equal number of cells to those enclosed by the corresponding symmetrical update boundary; hence, for a threshold of $D$, the number of cells included will be $3D^2 + 5D + 5$ [72]. In general, given that the level of correlation is quantified by $f_0$, the dimensions of the optimal location area is selected through a lookup table.

As a future extension, the size of the update area can also be adjusted according to the observed CMR levels. Specifically, assuming that initially $D_T$ cells are included in each registration area; the number can be increased if no call arrives during the time that leads to the update. Subsequently, further reductions in the costs due to the update can be achieved.
5.3.4 PERFORMANCE TRADEOFF

As a result of the asymmetrical update boundary definition, the mobile is required not only to keep track of the movement patterns, but also to record the set of update centers. Consequently, it is expected that a slight increase in database size at the mobile nodes becomes necessary in order to record the update cell identities. On the one hand, this appears to be the tradeoff for an improved performance when compared to the existing symmetrical update boundaries. On the other hand, given that cells are in reality randomly shaped (as opposed to hexagonal), even with a symmetrical update boundary, the mobile needs to explicitly record the information about individual cell identities. From this perspective, the additional database required will not cause any extra disadvantages than those already presented from the existing solutions.

5.4 SIMULATION FRAMEWORK

In this work, a Gauss-Markov distribution is used to model the traveling directions at each time slot. Thus, with an underlying assumption of a constant traveling speed (i.e. mobility rate=1), the complete operational cost can be determined.

To smooth out the variations caused by the probabilistic nature of the Gauss-Markov distribution model, 3000 movements have been generated to simulate a specific mobility pattern for each specifically defined correlation level.

Given that a unit location update often costs more than a unit paging, the main focus of the chapter is to demonstrate how, with the involved predictions based on the level of directivity, it is possible to reduce the necessary update load by assigning a location area of optimal shape. From an earlier analysis, it is clear that the total paging cost depends on the allowable delay constraints. Thus, for a delay constraint of one, all cells within the update boundary will be paged in one polling cycle upon call arrivals. Given that the number of cells within each update boundary is the same, the resultant paging cost for both update schemes will equate. Consequently, in quantifying the significance of the proposed location tracking scheme, it seems justified to only consider the resultant updating costs for varying threshold values. As a future extension however, a non-unity paging delay should be assigned to demonstrate further improvements made possible by the predictive update scheme. Given that the complete framework can be divided into different subsections, the roaming mobile can be paged sequentially in the order of decreasing probability according to the predictions of residency based on the elapsed time since the last update upon call arrival(s). The implementation of such selective paging scheme is, however, outside the scope of this chapter.
At the end of the iterations, we will have two matrices, one records the number of updates for the conventional distance-update scheme and the other counts the number of registrations for the proposed predictive update scheme. From these, the relation between the actual traveling patterns and the optimal update boundary can be determined. We will address two issues: (1) the inter-dependency between the traveling patterns and the predictive capabilities of Kalman filter, (2) the impacts of the correlation between transitional directions on the selection of optimal threshold boundary. Figure 5.8 illustrates the operation of the simulation.

**Figure 5.8** Steps for the evaluation of the updates number

### 5.5 RESULTS DISCUSSIONS

With the selection of an optimal update area exploiting the directional information, a substantial reduction in the update cost is expected for mobiles inheriting a granularly traceable direction in the movement patterns. Such directivity might simply be due to individual user’s profiles or might have resulted from specific geographical conditions. Discussion on the simulation results is divided into two parts. In the first part, simulation results are included to demonstrate how well the Kalman filter can be used to predict mobile’s future probability residence based on past movement history.

Recall that the precision of the prediction made by Kalman filter depends on the frequency at which registration locations were sampled. Intuitively, the higher the sampling rate (i.e. location updates), the more frequent feedback are made available for the filter to adjust its prediction weight. Consequently, an estimation of greater accuracy can be achieved. In order
to monitor the variations in the degree of prediction caused by the definition of sampling rate, mobility patterns with different level of directivity were simulated and the results compared.

Consider firstly a relatively directive movement pattern where there is a strong correlation between mobile's transverse directions at boundary crossings. Figure 5.9 illustrates three examples of such mobile movements independently generated in a two-dimensional framework.

![Figure 5.9](image)

**Figure 5.9** Scenario $1(f_D = 0.015)$ — a relatively directive movement pattern shown on a two-dimensional framework

Take the middle path for example, Figure 5.10 shows the actual traveling direction at each boundary crossing. The prediction made by Kalman filter at a sampling rate of 15 steps in advance is also plotted in the same figure.

![Figure 5.10](image)

**Figure 5.10** Scenario $1(f_D = 0.015)$ — the actual travelling directions at each boundary crossing (a sampling period of 15 steps)

It is clear that when the movement pattern is relatively directive, the precision of prediction remains appealing even with a low sampling time. Intuitively, a close approximation of the movement pattern ensures that the definition of update boundary
captures the specific mobile characteristics. This, in turn, increases the time the mobile spends in the registered area before its exit. Hence, the argument follows that the longer the time a mobile stays in the update area, the less likely another registration will be due in the time that leads to the next call arrival. The savings in bandwidth consumption are thus achieved when the number of location updates is minimal. With such an observation indicating that significant performance gains can be achieved without much incorporated computation power, the application of Kalman filter is especially recommended for mobiles inheriting a granularly traceable pattern.

Similar conclusions, however, cannot be drawn for movement patterns where the observed measure of directivity is low. For such mobile characteristics, the number of sudden fluctuations in the traveling directions increases, and hence, correlation between adjacent traveling directions decreases. To visualize the differences, for a simulated movement pattern with frequent abrasive directional changes, Figures 5.11 and 5.12 illustrate how the precision of predictions degrades when the sampling frequency varies.

![Figure 5.11](image)

**Figure 5.11** Scenario 2 ($f_0 = 0.04$) — a sampling period of 8 steps

As expected, the simulation confirms that with mobiles movements inheriting frequent distinct directional changes, infrequent feedback is not adequate for precise predictions to be made about future movement patterns. Although the fluctuations are still reasonably tracked by the Kalman filter when the sampling rate is set at an extreme level of one feedback per movement, the approximation then worsens when the sampling rate decreases. In extreme cases where there is absolutely no correlation between adjacent traveling directions in time, the chance of making a reasonable prediction is very low. In such scenarios, the implementation of a conventional distance-based update scheme is a better option for minimizing the operational cost.
Figure 5.12  Scenario 2 ($f_d = 0.04$) — a sampling period of 15 steps

Finally, Figure 5.13 gives a comparison of the incurred update numbers for increasing randomness in the movement patterns.

Figure 5.13  Performance comparisons between the proposed scheme and the conventional distance-based scheme

The purpose is to identify the Kalman filter's capability of predicting future movement patterns and thus assigning an optimal update boundary when the correlation in a mobile's traveling direction reduces. Thus, while the update boundary for the conventional distance-based scheme will have an equal threshold of 10 in all directions from the last updated center, the shape will be dynamically adjusted in the proposed Kalman filter based scheme according to the predictions about future movements. If the sampling frequency is
kept at a fixed rate however, the achievable performance gains decrease when fluctuations in traveling directions increase.

Recall that a lower value of $f_D$ indicates a greater correlation in traveling directions. On the one hand, with the update threshold fixed at 10, the conventional distance-based scheme has worsened performances when the correlation in traveling directions increases. On the other hand, the significance of the proposed Kalman filter intensifies when the correlation in traveling directions, quantified by $f_D$, increases.

Consider the scenario where a high correlation in traveling directions is assumed to have observed from past movement history ($f_D < 0.02$). For the same movement patterns, while the conventional distance-based update scheme would incur a total of 480 updates, the corresponding number of updates incurred by the proposed predictive scheme is significantly smaller. Although it appears that an optimal update boundary will have a rectangular approximation of a dimension 12 x 25, a significant performance gain is achievable with all other asymmetrical update boundaries. It is observed that a maximum gain of 40% reduction in the update load is achievable for this particular movement pattern generated. The saving is remarkable given a unit update cost is often more costly than a unit paging.

Conversely, when the correlation in traveling directions is found to be small (i.e. $f_D = 0.04$), the possible performance gain achievable by the incorporation of the Kalman filter becomes more restricted. In fact, given that the movement pattern is now less correlated, it will be difficult trying to predict the future movement patterns based on the previous feedback information. For this particular movement pattern generated, it is found that an optimal saving of 20% can be obtained when the boundary has a dimension of 8 x 40. Clearly, compared to the previous scenario, an update boundary will generally have a larger width than length in order to show some improvements in performances. Consequently, as the boundary width becomes narrower, the less likely the prediction will approximate the actual residency distribution. As a result, the mobile will cross update boundary more frequently, and thus incur a higher number of updates. For example, it is shown in the simulation that when $f_D = 0.04$, any location area which has a length greater than 12 will incur more update costs than in the conventional distance-based scheme. Clearly, the application of the Kalman filter in such scenario would be less suitable and should be avoided.

Lastly, it should be noted that with the improved performance gain, the tradeoff is the additional prediction power required from the implementation of a Kalman filter; a requirement that translates directly to an increased computation load. It is however important to note that the amount of extra resources vary according to the specific network environment and the desired QoS measures. In scenarios where correlation in traveling directions was shown to be high, a reliable prediction can be made even when movement patterns were projected on samples taken between long time frames. Simulation results clearly indicate
the Kalman filter can project future movement patterns with reasonable precision as long as the past traveling patterns are in a granularly traceable form. In other words, for mobiles that demonstrate a rough direction of traveling in their roaming characteristics, a minimal extra computation load can be easily justified by the significant performance gains. However, when the correlation observed between movement patterns becomes low, the precision of the prediction degrades dramatically unless the sampling frequency is appropriately increased. With the required computation load thus having become significant, the suitability of applying the Kalman filter should then be reconsidered to ensure a satisfactory tradeoff.

5.6 FUTURE WORK

Even with an asymmetrical update boundary in place, it is possible to further reduce the paging load by anticipating the distribution of residence likelihood at the arrival time of call requests. When the elapsed time between call arrival and the last registration is also available, further reductions in paging can be realized through the implementation of a sectional paging. One possible alternative is to send page signals onwards from midway instead of from the last updated cell. However, for such an additional mechanism to be viable, information collected will have to be of reasonable accuracy. It is also advantageous to consider the option of incorporating ideas from predictive call admission control techniques in order to enhance the precision of the predicted residency distributions.

In addition, to further enhance the operations of Kalman filter, it is worth exploring the possibilities of assigning groups of mobiles categorized by mobility characteristics different prediction frequencies. The ultimate goal is to design a location tracking scheme that minimizes the desirable computation load while being sensitive enough to distinguish the mobile characteristics of each roaming user. This in turn, will require the network to be capable of analyzing the general characteristics of mobile users and to assign appropriate groups for optimal selection of the prediction period.

5.7 CONCLUSIONS

In this chapter, we proposed a predictive Kalman filter based update scheme as an alternative to the conventional distance-based update strategy for application in cellular networks. The scheme utilizes additional information made available by Kalman filter about traveling directions, and eliminates unnecessary update costs through appropriate predictions of the mobile’s residence likelihood. With the number of cells to be included in each location area fixed, the dimension of the area reflects the correlation between the mobile’s traveling directions in time, and thus reveals the level of certainty the network has about the mobile’s
movement patterns. As a result, while restricting the maximum paging load, the total number of updates having to be registered can be significantly reduced. Simulation results have shown that the scheme performs best when there is a high correlation between traveling directions at boundary crossings. In general, the greater the directivity, the fewer the number of location updates necessary to maintain the desired QoS.

As the mobile has a greater tendency of traveling in a granularly traceable direction predicted by the Kalman filter, a more user-specific update boundary can be assigned to capture the movement patterns of individual users. Generally, the more certain the prediction is, the narrower the width of the update area needs to be for optimal performances. On the other hand, a symmetrical update boundary is the solution when the mobile patterns show a lack of specific traveling directions. As explained in section 5.3.2, a decision over the application of actual updating scheme can be made based on the predicted level of correlation $f_D$. Simulation results provided not only significant insights into the mobility problem, but also new directions for designing better strategies for managing mobility in future wireless networks.
Chapter 6
Sectional Paging

In this chapter, we propose a new paging technique, sectional paging, which reduces the paging cost while complying with the delay constraint for mobiles roaming with traceable patterns. Without having to install much additional intelligence, the developed scheme predicts the likelihood of residence and assigns the optimal paging boundaries. Thus, while complying with the required delay constraints, QoS measures will not need to be sacrificed as a result of increasing the update threshold. Under the same network conditions and mobile characteristics, simulation results reveal that sectional paging outperforms the ring-structured paging scheme as long as the mobile has a granularly traceable movement pattern. It is concluded that the usage of the newly proposed scheme is most suitable when the roaming pattern is either traceable or can be predicted with reasonable precision. As an additional advantage, it is noted that the designed technique could be used for all basic update schemes.

6.1 INTRODUCTION

From our previous work on the selection of optimal update threshold utilizing information about the directivity index, it is observed that the more "focused" the traveling pattern is in terms of directivity, the higher the optimal threshold can be. In other words, the higher the
directivity, the more precisely the location can be predicted and consequently, the update threshold can be increased to minimize the required registration cost. This is particularly advantageous in reducing the update portion of the operational cost especially when the mobile is traveling fast. However, acknowledging the tradeoff between updating and paging costs evident in most existing solutions [7-11], it is certain that as the update threshold increases, the paging cost will grow in direct proportion.

This observation motivates this research work; primarily, we intend to develop a corresponding paging scheme that would minimize the necessary paging cost without having to sacrifice the merits of allowing for a higher update threshold. Effectively, it will be possible to assign each mobile user an optimal threshold, which, while being large enough to promote savings from registration, is also small enough to avoid having to incur too much paging expense(s).

Section 6.2 begins by highlighting very briefly the developments of other related work. Followed by a concise literature review, the key design philosophy and distinct features that were incorporated in the proposed sectional paging scheme are indicated in Section 6.3. Section 6.4 gives a short description of the theoretical model used to quantify the performance improvements of the simulated paging schemes. Based on the simulation results, discussions are included in Section 6.5 to examine the significance and effectiveness of the proposed paging scheme. The study comprises two parts; to examine the impact various system parameters (e.g. directivity measures, update thresholds etc) have on performances, and to identify the strength of each technique, sectional and ring, in order to explore their most promising areas of application. Section 6.6 concludes our work on the development of the sectional paging scheme by summarizing the main contribution of this chapter.

6.2 RELATED WORKS

Despite the fact that an efficient paging algorithm plays an equally important role to the update algorithm for the satisfactory operation of location management, specific research devoted to paging algorithms is less commonly evident. Such observation is largely due to the increased restriction associated with design flexibility; not only because performance gain is greatly influenced by non-negotiable delay constraints, but a unit paging cost is often less costly than a unit update cost as the bandwidth requirement is a lot smaller. To facilitate the discussions, Figure 6.1 summarizes the inter-relations between different paging algorithms. The classification is made according to where and how polling signaling is sent during the paging operations. Thus, paging schemes are discussed to determine the locations and order at which the paging information is distributed in searching for roaming mobiles upon call arrivals.
The operations of each paging scheme are discussed in more detail in the following sections.

6.2.1 BLANKET POLLING

One of the main parameters used to quantify the performance of location tracking is the delay incurred in locating the roaming user. Depending on the nature of communications, the allowable delay constraints vary for real-time and non real-time connection requests. Take multimedia transmissions for example: the range of acceptable delay is normally smaller, and hence, the mobile will be expected to be located within a shorter timeframe. Upon call arrivals, the whole location area where the mobile last registered is paged in one go. Assuming the mobile has been updated according to the defined technique, it is certain that its location is determined with a minimal delay. Such real time services however come at the expense of large polling signaling load. Given that the complete area is to be paged during a single polling cycle, the paging cost grows proportionally to the size of the defined location area. Consequently, the operation defines the upper bound of the incurring paging load. Thus, in order to achieve optimal performances, the actual boundary of the location area needs to be more carefully selected during the phase of location update. It should be obvious that a higher computational load can be expected due to the increasing precision found necessary. Nevertheless, blank polling is the technique that has been standardized for the GSM and GPRS systems. The algorithm results in the least delay and has the advantage of having the simplest implementation [38-39,56].
6.2.2 SEQUENTIAL PAGING

The first possible savings in paging come when multiple polling cycles are allowed during the paging operations. Once the delay constraint is relaxed, the complete location area can be partitioned into smaller areas, with polling signals sending sequentially to each of these areas. Through the assignation of a paging area, the boundary within which polling signals should be sent is defined. In general, sequential paging allows the size of the paging area during each polling cycle to be reduced, and hence, promotes possible bandwidth reservation. Clearly, the bigger the area, the higher the paging load will be during each polling cycle. In addition, it is possible to restrict paging load further by appropriately assigning the order in which to send the polling information. Specifically, as long as the roaming mobile is located before the last set of the polling cycle is used, some savings in the resource utilization can be achieved. The reduced cost however is likely to result in a longer delay before the mobile's location is determined. Therefore, when delay constraint is set to unity, the evaluation represents an upper bound on the average paging cost. Conversely, when delay constraint is set to infinity, the estimation gives a measure to the lower bound of the paging cost.

In this section, several sequential paging techniques that are available in the literature are described in more details. Depending on the actual movement patterns for individual mobiles, it is clear that there are certain directions where the probability of locating the roaming mobile is higher than in others. In fact, the better the prediction, the more savings can be achieved from the operation of sequential paging [39].

6.2.2.1 SHORTEST DISTANCE FIRST PAGING

Shortest distance first paging has the simplest operational concept among all available alternatives. The paging starts from the cell where the last registration is performed, and progresses sequentially to cells of increasing separation from the last updated center. In a one-dimensional network topology, the polling signals are sent in linear fashion symmetrically from the last updated cell [58]. In a two-dimensional network topology, the term ring-structured paging is more commonly used to symbolize the sequence at which cells are paged. The scheme incorporates virtually no predictions, it simply assumes that the mobile has the greatest possibility of locating at the last update center, and the certainty reduces in cells that are further away from the center [66]. For a distance-based or a movement-based update scheme, the maximum delay that can possibly incur will be equal to the size of the threshold. In scenarios where unlimited paging delay is not allowed, rings of cells will be grouped to comply with the paging constraint. Variations for the actual grouping should be appropriately assigned to optimize the performance gains.
6.2.2.2 TIME-ELAPSED PAGING

It is also possible to consider the elapsed time since the last registration upon call arrivals as in the time-elapsed paging scheme. Assuming the mobile is traversing at a constant speed, it is expected that the longer the elapsed time, the further away the mobile will be from its last updated cell. Thus, the sequence of paging is adjusted to reflect the anticipation. Although the underlying assumption is general and has only little modification from the basic shortest-distance-paging scheme, performance gain is possible depending on the actual mobility patterns as demonstrated by [10].

6.2.2.3 VELOCITY PAGING

Velocity paging attempts to reduce the paging signaling by grouping users into different velocity classes according to the traveling speed registered at the time of update. Upon call arrivals, paging areas are generated from information based on individual's velocity characteristics and the elapsed time since last registration [11].

6.2.2.4 HIGHEST PROBABILITY FIRST PAGING

Given that the time-varying probability distribution of user location is known, cells within the registration area are sequentially paged in the order of decreasing residence likelihood [41,55,71]. Upon call arrivals, the precise location of the roaming mobile should be determined within constrained delay times. Even with an asymmetrical update boundary in place, it is possible to further reduce the paging load by anticipating the distribution of residence likelihood at the arrival time of call requests. Generally, the mean of the distribution defines the most likely location of the roaming mobile, the probability of residence then decreases symmetrically in both sides away from the mean [41,55]. It is thus justified to divide the complete update area into smaller paging sections, and to page each section in a sequential order of decreasing probability. However, for such sequential paging to be implemented, the mobile needs not only to evaluate the cost paging at the point, but also to acquire information about the conditional distribution of future locations.

6.2.3 SELECTIVE PAGING

The difference between sequential paging and selective paging is subtle, and thus, it is difficult to draw a clear distinction between the two operations. As was discussed earlier,
instead of restricting paging to be completed in one polling cycle, sequential paging allows only a partial area of reduced size to be searched during a single polling cycle. Depending on the actual formation of the partitions, it is likely that the roaming mobile is located before the complete location area is paged. The decreased signal requirement indicates a reduced operational cost. The saving, however, comes at the expense of increased paging delays. For certain mobiles where the corresponding movement patterns are highly predictive, the granular traceable areas of residency can be anticipated with relatively high accuracy. It is, therefore, not necessary to page the complete location area. Thus, instead of sending polling signals to every cell within the registered location area, only parts of the complete location area require paging. Hence the name selective, where partial areas are carefully selected to reflect the predicted information about residency probability. In some ways, it indicates a greater certainty about the prediction made in generating the dynamically adjustable paging areas [65,71).

Apart from those more general paging schemes that can be incorporated on top of any update schemes, there are some other more system-specific paging algorithms proposed in the literature. In [65], paging operations are designed based on the underlying assumption of a known update scheme and the corresponding shortest distance model. Thus, given the call arrival rate, the probability of having the mobile reside in any cell is calculated and the evaluation ranked. With the number of partitions fixed by the allowable paging delay, given the size of a location area, the number of cells to be included in each partition can be appropriately determined. The complete location area is thus comprised of smaller groups formed by associating cells in decreasing order of residence probability. Although simulation results have shown that the performance gain over conventional GSM standard is more significant than that achieved by other paging alternatives, the applications of such system-specific algorithms are more restricted.

6.3 SECTIONAL PAGING SCHEME

The sectional paging scheme proposed in this chapter eliminates consideration of speed and focuses only on direction. The scheme functions similarly to [50] in the sense that directivity aspects of the movement patterns in two dimensions are taken into consideration. However, we allow at the same time a certain degree of “freedom”; specifically, directivity being in the sense of a granular tracking rather than a precise directional information maintenance. Based on the underlying assumption that a mobile will continue to travel in the same “granular” direction as it traveled prior to an update registration, information regarding the mobile's previous movement patterns were used to predict its future residency distributions. Subsequently, decisions can be made on the user-specific definition of the paging areas along
with the paging order when a non-unity delay constraint is allowed. Therefore, the main difference between sectional paging and ring-structured paging comes from the way subsections are formed when the allowable paging delay is greater than unity. The ring-structured paging scheme groups cells according to their separation from the last updated cell with subsections paged in a sequential manner. Sectional paging groups cells with "angular" divisions in reference to the anticipated probability distributions. It aims to utilize the additional information made available about traveling directions, and to eliminate unnecessary paging costs via some appropriate predictions of the mobile's residence likelihood. The key issue is to minimize the total number of cells requiring paging before the location of a roaming mobile is found while keeping the number of polling cycles constant. Thus, provided that the possible section of residence can be predicted with reasonable accuracy, the proposed scheme will ensure that the paging cost incurred is always of a minimum possible value. Figure 6.2 briefly illustrates the differences in paging area formation between the ring-structured and sectional paging schemes.

![Cell division method for (a) ring-structured paging and (b) sectional paging](image)

**Figure 6.2** Cell division method for (a) ring-structured paging and (b) sectional paging

The purpose of this chapter is two fold: (1) to explore the capabilities in terms of reliable performances of the sectional paging scheme and (2) to examine the impact that variations in traveling patterns have on the overall paging cost. Thus, instead of restricting the analysis to be based on a specific mobility model, this chapter focuses on the methods of "detecting" directivity patterns between movements. We intend to demonstrate firstly how a measurement can be obtained from the traveling characteristics; and secondly, how feedback values can be used to decide the size of paging areas upon call arrivals. It is anticipated that a minimal intelligence will be sufficient to show some significant improvements in the operations.
6.3.1 DESIGN DETAILS

Upon entering a new cell, there are 6 directions that the mobile could possibly propagate into. Given that mobile communications are now used so universally, it seems non-realistic to assume that the probability to all sides will be the same. Instead, it is justifiable to assume some degree of correlation between the traveling directions to adjacent registration updates in time. In this chapter, we propose that the degree of "randomness" in the movement patterns could be modeled by defining two parameters; $P_f$ being the probability of moving forward and $P_d$ being the probability of continuing in the "dominant" direction. Figure 6.3 gives a graphical presentation to such parameter definitions.

![Diagram showing directions and probabilities](image)

**Figure 6.3** The model of directivity quantification

Obviously, the probability distribution of a mobile's location changes according to the level of directivity observed in the traveling patterns. Based on trigonometry concepts, it is possible to mathematically model the distribution of residence likelihood in direction $\theta$ after $m$ boundary crossings. For example, assuming the mobile travels in the directions of $z$ and $y$ with probabilities of $P_z$ and $P_y$ respectively, the resultant residency distribution in the area bounded by the two axes is thus

$$P(\theta) = P_z \left( \frac{z(\theta)}{m} \right) \times P_y \left( \frac{y(\theta)}{m} \right)$$  \hspace{1cm} (6.1)

with

$$z(\theta) = \frac{m}{(1 + \tan \theta)} \quad \text{and} \quad y(\theta) = \frac{m \tan \theta}{(1 + \tan \theta)} \hspace{1cm} (6.2)$$

To illustrate how varying values of $P_d$ and $P_f$ affect the prediction of the probability distribution, Figure 6.4 illustrates the variations observed for movement patterns with different directivity measures. Generally, the higher the value of $P_d$, the more obvious the granular traceable traveling direction.
The anticipated probability distribution functions for
(a) $P_d = 0.33$, $P_f = 0.5$  
(b) $P_d = 0.6$, $P_f = 0.8$  
(c) $P_d = 0.8$, $P_f = 0.8$

On one hand, when a random movement is inherited (i.e., case A), the mobile could reside anywhere in the circular area (of radius $D-1$) centered on the last updated cell labeled $(0, 0)$. In this case, it is obvious that the most economical strategy would be to page the roaming mobile symmetrically in all directions and hence, having the cost of paging varied depending on the update threshold $D$. On the other hand, if the mobile has a greater probability of going forward than backward as is illustrated by cases B and C, the residential function of probability distribution will not be symmetrical. Clearly, within the area bounded by the next reporting cells, there are certain directions where the mobile will have a greater probability of residing than others. Consequently, it becomes unnecessary to page the whole circular area for the roaming mobile. In such scenarios, the application of a selective paging scheme would ensure that the required polling load is strictly controlled at a minimum.

The complete operation of the proposed paging scheme is comprised of two parts. Firstly, the mobile is required to statistically analyze the pattern of movements upon registration and secondly, the network is given the responsibility to appropriately define the size of paging areas upon call arrivals. Figure 6.5 gives a concise illustration of the design details.

**Upon registrations**

- Analyze motion patterns
- Estimate directivity
  1. Identify possible correlation in time
  2. Compare with the previously cached values to obtain an overall average

**Prediction**

- Estimate the probability of residence in the circular area of radius $D-1$ centered at last updated cell
- Identify the section of paging
- Dynamically adjust the size of the paging area

Figure 6.5  An overview of the sectional paging scheme
Briefly, once it is detected that an update has become necessary, the mobile will give an estimation of the granular direction of traveling or trend of moving. Statistically, this is an averaged quantity of all previous movements leading to the registration. To more explicitly quantify the correlation in movement patterns between chronologically adjacent registration updates, this measure will then be compared to a previously cached value, statistically analyzed for past updates, to give a better prediction of the traveling patterns. Effectively, a matching definition of $P_d$ and $P_f$ will be projected to give an alternative presentation of the traveling trends.

In the second stage, the network performs a one-off evaluation mapping the motion characteristics to the width of the sectional area according to various geographical conditions and specific user behaviors. Thus, for each evaluated value of $P_d$ and $P_f$, a section boundary is “selected” to indicate the most probable residency distribution for the mobile in search. Upon call arrivals, where the location information is not already available, a search is carried out concurrently in all cells of the selected area. Generally, the greater the precision of the prediction, the fewer the number of cells required for paging and thus the better the system performance from the perspective of cost minimization.

It has been decided that one of the main design criteria of the sectional paging scheme is to minimize the required computations at the network while having performance improvement guaranteed. Given that only a simple look-up procedure is required at the network during call arrivals, the proposed strategy will not only ensure the additional system intelligence is minimized, but also allow large fluctuations in mobile population to be accommodated.

[8] has successfully demonstrated that even a slight relaxation of the paging delay from unity to two would be sufficient to give some significant improvements to the paging performance in terms of cost minimization. Consequently, it is justified to consider only the simplest case where the complete framework is partitioned into two sections in quantifying the improvements achievable by the sectional paging scheme. Figure 6.6 gives an illustration of the boundary selection concept for the sectional paging scheme.

Briefly, the bounded area will define the number of cells to be paged during the first polling cycle with the rest of the cells in the plane to be paged in the second polling cycle. The simulation is intended to show how, by adjusting the size of paging areas according to the online measures of the directivity information, the number of cells having to be included in the first polling cycle is minimized. Theoretically, the more directive the traveling pattern is, the higher precision the residency prediction will have, and hence, the performance will be optimized.
Finally, in order to increase the practicality of the developed paging scheme, note that there is no restriction on the specific mobility model to be selected for simulation. Therefore, various existing models (e.g. random walk mobile, velocity fluid-flow mobility model, Brownian model and Markov-Gauss mobility model) can be applied freely to generate the desired movement patterns.

6.4 ANALYTICAL MODEL

In evaluating the paging cost, a similar analysis framework to that of an ordinary distance-based scheme [8-9] is found applicable. Generally, given a per unit paging cost of \( V \), the calculation comprises two parts—to estimate the probability of having the roaming mobile reside in subarea \( j \) (i.e. \( A_j \)) and to predict the number of cells (\( \varphi \)) to be paged before the user is located. Specifically, the evaluation for such a cost can be defined by:

\[
C_v = V \sum_{j=1}^{n} \rho_j \varphi_j
\]

where \( \eta \) is a measure of the tolerable paging delay and it defines the total number of polling cycles permitted during the paging operation. Conceptually, this simulates a hierarchical structure where multiple cell-rings are managed as a big subarea. Assuming paging delay is constrained, the probability of having the mobile reside in subnet \( \rho_j \) is defined by

\[
\rho_j = \sum_{i \in A_j} \left( \sum_{m=0}^{\infty} \alpha(m) \gamma(i,m) \right)
\]

where \( \gamma(i,m) \) is a measure of the probability of mobile residing in subarea \( i \) after \( m \) movement. It should be obvious by this point that the derivation of \( \rho \) and \( \varphi \) will be similar but different for the two types of paging strategies under consideration. For example, in the case of
ring-structured paging, the quantity of \( \varphi \) will depend only on the actual ring \( i \) that the mobile resides at the time of call arrival.

\[
\varphi_j = \sum_{k=0}^{i} \left( \sum_{i \in \mathcal{A}_j} 6i \right)
\]

(6.5)

On the other hand, for the sectional paging developed in this chapter, the probability of location is considered according to the residence in each angularly-divided section. Thus, the total number of cells requiring paging will be a measure of certainty in predictions. Such a quantity can be evaluated by

\[
\varphi_j = \sum_{k=0}^{\infty} \left( D + 2\sum_{n=1}^{\infty} D - n \right)
\]

(6.6)

where \( n \) is the defined boundary size used to predict the most likelihood of residency with \( x \) and \( D \) respectively representing the maximum boundary size and the defined threshold. Clearly, the specific boundary size selected for the sectional area will have a significant impact on the overall performance assessment.

Generally, as some degree of prediction is introduced in the sectional paging to select the size of the paging boundary, for the same number of cell transitions during call arrival, it is expected that less paging costs will be incurred.

### 6.5 RESULTS DISCUSSION

With the selection of an optimal paging area exploiting the directional information, a substantial reduction in the paging cost is to be expected for mobiles inheriting a granularly traceable direction in the movement patterns. Such “directivity” might simply be due to individual user's behaviors or might have resulted from specific geographical conditions. Thus, provided that the mobile has significantly moved away from its last updated cell during call arrival, sectional paging would give a more satisfactory performance than ring-structured paging under the same delay constraints. However, when the mobile appears to have inherited a random movement pattern in its motion (i.e. an equal transitional probability to all six sides), the implementation of a ring-structured paging would be a safer option in minimizing the paging cost. Figure 6.7 gives a comparison of paging costs incurred for the sectional paging scheme with varying bounded size definitions.

As was pointed out in the previous section, the more certain the prediction is, the smaller the bounded size of the sectional area needs to be. Thus, minimal paging signaling is required to locate a roaming mobile. As a benchmark, the performance of a ring-structured paging scheme is also included to quantify the actual improvements of the proposed scheme.
As expected, the application of ring-structured paging is advantageous when random movement patterns ($P_d = 0.2$) are observed for a roaming mobile. Clearly, when the mobile’s location is unpredictable, the uni-directional paging mechanism would give the most effective alternative in combating such uncertainties in residency predictions. However, as the tendency of moving in one dominant granular direction becomes clear (i.e. $P_d$ increases), sectional paging started to show significant improvements. It is evident that with the application of sectional paging, a greater saving in the total paging cost is obtainable when the directivity measures of the roaming mobile increases. The crossover points in the figure represent the minimum requirement of the directivity measures before the application of sectional paging is justified.

In the worst case where the mobile resides in an area outside the prediction, it will then be located during the second polling cycle when the paging signals are sent to the rest of the cells. Generally, as can be readily evident in the figure, the greater the bounded size, the more significant is the improvement over the ring-structured paging scheme. For example, when the bounded size is equal to 5, the application of sectional paging has a superior performance even for a directivity level $P_d$ as low as 0.38. For an update threshold of 8 as was simulated in this particular case, the maximum bounded size of the sectional area will be 7. Thus, the crossing point will decrease further as the size of the sectional area (in terms of width) increases. To give another illustration of the possible improvements achievable by sectional paging, Figure 6.8 shows the simulation results when the definition of update threshold varies. In this example, the directivity measure is fixed at 0.65 with a bounded size of 3.
Clearly, for a fixed section width, paging cost reduces with a decreasing update threshold. Thus, as a general design rule, the higher the update threshold, the wider the size of the sectional paging area should be in order to ensure a reliable prediction of the mobile’s residency probability distribution.

![Graph showing performance comparisons for varying update thresholds]

**Figure 6.8** Performance Comparisons for varying update thresholds

### 6.6 CONCLUSIONS

In this chapter, a sectional paging scheme was proposed as an alternative to ring-structured paging for application in cellular networks. The scheme utilizes additional information made available about traveling directions, and eliminates unnecessary paging costs via some appropriate predictions of the mobile’s residence likelihood. As a result, while complying with the required paging delay constraints, the total number of cells having to be paged can be significantly reduced. Simulation results reveal that the scheme will perform best when there is a high tendency for the mobile to be traveling in a granularly traceable direction. Generally, the more certain the prediction is, the smaller the bounded size of the sectional area needs to be and thus incurring fewer paging load. On the other hand, ring-structured paging would be the solution to go with when the mobile patterns seem to have shown a lack of specific traveling directions. Effectively, switching between the two paging strategies is determined via the projected values of individual mobile’s traveling trends, $P_d$ and $P_b$, as detailed in the previous sections.

Further reductions in paging can be realized given the traveling speed of a mobile is also available. One of the possible alternatives is to send page signals onwards from midway instead of from the last updated cell. However, for such additional mechanism to be viable, collected information should be of reasonable accuracy. It could also be advantageous to
consider the option of incorporating ideas from predictive call admission control techniques in order to enhance the precision of the predicted residency distributions. All in all, the simulation results were satisfactory. They have not only given some significant insight into the mobility problem, but also provided some new directions for designing better strategies for paging operations.

SMILE! YOUR HAVING A GREAT DAY
A man of destiny knows that beyond this hill
Lies another and another.
The journey is never complete.

F. W. De Klerk (1936-)

Chapter 7
Mobility Management in IP Networks

As opposed to mobility between heterogeneous networks, mobility in a homogeneous network is concentrated at a subnet level. Mobility management techniques designed to deal with this aspect mainly handle mobile movements between different base stations within a subnet [75]. With mobile movements confined within a single radio access network, a finer scale of movement patterns results. In the first part of the thesis, an emphasis was on the development of location tracking schemes for applications to cellular networks. All the strategies proposed in the previous chapters were thus for managing mobility in homogenous networks. In this second part of the thesis, starting from this chapter, the design focus is on mobility management for heterogeneous networks.

7.1 MOBILITY BETWEEN HETEROGENEOUS NETWORKS

Currently, there are many radio access networks (RANs) which coexist. Whether it is geographical coverage, bandwidth or delay, each access technology has its own distinct characteristics and is optimized for a specific set of applications [76]. It would then be up to
the users to decide when and how to switch from one access network to another depending upon availability and appropriate cost/performance considerations. Consequently, it will be necessary to implement a preferably universal solution that prompts transparent user roaming among different wireless networks while delivering the widest possible range of cost effective services anywhere/anytime. This results in a vertical handoff scenario where networks are assumed to overlap. Therefore, the designed mobility management technique should be independent of the network topology. Note that when a handoff is between BSs/APs (i.e. base stations and access points) that belong to the same RAN, the term horizontal handoff is used to describe the transition, otherwise, vertical handoff is used to describe the occurrence of roaming between heterogeneous networks; the change could be either in the technology or in the administrator [77]. Figure 7.1 illustrates the co-existence of various technologies.

![Diagram of vertical handoff]

**Figure 7.1** Inter-technology mobility scenarios

Generally, depending on the level of network stack from which a movement is considered, “mobility” can be classified into three categories, namely air-interface mobility, link-level mobility and network-level mobility [78], which then constitute the three stacks forming a subnet.

Air-interface mobility is perhaps the most common case where a “handoff” takes place between two adjacent BSs (or APs) within a radio access network. One can envisage this scenario as a pedestrian walking across micro-cell boundaries while being engaged in a “conversation” whether through voice or data transmissions. Link-level mobility goes one level up in the network hierarchy, and is concerned with maintaining a point-to-point protocol (PPP) context across multiple radio access networks. The transitions however would still be within the same domain and technology. On the highest level of the three categories, IP provides network level mobility between different access networks (including wireless).
Basically, this involves a change in a mobile’s subnet-related IP address due to either (1) a change in radio access technologies (e.g. from a CDMA2000 to indoor wireless LAN) or (2) a transition between two network operators. Note that in the latter case, the two involved networks might be implemented by the same access technology even though they are under the management of different operators. Generally speaking, as long as movements occur between points of attachment on the same IP subnet, mobility management at the link-layer may offer faster convergence and far less overhead than mobility management at network level [79]. As a simple illustration, Figure 7.2 considers a scenario where a MN moves from a wireless RAN to a GPRS RAN. Intuitively, such a change of subnet technology needs to be registered at network level or more specifically at the IP backbone. The goal is to have a general network layer solution to handle different characteristics of various systems.

The research focus is thus shifted to the scenarios where movements occur between points of attachments between different IP subnets. Depending on whether the subnets belong to the same administrative domains, two further terms, macro- and micro-mobility scenarios, are used to describe the movements [76-77, 80-83].

In the second part of the thesis, starting from this chapter, the research will be focused on proposing an extension to the operations of Mobile IP such that movements between heterogeneous networks are more efficiently managed. Section 7.2 defines the operations between macro-(the intra-domain) and micro-(the inter-domain) mobility scenarios. To better understand the rationale behind the developments of existing micro-mobility proposals, section 7.3 summarizes the main concepts that have been frequently used in designing location tracking strategies for these mobility scenarios. With the ultimate goal of developing a more efficient alternative, section 7.4 explains why it is worth investigating the possibility of applying cellular-based operations to the IP challenges.
Figure 7.2  A possible scenario of network overlapping
7.2 MACRO-MOBILITY VERSUS MICRO-MOBILITY

Macro-mobility can be looked at as *inter-domain mobility*. Basically, it is about how to maintain the established communication for a mobile roaming between two distinct network domains. Global mobility has often been considered as a special case of macro-mobility scenarios. It is true that global mobility often involves longer time scales, however, the need for maintaining an established connection is still desirable. Moreover, it was pointed out in [75] that global mobility is more of an issue for re-establishing the connection rather than for providing a continuous connectivity.

Micro-mobility on the other hand emphasized *intra-domain mobility*. More specifically, the task is to maintain an ongoing connection while having a mobile roam between subnets within a domain. Generally, movements between subnets within a domain do occur more frequently than those between domains, and thus, it seems appropriate to separate the operations between the two [84]. Figure 7.3 gives a brief illustration of the macro- and micro-mobility scenarios.

![Figure 7.3](image)

*A brief illustration of the relations between different levels of mobility*

In this diagram, the thick arrow indicates a macro-mobility scenario while thinner arrows are used to illustrate some micro-mobility scenarios of a mobile node. The dotted lines used to connect various FAs to MA (used in TeleMIP or GFA used in regional registration schemes) show the inclusion of another FA hierarchy for the purpose of more efficiently handling mobile movements between subnets. Thus, all the registrations for movements between subnets will be sent to a local manager rather than to HA which is possibly situated
in a distant domain. Note that subnets within domain 2 are implemented by different access technologies. Hence, it should be clear by this point that the provisioning of mobility between heterogeneous networks is independent of the actual radio technology.

7.3 MICRO-MOBILITY SOLUTIONS

Having identified the major deficiencies in the Mobile IP protocol, the question now is whether there are good alternatives currently available for better protocol performances. New proposals that are specifically designed to enhance the Mobile IP operations continue to emerge. In this section, only a selection of such algorithms is included for discussion.

7.3.1 AN OVERVIEW

In proposing algorithms for better micro-mobility management, the design focus is on the provisioning of faster handoffs while achieving optimal bandwidth utilization. Despite the evolution of a large number of new techniques in the literature, all developments are oriented on a common set of design concepts.

The aim of this section is to provide an overview of such commonly observed operations and thus, to explore the possibilities of improving the performance further through appropriate modifications and extensions. Figure 7.4 summarizes the fundamental ideas of several popular techniques. Note that the comparison has been purposely kept at a general level. With such an overview becoming available, the commonalities between individual schemes can be easily seen.

![Diagram of micro-mobility strategies]

**Figure 7.4** A list of design concepts frequently used in designing location management techniques
As was previously defined, micro-mobility managements provide tracking solutions to movements between two access points within the same administrative domain — possibly inter-technology movement depending on the type of RAN in operation.

In this respect, the idea of proxy agent architecture is held in high regard. The main concept is to somehow group RANs such that inter-subnet movements within such divisions are also regarded as a semi-local movement. Effectively, this forms a hierarchical structure of some sort.

The other popular approach is to create temporary HAs. In this case, while recognizing the need of updating the subscribed HA, the actual path of signaling is altered to minimize the possible long delay that would otherwise arise. Variations occur in as to where to place the HAs such that maximum operational efficiency (in terms of security and maintenance cost) can be obtained. Currently, the emphasis is on developing algorithms for dynamic home address assignment. Given research studies in this area appear to be more uni-directional, no further consideration of this aspect will be included in the presentation of this thesis. The following sections will have more elaborated discussions on the concept of proxy agent architecture.

7.3.2 PROXY AGENT ARCHITECTURE

In the realm of proxy agent architecture, further categorizations are possible due to variations in the actual interpretation of the concept. Among them, the three most distinct proposals are hierarchical management [85], regional management [86], and the vision of mobile networks [87]. Despite frequent usage of these terms in the literature, it seems that the understanding of the protocols has been somewhat arbitrarily judged from the frequent redefinitions evident within the IETF research community. Appreciating the importance of having the operations of each protocol clearly distinguished, Figure 7.5 attempts to summarize the applications of each idea. This illustration helps to clarify the fact that the concepts are not as similar after all.

Furthermore, it might appear that the inclusion of mobile networks is somewhat inappropriate as it represents a different scenario in which the whole network moves. However, from the viewpoint of the roaming mobile itself, since an intra-domain movement is involved, this is considered as a classic example of the micro-mobility scenarios. Figure 7.6 provides a brief illustration of the main concepts under consideration.
Figure 7.5 An illustration of the correlation between different location management techniques.

Figure 7.6 A graphical presentation of hierarchical management and mobile network concepts.
7.3.3 MORE EXAMPLES

The strategies, Cellular IP, HAWAII and TeleMIP, not only represent pioneer work in this research area, but their initial design concepts are still found to be significant even at this date [75, 88-92]. In the following paragraphs, the feasibility of each of these protocols is briefly discussed in addition to a concise summary of their respective operations. Improvements that have overcome the inefficiencies identified in Mobile IP are also examined.

7.3.3.1 CELLULAR IP

Cellular IP is specifically designed for highly migrating mobiles during active data transfer where a minimum disturbance to ongoing data sessions becomes essential. To achieve optimal system operations, Cellular IP maintains two parallel mappings, Paging Cache (PC) and Routing Cache (RC) to separate paging and routing process. PC, maintained only in selected nodes, is used for all MHs that could be either in the idle mode or in the active mode. Even though this only coarsely tracks the movement of a mobile, the stored information is useful for searching the roaming mobile. Routing cache, on the other hand, maintains an updated location database for active hosts. When a timeout interval approximates packet time scale, the stored information is capable of routing high bit rate data [88,89].

Once a packet of a roaming host arrives at the gateway router (GW) in the Cellular IP network, it is routed according to the routing information maintained in RC. In cases where only PCs are available for the destination mobile, arriving packets are queued in GW while a paging packet, with zero payload, is sent to locate the roaming mobile as shown in Figure 7.7.

![Cellular IP Operations](image)
The default route from each BS to GW can be determined through various shortest path algorithms. Those include Bellman-Ford and Dijkstra’s algorithm, just to name a few. For an active mobile, RCs are maintained at all the nodes along the default route. For an idle mobile on the other hand, in order to save unnecessary resources used for storing “precise” location information, PCs are updated only at selected nodes. To ensure sufficient nodes are used to keep at least a coarse track of the roaming mobile, nodes that have “branched” out-ports and are located at remote areas are often included in the selection (e.g. Nodes N1 and N2 in Figure 7.7).

Cellular IP has a flexible handoff support. There is no need for advance information for each newly arriving MH to create or maintain PCs and RCs. Similarly, for every departing MH, no prejudice notifications are required. In short, this is simply a “plug-and-play” solution.

Having said that, the selection of PC timeouts, RC timeouts, and also the repetition rates of paging and routing updates are crucial not only to minimize unessential waste of resources, but also to ensure a seamless handoff as illustrated in Figure 7.8.

<table>
<thead>
<tr>
<th>Number of nodes maintaining paging cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less</td>
</tr>
<tr>
<td>More</td>
</tr>
<tr>
<td>More paging load (suitable for long continuous frequent data transmission)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Paging cache timeout</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shorter</td>
</tr>
<tr>
<td>Longer</td>
</tr>
<tr>
<td>Few PU’s packets' time</td>
</tr>
<tr>
<td>More paging load (more frequent PU)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Routing cache timeout</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shorter</td>
</tr>
<tr>
<td>Longer</td>
</tr>
<tr>
<td>Lower than RU repetition rate</td>
</tr>
<tr>
<td>More RU packets</td>
</tr>
</tbody>
</table>

Figure 7.8 Parameter selection criteria in Cellular IP

There are quite a few advantages associated with the Cellular IP protocol. Among them, simplicity in implementation with cheap BSs infrastructure, seems to be particularly attractive. In addition, the superb network scalability (which allows Cellular IP to be deployed in distinct
Chapter 7 Mobility Management in IP Networks

environments), gives network operators the extra capability to dynamically extend network sizes according to the growth in demand.

By and large, under the assumptions that MHs move frequently, Cellular IP does seem to give a better alternative than Mobile IP to accommodate local mobility. However, as MHs become less mobile, systems performance may be marginal due to the need to send periodic PU packets.

Furthermore, there are other obvious deficiencies with Cellular IP. As all routing and paging information are cached with respect to the default route between the mobile node and the gateway router, packets from a roaming mobile are always sent to the GW no matter what the actual destination address is. This operation becomes rather inefficient when two communicating parties are actually mobiles located within the same Cellular IP network. That is, in spite of the fact that two mobiles are physically close to each other, the transmitting packet between them still needs to be routed to the GW (located several hops away), before it "turns back" and is routed as an "ordinary" downlink packet. Figure 7.9 illustrates such a scenario.

![Figure 7.9](image)

Figure 7.9 Routing path for mobiles located in the same network

Moreover, as the only information that needs to be included in various control packets (mainly PUs and RUs), is simply the mobile's IP address, various authentication processes will have to be realized to ensure updates are flawless before the protocol can be practically implemented.

7.3.3.2 HAWAII

Another similar alternative, but with a further reduction in paging load, is seen in HAWAII (Handoff-Aware Wireless Access Internet Infrastructure). As before, paging updates are
sent to the domain root router (DRR), except this time, at a much reduced frequency. This is achieved by further dividing the location area into smaller routing areas. Thus, instead of having to send PU upon moving to a new BS (~1min), this requirement is relinquished in HAWAII only in movements across routing areas’ boundaries (~30mins) [91].

It is, however, interesting to note that in HAWAII, an increase in paging update time does not necessarily mean a reduction in signaling messages. In this case, although a paging update is sent only when shifting between routing areas, normal Mobile IP messages are still required from the mobile during each handoff. This, in comparison to the zero-payload control packets used in Cellular IP, and to frequent Mobile IP messaging packets to some extent, still wastes radio resources (MN-BS) that are already scarce.

On the other hand, by splitting the registration process into two parts (from mobile to base station, and from base station to mobile), the existence of the HAWAII network remains “invisible” to the roaming mobile. This is particularly significant from the mobile’s point of view, for there is no differentiation in a mobile’s operations no matter what network, or what sort of mobility is undertaken. In short, no additional “detecting” intelligence is required to be built into mobile terminals [92].

In addition to reducing updates frequency to HA, and minimizing disruption during handoff, HAWAII has also given better security and authentication measures for the mobile’s intra-domain movement. Figure 7.10 illustrates handoff operations of the HAWAII protocol.

![Default route from BS2 to DRR](image)

**Figure 7.10** Handoff management in HAWAII
Moreover, HAWAII divides the access network into domains; a mobile’s mobility movement varies depending on Home/Foreign Domain roaming. Figure 7.11 illustrates a scenario where the mobile roams within the home domain. In this scenario, packets reaching DRR are delivered to the destination mobile via the previously established paths. Without the involvement of HA in the procedure of call delivery, a more efficient routing is implemented in addition to an enhanced reliability.

Figure 7.11  Movement within home domain (with no HA involvement)

Conversely, when a mobile roams into a foreign domain as shown in Figure 7.12, a collocated CoA (CCoA) is assigned by the foreign agent. Subsequently, packets are intercepted and tunneled by HA in the home domain. However, the retained CoA remains unchanged for the time the mobile moves within the same foreign domain.

Figure 7.12  Movement within a foreign domain
In fact, as the protocol is built on top of IP, HAWAII is more like "the improved cellular version" of the Mobile IP solution. As summarized in Table 7.1, there is clearly a strong similarity of control signals between Mobile IP and HAWAII; although "handoff" refers to changes in BSs in HAWAII as opposed to changes of links in Mobile IP.

**Table 7.1** Exchange of control messages

<table>
<thead>
<tr>
<th>Functions</th>
<th>Mobile IP</th>
<th>HAWAII</th>
<th>Cellular IP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Advertisement</strong></td>
<td>Agent Advertisement indicating FA's CoA and its lifetime</td>
<td>Agent Advertisement indicating the NAI and the netmask</td>
<td>Beacon Messages specifying the cellular IP network identifier and the gateway IP address</td>
</tr>
<tr>
<td><strong>Register</strong> (during power up and after handoff)</td>
<td>Renewal registrations</td>
<td>Aggregate refresh messages</td>
<td>paging updates or routing updates depending status</td>
</tr>
<tr>
<td><strong>Acknowledgment</strong></td>
<td>MIP registration reply</td>
<td>Ack from DRR to BS and MIP registration reply from BS to mobile</td>
<td>No Ack</td>
</tr>
</tbody>
</table>

Having said that, there are also similarities between HAWAII and Cellular IP. From the system’s point of view, while a Cellular IP network is distinguished from the Mobile IP operations by *Cellular IP network identifier*, HAWAII uses *Network Access Identifier (NAI)* to recognize different HAWAII domains. In a similar manner, from the operation’s point of view, both establish and update host-based routing entries in only selective routers. Table 7.2 more specifically outlines the similarities and differences between Cellular IP and HAWAII.

**Table 7.2** Cellular IP versus HAWAII

<table>
<thead>
<tr>
<th></th>
<th>Cellular IP</th>
<th>HAWAII</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Similar Goals</strong></td>
<td><em>Maintenance of end-to-end connectivity with minimum disruption as MH moves</em>&lt;br&gt;<em>Reduced registrations to HA</em>&lt;br&gt;<em>Better tolerance of router and link failures within network</em></td>
<td></td>
</tr>
<tr>
<td><strong>Differences</strong></td>
<td>Operations remain the same regardless of MH location</td>
<td>Different treatment for roaming within HD and FD</td>
</tr>
<tr>
<td></td>
<td>No distinction between handoffs and normal operations</td>
<td>Path setup mechanism varies for power up and handoff operations</td>
</tr>
</tbody>
</table>

As HAWAII has advanced further along the developmental progress than Cellular IP, a lot
more issues (e.g. QoS) have been taken into consideration. Thus, it is still hard to conclude whether one protocol is better than the other. It is clear, however, that HAWAII as a whole is much more complicated than Cellular IP, but without solid performance measurements in hand, it is difficult to judge whether excessive simplification as is seen in Cellular IP is in fact feasible.

7.4 EXTENSIONS FROM CELLULAR CONCEPTS

Due to the difference in their underlying technologies (circuit- versus packet-switched) and applications (voice versus data), some operational differences are unavoidable between Cellular systems and IP networks. There are however certain aspects, particularly in terms of managing mobility, where areas requiring special attention do coincide. Table 7.3 gives a summary of such phenomena [93, 94].

Table 7.3 A comparison of design criteria between IP and Cellular networks

<table>
<thead>
<tr>
<th>IP networks</th>
<th>Cellular Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Location registration</strong></td>
<td></td>
</tr>
<tr>
<td>✦ Concentrate loads on HA</td>
<td>✦ Great reliance on HLR</td>
</tr>
<tr>
<td>✦ Prolonged transmission delay</td>
<td>✦ High latency (for MNs away from HLR)</td>
</tr>
<tr>
<td>✦ Waste network resources</td>
<td>✦ Unnecessary signaling traffic</td>
</tr>
<tr>
<td>**Main issue of concern — Resource utilization (\propto (\text{required location updates}))^†</td>
<td></td>
</tr>
<tr>
<td><strong>Call Delivery</strong></td>
<td></td>
</tr>
<tr>
<td>✦ All packets route via terminal's HA</td>
<td>✦ Query requests sent to HLR by default</td>
</tr>
<tr>
<td>✦ Additional functionality required at CN</td>
<td></td>
</tr>
<tr>
<td><strong>Main issue of concern — Delay (\propto \text{total data base accessing time} )</strong></td>
<td></td>
</tr>
</tbody>
</table>

Undoubtedly, it is worth investigating the possibility of applying operational concepts from techniques developed for cellular networks to that of IP networks. However, with mobile movements now between heterogeneous networks, considerations should be incorporated to ensure that the performance is optimized.

Based on an acknowledgement of the similarities between IP and Cellular systems, we were motivated to perform a detailed comparison study of several evolving techniques in the sphere of cellular networks. Through simulations, not only can the distinct characteristics of individual schemes be identified, but it is possible to identify whether one scheme is better than the other.
7.5 CONCLUSIONS

Looking at the rapid growth in mobile communications in the last decade, there is a trend towards ever increasing traffic. Consequently, the currently used channel (which has limited bandwidth) will require a special protocol limiting the total capacity for each user. While Mobile IP does seem to provide satisfactory performance for mobiles that are roaming among various networks, the cumbersome handoff operations become rather inefficient when the scale of movement is small. As a result of the inefficient location update procedures during handovers, seamless mobility is not supported as a result of the incurring high delay. Thus, while the Mobile IP provides a solid basis for enabling elementary mobility in the Internet, there is a demand for an alternative scheme that more efficiently manages local mobility of mobile hosts.

Consequently, it seems logical to simplify the operations by separating the mobility management tasks according to the nature of the mobile movement. The research focus is thus shifted to the scenarios where movements occur between points of attachments of different IP subnets, a scenario that is commonly referred to as micro-mobility. This chapter highlights some of the existing developments and suggests possible approaches for further enhancement of the performance.
Chapter 8

More Studies on Cellular Networks

With the ultimate goal of proposing a new location management technique(s) in mind, this chapter is a report on the comparison studies (via simulations) that were carried out for various location management techniques in a cellular network. Two mobility scenarios (macro-cell and micro-cell) are simulated in the mobility framework. From these, the impact of various factors have on the performance of different location management techniques was analyzed for different call arrival rates and mobility characteristics. The goal is to determine whether there is in fact one most proficient location managing algorithm that would have the least combined cost for location registration and call delivery. The three quantifying parameters are updating loads, database accessing loads and paging loads.

8.1 INTRODUCTION

As far as mobility management is concerned, a close review of the existing techniques in IP networks reveals great similarities in the general problems encountered in cellular networks. The problems come down to when and where to perform the updates. However, while the issue of “when” to update often plays a dominant role in the operations of location tracking in cellular networks, for roaming between heterogeneous networks, the issue of “where” to send

Love our principle, order our foundation
Progress our goal.

Auguste Comte (1798-1857)
the registrations becomes the main concern. For roaming between heterogeneous networks, the design is centered on determining the locations to which to send the registration signals, hence the issue of where to update. Clearly, given that developments in cellular networks are in general more mature, it is highly viable to re-apply certain aspects of the cellular systems to the IP framework. This need underpins the research work outlined in this chapter.

Recall that in the process of location registrations, decisions over when and where to send the update signaling are required. Various developments detailed in Chapters 3 to 6 look at one part of the problem of when to update with the underlying assumption that all updates were made directly to the home agents, being HLR in cellular networks. Specific updating and paging strategies were proposed to provide solutions determining an optimal time to perform update to minimize the necessary operational cost. Chapter 8 goes one step further in order to determine whether it is possible to reduce the operational cost by seeking alternative locations as to where to send the registrations; more specifically, can updates at the home network (HLR or HA) always provide optimal performance or could alternative centers for update achieve more appealing results? Through simulations, this chapter attempts to provide justifications for the applications of different location management schemes under distinctive network conditions.

8.2 STUDY OF THE CELLULAR SYSTEMS

For each call request, operational costs are associated with both the network and with the callers. On the one hand, it is important from the network’s perspective to minimize the necessary searching cost in locating the roaming mobile within the registered update area. On the other hand, the caller also incurs expenses for lodging such connection requests; specifically, a locally accessed database will result in a shorter delay and consequently, a smaller cost than that incurred when a distant database is accessed. Although much attention has been focused on the specific method of sending polling signals within the LA that is registered in the network, the importance of minimizing the time to find the relevant LA to page in the first place should not be underestimated. Similar to what is experienced in the update operation, the question of where to send the query signals becomes crucial where the mobile is away from its subscribed network but is “literally” close to the caller mobiles.

Despite the fact that there exist several potential location management proposals in the literature, it is difficult to determine the effectiveness, or perhaps the limitations each has upon the application of a common set of system parameters. Thus with the ultimate goal of proposing a new location management algorithm in mind, the present simulation work aims to draw up some comparison guidelines (via qualitative measures) between different existing mechanisms. This is, by numerically quantifying the impacts mobility and call arrival
characteristics have on system performance (in terms of QoS measures, mainly delay and throughput), it is envisaged that the newly proposed location management technique would give the best resource utilization while maintaining possible signaling delay at a minimum level.

It should be clarified that delay is, in this framework, quantified by the amount of necessary querying load in providing the relevant location information of the called mobile. Specifically, the more the required database access, the longer the call connection delays to be encountered by the calling party. Furthermore, with the general anticipation that the longer the signaling load is in progress, the less the remaining bandwidth that is available for actual data transmissions, throughput is calculated in terms of the total distance “traveled” by the necessary signaling loads due to location updates. Hence a longer delay leads to a reduced throughput.

Having discussed specific update techniques used to determine the time when an update becomes necessary, it is important to consider where the update signals should be sent. For all the above strategies, it is assumed that a registration is made directly to the HLR where the updated user profile is maintained. Depending on where the roaming takes place, the time it takes to get a new update registered grows when the mobile is far away from the subscribed home network. To minimize possible delays, alternative locations to have the update registered were suggested. The rest of the chapter will have a brief description of the available techniques.

To increase the practicability of our simulation results, it is certainly important to consider the specific location management techniques used in current cellular technologies, namely GSM, GPRS. In addition, 3 potential location management algorithms (all based on centralized database architecture) will be considered. Those comprise the user profile replication (UPR), local anchoring scheme (LAS), and pointer forwarding strategy (PFS) [34,64,67,95,96].

The chapter begins by summarizing the operations of UPR, LAS and PFS in section 8.2. Section 8.3 provides an overview of the simulation framework with the actual methodologies of implementation described in section 8.4. Section 8.5 discusses the simulation results; from which, the main applications of individual schemes are summarized in section 8.6. Future work is suggested in section 8.7 and is followed by the chapter conclusions in section 8.8.

8.2.1 USER PROFILE REPPLICATION (UPR)

The fundamental concept of UPR is straightforward — by replicating the location information at multiple databases, a faster location search can be achieved, resulting in a shorter call setup delay [64]. In the simplest case, whenever an update is made at the home location register
(HLR), the network will update all the replicated databases. Thus, any database inquiry for a roaming mobile can be made locally before a possibly long-distant signaling message is sent to the HLR as illustrated in Figure 8.1. With appropriate modifications, it is envisaged that the proposed scheme should be able to successfully avoid the problem of triangle routing evident in the Mobile IP standard, and thus, serve as a better alternative to the management of mobility in heterogeneous networks.

![Diagram showing signaling transfers](image)

**Figure 8.1** A graphical representation of the signaling transfers

Generally, depending on the mobility rate and the call arrival rate, the method may significantly reduce the signaling and database access overhead to compensate for the higher signaling overhead incurred during the phase of location registration. This improvement over delay is however at the expense of increased cost for database maintenance.

### 8.2.2 LOCAL ANCHORING SCHEME (LAS)

The local anchoring scheme allows location changes to be reported to a local anchor that is periodically selected from a set of locally available VLRs. Consequently, the necessary signaling traffic due to location updates to the HLR is reduced [95,96]. In the static scheme, the serving VLR when the last call arrived is selected to be the ‘anchor’. Alternatively, a dynamic scheme can be used to select the anchor VLR. In such a case, the network also makes decisions on whether to change the local anchor every time the mobile makes a movement in addition to what has implemented for the static system. The resulting improvement varies according to the mobility and call arrival characteristics.

### 8.2.3 POINTER FORWARDING SCHEME (PFS)

The Pointer forwarding strategy eliminates the procedure of reporting to a HLR by setting up
a forwarding pointer from the old visitor location register (VLR) to the new VLR subject to a location change [34,67]. It is noted, however, that depending on mobility, call arrival and the length of pointer chain $K$, an improvement in system performance over the conventional system is not always evident.

Table 8.1 provides adequate summarizing information to highlight the variations observed in different techniques.

<table>
<thead>
<tr>
<th>Location updating</th>
<th>Database accessing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GSM (LM)</strong></td>
<td>For active cells, update HLR at each cell boundary, otherwise LAs are used</td>
</tr>
<tr>
<td><strong>EGPRS (LM)</strong></td>
<td>For active cells, update HLR at each RA boundary, otherwise LAs are used</td>
</tr>
<tr>
<td><strong>UPR</strong></td>
<td>Update HLR and replicates users' records in multiple databases</td>
</tr>
<tr>
<td><strong>PFS</strong></td>
<td>Setting a forwarding pointer from the old VLR to the new VLR (if number of movements is less than the predefined pointer length $k$ )</td>
</tr>
<tr>
<td><strong>LAS</strong></td>
<td>Report to a local anchor which can be selected either statically or dynamically</td>
</tr>
</tbody>
</table>

Table 8.1  Brief summary of the simulated location management techniques

\* Geographical-wise

### 8.3 SIMULATION FRAMEWORK

Figure 8.2 summarizes the underlying philosophy employed in this simulation work. Very simply, the complete system framework can be looked at as a big block comprised of three sub-sections being the input, the actual system, and the output. Each, as its name suggests, represents a different dimension of this work. Provided the entire system architecture is defined, when given subscribers' movement characteristics at the input, various QoS measurements can then be calculated at the output. This in turn permits a comparison to be made on operational efficiencies between different systems. More detailed descriptions of each are discussed for clarification.
The Simulation Framework

**Subscribers' movement characteristics**

1. initial parameters  
   location/speed/direction

2. subsequent parameters  
   rate (and range) of speed change  
   rate (and range) of direction change  
   → probability and distribution

**Assumptions**

- No distinctions of new/handover calls (all calls get delivered)

**Input Parameters**

**Simulation Scenarios**

1. Macro-mobility models (suburban areas)  
2. Micro-mobility models (city areas)

**Movement patterns**

1. Random  
2. Straight-line motion  
3. Two-dimensional random motion  
4. Biased markovian motion

**System Layout**

1. Detailed descriptions of cell shape, cell size and the system framework  
   → e.g. Manhattan-like structure mobility model

2. Assign location of HLR & VLRs

3. Define a particular location management technique  
   - Define path of signaling  
   - Implement additional components/parameters

**Evaluation parameters**

1. End-to-end QoS matrix  
   - Delay  
   - Throughput

- Blocking probability  
- Forced termination probability

2. Operator's perspective  
   - Cost

3. Reliability  
   - The recovering process when VLR (and/or other relevant parameters) fails

**Output Parameters**

**Figure 8.2** The simulation framework
A. System Layout

To run meaningful simulations, it is necessary to clearly define various assumptions imposed on the system framework. More specifically, it should be unambiguous as to how various parameters, such as the cell shape and the cell size, are characterized.

The simulated cellular network consists of 25 equal-sized location areas. Each is managed independently by a MSC/VLR database. Within every LA, 4 routing areas RAs (made of 9 cells each) are implemented to aid the paging operations that are necessary later on for GPRS and EGPRS systems. Note that cells (each has a separate base station (BS)) are shown in Figure 8.3 as square shaped only to simplify illustrations.

![Network Architecture Diagram](image)

**Figure 8.3** A brief illustration of the simulated network architecture

It is necessary to remember that a MSC/VLR database is responsible for managing user's mobility and each is assumed to be located at the center of the corresponding LA (or RA for GPRS systems) for simplicity reasons.

Intuitively, the updating loads saved in GPRS are on the dimension of wireless connections, thus instead of having 9 BS→MSC/VLR updates in this simulation for example, there will only be 1 BS→MSC/VLR at the boundaries of RA crossings as shown in Figure 8.4. For this simulation network, we focus on the conservation of wire-line resources (i.e updates necessary from VLR to HLR), thus, there should be no distinguishable differences between GSM and GPRS. On the other hand, additional paging loads will need to be accounted for.
B. Simulation Scenarios

Having framed out the system, the next step is to model the anticipated movement patterns of mobile subscribers in various scenarios. In order to obtain some lower-bound performance, it was decided to maintain the framework at a level as general as possible for the time being. Thus, at this stage, only randomly generated movements will be implemented for both the macro- and micro-cell scenarios considered. However, it could be interesting to extent the research further by considering a more distinct movement patterns that considers the specific geographical conditions and user profile. In addition, although previous research has shown that the actual protocol performances are not restricted by residence time distributions, it could be advantageous to model in the mobility-related traffic parameters as a part of the further extensions. The implementation of such network feature will help to confirm the negative impacts.

C. Macro-cell Models (suburban areas)

To ensure that a wide range of possible trajectories is covered in the simulation, there are $n$ uniformly distributed traveling paths/trajectories allocated in the system framework. For macro-cell scenarios (often used to model sub-urban areas), it is justified to assume constant traveling directions, hence, the straight line illustrations. Figure 8.5 gives a graphical presentation of the modeling concept (assume $n = 5$). Note that due to symmetry, it is not required to have HLR located at the center. In fact, by setting the location to be at the left-up corner, the maximum number of trajectory patterns can be acquired, as there is only one quadrant being considered in such a case.
Figure 8.5  The macro-cell mobility scenario \((n = 5)\)

For this simulation work, we set \(n\) to 20. In other words, there will be a trajectory branch out from LA0 (i.e. HLR) for every \(3.5^\circ\) of separation in the radial direction. Table 8.2 lists the 20 generated trajectories specifying the visited VLRs in sequence along each course. The numbering of VLRs is assigned in the direction from left to right and from up to down as is exemplified in Figure 8.5.

<table>
<thead>
<tr>
<th>Path No</th>
<th>Trajectory (numbers indicate the LA locations)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1 (\rightarrow) 2 (\rightarrow) 3 (\rightarrow) 4 (\rightarrow) 5</td>
</tr>
<tr>
<td>1</td>
<td>1 (\rightarrow) 2 (\rightarrow) 3 (\rightarrow) 4 (\rightarrow) 5</td>
</tr>
<tr>
<td>2</td>
<td>1 (\rightarrow) 2 (\rightarrow) 3 (\rightarrow) 4 (\rightarrow) 5</td>
</tr>
<tr>
<td>3</td>
<td>1 (\rightarrow) 2 (\rightarrow) 3 (\rightarrow) 4 (\rightarrow) 9 (\rightarrow) 10</td>
</tr>
<tr>
<td>4</td>
<td>1 (\rightarrow) 2 (\rightarrow) 3 (\rightarrow) 8 (\rightarrow) 9 (\rightarrow) 10</td>
</tr>
<tr>
<td>5</td>
<td>1 (\rightarrow) 2 (\rightarrow) 3 (\rightarrow) 8 (\rightarrow) 9 (\rightarrow) 10 (\rightarrow) 15</td>
</tr>
<tr>
<td>6</td>
<td>1 (\rightarrow) 2 (\rightarrow) 7 (\rightarrow) 8 (\rightarrow) 9 (\rightarrow) 14 (\rightarrow) 15</td>
</tr>
<tr>
<td>7</td>
<td>1 (\rightarrow) 2 (\rightarrow) 7 (\rightarrow) 8 (\rightarrow) 9 (\rightarrow) 14 (\rightarrow) 15 (\rightarrow) 20</td>
</tr>
<tr>
<td>8</td>
<td>1 (\rightarrow) 2 (\rightarrow) 7 (\rightarrow) 8 (\rightarrow) 13 (\rightarrow) 14 (\rightarrow) 19 (\rightarrow) 20</td>
</tr>
<tr>
<td>9</td>
<td>1 (\rightarrow) 2 (\rightarrow) 7 (\rightarrow) 8 (\rightarrow) 13 (\rightarrow) 14 (\rightarrow) 19 (\rightarrow) 20 (\rightarrow) 25</td>
</tr>
<tr>
<td>10</td>
<td>1 (\rightarrow) 6 (\rightarrow) 7 (\rightarrow) 12 (\rightarrow) 13 (\rightarrow) 18 (\rightarrow) 19 (\rightarrow) 24 (\rightarrow) 25</td>
</tr>
<tr>
<td>11</td>
<td>1 (\rightarrow) 6 (\rightarrow) 7 (\rightarrow) 12 (\rightarrow) 13 (\rightarrow) 18 (\rightarrow) 19 (\rightarrow) 24</td>
</tr>
<tr>
<td>12</td>
<td>1 (\rightarrow) 6 (\rightarrow) 7 (\rightarrow) 12 (\rightarrow) 17 (\rightarrow) 18 (\rightarrow) 23 (\rightarrow) 24</td>
</tr>
<tr>
<td>13</td>
<td>1 (\rightarrow) 6 (\rightarrow) 7 (\rightarrow) 12 (\rightarrow) 17 (\rightarrow) 18 (\rightarrow) 23</td>
</tr>
<tr>
<td>14</td>
<td>1 (\rightarrow) 6 (\rightarrow) 11 (\rightarrow) 12 (\rightarrow) 17 (\rightarrow) 22 (\rightarrow) 23</td>
</tr>
</tbody>
</table>
In addition, it is noted that all parameters are made variable. Thus, they can be adjusted according to the characteristics of each individual system.

D. Micro-cell Models (city areas)

For micro-cell scenarios, again, an arbitrary $n$ is chosen to represent the number of simulated trajectories. However, as a much more restricted roaming area is now considered, cells will have small dimensions (i.e. 1km instead of 10km as for macro-cell scenarios) in addition to a smaller system framework (18 x 18 cells).

Different from that modeled in the macro-mobility scenario, it is not realistic to assume constant traveling direction in this case. As the whole system is now a lot more restricted, it is necessary to incorporate change of directions into the simulation. Hence, the introduction of a new parameter $cd$, the directional change to indicate the number of times a change of direction that will incur after the initial traveling direction is specified. This in turn, allows an arbitrary number of directional changes to be implemented. Figure 8.6 gives a visual aid to this model. Here for the simplicity of illustration, we set $n = 2$ and $cd = 6$.

![Cell Size = 1km...](image)

**Figure 8.6** The micro-cell mobility scenario ($n = 2$)
When implementing the trajectories in the simulation, an array of random times is generated to indicate the times at which a change will occur in mobile's traveling direction. It is assumed that for each trajectory, apart from the initial traveling direction assigned, each mobile can make another 12 directional changes. At the time randomly scheduled for a directional change, the mobile can move freely to any of the pre-selected 20 possible directions.

Figure 8.7 briefly illustrates the flow of how the trajectories in the micro-mobility scenarios are generated.

![Flow chart showing the generation of the trajectories in a micro-mobility scenario](chart.png)

Note that as the roaming area is now more restricted, an additional constraint imposed on the system is that the calculation loop aborts automatically whenever mobiles move out of the boundary.

Without actually including the detailed trajectory descriptions in the thesis, a brief example of the output is shown in Table 8.3 for illustration purposes.

<table>
<thead>
<tr>
<th>Number of directional change</th>
<th>Time (seconds)</th>
<th>New direction (radians)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.000000</td>
</tr>
<tr>
<td>1</td>
<td>2749</td>
<td>1.256637</td>
</tr>
<tr>
<td>2</td>
<td>4086</td>
<td>0.000000</td>
</tr>
<tr>
<td>3</td>
<td>5051</td>
<td>1.570796</td>
</tr>
<tr>
<td>4</td>
<td>5627</td>
<td>0.942478</td>
</tr>
<tr>
<td>5</td>
<td>5758</td>
<td>2.827433</td>
</tr>
<tr>
<td>6</td>
<td>7419</td>
<td>0.942478</td>
</tr>
<tr>
<td>7</td>
<td>10113</td>
<td>5.340708</td>
</tr>
<tr>
<td>8</td>
<td>12767</td>
<td>1.884956</td>
</tr>
<tr>
<td>9</td>
<td>16212</td>
<td>1.884956</td>
</tr>
<tr>
<td>10</td>
<td>16838</td>
<td>5.654867</td>
</tr>
<tr>
<td>11</td>
<td>17515</td>
<td>4.712389</td>
</tr>
<tr>
<td>12</td>
<td>23010</td>
<td>3.455752</td>
</tr>
</tbody>
</table>
Having defined both the system and the mobility patterns, it is now time to explain the simulation philosophy determining the methodology. The idea is firstly, to draw up different trajectories, and secondly, to calculate the associated location and paging loads for each trajectory. The performance will be quantified by three parameters: $U_x(X)$, being the information load measured in terms of number of location updates and traveled distance; $A_x(X)$, representing the number of necessary intermediate database accesses and $S_x(X)$, denoting the search load in terms of number of querying signaling load.

The range of the mobility rate is chosen such that it can give some realistic measures to the impact different mobility characteristics (in terms of the relocation frequency) have on the efficiencies of system performances. An implementing mobility rate varying from 1 to 0.000031 is found adequate in representing the desired range of mobility scenarios. This range conveniently represents the high mobility rate to be as frequent as one movement per second while restricting the low mobility rate to be as rare as one movement in every 8 hours. In addition, intermediate values (such as movements in every 10 seconds, 1 minute, 10 minutes, 30 minutes, 1 hour and 3 hours) are also modeled to better realize the wide range of possible mobility variations.

It has been assumed that one message is comprised of only one packet with arrival rates being modeled by random distribution. However, it will be worth while to examine the effects of traffic load in a more realistic scenario where one message is made of multiple packets. In such a case, an additional parameter will be introduced to separate the inter-arrival times of packets that belong to the same message from those that belong to different messages.

Calculations of the respective update loads and database query loads are tabulated for each of the considered 20 trajectories. In addition, a statistical summary is included at the end of each simulation run at a different mobility rate for the two distinct mobility scenarios. The procedure is repeated for all 5 specified location management techniques. Comparisons on system efficiencies are then possible based on various plotted graphs.

### 8.4 PROGRAMMING PHILOSOPHY

Figure 8.8 gives an overview of the network indicating the relations between the various implemented location management techniques, the assumed simulation scenarios, and also the expected outputs.
Figure 8.8  An overview of the simulated system

It is anticipated that the final outputs (the updating loads and database querying loads in particular) greatly depend on various system parameters. Those include, but are not limited to:

i) Number of sampling trajectories
ii) Size of system framework (cells, PAs, LAs)
iii) Simulation time
iv) Size/number of local databases
v) Number of calls—time & geographical distributions
vi) Possible traveling speeds

OPNET is known to have limited suppleness in scaling freely (up/down) the system size (in terms of cell numbers), which is an extremely important criterion for modeling cell boundary crossings. Thus, due to this scalability reason, we found that C programming will be the most promising simulation tool for its great capability in providing a huge range of flexibility to various system parameters, a requirement that is much needed in this simulation work.

Appendix A gives a brief illustration of the relations between the different loops that are implemented in the simulation work. In addition, brief descriptions of some major steps in
performing location management are shown in the flow chart to aid better understanding of the system framework.

The three major loops implemented in the system are 1) the mobility rate, 2) the path loop and 3) the time loop. Those are listed in order from the outermost to the innermost.

As the name suggests, the time loop specifies the time at which measurements are taken to determine the most recent locations of the mobile node. This indirectly represents the call arrival rate. Thus, the higher the call arrival rate, the more frequent the measurements are taken. The path loop allows a different number of paths to be generated in accordance to the specified mobility scenario. For each simulated path, calculations are made to determine the number of location updating loads and database querying loads. The mobility rate loop actually defines the rate at which movements take place. It is expected that the greater the mobility rate, the larger the number of location updates. As mentioned earlier, mobility rate \( v \) is implemented in the range of 0.000031 and 1 (i.e 0.000031 < \( v \) < 1). To help better relate to the practical scenario, Table 8.4 lists each of the simulated mobility rate with its corresponding frequency of movement.

<table>
<thead>
<tr>
<th>Mobility rate ( v )</th>
<th>CMR(^1)</th>
<th>Movement frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.02</td>
<td>one movement per 1 seconds</td>
</tr>
<tr>
<td>0.5</td>
<td>0.04</td>
<td>one movement per 2 seconds</td>
</tr>
<tr>
<td>0.25</td>
<td>0.08</td>
<td>one movement per 4 seconds</td>
</tr>
<tr>
<td>0.125</td>
<td>0.16</td>
<td>one movement per 8 seconds</td>
</tr>
<tr>
<td>0.0625</td>
<td>0.32</td>
<td>one movement per 16 seconds</td>
</tr>
<tr>
<td>0.03125</td>
<td>0.64</td>
<td>one movement per 32 seconds</td>
</tr>
<tr>
<td>0.015625</td>
<td>1.28</td>
<td>one movement per 64 seconds (1 minute)</td>
</tr>
<tr>
<td>0.0078125</td>
<td>2.56</td>
<td>one movement per 128 seconds (2 minutes)</td>
</tr>
<tr>
<td>0.00390625</td>
<td>5.12</td>
<td>one movement per 256 seconds (4 minutes)</td>
</tr>
<tr>
<td>0.001953125</td>
<td>10.24</td>
<td>one movement per 512 seconds (8 minutes)</td>
</tr>
<tr>
<td>0.000976563</td>
<td>20.48</td>
<td>one movement per 1024 seconds (16 minutes)</td>
</tr>
<tr>
<td>0.000488421</td>
<td>40.96</td>
<td>one movement per 2048 seconds (32 minutes)</td>
</tr>
<tr>
<td>0.000244141</td>
<td>81.92</td>
<td>one movement per 4096 seconds (1 hour)</td>
</tr>
<tr>
<td>0.00012207</td>
<td>163.84</td>
<td>one movement per 8192 seconds (2 hours)</td>
</tr>
<tr>
<td>0.00006103</td>
<td>327.68</td>
<td>one movement per 16384 seconds (4 hours)</td>
</tr>
<tr>
<td>0.00003058</td>
<td>655.36</td>
<td>one movement per 32768 seconds (8 hours)</td>
</tr>
</tbody>
</table>

To give some further clarifications to the simulation philosophy, details of the implemented local anchoring scheme algorithm are visualized by a number of flowcharts with each representing a different dimension of location management technique. Similar programming concepts are applicable for all other simulated algorithms. The generation of a local anchor is shown in Appendix B. It is necessary to remember that in the case of LAS, the performance greatly depends on the availability of a local anchor, which is generated upon

\(^1\) Applicable to LAS (calculated based on a call arrival rate of 0.02)
call arrivals. In this case, upon a call arrival, the location of the currently serving VLR will be assigned as the temporary local anchor which requires updates to the HLR.

Appendix C provides a detailed illustration of managing mobility. Basically, at each instant of sampling time, the system will determine whether a change in location has been occurred and if it has, whether there is in fact a local anchor available. This then decides where the location update should be sent. Relevant calculations are made for analysis and comparison purposes at a later stage.

It should be stated here that the database accessing load (or querying load) is a calculation of the required number of database accesses before the location of a called mobile can be identified. Take GSM system for example, as the information is always (and is only) stored in the HLR, upon the creation of a connection request, the signaling information goes straight to the HLR, where the most recent location information is obtainable, hence an access load of 1. All other studied location management systems use different mechanisms in managing mobility, hence, a dissimilar database querying load will result. The simulation programs written for each of the systems will need to cater for these variations. Appendix D provides the detailed simulation method for the LAS scheme for calculation of the database querying load. Note that the calculation is dependent on the rate at which calls arrive.

Based on a similar philosophy, Appendix E demonstrates the calculation of the database querying load occurring for the user profile replication scheme. Different from that seen in the case of LAS, UPR maintains the database access load to be at unity (i.e. only one database access is required before the location of the mobile is identified). However, the actual queried database might not be the HLR as is always the case for GSM. Thus, depending on the geographical distribution of the arriving call, savings can be easily expected from the UPR operations with the actual quantity of improvement varying according to the selections in locating the replicating databases. Obviously, the shorter the distance traveled by the querying signaling load, the less delay that can be expected in the process of call delivery, hence, the better QoS measures.

8.5 RESULTS AND DISCUSSIONS

It is reiterated that the motivation for doing this simulation work is to determine a location management technique with two-folded design goals: one to reduce the loads at HLR (for both cost reductions and better risk management), and two, to reduce the total number of necessary updates in managing mobility. While Figures 8.9 and 8.10 give a measure to the load being imposed solely on HLR respectively for macro- and micro-cell scenarios, Figures 8.11 and 8.12 compare the total updates introduced into the system by various location management techniques.
Figure 8.9  Impact of mobility on HLR updates (macro-cell scenario)

Figure 8.10  Impact of mobility on HLR updates (micro-cell scenario)

In terms of updates at HLR, it appears that LAS imposes the most loads among all simulated schemes for both mobility scenarios. Although the actual loads would vary according to the mobile's trajectory and the call arrival characteristics, this general trend of
increasing load is due to the additional updating loads to HLR upon call arrivals. PFS on the other hand, has, as expected, imposed a lessened load to the HLR as compared to GSM/GPRS, which corresponds to the special case of PFS with a pointer length of $K=0$. In fact, it is readily evident that in the macro-cell scenario, PFS ($K=2$) outperforms GSM by almost twice the efficiency. Similar improvement is also observed for the micro-cell scenario despite the irregularity in the implemented trajectories that is intentionally introduced into the system.

![Figure 8.11](image1.png)  
*Figure 8.11  Impact of mobility on total updates (macro-cell scenario)*

![Figure 8.12](image2.png)  
*Figure 8.12  Impact of mobility on total updates (micro-cell scenario)*
Clearly, the loads necessary for updating the duplicated databases has put UPR on the top of the list for creating the highest total location loads. As mentioned earlier, this additional load is proportional to the number of duplicated databases available in the system. In other words, while the number of HLR updates remains the same as that obtained from conventional GSM or GPRS systems, the replicated signaling loads grow in accordance with the frequency of location updates. For this reason, UPR would be most suited in scenarios where a high call arrival rate is observed in conjunction with a relatively lower mobility rate.

On the other hand, though it does seem that LAS has higher updating loads than GSM, the difference becomes almost negligible for higher CMRs (in the range of 2.5 and 5) for simulated scenarios. In other words, given a mobile environment with CMR in the range of 2.5 and 5, the feasibility of employing LAS would be most promising.

To enable a meaningful comparison, firstly, it is necessary to clarify the definition of database access (or querying) load. In this work, we define the access load as the number of database accesses that are required before information on the currently serving VLR is obtained upon a call arrival. Take GSM for example, it will always use only one database (being HLR) for maintaining the most recent location information about the mobile node.

As PFS has demonstrated the advantage of reducing updates by factors proportional to the assigned pointer length $K$, one would expect that a lot more database access load would result. From the simulation results shown in Figures 8.13 and 8.14, it is however remarkable to note that such an expectation is not always true.

![Impact of mobility on database querying loads](image)

**Figure 8.13** Impact of mobility on database querying loads (macro-cell scenario)
Figure 8.14  Impact of mobility on database querying loads (micro-cell scenario)

Although it is acknowledged that when the mobility is high, PFS does have the highest database access load, when the mobility rate is dropped to approximately 0.0001 (corresponding to a movement every 2 hours), the averaged database access load significantly reduces to a level that the resultant load becomes comparable to all other schemes. If the mobility rate is further decreased, the viability of LAS starts to become appealing as it gives the lowest database access load among all simulated algorithms.

All in all, the performances of all simulated location management techniques are dependent on the mobile's travel trajectory and the rate of mobility. Generally, the lower the mobility, the less the number of updates is necessary. This would indirectly reduce the averaged value of the database access load.

Figures 8.15 and 8.16 give a comparison of the total costs incurred in managing mobility. Here, it is assumed that the updating cost equals the database querying load on a per unit basis. In reality however, this parameter would vary according to the method used to define different infrastructure costs — often, the expense for sending location updates is found to be greater than that resulting from database accessing (or querying). In other words, if the updating load is assumed to be twice as expensive as that of the querying load, the differentiation between different location management schemes would be even more significant. This indicates the importance of the careful selection of various parameters in order to achieve an optimal operation.
At first glance, the high location updates incurred by UPR location updates do not seem promising at all. However, when comparing the savings (in terms of traveled distance) obtained by UPR to other schemes during the phase of database querying, it seems justifiable
to employ the UPR scheme. Appendix F shows the simulation results obtained for the 25 randomly generated calls. The intention is that given a randomly generated call arrival distribution, the combined traveled distance (for both the updating and database querying loads) can be compared to provide indications as to which specific location management technique would give the least end to end delay. This is important particularly for UPR as it determines whether the savings in database querying would actually justify the additional delay incurred to update the replicated databases.

It is found that more often than not, the necessary distance for accessing one of the replicated databases is shorter than that required to access HLR. This reduced distance implies a shorter delay during call delivery, which has direct impact on the QoS measures.

Moreover, as the locations of calls are randomly generated, the QoS measures do greatly depend on the location and the number of the pre-selected replicated databases. Hence, providing that the geographical distributions of call arrivals are traceable, relevant parameters can be dynamically adjusted to achieve better system performances. Note also that during this phase of the operation, the performance is independent from the mobility rate.

8.6 APPLICATION SUMMARY

To sum up, Table 8.4 summarizes the main parameters that would have direct impacts on the system performances for each distinct location management scheme.

Table 8.4  Main factors that would influence the efficiency of the system operations

<table>
<thead>
<tr>
<th></th>
<th>Update load</th>
<th>Database access load</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPR</td>
<td>♦ Geographical distributions of the replicated databases ♦ Geographical distributions of call arrivals</td>
<td></td>
</tr>
<tr>
<td>PFS</td>
<td>♦ Pointer length                                 ♦ Pointer length</td>
<td></td>
</tr>
<tr>
<td>LAS</td>
<td>♦ Inter-arrival time distribution of call arrivals ♦ Inter-arrival time distribution of call arrivals</td>
<td></td>
</tr>
</tbody>
</table>

Overall, the improvement achievable by the simulated location management techniques with reference to GSM systems can be summarized in Table 8.5.

Table 8.5  A comparison of the specific savings seen for various location management techniques

<table>
<thead>
<tr>
<th></th>
<th>Update load</th>
<th>Database access load</th>
<th>Call setup delay</th>
<th>Paging load</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPRS</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>UPR</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>PFS</td>
<td>+</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>LAS</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>0</td>
</tr>
</tbody>
</table>

0 no improvement + better performances - worse performances
Note that the update load comprises all updates from MN to HA and from MN to all local mobility agents. From this comparison, it is possible to derive a list of criteria for the design of our own location management algorithm. Again, two aspects should be considered: the location updates and database querying. In terms of location update, dependence on HLR should be reduced for better efficiency and utilization. In terms of database accessing load, the number of locations of the replicated databases should be selected according to the rate and geographical distributions of the call arrivals.

8.6.1 UPR

UPR would be the most efficient algorithm in scenarios where calls are arriving at a greater rate than the rate of mobility. The usage of this scheme is particularly promising if the replicated databases can be dynamically relocated according to the geographical distributions of call arrivals. The total number of updates is the summation of updates from the mobile node to the subscribed HLR and the subsequent updates sent from HLR to all replicated databases. While the number of HLR updates remains the same as that arising from the traditional GSM operations, the additional updates from the act of location information duplications increase according to the number of assigned location replications. Basically the more replicated databases, the greater the resultant location updates, and thus, the less delay in locating a mobile user upon a call arrival.

It was assumed in the simulation that the locations of replicated databases are evenly distributed in the system framework. This, however, might not always be the optimal implementation in a real-life scenario. Although it is possible to implement the replicated databases at randomly selected VLRs, it does not appear to be a feasible or even a practical assumption. It is obvious that to a great extent the performance characteristics of this particular scheme depend on both the locations of the replicated databases, and also the geographical distributions of the arriving calls. Consequently, provided some on-line feedback is available for analyzing the geographical distributions of mobiles' movement, better performance characteristics can be expected by dynamically adjusting the locations of the duplicating local databases. It might also be advantageous to model the incoming call arrivals (for each mobile terminal) to be geographically distributed with an arriving probability of \( c \) at each discrete time \( t \).

8.6.2 PFS

In contrast to UPR, PFS outperforms other algorithms for low averaging call-to-mobility ratios. As the rate of call arrivals is now small compared to the mobility rate, the major
disadvantage seen in PFS (i.e. the relatively higher database accessing load) is minimized while benefiting by significant reductions in the location updating loads.

The biggest tradeoff seen in this protocol is the amount of updating loads versus database accessing loads. It is readily evident that the number of HLR updates (for both macro- and micro-cell scenarios) reduces at a rate directly proportional to the size of pointer length. Intuitively, the less updates required at HLR, the more database accessing loads will result upon the arrival of an incoming call, thus, causing a longer delay during the phase of a call setup. Given that location updates would normally consume much more infrastructure cost than that contributed from the accessing signaling, careful selection of the pointer length is of paramount importance to ensure an optimal performance from the pointer forwarding scheme. The actual improvement varies depending on the actual traveling trajectories taken by the specific mobile.

The underlying concept for PFS is almost identical to that of the LAS. The differentiating factor is only the method that defines how the “replacement” location updating database is selected. In the case of PFS, a fixed pointer length $K$ for the system would automatically reduce the updates at HLR by a factor of $K$. Thus, the higher the mobility, the greater the savings achievable in terms of delay. This improvement is most significant when frequent updates are necessary to a distant HLR.

In areas where high mobility is often predicted, a larger pointer length can be appointed to further reduce the necessary HLR updating loads. Thus, upon the identification of various geographical divisions classified by the level of users' mobility, different pointer lengths are set to ensure an optimal operation. The two design rules identified for this purpose are (1) the higher the mobility, the larger the pointer length, and (2) the more obvious the “locality” pattern, the greater the pointer length.

8.7 Future Work

Although it would have been advantageous to more realistically design the size of the simulated framework in terms of the number and size of cells, the simulated results provided the necessary guidelines to highlight the distinct characteristics of individual schemes. As illustrated in Figure 8.17, in order to give the most satisfactory performances, one should consider the possibility of combining different location tracking strategies at various parts (or levels) of the complete system.
Based on the same philosophy, provided that online feedback on the system characteristics is available, to find out whether individual systems can be appropriately enabled/disabled to ensure better performance efficiencies. The main characteristics to be considered include not only the call arrival and mobility characteristics (or CMRs) but also the actual definitions for the system infrastructure [97].

8.8 CONCLUSIONS

This chapter gave some numerical comparisons of various location management techniques in reaction to different levels of mobility and call arrival rates under different mobility scenarios. The goal was to determine whether there is in fact one most proficient location managing algorithm that would have the least combined cost for location registration and call delivery. The three quantifying parameters are updating loads, database accessing loads and paging loads.

To ease the presentation complexity, mesh cells are illustrated throughout this chapter. In addition, it was decided that no directional restrictions should be imposed on the mobiles’ movements within cells in order to better simulate the irregularity of individual city scenarios. It is envisaged that through such a simulation study, better categorizations can be made regarding the performance characteristics of location management schemes, thus, giving some
insight into designing alternative methodologies that would better suit our specified and/or desired mobility environment.

It is found that the *user profile replication* scheme has a superior performance in terms of reduced database accessing loads. This improvement, however, is counteracted by the additional replication load imposed on the home network in sending duplicated database information to the pre-selected local databases. Thus, in a scenario where a more frequent mobility pattern is expected, the *pointer forwarding scheme* would be the best alternative for the implementation of location management.

All in all, the simulation results were satisfactory. They not only gave some preliminary analysis of the mobility problem, but also provided some new directions in designing better strategies for updating locations.
I can honestly say that I was never affected by the question of the success of an undertaking. If I felt it was the right thing to do, I was for it regardless of the possible outcome.

Golda Meir (1898-1978)

Chapter 9
Applications of Cellular Strategies onto IP Networks

In this chapter, the research work is revolved around three main concepts. Firstly, acknowledging the emerging market demand for network overlapping, a network-initiated handoff technique is developed. With combined operations of location and handoff management, it is envisaged that optimal system efficiency would be achievable. Secondly, recognizing the great success demonstrated by the specific techniques of user profile replication (UPR) and pointer forwarding scheme (PFS) for mobility management in cellular networks, a comprehensive discussion is presented to examine how, through appropriate modifications, the applications of the schemes suit well IP networks. Lastly, appreciating the importance of implementing flexibility into system operations, new location management schemes designed with a special focus on the provisioning of better scalability, backward-compatibility, simplicity, adaptability, and performances are presented. It is envisaged that a better trade between cost and performances can be achieved through the implementation of our dynamic schemes.
9.1 INTRODUCTION

In this part of the work, the goal is to develop a new location management technique capable of providing the required QoS measures in the world of wireless IP networks. In terms of basic design criteria, our preliminary search revealed a great similarity between cellular systems and IP networks in their respective paths of evolution. In fact, the resemblance was so strong that it seemed promising to re-apply some of the well-developed operational philosophies used in cellular technology to the vision of an all-IP network. This observation motivated the content of this chapter. It aims to demonstrate how it is conceptually possible even when the considered “mobility” is at different levels.

Firstly, a new network-initiated handoff scheme is proposed in section 9.2. Based on the definitions of mobility management, it appears promising to consider the possibility of combining the operations of location and handoff management to enhance overall operational efficiency. Section 9.2 also discusses the key merits and downfalls of the proposed scheme.

The rest of the chapter focuses on the applications of cellular strategies onto IP networks. Clearly, in addition to accommodate the specific requirements from the network overlapping scenarios, any proposed scheme should also comply with the basic criteria expected from tracking schemes in general. Consequently, section 9.3 summarizes the desired design criteria. The actual tasks to be defined are also tabulated in the section to better frame out the design problems. Section 9.4 provides an explanation for the implementation of UPR and PFS, the two well-developed cellular strategies in the context of IP networks. This is followed by, in section 9.5, a brief description of the actual operational proposals that were incorporated. Following a preliminary analysis that confirms the potential of the designed concepts, section 9.6 suggests a possible simulation platform to be incorporated into the research for obtaining a more complete analysis of the proposed schemes. Although the actual implementation is outside the scope of this chapter, it is expected that the level of technical detail included in this section will be sufficient to initiate further studies on this particular research work.

9.2 CONCEPT OF NETWORK-INITIATED HANDOFF

On the one hand, handoff management is about the appropriate reservation of resources (such as bandwidth) along the roaming path of a mobile user while engaging in a call connection. Its efficient operation is important to ensure that the various aspects of the QoS requirements (e.g., throughput versus forced call termination) are satisfactorily complied [44]. Location management, on the other hand, is mainly for users who are currently idle, but are expected to receive calls (or become active) while they are changing their points of attachment to the
network. In which, sufficient location information about the mobile should be maintained so that the network could track the mobile’s movement and subsequently incur a minimal paging load when the precise residency is required [39]. Clearly, the operations of location management and handoff management complement each other. Thus, a possible approach, extended from the current mobile-initiated handoff, is to combine the necessary location management operations with an appropriate handoff management technique.

It appears viable to introduce some sort of a handoff controlling entity with the function of determining the next access point for the roaming mobile’s handoff in the overlapping part of the networks. By appropriately incorporating intelligence, such controlling entities can be expected to recognize certain inherited characteristics of the running applications. Thereby, whenever a MN moves into one of those “controlling” cells (or possibly sectors), an automatic network transition is incorporated. The selection should be made in a way that it allows the mobile to obtain a best tradeoff between service and cost, and to keep the required signaling at a minimum at all times. The functions of this entity are to initiate handoffs upon determination of the best network for the roaming MN to be connected to, based on periodic feedback of the characteristics of the MN’s running applications. In addition, those network controlling entities can function independently from those of the co-existing domain administrators, and thus, have security issues more carefully handled.

Figure 9.1 briefly illustrates the proposed concept. Generally, at different parts of the “overlapping” sector, one pre-allocated BS/AP will be available to send appropriate advertisements notifying its capability of initiating handoff.

![Network-overlapping Scenario](image)

**Figure 9.1** A simple illustration of the network overlapping scenario

Upon receiving such special beacon messages, the MNs are required to send their general statistical information (e.g. call arrival, mobility rate about their past roaming history) to the network controlling entity, and thus, enable an up-to-date measure to be made on the operational efficiency of the current network connection.

Note that in order to reduce the computational load at those network-controlling points, the intelligence will be implemented at the MNs. Specifically, the MN will keep a record of the number of times a path setup request has been generated and sent to the corresponding
BSs/APs along with other statistical information about the mobile characteristics. Clearly, the introduction of a network-controlling entity would ensure that the best network services can always be provided. It takes the feedback information as the input, and produces at the output the appropriate decision as to which BS/AP (or network) to hand off to if required.

The essence of the designed operation is realized by the improved treatment of inter-technology roamings within the same administrative domain. Specifically, for movements within an administrative domain, under the management of a network controlling point, the type of handoff will be determined based on the characteristics of the MN. It then relies on the incorporation of another database (implemented at the network level) to store the necessary location information that are replicated among domain operators.

Furthermore, it should be possible to make the handoff proactive. In other words, the network should be able to somehow give predictions of mobiles' movement. Although the predictions might not be very precise, it should at least be able to give some helpful hints to ease the system operations. One disadvantage to this, however, is that once the necessary control signaling has been done to set up the connection, any “sudden” change of pattern in mobile's roaming would cause a waste of the bandwidth usage. Nonetheless, the possibility of such occurrence is likely to be small. It is expected that the benefits will still outweigh the use of resources, given an appropriate selection of the feedback period for analysis.

To evaluate the significance of the proposed network-initiated handoff scheme, the following discussions are included to highlight the pros and cons of the operations. Firstly, by implementing such intelligence in some pre-selected network controlling points, extra signaling load and a demand for processing memory might become unavoidable. Among other difficulties, this feature would require the controlling point to have a thorough understanding of the characteristics and behaviors of each individual network. From this aspect, a mobile-initiated handoff might appear to be a simpler alternative as there will be no need for the networks to have knowledge of the existence of each other [6,29,30]. However, as mobiles are no longer required to be equipped with multi-moded terminals before being allowed to commit seamless roaming among RANs of various technologies, the proposed network-initiated handoff scheme will still be a better alternative. Clearly, by freeing the restriction on the mobile terminal to be multi-moded, the proposed algorithm has an increased flexibility, an important measure of performance optimization.

In order to more precisely quantify the performance, further analysis will be necessary to compare the resultant signaling load with and without such controlling nodes in a vertical handoff scenario. Specifically, this procedure is required to determine whether the additional control signaling incurred during information transmission will have much impact on the overall measures.
9.3 LOCATION TRACKING SCHEMES FOR IP NETWORKS

It is evident that the TeleMIP protocol proposed as an extension to MIP incorporates a very similar idea to the usage of local anchoring scheme, LAS, in cellular applications [34,75,95,96]. The core concept of both lies in locally handling the mobility instead of having to send the registrations all the way to HA. In terms of implementation, an additional hierarchical layer was introduced in TeleMIP for the incorporation of such an “anchoring” point. This observation is critical as it helps to determine the specific network scenarios where the technique would be most effectively applied. In fact, from the simulations performed in Chapter 8, it is obvious that the static LAS would be best suited for applications where the system demonstrates a low CMR behavior. Given the great similarity observed between the two protocols (LAS and TeleMIP), it is envisaged that the improvement of the operations will be most significant when frequent movement occurs in an area far from the mobile’s subscribed home network. Mobility, in this aspect, refers to movements within the administrative domain. Consequently, the simulation results clearly suggest that TeleMIP, by itself, might not always provide optimal performance when the network conditions (and mobile characteristics) vary. By adapting certain characteristics of database replication and pointer forwarding, it seems most promising that better QoS measures can be achieved through dynamic multi-moded operations.

Given the strong resemblance in their evolutions, it seems promising to re-apply some of the well-developed operational philosophy used in cellular technologies to the vision of an all-IP network between the two technologies. As detailed in Chapter 8, our previous simulation indicates that the two techniques of particular interest are those employed in the user profile replication (UPR) and the pointer forwarding scheme (PFS) [34,64,67].

One interesting lesson learned from previous simulation works on PCS network was that despite the existence of several proposals, the performances obtained were not always good. In fact, there were scenarios (depending on the specific mobility and call arrival characteristics) where the proposed schemes would actually have worse performances than those achieved by the conventional GSM/GPRS systems. Consequently, apart from focusing on the provisioning of roaming between heterogeneous networks, special care will be taken to avoid the above-mentioned deficiencies. Furthermore, while keeping the operations flexible, the proposal will need to be compatible not only with legacy networks but also with the soon-to-be standardized Mobile Internet Protocols (for both version 4 and version 6). The latter suggests the need for the mobility solution to be architecture independent.

Lastly, it is noted that although the intention is to develop a new protocol, the option of proposing a new architecture framework is also kept open shall such a feature later prove to be essential.
9.3.1 DESIGN CRITERIA

Basically, the operation can be divided into two categories; from the perspective of the mobile and from that of the on-going communication. While the prime goal is to keep user movement transparent for valid data delivery, the mission of mobility management technique is two-fold: from the callee’s point of view, to efficiently determine the time and place to send the update registrations to; and from the caller’s point of view, to efficiently determine the place for sending the query signal to locate the roaming mobile. The vital question comes down to how to reach a balance between the number of updates and the polling load.

Furthermore, the new location management scheme(s) are designed with a special focus on the provisioning of better scalability, backward-compatibility, simplicity, adaptability, and performances. Along with the main design criteria, Table 9.1 summarizes the main operational requirement to be accommodated [19].

To impose some formal organization on the design process, Table 9.2 summarizes the operational details that will need to be considered. In addition to answering the basic questions of where and when to perform updates, the table outlines other issues that require closer consideration. The formation of this table not only sets boundaries for the tasks to be considered, but also provides the necessary design framework.

Other requirements such as having the algorithm adaptive to network parameters also need attention. Specifically, the set of parameters that needs constant online feedback would include, but not be limited to, the mobility rate, the state transition rates, and where attainable, the call arrival characteristics (e.g. arrival rate and geographical distributions). For this purpose, the use of counters at the terminal level could be a viable implementation alternative. Lastly, the scheme should be easy to implement in real-life mobile networks.

9.3.2 DESIGN PROCEDURE

Generally, in receiving a location request, the system should analyze the type of movement that has been encountered – whether it is an inter-/intra- domain handoff or an inter-/intra-technology handoff. Specifically, whenever certain network characteristics are detected, a different “mode” of operation can be activated to ensure the maintenance of an optimal system performance. A location update will then be issued at the local and/or home network according to the information derived from the signaling from roaming MN. Figure 9.2 gives an illustration of such a design philosophy. As a starting point, the design focus was on a scenario where roaming occurs between domains implemented in the same technology. Therefore, by appropriately modifying the signaling algorithms, other mobility scenarios of interest can be better accommodated.
<table>
<thead>
<tr>
<th>Table 9.1</th>
<th>A list of design criteria to be complied with in the protocol development</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Requirements for network-overlapping</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Applicability and flexibility</strong></td>
<td></td>
</tr>
<tr>
<td>✷ Registration should allow multiple levels of hierarchy</td>
<td></td>
</tr>
<tr>
<td><strong>Adaptability</strong></td>
<td></td>
</tr>
<tr>
<td>✷ Registrations should support auto-configuration capabilities for mobile agents/FAs, access routers</td>
<td></td>
</tr>
<tr>
<td><strong>Backward-compatibility</strong></td>
<td></td>
</tr>
<tr>
<td>✷ Registration should not impose any additional requirements on fast handoff beyond what is required by basic MIPv6</td>
<td></td>
</tr>
<tr>
<td><strong>Performance improvements (From service providers' perspective)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Reliability</strong></td>
<td></td>
</tr>
<tr>
<td>✷ Avoid single point of failures:</td>
<td></td>
</tr>
<tr>
<td>✷ Connectivity to the mobiles should not be interrupted in the presence of the failure of regional registration agents</td>
<td></td>
</tr>
<tr>
<td><strong>Signaling load</strong></td>
<td></td>
</tr>
<tr>
<td>✷ Registration should be introduced to minimize the signaling traffic to the HA or CN for intra-domain mobility</td>
<td></td>
</tr>
<tr>
<td><strong>Resource consumption</strong></td>
<td></td>
</tr>
<tr>
<td>✷ Registration should not introduce new overhead on links between the mobile and the regional registration agents</td>
<td></td>
</tr>
<tr>
<td><strong>Delay</strong></td>
<td></td>
</tr>
<tr>
<td>✷ Scheme should provide optimized routing for mobile-to-mobile communication</td>
<td></td>
</tr>
<tr>
<td><strong>Performance improvements (From operators' perspective)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Packet loss</strong></td>
<td></td>
</tr>
<tr>
<td>✷ Scheme should minimize the number of network nodes affected by handoff signals</td>
<td></td>
</tr>
<tr>
<td><strong>Performance improvements (From customers' perspective)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Scalability</strong></td>
<td></td>
</tr>
<tr>
<td>✷ Scheme should be scalable in terms of the number of connected users and the number of users on the move (in order of millions)</td>
<td></td>
</tr>
<tr>
<td>✷ Registration should scale to support millions of nodes in a visited network</td>
<td></td>
</tr>
<tr>
<td>✷ Registration should not introduce host routes in routing tables</td>
<td></td>
</tr>
<tr>
<td><strong>Simplicity</strong></td>
<td></td>
</tr>
<tr>
<td>✷ Registration should not require changes to MN, HA or CNs</td>
<td></td>
</tr>
<tr>
<td>✷ Scheme should simplify the network design and enable progressive LMM deployment</td>
<td></td>
</tr>
<tr>
<td><strong>Performance improvements (From service providers' perspective)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Security</strong></td>
<td></td>
</tr>
<tr>
<td>✷ registration should be secure against malicious behavior from visiting mobiles</td>
<td></td>
</tr>
</tbody>
</table>
Table 9.2  A summary of the complete design problem

<table>
<thead>
<tr>
<th>Issues of consideration — Location wise</th>
</tr>
</thead>
<tbody>
<tr>
<td>✷ Where to send the location updates</td>
</tr>
<tr>
<td>✷ Static/dynamic</td>
</tr>
<tr>
<td>✷ Locally/sequentially</td>
</tr>
<tr>
<td>✷ Duplicating the functions of HLR at a local level</td>
</tr>
<tr>
<td>✷ Places for storing information</td>
</tr>
<tr>
<td>✷ Centralized/distributed</td>
</tr>
<tr>
<td>✷ Load balancing algorithm (for both registration and query loads)</td>
</tr>
<tr>
<td>✷ Replication scheme (centralized/distributed database architecture)</td>
</tr>
<tr>
<td>✷ Per-user basis</td>
</tr>
<tr>
<td>✷ Based on the geographical distributions of activity centers</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Issues of consideration — Frequency wise</th>
</tr>
</thead>
<tbody>
<tr>
<td>✷ Time for sending location updates</td>
</tr>
<tr>
<td>✷ Static (i.e. define network infrastructure)</td>
</tr>
<tr>
<td>✷ Change of IP addresses</td>
</tr>
<tr>
<td>✷ Dynamic (i.e. movement dependency)</td>
</tr>
<tr>
<td>✷ Time-based, distance-based, movement-based</td>
</tr>
<tr>
<td>✷ Mobile’s activity status dependent (highly-mobile/relaxed or hot/cold mobiles)</td>
</tr>
<tr>
<td>✷ Lifetime of a stored location information (time for a specific record deletion)</td>
</tr>
<tr>
<td>✷ Cache</td>
</tr>
<tr>
<td>✷ Pre-select the lifetime of the stored location information</td>
</tr>
<tr>
<td>✷ Possibility to consider a range of acceptable values (as opposed to a precise one)</td>
</tr>
</tbody>
</table>

Additional functionality

✷ Algorithm is fully adaptive to changes in the network parameters
   ➢ Call arrival\(^1\)
      ✷ Rate/frequency
      ✷ Traceable geographical distributions/ characteristics
   ➢ Mobility rate
   ➢ State transition rates

Special considerations

✷ Specific scenarios to be catered for
   ➢ ping-pong effects
   ➢ tolerance and robustness to network/link failures

\(^1\) The call arrival rates will only be significant in the case of multimedia transmissions
Figure 9.2  Illustration of the overall operational philosophy

The rest of the chapter focuses on the two shaded boxes that indicate the possible implementation of PFS- and UPR- based updating techniques.
9.4 INSPIRATION FROM CELLULAR TECHNOLOGIES

Considering the great advancements in wireless technologies, it seems reasonable to assume that a mobile device will be the main mechanism to accommodate the daily needs of communications. Consequently, an efficient call delivery operation would be essential to guarantee system performances. While a significant effort has been reported in the literature for developing efficient location updating techniques, relatively less attention has been paid to the developments of the corresponding call delivery process; more specifically, the cost associated with database access.

The development of UPR avoids such a deficiency. Its fundamental concept is straightforward — by replicating the location information at multiple databases, a faster location search can be achieved, resulting in a shorter call setup delay. With appropriate modifications, it is envisaged that the proposed scheme should be able to successfully avoid the problem of triangle routing evident in the Mobile IP standard, and thus, serve as a better alternative to the management of mobility in heterogeneous networks.

Generally, the user replicated databases scheme would be most suited to scenarios where a high call arrival rate is observed in conjunction with a relatively lower mobility rate. The focus is to distribute the request signaling, thus reducing the corresponding query delay or database accessing delay. An important goal would be to locate the replicating databases according to the geographical distributions of the call arrivals.

In addition, UPR is also capable of reducing undesired packet loss caused by the sudden loss of a maintained database during call delivery. By effectively making “backups” of the maintained mobile location information, a better management is put in place to combat high risks of single point failures.

On the other hand, with increasing mobility, the concern with operational inefficiency becomes apparent specifically with excess location updates during the phase of registrations. It would then be necessary to focus on minimizing the signaling loads during the “control” phase and thus save the scarce bandwidths for actual data transmissions.

As an illustration, consider the simplified network shown in Figure 9.3.

![Figure 9.3](image-url)
Here it is assumed that only packets intended for the mobile users in the network reach the domain. While all subnets are managed by the same administrative domain, movements within individual subnets are controlled by an associated foreign agent FA.

For a pointer length of 2 defined in PFS, assume an update at the home network is registered when the mobile resides in FA1. Upon the detection of movement to subnet 2, MN will first notify FA2 which then notifies FA1, resulting in a scenario where both FA1 and FA2 maintain the required location information. In general, it is noted that the implementation of PFS is relatively straight-forward. Other than having to incur a greater call setup delay, there seems to be less concern about the basic PFS operations. Although the issue of inconsistency between what is stored in HA and the local PFS could be something to consider, it does not seem to be a major problem. After all, mobile's location information can always be determined either directly (i.e. through PFS, the previous database) or indirectly (i.e. through HA).

To better illustrate the potential implementations of cellular concepts, Table 9.3 shows the associated signaling transfer of the three techniques that are currently under consideration; UPR, PFS and LAS. The idea is to draft out a comparison table and by listing the importance of each parameter, it will become clear as to how effective each concept is in solving various problems encountered in establishing data connections in a mobile infrastructure. It is reminded that the main factors that were identified to have impact on the actual system performances are summarized in Table 8.4.
### Table 9.3  A comparison of the signaling transfers: assuming MN moves from FA1 to FA2 and then to FA3

<table>
<thead>
<tr>
<th>Assumptions</th>
<th>Control MIP</th>
<th>Scheme one PFS [34,67]</th>
<th>Scheme two LAS [34,95,96]</th>
<th>Scheme three UPR [34,64]</th>
</tr>
</thead>
</table>
| Location registrations | (1) MN -> FA2 -> HA  
(2) MN -> FA3 -> HA | k = 2  
(p) FA1 -> HA  
(1) MN -> FA2 -> FA1  
(2) MN -> FA3 -> FA2 | FA1 is the local anchor selected by a recent call arrival  
(p) FA1 -> HA  
(1) MN -> FA2 -> FA1  
(2) MN -> FA3 -> FA1  
It is always the FA1 gets updated | FA3 is pre-selected according to the historical call arrival information  
(1) MN -> FA2 -> HA  
(2) MN -> FA3 -> HA  
(3) HA -> FA4 (replicates) |
| Request signaling upon CN arrivals | CN -> HA -> FA3 | CN -> HA -> FA1 -> FA2 -> FA3 | CN -> HA -> FA1 -> FA3 | CN -> FA4 -> FA3 |
| Comments | The operation will depend on the size of the pointer length k. When k=1 approximates to that of LAS | Selection of the anchor database | Saving in call setup delay if the replication database is in the vicinity of the calling mobile |
| Main issues | Selection of the pointer length | | Selection of the replicated database (number & locations) |
| Benefits (providing optimum tradeoff is obtainable) | Location updates at HA is reduced (saving in registration numbers ≪ k) | Location updates at HA is reduced (saving in number 1/∝ call arrival rate) | Call setup delay is minimized |
| Areas of applications | MNs encounter frequent movements | Call arrives from adjacent areas of MN's roaming area | Some distribution patterns are obtainable from analysis of the call arrivals (mainly geographical-wise) |
| Main downfalls | Triangle routing and also Long delay during location registrations | Suffers from longer delay during call setup (***) | Suffers from longer delay during call setup (**) | Additional control signaling when setting up the replicated information |

Note: The more * indicates a more severe delay is associated with the operation
9.5 THE PROPOSED DESIGN

By now, having determined what to implement, the aim of this section is to determine where and how to implement the designed location management techniques. A top-down approach is adopted in revealing the set of potential operations. To start with, it is logical to determine how to have the ideas of PFS and UPR actually included in the system operations. The goal will be to optimize the efficiency of the location tracking operations.

9.5.1 AN OVERVIEW

Clearly, by implementing the concepts of UPR and PFS as two alternatives of the same network layer, a multi-mode scheme becomes more promising in terms of increased flexibility. The key issue is to ensure that network structure allows dynamic “switching” to be made from one technique to another such that an optimal operation is achievable. Furthermore, it is also essential to design an operation which requires minimal signaling loads and computational requirements. To summarize, Figure 9.4 demonstrates a possible incorporation of PFS and UPR.

![Diagram showing Low CMR and High CMR thresholds with Low call arrivals and high mobility, and High call arrivals and low mobility with Pointer forwarding, Schemes with variable parameters, and Optimal replication]

**Figure 9.4** The design philosophy of a multi-mode location management technique

Basically, depending on the mobility and call arrival rate of the specific network under consideration, the operational characteristics can be divided into three major sections with details of each as listed.

At the one extreme where the call arrival rate outweighs the mobility rate, the application of a replication scheme provides the best performance. With the possible incorporation of load-balancing mechanisms into the available database, it is envisaged that the resultant query delay can be minimized. Generally, the replicated databases can employ either centralized or distributed database architecture. To optimize the effectiveness, different replication factors can be chosen for hot and cold mobiles. Finally it is noted that although replicated databases can be implemented in many locations, it is not necessarily essential to update all databases at any given time. The tradeoff of such flexibility, however, will experience additional
difficulties with inconsistent database management. Note that in all cases, the decisions are made on a per-user basis. However it is also possible to distribute the activity centers on a geographical basis when the distinct geographical conditions show such implementation to be promising.

At the other extreme when the transitioning rate of mobiles outweighs that of the call arrivals (i.e. the CMRs fall below a certain threshold), the implementation of PFS provides the necessary services to minimize the operational cost.

In cases where neither of these two extreme cases applies, the proposed design allows the system to divert to the mode where various parameters are adjusted to optimize the performance. In terms of the implementation, the same network entity will utilize the functions of both techniques combined in various proportions. As an example, given that the concept of UPR is implemented as a special database buffer, the concepts of PFS will be employed to determine how to have a selection made about the location to replicate the database. More details of such operations will be discussed in later sections.

In general, the transitioning between different operational "modes" (and thus location tracking procedures) can be invoked upon the detection of changes in network and mobile conditions. As a simple illustration, Table 9.4 lists some of the possible activation criteria for such mode changes.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>When certain &quot;localities&quot; are identified in the movement patterns</td>
<td>3</td>
</tr>
<tr>
<td>- Detect frequent forward/backward movements between certain LAS</td>
<td></td>
</tr>
<tr>
<td>When the activity state of a mobile changes</td>
<td>1</td>
</tr>
<tr>
<td>When the definition of network structure dependent parameters changes</td>
<td>2</td>
</tr>
<tr>
<td>- e.g. HA-HA signaling cost and FA-FA signaling cost</td>
<td></td>
</tr>
<tr>
<td>- Significant when deciding whether to update the HA or FA (database access delay and transmission cost)</td>
<td></td>
</tr>
</tbody>
</table>

In selecting the specific models for operation, the principle is to achieve an optimal tradeoff between the infrastructure cost and the flexibility measures.

9.5.2 SPECIFIC DESIGN DETAILS

In terms of sending location registrations, the operation would be more efficient if the roaming mobile is not asked to send updates to a possibly distant HA when some sort of "locality" movement is detected. In such a scenario, a "larger location managing area" can be dynamically formed to handle the registrations. Alternatively, it might be possible to give a
FA (assuming an IPv4 architecture) the authority to update the nearby replicated databases. Principally upon a call arrival, a forwarding action would only be needed if the query is made directly to HA. Otherwise, a simple table lookup at the local UPR will be all that is necessary.

In summary, as far as the implementation of UPR is concerned, there are mainly two aspects to be considered: (1) whether or not to register the movement at HA and (2) when an update is received at HA, where and when to perform the replicates. As was mentioned earlier, the locations to implement UPR need to be carefully considered in order to fully enjoy the benefits of reduced call delivery cost through a minimization of the replication cost. The key challenge is not only to maintain the required databases at a reasonable size, but also to restrict the resultant signaling load. Thus the goal is to achieve an optimal tradeoff between the operational cost (including all necessary feedback and computation load) and the controlling signals due to updates at the replicated databases.

Several possible options for replications were considered feasible for such purposes. A more in-depth description is now given for each of the proposals with a special focus on the feasibility of the applications.

The first option assumes that the control signaling is sent through HN to the replicated databases. The selected replicated databases can be either fixed or dynamic; when the locations of the replicated databases are fixed, the intended databases are pre-selected prior to the operations.

At one extreme, the most primitive form of the replicated scheme would be to have all databases updated when a replication is due. However, learning from the previous simulation work, there would be two main problems arising from such operations. Firstly, the amount of controlling signals required to update all the replicated databases becomes tremendous. Secondly, the amount of memory required at each of the replicated databases and also at HA will be enormous for maintaining the desired location information. As the proposed replications will push the desired infrastructure cost to be many times that of the actual subscribers, it is clearly not feasible to have the complete databases updated in all foreign domains.

It is obvious that any system improvements will involve a tradeoff between the infrastructure cost and the simplicity of system implementations. In particular, the method of database replication plays a crucial part in ensuring that the savings achieved in call setup delay and throughput will outweigh the combined cost of update signaling and memory requirement. Therefore, unless some sort of logical distribution mechanism (and hence load-balancing algorithm) is implemented, the implementation of such replication mechanism is not an optimized option.

Now, consider the other extreme where the number of replicated databases is limited to one. Without having to maintain complete replicated databases at all FDs, applications of
such a scheme provide a basic mechanism for describing the distribution of the location information. As demonstrated in Figure 9.5, this effectively bypasses the requirement of having large memory size for database replications.

![Diagram of location distributions]

**Figure 9.5** The concepts of location distributions

As each database maintains information only for a partial number of MNs, the size of individual local databases will be relatively smaller when compared to those of the original home network. As a consequence however, an increasing demand for reliable security associations will result in a growing difficulty in keeping track of each mobile’s current location. Furthermore, the improvement in performance will be significant only when the majority of calls are coming from the foreign domain where a replicated database was maintained.

Recall that the purposes of replication are two-fold: (1) to minimize possible database access delay and (2) to reduce the possibility of single point failures. Intuitively, the more distributive the replicated database, the greater the possibility of reducing the delay associated with connection setup. In the extreme case where only one replication is made, the most suitable place will be where most new calls seem to have arrived from. When such information is not available, the next compromise point will be to have the replicated databases maintained in the vicinity of the domain that the mobile is currently roaming into. In such a scenario, the database will not be implemented using load balancing algorithms. Figure 9.6 considers a scenario where global coverage can be realized by the co-existence of five domains.
It is to be noted here that the improvement in the replications will be particularly significant when a mobile is roaming in an area away from the home network. As a compromise, finding a tradeoff will be a feasible alternative to overcoming such a deficiency. For example, we could make the directly adjacent databases of the currently connecting FD updated by the implementation of appropriate authentication procedures. In this way, once a replication is required, HA selects the nearest UPR database to update the latest location information of the roaming mobile. Given UPR periodically advertises its role as a local "replicated" database, a correspondent node closer to the MN's current roaming area can then send a connection request directly to the replicated database. Consequently, by not having to go through the HA, it is envisaged that a shorter query delay will result, and thus prompt a faster call setup.

As mentioned earlier, the merits of implementing UPR are in avoiding unnecessary delays during the phase of call setup. Consequently, there arises a fundamental problem: what happens if the calls do not arrive from the area where the replicated databases are situated? In this case, in addition to a minimal benefit of implementing UPR, a waste of resource is introduced. Given the importance of the tradeoff, a more significant treatment is needed to identify the optimal locations for replications. Specifically, this is where an additional "AND" function will be advantageous for determining when and where to perform such replications. Figure 9.7 briefly illustrates such an idea.
Generally, two independent parameters will be enlisted to determine the distribution characteristics of the databases. While one input gives a sense of the time taken to perform the “replicate” action, the other indicates the location at which to implement the replicated database(s). This effectively simulates an optimized “distribution” scenario. Firstly, in selecting the times at which a replication is required, the usage of the “pointer length” concept could be useful. Specifically, by defining a threshold (whether in terms of time, movement, or distance) according to the roaming mobile’s characteristics, an estimation can be made of the best frequency for performing the replicates. Thus, given that the mobile status has changed, the pre-selected threshold can be dynamically adjusted to the rate at which replications are initiated, thus promoting a better operational flexibility in the system. For example, if a mobile is roaming frequently between boundaries of two subnets, a threshold in “time” might be the best choice as the rate at which replications are required can be most efficiently managed.

Secondly, based on the call arrival characteristics, including geographical distributions and call arrival rates, two operational alternatives are considered in determining the locations for implementing the replicated databases. The first alternative is to have several databases pre-assigned (and thus, have those implemented differently from the “common” databases) and to select the geographically “nearest” database to update. In this case however, although dynamics and adaptability are implemented, the operations might not be optimized. Specifically, as the locations of the pre-assigned databases become important, the resultant flexibility advantages could be more restricted. The second alternative is to have a fully “dynamic” scheme. In other words, all databases are capable of maintaining the location information without having to be implemented with distinct features (they simply behave like routers). Subsequently, the number and locations of replicated databases will be dynamically assigned depending on the MN’s movement. The selection process can be made as a variable parameter in response to the specific network scenario and user characteristics at any instant of time. However, while enjoying the benefits of being “totally” flexible, the introduced possible security flaws might be large. Thus, a careful analysis of the feasibility of such a feature is necessary to ensure that the additional concerns over security would not upset the system performances.

Finally, given the replication is done through HA, not only is the problem of single point failures still present, but additional intelligence is necessary at HA to determine the best location for replicating the databases. Although, the intelligence can be easily realized at the mobile levels (through the implementation of a counter of the call numbers), a high memory requirement is needed to record the locations for replicates.

It is thus necessary to clarify whether an update at home domain is in fact essential. It seems that other than giving more direct information to callers from mobile’s subscribed HN,
there are few other obvious advantages. As the calling MN will always have to access the location databases for the MN’s location anyway, it doesn’t seem necessary to access the MN’s HD as a result of implementing UPR. Meanwhile, there is evidence to suggest that further improvements could be made possible by cutting off the required number of updates at HA. As a potential option, it might be worth exploring the possibility of implementing a scenario where location databases will have the authorization to update other replicated databases. Hence, the replication could be done without having to have the signals sent through HA. A careful security management design will be required before any such implementations become possible.

In order to realize such implementation, it is however necessary to clarify what levels of security associations between subnets and administrative domains are considered to be acceptable. Specifically, would it be reasonable to implement a separate “layer” of network-controlling entities which somehow operate “independently” from the specific domain operators? Another issue to consider would be who has the authorization to update the network? It seems that if it were the home domain performing the replications, there would be less chance for false updates. However, given that multiple domains could be geographically far apart, it is debatable as to whether it is realistic to offer HD the sole authorization of updating which could easily result in high delay and signaling loads.

9.6 A PRELIMINARY FEASIBILITY STUDY

Before actually implementing the proposed tracking strategies, it is advantageous to conduct some preliminary analyses to quantify the feasibility of proposing this scheme in comparison to other existing developments in the literature. There are a number of possible ways to carry out an analysis to obtain such initial measures. As an example, while having all the possible mobile scenarios listed, the amount of signaling required to handle mobility can be evaluated and compared for each of the proposed alternatives.

In this section, the impacts that various network parameters have on system performances are determined using measurements taken from the number of location updates and the signaling loads due to call delivery. The evaluation is based on the assumption that the longer it takes to have the new location information set up, the greater the handoff cost will be in buffering the arriving packets at the old databases to preserve an ongoing connection. In addition, it is noted that in such a preliminary comparison, measurements do not include signaling loads for database replication because they would vary depending on the actual scenario under consideration. As far as the adaptability criterion is concerned, since a dynamic scheme will need to be incorporated, the evaluation of this particular criterion has also been left out for the time being.
Tables 9.5 and 9.6 show a preliminary calculation of the system performances. As was indicated in the tables, while location updates are measured in terms of numbers, the quantification of propagation delay is based on the number of traveled hops in registering the location update.

**Table 9.5** A comparison of the registration load

<table>
<thead>
<tr>
<th>Micro-mobility scenarios (inter-subnet movements)</th>
<th>Location updates</th>
<th>Propagation delay</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Local</td>
<td>Global</td>
</tr>
<tr>
<td><strong>Within HD</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subnets within HD</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Away from HD</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subnets between HD &amp; FD</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Subnets between attaching FD</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Subnets between FDs</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

In table 9.6, the call delivery load is a measure of the number of traveled hops before information becomes available to the caller. The varying values depend on whether the mobile is connecting to a PFS-aware database or if a direct consultation from HA is required.

**Table 9.6** A comparison of the call delivery load

<table>
<thead>
<tr>
<th>Distribution of the callers</th>
<th>Number of database accessed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Local</td>
</tr>
<tr>
<td><strong>Within HD</strong></td>
<td></td>
</tr>
<tr>
<td>MN’s subscribed HN</td>
<td>1</td>
</tr>
<tr>
<td>Different subnets within HD</td>
<td>$1-k^*$</td>
</tr>
<tr>
<td><strong>Away from HD</strong></td>
<td></td>
</tr>
<tr>
<td>MN’s attaching FN</td>
<td>1</td>
</tr>
<tr>
<td>Subnets from the same FD</td>
<td>$1-k$</td>
</tr>
<tr>
<td>Subnets from other FDs</td>
<td>1</td>
</tr>
</tbody>
</table>

$k$ represents the size of the pointer length

### 9.7 PERFORMANCE QUANTIFICATION MATRIX

In the light of the positive performance improvements from the preliminary analysis, the development of a more complete simulation platform would be advantageous to consider additional QoS parameters. Regarding future research work, examinations could be carried out to determine how well the actual performances comply with the anticipated performances and thus, to more precisely quantify the effectiveness of the proposed operations.

Extending from the design criteria listed in section 9.3, Table 9.7 gives a more technical interpretation of the QoS quantification in terms of the actual evaluation parameters. Some
of the criteria could also be determined from multiple measurements taken from various system aspects. Included also in the summary table is the level of difficulty at which individual measurements are obtained.

Table 9.7 A matching between QoS parameters and corresponding measurements

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Descriptions</th>
<th>Note†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scalability</td>
<td>Relation between database size and number of roaming mobiles</td>
<td>1</td>
</tr>
<tr>
<td>Backward-compatibility</td>
<td>Stand-alone operations/Implementations as an extension</td>
<td>3</td>
</tr>
<tr>
<td>Simplicity</td>
<td>Minimization of the infrastructure cost and implementation complexity</td>
<td>2</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delay</td>
<td>Propagation delay and database access delay</td>
<td>2</td>
</tr>
<tr>
<td>Signaling delay</td>
<td>Minimization of signaling load</td>
<td>1</td>
</tr>
<tr>
<td>Packet loss</td>
<td>Dropped packet per handoff</td>
<td>1</td>
</tr>
<tr>
<td>Adaptability</td>
<td>Dynamic response to characteristics of the existing systems</td>
<td>3</td>
</tr>
<tr>
<td>Security</td>
<td>Minimize possible security flaws</td>
<td>1</td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location privacy</td>
<td>Maintain confidentiality of user profiles†</td>
<td>3</td>
</tr>
<tr>
<td>Reliability</td>
<td>Avoid single point of failures</td>
<td>2</td>
</tr>
<tr>
<td>Support of mobile networks</td>
<td>Support not only the MNs but also the whole mobile network that attaches to it</td>
<td>3</td>
</tr>
</tbody>
</table>

† The difficulty increases with ascending numbers
‡ Difficult to achieve complete location privacy with optimal routing.

Also as part of any future work, the effectiveness of PFS and UPR need to be more precisely quantified. In fact, it might be beneficial to provide some quantitative measures as to how fast the performance would degrade in response to various MN's behaviors and network characteristics in terms of QoS metrics. Thus, in taking various system measurements, instead of quantifying the performances by one precise value, the performances would be compared in terms of ranges. As an illustration, Table 9.8 briefly lists some of the possible parameters to be considered in conjunction with their contributing variables. By categorizing the design criteria into vendors', service providers' and users' perspectives, high level performance expectations are matched to their corresponding lower-level QoS parameters [82,83]. This table will be particularly helpful when a more comprehensive set of parameters is selected for quantifying the effectiveness of the proposed location tracking alternatives.
Table 9.8  A list of the ultimate system requirements

<table>
<thead>
<tr>
<th>Performance Expectations</th>
<th>QoS Parameters</th>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vendors view</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complexity</td>
<td>▪ Sensitivity of complexity</td>
<td></td>
</tr>
</tbody>
</table>
| Cost effectiveness       | ▪ Location updates (up to HN)  
                         | ▪ Location updates (within FN)  
                         | ▪ Host specific routing updates (within FN)  
                         | ▪ Handover protocol cost |  
                         |                                                                  |
| Scability                | ▪ Scale to support hundreds of millions of subscribers |              |
| Service providers         |                |              |
| Seamless mobility        | ▪ MH’s ability to receive from multiple BSs  
                         | ▪ The ability of quick switching to the new route |  
                         |                                                                  |
| Reliability              | ▪ Rendezvous time | The elapse period between the time a MN roams out of the coverage area of its old BS and the time its new BS received the setup update information |
| Robustness               | ▪ Call dropping probability | Lack of location information |
| Signaling delay reduction| ▪ Update latency  
                         | ▪ Handoff latency | Support of real-time service applications |
| Total Packet loss        | ▪ Dropped (& lost) packets per handoff versus playout delay | During the time location information is updated in the controlling network entity |
| Customers                |                |              |
| Service quality          | ▪ QoS compliance |              |

Other requirements such as having the algorithm adaptive to network parameters should be fulfilled. Specifically, parameters that need constant online feedback would include the mobility rate and the state transition rates (i.e. the frequency at which a mobile alters its activity status). A geographical distribution of call arrival characteristics is also important to determine the optimal locations for implementing databases. For this purpose, the usage of counters at the terminal level could be a possible implementation alternative. Lastly and most importantly, the scheme should be simple to implement before it is a feasible choice for use in real-life mobile networks.
9.8 CONCLUSIONS

As explained earlier, the purpose of the second part of the thesis is to propose an alternative solution to Mobile IP in managing intra-domain movements, where a mobile is assumed to roam between subnets that are managed by the same administrative domain. It is acknowledged that there are two major problems associated with the existing operations of Mobile IP that require further consideration. Firstly, large bandwidth consumption is evident when the roaming mobile is closer to the CN than to its corresponding HA. The impact thus magnifies a waste of resource which becomes even less tolerable when dealing with small data packet transmission. Secondly, there are reliability issues arising from the great signaling loads that are imposed on HA due to a mobile's frequent movement.

In this chapter, several alternative tracking solutions were proposed based on the incorporation of similar operational concepts utilized in the cellular-based techniques. To solve inefficient operations, it is necessary to look into the possibilities of replicating user's location information at multiple databases. In other words, before a query is sent to the HA by the calling mobile, a search should be made to determine the availability of a nearer local database (i.e. the applications of user profile replication).

As for the described reliability issue, an alternative strategy is to reduce, or even abandon if possible, updates at HLR. By doing so, the risk of having a complete malfunctioned system would be reduced to a minimum as the sole reliance on any one particular database was distributed among several alternatives — an achievement underlying the application of the proposed design.
To burn always with this hard, gemlike flame
To maintain this ecstasy
Is success in life

Walter Pater (1839-1894)

Chapter 10
Summary of Contributions

Developing an efficient location management technique is an important step towards an optimal solution to managing mobility. Given the irregular nature of cell sizes in a cellular network, the behavior of mobile movement changes from cell to cell and from user to user. Thus, the need for designing an adaptive algorithm for tracking a roaming mobile becomes imperative. The biggest challenge in framing location management is to find the most favorable tradeoff between the location updates load and the searching load—the two parameters that are frequently referred to as location registration cost and call delivery cost.

In this thesis, specific strategies were developed to enhance the operations of the existing location tracking mechanisms. Alongside a growing interest in other wireless technologies, the need for efficient operation in personal communications services (PCS) networks is actively present, given that the number of subscribers has been continuously increasing. Thus, in the first part of the thesis, the focus was on improving existing operations in the cellular networks. After a detailed study of the literature, two adaptive updating strategies were proposed in addition to a corresponding paging scheme. Moreover, as far as the Markov model is concerned, a simplified model was also proposed to reduce the number of states required in simulating such memoryless movements. In order to realize the ultimate goal of providing an anytime/anywhere multimedia communication environment to an ever
increasing number of mobile subscribers, in the second part of the thesis, a more detailed study was performed on the IP networks. New location tracking strategies were proposed to improve the deficiencies that are inevitable in the existing Mobile IP operations. This chapter summarizes the main contributions of the research work described in this thesis.

Firstly, in order to capture the full transitioning characteristics of a roaming mobile following Markov movement, early research studies showed that a large number of $3D^2 + 3D - 5$ states would be necessary for a $D$-layer cluster of cells. In this thesis, a new model was developed, and the number of states can now be reduced to $2(D-1)$. Clearly, with the development of the new model, the computational complexity required to model the transitional probabilities significantly decreases. It is proposed that the new random walk model will find applications in cellular networks mainly to determine the number of units that the mobile visits before moving out of the location area. For example, it can be used to model the cell residence time distributions in macro-cells overlapping with micro-cells. It is thus possible to determine the number of smaller routing areas that the mobile visits before it moves to a new location area.

The main purpose of update is to restrict the area of paging upon call arrivals. It is expected that the update boundary is defined such that it reflects the network's confidence in locating the roaming mobile. It becomes apparent that the assignment of location areas should not only be specific to each mobile's traveling rates, but also be sensitive to their actual movement patterns. However, despite its importance, there seems to be little research on the potential of utilizing such direction information in the existing literature.

With the incorporation of a transitional directivity index, simulation results confirmed that the selection of an optimal threshold in the sphere of location management should no longer be a function of only the call-to-mobility ratio. As far as mobility characteristics are concerned, the actual transitional direction of roaming mobiles also plays a significant role in selecting the optimal threshold particularly for distance-based update schemes. The directivity index, the newly introduced parameter that is capable of providing measures of the mobile's traveling patterns, has successfully demonstrated its ability in determining an optimal update threshold. Given that it is difficult to reliably anticipate call arrival characteristics, the proposed scheme is of particular importance. Its applications are even more significant when the theoretically determined "ideal" optimal threshold is not obtainable due to certain restrictions imposed by the network during times of high system loading. Simulation results reveal that the additional information made available about a roaming mobile's transitional directivity is critical to ensure that the best available sub-optimal threshold is realizable.

However, even with the introduction of a directivity index, the range of mobility patterns that can be simulated would still be restrictive with the assumption of a Markov movement
model. In addition, while the size of the threshold is selected adaptively to the mobile’s traveling directions, the shape of the location area remains symmetrical to the last update cell.

A second proposal, the Kalman-filter based location tracking scheme, was developed to overcome such a deficiency. Not only is the Gauss Markov model used to simulate the varying degrees of correlation between the mobile’s traveling directions, the application of a Kalman filter allows predictions to be made about the mobile’s future movement patterns from its past roaming history. Most importantly, as the shape of the update area is now assigned asymmetrical dimensions, the level of certainty the network has about a mobile’s residency is reflected by the dimensions of the registration boundary.

It is to be emphasized that the complete tracking operation is comprised of two processes; that of the update and of the paging. However, given a unit update costs more than a unit paging, while much research effort has been evident in the development of updating schemes, the problem of paging has received little attention. As a result, with the existing paging schemes, paging cost increases when the update frequency decreases; more specifically, reductions in the update load are often compromised by the increasing paging load.

Having proposed two tracking algorithms, work then focused on the development of a corresponding paging scheme. Clearly, with directional information taken into consideration, it is not realistic to assume equal likelihood of residence in all cells of the location area as is the usual assumption in the literature. The fundamental idea was to develop a new paging strategy such that even taking into account the effects of traveling characteristics when assigning paging areas, the proposed strategy should remain simple; simple to maintain (i.e. minimal complexity) and simple to implement (i.e. minimal additional intelligence). The development of sectional paging achieves exactly that; without having to obtain detailed traveling patterns, a granular direction identified is sufficient to determine the section in which the network is most confident of locating the roaming mobile. Simulation results confirm satisfactory performances. In comparison to the basic shortest-distance-based paging, the sectional paging scheme showed superior improvements with significant reductions in the operational cost.

So far, all of the above mentioned tracking strategies were proposed with applications to cellular networks in mind. With simulations confirming satisfactory performances, in the second part of the thesis, an attempt was made to explore the possibilities of extending cellular-applications onto IP networks.

Mobile IP has gained significant attention since its initiation in 1993 for its capability of handling mobile movements between heterogeneous networks. However, the current operation requires a registration to be performed at the home network for every movement that involves a change of IP address while in communication. Given that the lowest network entity in an all IP-based system is an IP-based access point (or base station), such “change of
IP address” will incur at every change in network point of attachment. With mobility no longer an exception but a criterion, it is likely that frequent movements would occur in an area far from the mobile's subscribed home network. As a result, not only is the complex signal transferring unacceptable, but large delays associated with the operations of registration and call setup are all important indications of the need for a better location management alternative. Foreseeing the presence of some fundamental problems with the soon-to-be standardized protocol, the goal was to develop an alternative that would enable mobility at the local level to be more efficiently managed.

Preliminary analysis of the existing solutions indicates a great resemblance in the basic design criteria between cellular systems and IP networks. Although additional problems and design criteria will emerge, for example, increasing security concerns in the operation of an all-IP network, in terms of location management, the underlying framework is the same. The questions come down to when, where, and how to perform location update. However slightly different from the first part of the thesis where the focus was on when to perform an update, the question that requires more attention in IP networks is where to send the registrations.

It appears promising to re-apply some of the well-developed techniques from PCS networks onto operations of IP networks. In this way, while achieving an optimal efficiency, the applicability of the protocol is stretched to a maximum. The fundamental idea is to determine whether registrations at the home agents can provide an optimal performance; and if not, how to enhance the existing operations with minimal signaling transfers.

Clearly, without a common evaluation platform, it is difficult to consider whether one scheme would actually outperform the other. Simulation work was thus carried out to examine and identify the main characteristics of the potential strategies. Among other measures, individual schemes were evaluated along the lines of registration delay, query delay and database access delay, and results were compared with those obtained from registrations at HA.

With the clear feasibility of PFS and UPR, potential location management schemes specifically designed for IP applications were proposed in this thesis. While a more detailed design is still required for conducting a more comprehensive evaluation, preliminary analysis suggests positive improvements with the proposed strategies. To prompt continuation of this research work, an overview of the design criteria was derived in addition to a matching set of quantification parameters. It is envisaged that actual implementation of such schemes will enable a more precise quantification of the performance improvements. Although further analysis is beyond the scope of the thesis, the initial proposal outlined in this research is novel. Subsequently, extensions from this thesis work are likely to lead to promising discoveries.

Finally, although this thesis has given substantial analysis to the location tracking aspects of mobility management, a complete solution to the challenge of managing mobility is still
unfinished. Issues such as security management and billing strategies all require more attention from the research community as a whole. With the problem having been re-defined, the way is clear for future research work enhancing other aspects of the managing process with the aim of achieving the ultimate goal of any/anywhere multimedia communications.
Bibliography


Appendix A

An overview of the simulation philosophy used in Chapter 8

Implement different mobility rate (ranging from 0.000031 to 1)

Path loop

Take measurements at different times

Time loop

- Determine the location of MN
  > {xCoordinate, yCoordinate}
  > cell number/VLR number

macro-mobility

- cell\textsubscript{new} \neq cell\textsubscript{old} ?

Micro-mobility

no

Paging Loads

PA\textsubscript{new} = PA\textsubscript{old} ?

yes

The GPRS scheme

no

LA\textsubscript{new} = LA\textsubscript{old} ?

yes

no

GSM/GPRS

UPR

PFS

LAS

Location Updates

Database Query Loads
Appendix B

The generation of various call arrival rate for LAS

\[ \text{For } (\text{time}\_1 = 0; \text{time}\_1 \leq \text{TIME}; \text{time} = \text{time} + 200) \]

<table>
<thead>
<tr>
<th>Time loop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determine the location of the serving VLR</td>
</tr>
<tr>
<td>[ [\text{xCoordinateVLRserve}, \text{yCoordinateVLRserve}] ]</td>
</tr>
<tr>
<td>VLRserve number</td>
</tr>
</tbody>
</table>

\[ \text{VLRServe} = \begin{cases} \text{VLRServe}\_\text{old} \text{ if yes} \\ \text{no} \end{cases} \]

Update VLRla
\[ \text{VLRLa} = \text{VLRServe} \]

Update HLR
\[ \text{HLR\_Info = VLRLa} \]
\[ (++ \text{rcvd\_updates\_VLRLa}) \]

Calculate distance between VLRLa and HLR!
\[ \text{(total\_distance\_VLRLa} \]
\[ += \text{update\_distance\_VLRLa}) \]

// make the serving VLR // (when the last call // arrived) to be the local // anchor
Appendix C

The implementation of LAS

**Time loop**

- Determine the location of MN
  > \([x\text{Coordinate}, y\text{Coordinate}]\)
  > VLR number

**For** \((time=0; \ time<=TIME; \ time=time+100)\)

- \(\text{speed} \times \text{distance} / \text{cellsize}\)

**yes**
- \(\text{VLR}_{\text{new}} = \text{VLR}_{\text{old}}?\)
  - // MH stay in the same LA

**no**

- \(\text{VLR}_{\text{R}}a = 0?\)

**yes**
- Update HLR
  - \(\text{HLR}_\text{Info} = \text{VLR}\text{number} \ (++ rcvd\_updates) \odot\)

**no**

- Update VLRla
  - \(\text{VLR}_{\text{la}}\_\text{Info} = \text{VLR}\text{number} \ (++ VLR_{\text{la}}\_\text{updates})\)

- Calculate distance between \(\text{VLR}_{\text{new}}\) and \(\text{VLR}_{\text{la}}\)!
  - \((\text{total\_distance} + = \text{update\_distance}) \oplus\)

**Total update distance**
- \((\text{distance\_total} = \text{total\_distance} + VLR_{\text{la}}\_\text{total\_distance})\)

**Output parameters (for one path)**

1. Updates to HLR \((\text{rcvd\_updates} \_\text{HLR}\_\text{total}) \oplus \odot\)
2. Total updates (at HLR & VLRla) \((\text{total\_Updates\_onopath})\)
3. Distance to HLR \((\text{distance\_to\_HLR}) \odot \oplus\)
4. Total update distance (to HLR & VLRla) \((\text{distance\_total\_update})\)
Appendix D

The calculation of database querying load for LAS

\[ \text{For } (\text{time}=0; \text{time}<\text{TIME}; \text{time}+%100) \]

\[ \text{VLR}_{\text{new}} = \text{VLR}_{\text{old}}? \]

\[ \text{no} \]

++ \text{VLR}_{\text{crossing}}

\[ \text{VLRLa} = 0? \]

\[ \text{yes} \]

Access HLR only
(i.e. Accessload = 1)
(++accessload\_HLR)

\[ \text{no} \]

Access VLRLa
(++accessload\_VLRLa)

Average number of access load (for one path)

\[ \text{average\_accessload\_onopath} = (\text{accessload\_HLR} + \text{accessload\_VLRLa})/\text{VLR}_{\text{crossing}} \]

Average number of access load (averaged over all paths)

\[ \text{average\_accessload\_allpath} = (\text{total\_accessload\_allpath})/(\text{pathNo}) \]

\[ +/- \text{Average\_accessload\_onopath} \]

Note that Access load is defined as the number of databases being accessed before the current serving VLR is found (e.g. for GSM: the access load = 1, being the HLR)
Appendix E

The calculation of database querying load for UPR

For (time=0; time<=TIME; time=time+100)

- Determine the location of MN (speed * distance / cellsize)
- [xCoordinate, yCoordinate]
  - VLR number

Time loop

VLR_{new} = VLR_{old} ?

yes

// MH stay in the same LA

no

Update HLR
HLR_Info = VLRNumber
(++) rcvd_updates

Calculate distance between VLR_{new} and HLR!
(total_distance
  += update_distance)

Update VLRmd[PRno]

for (k=0; k<PRno; k++)
VLRmd_Info[k] = HLR_Info

distance between HLR and VLRmd[k] (total_distanceVLRmd)
(++) updateVLRmd

Output parameters (for one path)

1. Updates to HLR (rcvd_updates)
2. Total updates (to all VLRmd ≠ VLRnumber) (updateVLRmd)
3. Distance to HLR (total_distance)
4. Total update distance (to all VLRmd ≠ VLRnumber) (total_distanceVLRmd)
Appendix F

An Illustration of Querying Delay Reductions for UPR

<table>
<thead>
<tr>
<th>Call</th>
<th>(nearest) VLRmd</th>
<th>VLRmd_query_dis</th>
<th>HLR_query_dist</th>
<th>(HLR-VLRmd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VLRcalling</td>
<td>VLRmd_query_dis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>13</td>
<td>12</td>
<td>60000.000000</td>
<td>258069.758011</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>2</td>
<td>84852.813742</td>
<td>228473.193176</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>12</td>
<td>60000.000000</td>
<td>258069.758011</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>10</td>
<td>60000.000000</td>
<td>212132.034356</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>60000.000000</td>
<td>94868.329805</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>0.000000</td>
<td>152970.585408</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>10</td>
<td>0.000000</td>
<td>152970.585408</td>
</tr>
<tr>
<td>7</td>
<td>19</td>
<td>14</td>
<td>60000.000000</td>
<td>342052.627530</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>12</td>
<td>0.000000</td>
<td>212132.034356</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>10</td>
<td>60000.000000</td>
<td>174928.556845</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>14</td>
<td>120000.000000</td>
<td>381837.661841</td>
</tr>
<tr>
<td>11</td>
<td>17</td>
<td>12</td>
<td>60000.000000</td>
<td>258069.758011</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td>14</td>
<td>60000.000000</td>
<td>284604.989415</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>10</td>
<td>0.000000</td>
<td>152970.585408</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>42426.406871</td>
<td>42426.406871</td>
</tr>
<tr>
<td>15</td>
<td>18</td>
<td>12</td>
<td>84852.813742</td>
<td>296984.848098</td>
</tr>
<tr>
<td>16</td>
<td>14</td>
<td>14</td>
<td>0.000000</td>
<td>308868.904230</td>
</tr>
<tr>
<td>17</td>
<td>8</td>
<td>2</td>
<td>84852.813742</td>
<td>228473.193176</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>12</td>
<td>0.000000</td>
<td>212132.034356</td>
</tr>
<tr>
<td>19</td>
<td>16</td>
<td>10</td>
<td>84852.813742</td>
<td>228473.193176</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>2</td>
<td>60000.000000</td>
<td>212132.034356</td>
</tr>
<tr>
<td>21</td>
<td>20</td>
<td>10</td>
<td>120000.000000</td>
<td>271661.554144</td>
</tr>
<tr>
<td>22</td>
<td>11</td>
<td>10</td>
<td>60000.000000</td>
<td>174928.556845</td>
</tr>
<tr>
<td>23</td>
<td>17</td>
<td>12</td>
<td>60000.000000</td>
<td>258069.758011</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>2</td>
<td>120000.000000</td>
<td>271661.554144</td>
</tr>
<tr>
<td>25</td>
<td>6</td>
<td>2</td>
<td>84852.813742</td>
<td>127279.220614</td>
</tr>
</tbody>
</table>

Average querying distance traveled (to VLRmd) = 6.034629 cells
Average querying distance traveled (to HLR) = 22.56087 cells
Average savings in querying distance (per call) = 16.52628 cells
# Glossary of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARP</td>
<td>Address resolution protocol</td>
</tr>
<tr>
<td>BU</td>
<td>Binding updates</td>
</tr>
<tr>
<td>BSS</td>
<td>Base station subsystem</td>
</tr>
<tr>
<td>CCoA</td>
<td>Collocated care-of-address</td>
</tr>
<tr>
<td>CMR</td>
<td>Call-to-mobility ratio</td>
</tr>
<tr>
<td>CN</td>
<td>Corresponding node</td>
</tr>
<tr>
<td>CRT</td>
<td>Cell residence time</td>
</tr>
<tr>
<td>DHCP</td>
<td>Dynamic host configuration protocol</td>
</tr>
<tr>
<td>EDGE</td>
<td>Enhanced data rate for GSM evolution</td>
</tr>
<tr>
<td>FA</td>
<td>Foreign agent</td>
</tr>
<tr>
<td>FACoA</td>
<td>Foreign agent care-of-address</td>
</tr>
<tr>
<td>GGSN</td>
<td>Gateway GPRS support node</td>
</tr>
<tr>
<td>GPRS</td>
<td>General packet radio services</td>
</tr>
<tr>
<td>GSM</td>
<td>Global system for mobile communications</td>
</tr>
<tr>
<td>GTP</td>
<td>GPRS tunneling protocol</td>
</tr>
<tr>
<td>HA</td>
<td>Home agent</td>
</tr>
<tr>
<td>HLR</td>
<td>Home location register</td>
</tr>
<tr>
<td>IETF</td>
<td>Internet engineer task force</td>
</tr>
<tr>
<td>IMT</td>
<td>International mobile communications</td>
</tr>
<tr>
<td>IP</td>
<td>Internet protocol</td>
</tr>
<tr>
<td>LA</td>
<td>Location area</td>
</tr>
<tr>
<td>LAS</td>
<td>Local anchoring scheme</td>
</tr>
<tr>
<td>MA</td>
<td>Mobility agent</td>
</tr>
<tr>
<td>MH</td>
<td>Mobile host</td>
</tr>
<tr>
<td>MN</td>
<td>Mobile node</td>
</tr>
<tr>
<td>MS</td>
<td>Mobile station</td>
</tr>
<tr>
<td>MSC</td>
<td>Mobile switching center</td>
</tr>
<tr>
<td>PC</td>
<td>Paging cache</td>
</tr>
<tr>
<td>PCS</td>
<td>Personal communications system</td>
</tr>
<tr>
<td>PDP</td>
<td>Packet data protocols</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>PFS</td>
<td>Pointer forwarding scheme</td>
</tr>
<tr>
<td>RC</td>
<td>Routing cache</td>
</tr>
<tr>
<td>RA</td>
<td>Routing areas</td>
</tr>
<tr>
<td>SGSN</td>
<td>Serving GPRS support node</td>
</tr>
<tr>
<td>TLLI</td>
<td>Temporary logical link identity</td>
</tr>
<tr>
<td>UMTS</td>
<td>Universal mobile telecommunications system</td>
</tr>
<tr>
<td>UPR</td>
<td>User profile replication</td>
</tr>
<tr>
<td>UTRAN</td>
<td>the UMTS terrestrial access radio network</td>
</tr>
<tr>
<td>VLR</td>
<td>Visiting location register</td>
</tr>
</tbody>
</table>