Location Management and Network Architecture Design for GSM

By
SEIAMAK VAHID

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey

Unis

Centre for Communication System Research
School of Electronic Engineering, Information Technology and Mathematics
University of Surrey
Guildford, Surrey GU2 7XH, UK

October 2000
Summary

Recent years have witnessed a tremendous growth of research and development to provide mobile users with "seamless" communication through wireless media. Service provision in mobile networks is closely intertwined with network mobility management and therefore efficient management of subscriber mobility remains a challenging and important area of research. This thesis examines management of signalling traffic in mobile networks and packet network design approaches, with specific contributions relating to i) reduction of network database signalling costs in GSM; ii) minimization of mobility signalling over the air-interface based on a new technique for subscriber location management; iii) an efficient network design technique for packet communication networks with application to GPRS, based on a novel hybrid GA-Heuristic approach.

A novel location management technique is proposed to reduce signalling costs in the core network, by reducing the rate of HLR-related transactions. By intelligent distribution of the HLR related signalling over a number of VLRs, the HLR is prevented from being a potential bottleneck and a single point of failure in the signalling network. Based on the analytical model developed, it is demonstrated that within the call-to-mobility range considered, the total network signalling cost as well as location management costs can be significantly reduced, compared to current strategy adopted in GSM.

To address location management signalling in the access network, a number of dynamic schemes are considered and compared with the proposed adaptive multilayer technique. Through detailed simulations under various scenarios, the superiority of the proposed scheme, in terms of significant savings in the total signalling traffic (i.e. location update and paging) and ease of implementation, compared to other techniques is demonstrated. The location management scheme of GSM is used as baseline for comparisons.

Finally, efficient design of packet communication networks, with application to GPRS backbone architecture, is addressed. To meet the cost and traffic requirements whilst ensuring that the delay and reliability constraints are also satisfied, a hybrid GA-heuristic approach is developed as an alternative to purely GA based and heuristic-only approaches. Under a common set of parameters, the performance of various techniques are compared and it is shown that the hybrid approach is capable of producing solutions that are superior to other typical network design methods.
Acknowledgments

I would like to express my thanks and gratitude to my supervisors Prof. R. Tafazolli and Prof. B.G. Evans, of CCSR for their continuous support during the course of this work. I am also grateful to former and present members of the centre for their friendship, lively discussions and insightful comments. I am also indebted to the many of the staff at the center. Special thanks go to my family and parents in particular, for their encouragement, kindness and patience throughout the years. This thesis is dedicated to them.

Seiamak Vahid

October 2000
# Table of Contents

Summary ........................................... ii  
Table of Contents ................................ viii  
List of Figures ..................................... vii  
List of Tables ...................................... xi  

## CHAPTER 1 Introduction  
1.1 Motivation and Objectives 13  
1.2 Main Contributions 16  
1.3 Organization of the Thesis 17  

## CHAPTER 2 Signalling Traffic in GSM Networks  
2. Introduction ................................ 20  
2.1 Signalling Across the User Interface (Um) 21  
2.1.1 Layer 1 Signalling (MS-BTS interface) 21  
2.1.2 Layer 2 Signalling 23  
2.1.3 Layer 3 Signalling – RR, MM and CM 24  
2.2 Signalling Transport Protocols in GSM 26  
2.2.1 Message Transfer Part 26  
2.2.2 Signalling Connection Control Part 27  
2.2.3 Transaction Capabilities Application Part 28  
2.2.4 Integrated Services Digital Network User Part 29  
2.2.5 Mobile Application Part 29  
2.3 Mobility Management and SS7 Traffic in GSM Networks 30  
2.4 Analysis of SS7 Signalling Load in Cellular Networks 35  
2.4.1 Mobility Model 37  
2.4.2 Determination of Optimum Location Area Size 37  
2.4.3 Transaction Rates 39  
2.4.4 Total Signalling Load 41  
2.5 Radio Resource Occupancy Due to Mobility 44  
2.5.1 MSC/VLR Mobility Related Transaction Load 45  
2.6 Concluding Remarks 46  

## CHAPTER 3 The Anchoring Scheme for Reducing Signalling costs in GSM Networks  
3. Introduction ................................ 48  
3.1 Literature Survey 50  
3.2 Reference Network Architecture 58  
3.3 Anchoring Technique in GSM 59  
3.3.1 Location Registration and Call Delivery 59  
3.3.2 Detach/Attach Procedures 61  
3.4 Cost Evaluation Procedures 62  
3.4.1 Determination of the Call Routing Cost 63  
3.4.2 Determination of Location Updating Cost 64  
3.4.3 Determination of Attach/Detach Cost 65  
3.5 Analytical Model of the Basic Anchoring Technique 66  
3.6 Analytical Model of the Modified Anchoring Technique 67  
3.6.1 Determination of the Signalling Costs in GSM 70  
3.6.2 Total Cost of the GSM Scheme 71  
3.7 Analytical Results of the Basic Scheme 71  
3.7.1 Database Impacts 74  
3.8 Optimization of the Basic Anchoring Technique 75
Table of Contents

3.8.1. Results of the Proposed Optimisation Procedure 79
3.8.2. Database Implications 81
3.9. Conclusions 83

CHAPTER 4 Performance Evaluation of Location Management Techniques

4. Introduction 86

4.1. Literature Survey 89
4.1.1. Location Update and Paging Schemes 90
4.1.2. Memory-based Methods 96
4.2. Performance Comparison of LA Optimization Techniques 102
4.3. Simulation Parameters and Assumptions 110
4.4. Numerical Results 111

4.4.1. Results with FF movement (All Vehicular) – AVFF 111
4.4.2. Results with FF movement (All Vehicular) – AVFF w/RIP 114
4.4.3. Results with FF movement (Pedestrians & Vehicular) – PVFF 115
4.4.4. Results with FF movement (Pedestrians & Vehicular) – PVFF w/RIP 117
4.4.5. Results with MG movement (All Vehicular) – AVMG 119
4.4.6. Results with MG movement (All Vehicular) – AVMG w/RIP 121
4.4.7. Results with MG movement (Pedestrians & Vehicular) – PVMG 123
4.4.8. Results with MG movement (Pedestrians & Vehicular) – PVMG w/RIP 125
4.4.9. No. of Layers and Threshold in the Adaptive Multilayer Scheme 127

4.5. Summary of Results 129

4.5.1. Over the BH Period 129
4.5.1.1. The AVFF case 129
4.5.1.2. The AVMG case 130
4.5.1.3. The PVFF case 130
4.5.1.4. The PVMG case 130
4.5.2. Over the 24Hr. Period 131

4.6. Conclusions 131

CHAPTER 5 A Hybrid GA-Heuristic Approach to Packet Network Design

5. Introduction 136

5.1. Literature Review 141

5.1.1. Decomposition of the Network Design Problem 142
5.2. Design of Mesh Communication Networks 143

5.2.1. Requirements, Constraints and Costs 143

5.2.1.1. Traffic Requirements 144
5.2.1.2. Routing Policy 144
5.2.1.3. Link Flows 145
5.2.1.4. Communication Cost 145
5.2.1.5. Delay Analysis 146
5.2.1.6. Capacity Constraint 147
5.2.1.7. Reliability Considerations 147

5.3. A 10-Node Network Design Problem 147

5.3.1. Design Parameters 149
5.3.2. The MAN-GA Approach 150
5.3.3. The Proposed Approach 153

5.3.3.1. Main GA process – Topology Optimisation 155
5.3.3.2. Mapping Objective Values to Fitness Form 156
5.3.3.3. Selection Scheme 158
5.3.3.4. Crossover 158
5.3.3.5. Mutation 159
5.3.3.6. Fitness Evaluation 159
5.3.3.7. Reliability Considerations – Biconnectivity 160
5.3.3.8. Implementation of the Proposed Technique 162
5.4. Numerical Results

5.4.1. GA Parameter Setting and Process Performance

5.4.2. Comparison With Other Methods

5.5. Conclusions

CHAPTER 6 Conclusions and Future Work

6. Research Contributions

6.1. Future Research Directions

CHAPTER 7 Bibliography & References

Appendix A – GSM Signalling Message Charts & SS7 Tables

Appendix B – Modelling Communication Networks

Appendix C – The Bertsekas-Gallager Algorithm

Appendix D – On Genetic Algorithms
List of Figures

Figure 2-1: Layer 1 service interfaces ................................................................. 21
Figure 2-2: Location update procedure .............................................................. 32
Figure 2-3: Call delivery procedure ................................................................. 33
Figure 2-4: SS7 location update traffic ............................................................. 37
Figure 3-1: Per-user location caching scheme .................................................. 51
Figure 3-2: Pointer forwarding strategy .......................................................... 53
Figure 3-3: Distributed hierarchical tree-based database architecture ............. 55
Figure 3-4: Registration in the distributed hierarchical database architecture .... 56
Figure 3-5: Partitioning scheme ........................................................................ 57
Figure 3-6: Distributed database architecture ................................................ 58
Figure 3-7: Reference network architecture .................................................... 58
Figure 3-8: (a) Location registration and (b) Mobile Terminated (MS) call in the proposed scheme .... 59
Figure 3-9: signalling exchange during DETACH in GSM .............................. 61
Figure 3-10: signalling exchange during DETACH in the anchoring technique .... 61
Figure 3-11: Embedded markov chain model .................................................. 66
Figure 3-12: The modified markov chain model .............................................. 67
Figure 3-13: (a) Call routing cost (MM: 20% - FM: 80%) (b) Call routing cost (MM: 80% - FM: 20%) ................................................................. 72
Figure 3-14: (a) HLR access cost , (b) VLR access cost \( \lambda_A = 0.1 \cdot \lambda_C \) ................................. 74
Figure 3-15: Call routing vs location update cost : (a) and (b) set 4 , (c) and (d) set 6 results .................. 76
Figure 3-16: Call routing vs location update cost : (a) and (b) set 4 , (c) and (d) set 6 results .................. 77
Figure 3-17: Optimised (a) Call routing (b) Location updating and ...................... 80
Figure 3-18: (a) HLR access cost - set 1 , (b) VLR access cost - set 1 ................. 82
Figure 4-1: (a) Time-Based Location Update Scheme ........................................ 90
Figure 4-2: Distance-Based Update Scheme .................................................... 91
Figure 4-3: LA overlapping example with Okasaka's method ......................... 94
Figure 4-4: Location update strategy of [Hu95] ................................................. 97
Figure 4-5: Illustration of classical registration strategy with fixed location areas. Small white circles represent location updating ................................................ 98
Figure 4-6: Depiction of personal location areas, recalculated after every contact with the system. The shape of LA may change in time, depending on mobility & call characteristics of the user ........ 98
Figure 4-7: The structure of decision technique ................................................. 106
Figure 4-8: Trajectory of a mobile using the distance-based strategy ............... 107
Figure 4-9: The structure (pseudo code) of the adaptive multilayer technique .... 109
Figure 4-10: Signalling cost vs. velocity (km/hr.) (a) 2 calls/hr. (b) 6 calls/hr. ........................................ 112
Figure 4-11: Signalling cost vs. call rate (a) 10 km/hr. (b) 30 km/hr. ...................................................... 112
Figure 4-12: Signalling cost per user vs. time of day (a) call cost (b) total cost ........................................... 112
Figure 4-13: Signalling cost vs. velocity (km/hr.) (a) 2 calls/hr. (b) 6 calls/hr. .............................................. 114
Figure 4-14: Signalling cost vs. call rate (a) 10 km/hr. (b) 30 km/hr. ............................................................ 114
Figure 4-15: Signalling cost per user vs. time of day (a) call cost (b) total cost ............................................. 114
Figure 4-16: Signalling cost vs. velocity (km/hr.) (a) 2 calls/hr. (b) 6 calls/hr. .............................................. 116
Figure 4-17: Signalling cost vs. call rate (a) 10 km/hr. (b) 30 km/hr. ............................................................ 116
Figure 4-18: Signalling cost per user vs. time of day (a) call cost (b) total cost ............................................. 116
Figure 4-19: Signalling cost vs. velocity (km/hr.) (a) 2 calls/hr. (b) 6 calls/hr. .............................................. 118
Figure 4-20: Signalling cost vs. call rate (a) 50 km/hr. (b) 10 km/hr. ............................................................ 118
Figure 4-21: Signalling cost per user vs. time of day (a) call cost (b) total cost ............................................. 118
Figure 4-22: Signalling cost vs. velocity (km/hr.) (a) 2 calls/hr. (b) 6 calls/hr. .............................................. 120
Figure 4-23: Signalling cost vs. call rate (a) 30 km/hr. (b) 10 km/hr. ............................................................ 120
Figure 4-24: Signalling cost per user vs. time of day (a) call cost (b) total cost ............................................. 120
Figure 4-25: Signalling cost vs. velocity (km/hr.) (a) 2 calls/hr. (b) 6 calls/hr. .............................................. 122
Figure 4-26: Signalling cost vs. call rate (a) 30 km/hr. (b) 10 km/hr. ............................................................ 122
Figure 4-27: Signalling cost per user vs. time of day (a) call cost (b) total cost ............................................. 122
Figure 4-28: Signalling cost vs. velocity (km/hr.) (a) 2 calls/hr. (b) 6 calls/hr. .............................................. 124
Figure 4-29: Signalling cost vs. call rate (a) 10 km/hr. (b) 30 km/hr. ............................................................ 124
Figure 4-30: Signalling cost per user vs. time of day (a) call cost (b) total cost ............................................. 124
Figure 4-31: Signalling cost vs. velocity (km/hr.) (a) 2 calls/hr. (b) 6 calls/hr. .............................................. 125
Figure 4-32: Signalling cost vs. call rate (a) 30 km/hr. (b) 10 km/hr. ............................................................ 125
Figure 4-33: Signalling cost per user vs. time of day (a) call cost (b) total cost ............................................. 126
Figure 4-34: Signalling cost per user vs. time of day (a) call cost (b) total cost ............................................. 126
Figure 4-35: Signalling cost per user vs. time of day (a) call cost (b) total cost ............................................. 127
Figure 4-36: Signalling cost per user vs. time of day (a) call cost (b) total cost ............................................. 128
Figure 4-37: Signalling cost per user vs. time of day (a) call cost (b) total cost ............................................. 128
Figure 5-1: GPRS system architecture ........................................................................................................ 138
Figure 5-2: The ten nodes (cities) to interconnect [Man97] ............................................................................ 148
Figure 5-3: 3-level optimization process ........................................................................................................ 149
Figure 5-4: The optimisation strategy of the proposed technique ................................................................. 154
Figure 5-5: Example of topology representation ............................................................................................ 155
Figure 5-6: Network Optimisation Problem ; Pareto-Optimal set of solutions ................................................ 156
Figure 5-7: The Biconnectivity Testing Routine ............................................................................................. 162
Figure 5-8: Network Solution.1 .................................................................................................................... 163
Figure 5-9: Network Solution.2 .................................................................................................................... 163
Figure 5-10: Network Solution.3 ................................................................. 163
Figure 5-11: Network Solution.4 ................................................................. 163
Figure 5-12: Network Solution.5 ................................................................. 164
Figure 5-13: Convergence of the GA process for different sizes of the population ................................................................. 167
Figure 5-14: Branch Exchange Solution ........................................................ 169
Figure 5-15: Man-GA method Solution .......................................................... 169
List of Tables

Table 2-1: GSM and PCS traffic parameters ........................................................................ 35
Table 2-2: SS7 traffic by network element ......................................................................... 36
Table 2-3: aggregate network message traffic (SS7 bytes per transaction type) ................. 36
Table 2-4: List of calculated parameters .............................................................................. 39
Table 2-5: SS7 bytes per call .............................................................................................. 41
Table 2-6: Call related SS7 traffic ....................................................................................... 42
Table 2-7: relative (PCS-to-GSM) signalling costs .............................................................. 42
Table 2-8: PCS/GSM signalling transaction rates ............................................................... 42
Table 2-9: per PLMN HLR signalling transaction rates ....................................................... 43
Table 2-10: Call and LU related signalling load as percentage of total SS7 load ................ 43
Table 3-1 Classification of Call Routing cost based on current location of the called MT .... 64
Table 3-2 Classification of Location Updating cost based on movement of MT ................. 65
Table 3-3 Classification of Detach costs in the anchor scheme ........................................... 65
Table 3-4 Classification of Attach costs, based on MT current location ................................ 65
Table 3-5. Cost parameters ................................................................................................... 71
Table 3-6. Classification of Location Updating cost based on movement ......................... 78
Table 4-1: Mobility-related simulation parameters .............................................................. 110
Table 4-2: Total signalling cost, velocity range and call rates (AVFF) .................................. 113
Table 4-3: Performance comparison of various schemes , % improvement over GSM (AVFF) ...................................................... 113
Table 4-4: Total signalling cost, velocity range and call rates (AVFF w/RIP) ....................... 115
Table 4-5: Performance comparison of various schemes , % improvement over GSM (AVFF w/RIP) .............................................................. 115
Table 4-6: Performance comparison of ALL schemes , % improvement with RIP (AVFF) .... 115
Table 4-7: Total signalling cost, velocity range and call rates (PVFF) .................................. 117
Table 4-8: Performance comparison of various schemes , % improvement over GSM (PVFF) .............................................................. 117
Table 4-9: Total signalling cost, velocity range and call rates (PVFF w/RIP) ....................... 119
Table 4-10: Performance comparison of various schemes , % improvement over GSM (PVFF w/RIP) .............................................................. 119
Table 4-11: Performance comparison of ALL schemes , % improvement with RIP (PVFF) .... 119
Table 4-12: Total signalling cost, velocity range and call rates (AVMG) .............................. 121
Table 4-13: Performance comparison of various schemes , % improvement over GSM (AVMG) .............................................................. 121
Table 4-14: Total signalling cost, velocity range and call rates (AVMG w/RIP) .................... 123
Table 4-15: Performance comparison of various schemes , % improvement over GSM (AVMG w/RIP) .............................................................. 123
Table 4-16: Performance comparison of ALL schemes , % improvement with RIP (AVMG) .... 123
Table 4-17: Total signalling cost, velocity range and call rates (PVMG) .............................. 125
Table 4-18: Performance comparison of various schemes, % improvement over GSM (PVMG) .......... 125
Table 4-19: Total signalling cost, velocity range and call rates (PVMG w/RIP) ............................... 126
Table 4-20: Performance comparison of various schemes, % improvement over GSM (PVMG w/RIP) ........................................................................................................................................... 126
Table 4-21: Performance comparison of ALL schemes, % improvement with RIP (PVMG) ............ 127
Table 4-22: Summary of the results over the 24 hr. period ............................................................... 131
Table 5-1: Reliability classes ........................................................................................................ 139
Table 5-2: Delay classes ............................................................................................................... 139
Table 5-3: Predicted traffic requirements between the 10 cities (nodes) in Mbps [Man97] ............... 148
Table 5-4: Distances between the 10 cities (nodes) in Km [Man97] .............................................. 148
Table 5-5: Design problem formulation .......................................................................................... 149
Table 5-6: Table of available cost effective capacity choices ......................................................... 154
Table 5-7: Adjacency matrix representation of the topology of Figure 5.3 ...................................... 155
Table 5-8: Capacity Flow Assignment Information for Solution 1 .............................................. 164
Table 5-9: Capacity Flow Assignment Information for Solution 2 .............................................. 164
Table 5-10: Capacity Flow Assignment Information for Solution 3 ............................................ 165
Table 5-11: Capacity Flow Assignment Information for Solution 4 ............................................ 165
Table 5-12: Capacity Flow Assignment Information for Solution 5 ............................................ 165
Table 5-13: Routing information for the best solution (Solution 1) ................................................ 166
Table 5-14: Routing information for the .......................................................................................... 166
Table 5-15: Summary of final topologies ..................................................................................... 169
INTRODUCTION

The unprecedented success and rapid growth of second generation cellular systems offering voice telephony and limited data services, has created the momentum towards introduction of different types of services and establishment of new service providers. Currently available circuit-based voice, fax, paging and email services, will give way to packet data transfer, video conferencing, image transfer, video delivery and a multitude of other packet data-based services, in the next generation of personal communication systems (PCS). The research efforts thus range from development of optimised mobility management schemes, to design of efficient multiple access and resource allocation techniques. The design and operation of protocols in support of seamless mobility and for provision of Quality of Service (QoS) in the next generation networks is posing significant challenges. However management of subscriber mobility i.e. locating and tracking, and in particular issues such as the rate of network database transactions, control traffic due to user location area/routing area or cell update and paging, remain fundamental aspects of mobility management in cellular mobile systems and will need to be addressed irrespective of the advances in the networking/transport technologies. Also up to now, the role of network design in providing QoS has often been overlooked. A poorly designed network architecture will not — regardless of the sophistication of the bandwidth allocation strategy used — be able to match the performance of a well designed network.

This thesis addresses the following:

- Reduction of network database signalling costs in GSM;
- Enhanced location update/paging mechanisms and reduction of signalling traffic in the access network;
- Network design and dimensioning for future packet based personal communication networks.

1.1 Motivation and Objectives

Cellular communications systems provide seamless and uninterrupted communication services to mobile subscribers. Irrespective of location and mobility patterns, mobile terminals can
receive services and communicate with remote terminals. However, unlike fixed/static networks such as the internet or PSTN, where the routing information is embedded in the address of each node, the current location of a mobile terminal can not be obtained from its identification number. Mobility/location management schemes and handover therefore become necessary in order to track and locate terminals for incoming calls and ensure continuity of communication with minimal interruption for mobile subscribers. The two commonly used standards for mobility management are GSM (European) and IS41 (US). Under these standards, the user location information is stored in location databases which are updated as the user moves to another location, and are queried when a call is to be delivered/initiated. The classical user location management strategies of GSM and IS41 are based on a two-tier hierarchy of HLR and VLR databases. The HLR is often a single centralised database containing user profiles (subscription info, user IDs, current location, billing info, etc.), whilst VLRs which are distributed throughout the network store information of the only the users currently roaming within their particular coverage areas. Thus the user profile may be replicated at its current serving VLR. The user performs a registration at the HLR and the new serving VLR, every time the user changes the registration area, and de-registers at the previous VLR. When a call arrives, the network simply routes the call to the last known/reported location of the mobile terminal. However, if for instance, the incoming call arrival rate is low compared to the mobility rate of the terminal the classical schemes are not cost effective and in many cases it should be possible to avoid these registrations at the HLR. Thus, efficient location management schemes may be characterized by their ability to:

- reduce the rate of database related transactions
- minimize fixed/access network signalling traffic
- reduce call set up delays
- support scalability of architecture
- applicability of scheme over a wide range of call-to-mobility ratios

Since location management of subscribers may generate significantly high levels of signalling traffic due to the reduction of cell sizes particularly in dense urban areas in response to capacity/coverage requirements, new methods for reducing the signalling traffic and database access (updates and queries) have become necessary. We analyse and evaluate the performance of the anchoring technique. The technique is intended to minimise the expensive HLR access, thereby reducing the overall network signalling load and preventing the HLR from becoming a bottleneck. The basic anchoring technique is shown to successfully reduce the location updating and the total network signalling costs compared to classical strategy of GSM. Although by
Chapter 1. Introduction

distributing the signalling load over a number of VLRs, the HLR is prevented from becoming a bottleneck in the signalling network, at low Call-to-Mobility Ratios (CMR), the basic anchoring scheme can result in rather high call routing costs compared to baseline scheme of GSM. To address this problem, an optimisation of the basic technique is proposed and evaluated. The optimisation procedure is based on identification of the optimal trade-off point between the call routing and location updating costs given the CMR, and analytical results indicate that the proposed optimisation strategy can significantly decrease the call routing and locating costs whilst maintaining the cost of HLR access well below that of GSM, compared to the original anchor scheme. The cost evaluations take into account the signalling loads generated due to registration and de-registration procedures as well as call routing and location updating thus providing a more complete picture of the total network signalling cost. Then the proposed optimisation procedure is applied and results are compared with the basic scheme. Finally, the impact of the proposed optimisation procedure on database access times are examined and compared to the original scheme. Call routing cost for the fixed-to-mobile and mobile-to-mobile calls are also evaluated separately.

Location management i.e. location update and terminal paging, schemes in the current standards are basically static techniques which can not be adjusted in response to variations in user mobility and/or call patterns. For example, in Location Area (LA) based schemes, the location area size most suitable for one user or a particular class of users, may not be suitable or optimal in terms of minimisation of signalling cost, for another user or other classes of users. Dynamic schemes allow for online adjustments based on mobility/call characteristics of users. We introduce a new adaptive multilayer scheme for location management, and use simulations to evaluate the performance of the proposed scheme in comparison to a number of other adaptive/dynamic location management schemes, under a uniform set of mobility and call related parameters. The objective is to quantify the cost saving under a common set of assumptions and compare these to the cost of the baseline scheme of GSM. Furthermore, the potential of recent interaction paging (RIP) - a two-step paging process - on reducing the paging signalling traffic is also examined.

Telecommunications networks can consist of mixes of many physical and logical layers, built using different technologies. A typical network can be viewed as a multi-layered structure, with each layer requiring different network design to account for different design objectives, constraints and technologies. Mesh and ring network topologies are good starting points for network design, as they embody various trade-off involved in almost all network design
problems. The role of network design in providing Quality of Service (QoS) is often overlooked; a poorly designed network architecture will not - regardless of the sophistication of the bandwidth allocation strategy used - be able to match the performance of a well designed network. Regular and random partially meshed network designs can overcome many of the problems associated with conventional network designs. These partially meshed networks have bounded hop counts, relatively few links and an even distribution of routes across the network. We investigate the application of genetic algorithms to the design of meshed packet-switched communication networks. The aim is to find optimal or near-optimal network topologies which could support the expected levels of transit traffic given the traffic requirements and pre-defined cost, delay and reliability constraints.

1.2 Main Contributions

The objective of this thesis is to address location management in GSM networks with particular emphasis on database related aspects, location management signalling traffic over the air-interface and cost-efficient design of packet communication networks. The main contributions are as follows:

- A technique for reducing the network database signalling traffic cost - the operation and implementation of the basic anchoring technique and an optimisation procedure for application in GSM networks, have been addressed and evaluated. Distribution of the HLR related signalling load is achieved through replacing of the expensive location registration messages between the serving VLR and the HLR by message exchanges between two MSC/VLRs. In response to subscriber call and mobility patterns, the optimised anchoring scheme removes the need to always report location changes to the anchor VLR whilst balancing the costs of both location update and call routing. As the benchmark for the comparison of the results, GSM standard signalling transactions and database traffic under the same traffic model and user mobility assumptions, are used. An analytical model for the proposed optimisation procedure is introduced. It is shown that within the call-to-mobility (CMR) range considered i.e. 0.01 to 100, the location registration signalling cost of the optimised scheme can be reduced to less than half that of GSM. It is also shown that HLR related signalling costs can be reduced by as much as 25-65%, compared to GSM. Also under typical operating conditions, reductions in the total network signalling cost of about 60-25% at low to high CMR respectively, can be achieved. Where as the cost of call routing in the original anchor scheme is about 2 - 3.5 times that of GSM (at low CMR
values), the optimisation procedure manages to reduce this cost to within 1.6 times that of GSM under a variety of network loading conditions.

- **Comparative assessment of dynamic/adaptive location management techniques** - the performance of a number of adaptive/dynamic location management schemes for reducing the location update and paging control traffic have been evaluated. It is shown that the type of environment, mix of subscribers and variations in velocity/call arrival rates can have a wide-ranging impact on the performance of various algorithms. The conclusions indicate that despite good performance and reasonable gains (i.e. signalling load reductions) compared to fixed LA scheme of GSM, the performance of dynamic/adaptive schemes are in fact subject to assumptions re. type of operating environment and classes/mixes of subscribers, whilst the proposed adaptive multi-layer scheme is shown out perform all schemes considered, in both environments. *It is shown that the proposed adaptive multilayer technique can reduce the total signalling cost (i.e. location update and paging signalling costs) by up to 30%, and with RIP enabled by as much as 37% compared to baseline scheme of GSM.* The impact of Recent-Interaction-Paging (RIP) is extensively studied through simulations and it is shown that it has a smoothing effect on the total signalling load, with the most significant impact observed in the Manhattan grid environments with a mix of fast and slow subscribers.

- **An efficient technique for packet communication network design based on a Hybrid GA-Heuristic approach** - the design of partial-mesh packet communication networks has been addressed. The proposed hybrid design methodology, has been based on a decomposition of the original network design problem into: Link Capacity Assignment (CA), Routing or Flow Assignment (FA), Capacity and Flow Assignment (CFA) sub-problems, given the capacity and reliability i.e. bi-connectivity, and delay constraints. A new efficient packet communication network design technique has been developed. *The proposed hybrid technique was found to be superior to the other two typical design methods i.e. Branch exchange and pure GA, in terms of network cost and delay.*

### 1.3 Organization of the Thesis

The remainder of this thesis is organised as follows.

The impacts of mobility-related signalling traffic in mobile networks is the main focus of chapter 2. Following a brief presentation of signalling across the user interface and the
signalling transport mechanism in GSM, the results of analysis of the signalling traffic based on analytical expressions for calculation of signalling load are presented. Numerical results obtained are used to highlight (a) the impacts of location management signalling in GSM networks and (b) the radio resource requirement in support of mobility.

Chapter 3 deals with optimisation of fixed network/database signalling and its impacts on the overall network signalling traffic and rate of database transactions. The focus of the chapter is on the impacts of location registration and call routing signalling on the network databases. In this chapter the anchoring technique is introduced and analysed. The technique is intended to minimise the expensive HLR access, thereby reducing the overall network signalling load and preventing the HLR from becoming a bottleneck. Later in the chapter, optimisation of the basic anchoring technique is presented and analysed. The basic anchoring technique is shown to successfully reduce the location updating and the total network signalling costs compared to classical strategy of GSM. Although by distributing the signalling load over a number of VLRs, the HLR is prevented from becoming a bottleneck in the signalling network, the basic anchoring scheme can results in rather high call routing costs compared to GSM at low Call-to-Mobility Ratio (CMR). To alleviate this problem, an optimisation of the basic technique is proposed and evaluated. The cost evaluations take into account the signalling loads generated due to registration and de-registration procedures as well as call routing and location registrations thus providing a more complete picture of the total network signalling cost. Then the proposed optimisation procedure is applied and results are compared with the basic scheme. Also more importantly, the effects of the proposed optimisation procedure on database access times are examined and compared to the original scheme. Call routing cost for the fixed-to-mobile and mobile-to-mobile calls are also evaluated separately.

Chapter 4 considers the signalling impacts of location updates and paging procedures over the air-interface. The control traffic generated and processing required in the LA-based location management methods most widely used in current cellular (e.g. GSM, IS-41, etc.) systems, may lead to congestion problems in high-density environments. One of the main concerns of system designers is to define methods allowing the system to reduce the overhead control traffic dynamically. In this chapter several location management methods proposed within the past few years that attempt to reduce the overhead signalling traffic, have been reviewed. Based on detailed simulations under various scenarios the performance of four representative dynamic/adaptive techniques are evaluated and compared to the fixed-LA strategy of GSM and
the proposed adaptive multi-layer strategy. Furthermore, the impact of recent interaction paging or RIP—a two-step paging technique—on all the above is also investigated.

Chapter 5 deals with problem of packet-switched communication network design. Designing a communication network often can be viewed as a complex multi-constraint optimisation problem, where the design must meet a set of possibly conflicting, interdependent requirements and invariably, the design must provide optimal assignments for link capacity, routing and topological connectivity, such that costs are minimised and yet the traffic requirements are satisfied and network delays kept within permissible bounds. In telecommunication networks, the common capacity and delay constraints have significant impacts on routing assignment, optimal design of topology and network cost minimisation. The objective of the work presented in this chapter, is to develop a new GA-based efficient network design method for partially-meshed communication network design problem. This objective implies modelling of the network design problem in a format suitable for GA application, and exploration of existing, conventional network design techniques to identify any useful component algorithms to incorporate into the GA optimisation process. Using simulations, the performance of the proposed method for the solution of the complete network design problem, is evaluated and compared with two other typical network design techniques, to show its superiority and capability in satisfying the traffic requirements as well as capacity, delay and reliability constraints.

Chapter 6 summarises the main contributions of this dissertation, and describes possible future research directions.
2 Introduction

Signalling traffic in GSM networks can be classified into two broad categories: Call-related signalling and non-call related i.e. mobility signalling. The call related signalling is primarily responsible for call establishment, maintenance and tear-down of connections, whilst mobility-related signalling covers terminal, personal and service mobility aspects. Mobility support in cellular networks can be further classified into radio mobility i.e. handover, and network mobility which consists of location management i.e. location updating and call delivery (paging) procedures. The recent exponential growth in user penetration across many GSM networks, is accompanied by a corresponding rise in the levels of signalling traffic (mobility and call related) and additional signalling load on the network databases (due in part to the introduction of new services). The integrity and proper functioning of the signalling network is perhaps the foremost consideration for network equipment designers and planners as well as operators. Later in the chapter, it is shown that support of subscriber mobility places a heavy additional signalling burden on network switching equipment and databases.

To ensure network integrity, maintaining safe operational limits, reducing the mean time between failures in the face of increasing levels of signalling traffic load, has focused research into more efficient and alternative methods for location management and tracking of mobile subscribers. The contention that mobility signalling is likely to have significant impacts on quality of service provided by the mobile networks, is the main focus of this chapter. Following a brief presentation of signalling across the user interface and the signalling transport mechanism in GSM, the chapter presents results of analysis of the signalling traffic based on analytical expressions for calculation of signalling load and numerical examples. The impact of location management signalling in terms of signalling traffic generated is then assessed and the radio resource requirement in support of mobility is quantified and presented.
This chapter is organised as follows. In section 2.1, a detailed overview of the signalling across the user interface covering layers 2 and 3 signalling mechanisms is presented. The SS7 signalling transport protocol and the services provided by each of the SS7 layers are discussed in section 2.2. Section 2.3 covers mobility management and call related signalling procedures, and in section 2.4, based on simple analytical expressions for calculation of signalling load and numerical examples, the impact of location management signalling in terms of signalling traffic generated is assessed. Then in section 2.5 the radio resource requirement on the location area edge-cells in support of subscriber mobility are quantified. The chapter ends with a summary of the results and concluding remarks in section 2.6.

2.1 Signalling Across the User Interface (Um)

Signalling at the User-Network Interface in GSM is essentially concentrated in Layer 3. Layers 1 and 2 provide the mechanisms for the protected transmission of signalling messages across the air interface. Besides the local interface, they contain functionality and procedures for the interface to the BTS. The signalling of Layer 3 at the User-Network Interface is more complex and comprises protocol entities in the mobile station and in all functional entities of the GSM network (BTS, BSC, and MSC).

2.1.1 Layer 1 signalling - MS-BTS Interface

Layer 1 of the OSI Reference Model contains all the functions necessary for the transmission of bit streams over the physical medium, in this case the radio channel. GSM Layer 1 defines a series of logical channels based on the channel access procedures with their physical channels. The higher layer protocols access these services at the Layer1 service interface. The three interfaces of Layer 1 are schematically illustrated in Figure 2.3.
LAPDm protocol frames are transmitted across the service mechanisms of the data link layer interface, and the establishment of logical channels is reported to Layer 2. The communication across this interface is defined by abstract physical layer service primitives. A separate Service Access Point (SAP) is defined for each logical control channel (BCCH, PCH+AGCH, RACH, SDCCH, SACCH, FACCH). Between Layer 1 and the RR sublayer of Layer 3 there is a direct interface. The abstract service primitives exchanged at this interface mostly concern channel assignment and Layer 1 system information, including measurement results of channel monitoring. At the third Layer 1 interface, the traffic channels for user (payload) data are provided. The service access points (SAP) of Layer 1 as defined in GSM are not genuine service access points in the spirit of OSI. They differ from the PHY-SAPs of the OSI Reference Model insofar as these SAPs are controlled by Layer 3 RR sublayer (layer management, establishment and release of channels) rather than by control procedures in the link layer. Control of Layer 1 SAPs by RR comprises activation and deactivation, configuration, routing and disconnection of physical and logical channels. Furthermore, exchange of measurement and control information for channel monitoring occurs through service primitives. Layer 1 services of the GSM User-Network Interface are divided into three groups: Access capabilities, Error detection and Encryption. Layer 1 provides a bit transport service for the logical channels. These are transmitted in multiplexed format over physical channels which consist of elements defined for the transmission on the radio channel (frequency, time slot, hopping sequence, etc.). Some physical channels are provided for common (shared) use (BCCH and CCCH), whereas others are assigned to dedicated connections with single mobile stations (dedicated physical channels). The combination of logical channels used on a physical channel can vary over time, e.g. TCH+SACCH/FACCH replaced by SDCCH+SACCH.

The GSM standard distinguishes explicitly between access capabilities for dedicated physical channels and for common physical channels BCCH/CCHs. Dedicated physical channels are established and controlled by Layer 3 RR management. During the operation of a dedicated physical channel, Layer 1 continuously measures the signal quality of the used channel and the quality of the BCCH channels of the neighbouring base stations. This measurement information is passed to Layer 3 in measurement service primitives MPH. In idle mode, Layer 1 selects the cell with the best signal quality in co-operation with the RR sublayer based on the quality of the BCCH/CCH (cell selection). GSM Layer 1 offers an error-protected bit transport service and therefore also error detection and correction mechanisms. To do this, error-correcting and error-detecting coding mechanisms are provided. Frames recognised as faulty are not passed up to Layer 2. Furthermore, security-relevant functions like encryption of user data is implemented in Layer 1.
2.1.2 Layer 2 Signalling

The LAPDm protocol is the data link protocol for signalling channels at the air interface [Mou92]. It is similar to HDLC. It supports Unacknowledged and Acknowledged operational modes. In the Unacknowledged operation mode, data is transmitted in UI-frames (unnumbered information) without acknowledgement; there is no flow control or L2 error correction. This operational mode is allowed for all signalling channels, except for the RACH which is accessed in multiple access mode without reservation or protection. The Acknowledged operation mode provides protected data service. Data is transmitted in I-frames (information) with positive acknowledgement. Error protection through retransmission (ARQ) and flow control are specified and activated in this mode. This mode is only used on DCCH channels. In LAPDm, the Connection End Points (CEPs) of L2 connections are labelled with Data Link Connection Identifiers (DLCIs), which consist of two elements:

- The Layer 2 Service Access Point Identifier (SAPI) is transmitted in the header of the L2 protocol frame.
- The physical channel identifier on which the L2 connection is or will be established, is the real Layer 2 Connection End Point Identifier (CEPI). The CEPI is locally administered and not communicated to the L2 peer entity. (The terminology of the GSM standard is somewhat inconsistent in this case - what is really meant is the respective logical channel. The physical channels from the viewpoint of LAPDm are the logical channels of GSM, rather than the physical channels defined by frequency/time slot/hopping sequence.)

When a Layer 3 message is transmitted, the sending entity chooses the appropriate SAP and CEP. When the service data unit SDU is handed over at the SAP, the chosen CEP is given to the L2 entity. Conversely, when receiving an L2 frame, the appropriate L2-CEPI can be determined from the physical / logical channel identity and the SAPI in the frame header. Specific SAPI values are reserved for the certain functions:

- SAPI=0 for signalling (CM, MM, RR)
- SAPI=3 for SMS

In the control plane, these two SAPI values serve to distinguish and separate signalling messages from packet-oriented user data (short messages). Further functions needing a new SAPI value can be defined in future versions of the GSM standard. A LAPDm entity is established for each of the pertinent physical / logical channels. For some of the channel / SAPI combinations only a subset of the LAPDm functions is needed (e.g. unacknowledged
operation), and not all channel / SAPI combinations are supported. These LAPDm entities perform the Data Link procedure, i.e. the functions of the L2 peer-to-peer communication as well as the service primitives between adjacent layers. Segmentation and reassemble of Layer 3 messages is also included. Other Layer 2 procedures are the Distribution Procedure and the Random Access (RA) procedure. The distribution procedure is needed if multiple SAPs are associated with one physical / logical channel. It performs the distribution of the L2 frames received on one channel to the respective data link procedure, or the priority-controlled multiplexing of L2 frames from multiple SAPs onto one channel. For certain aspects of RR, the protocol logic of Layer 3 has to have direct access to the services of Layer 1. Especially, this is needed for functions of Radio Subsystem Link Control, i.e. for channel measurement, transmitter power control, and timing advance.

2.1.3 Layer 3 Signalling – RR, MM and CM

Radio Resource Management (RR) - The procedures for Radio Resource Management (RR) are the basic signalling and control procedures at the air interface. They handle the assignment, allocation and administration of radio resources, the acquisition of system information from broadcast channels (BCCH) and the selection of the cell with the best signal reception. Accordingly, the RR procedures and messages are defined for idle mode as well as for setting up, maintaining, and taking down of RR connections. In idle mode, the MS is continuously reading the BCCH information and conducts periodic measurements of the signalling strength of the BCCH carriers in order to be able to select the current cell. In this state, there is no exchange of signalling messages with the network. The data required for RR and other signalling procedures is collected and stored: the list of neighbouring BCCH carriers, thresholds for RR algorithms, CCCH configurations, information about the use of RACH and PCH, etc. Also important is the periodic monitoring of the paging channel (PCH) so that paging calls are not lost. For this purpose, the BSS is sending on all paging channels of a cell continuously valid Layer 3 messages (PAGING REQUEST) which the MS can decode and recognise if its address is paged.

Each exchange of signalling messages with the network (BSS, MSC) requires an RR connection and the establishment of an LAPDm connection between MS and BTS. Setting up the RR connection can be initiated by the network or the MS. In either case, the MS sends a channel request (CHANNEL REQUEST) on the RACH in order to get a channel assigned on the AGCH (immediate assignment procedure). There is also a procedure to deny a channel request (immediate assignment reject). If the network does not immediately answer to the channel
request, the request is repeated using the S-Aloha method with a random number controlled timer. In the case of a network-initiated connection, this procedure is preceded by a paging call (PAGING REQUEST) to be answered by the mobile station (PAGING RESPONSE). After an RR connection has been successfully completed, the higher protocol layers (CM, MM) can receive and transmit signalling messages at SAPI 0. In contrast to the setup of connections, the release is always initiated by the network (CHANNEL RELEASE). Reasons for the release of the channel could be end of the signalling transaction, too many errors, removal of the channel in favour of a higher priority call (e.g. emergency call), or end of a call. RR procedures fall into 6 main groups. These are Channel Establishment/Release, Ciphering, Paging, Handover, System Information and miscellaneous procedures.

Mobility Management (MM) - The main task of Mobility Management (MM) is to support the mobility of the mobile station. Another task of the MM sublayer is to offer MM connections and associated services to the CM sublayer above. All MM procedures presume an established RR connection, i.e. a dedicated logical channel must be assigned with an established LAPDm connection in place, before MM transactions can be performed. These transactions occur between MS and MSC, i.e. messages are passed through the BSS transparently without interpretation and forwarded to the MSC with the DTAP transport mechanism. The MM procedures are divided into three categories: Common, Specific, and Connection related. Whereas Common procedures can always be initiated and executed as soon as an RR connection exists, Specific procedures exclude one another, i.e. they cannot be processed simultaneously or during an MM connection. Conversely, an MM connection can only be set up if no Specific procedure is running. MM Specific procedures consist of IMSI Attach and Location Update procedures, whereas MM Common procedures consist of TMSI Allocation/reallocation, IMSI Detach, Authentication/Identity Request.

Connection Management (CM) - The CM sublayer terminates at the MSC and consists of Call Control (CC), SMS and supplementary Services (SS) entities. Once a MM connection is established, the CM can use it for information transfer. The CC entity is based on CCITT Q931 protocol, with minor modifications, for communication of call-related messages between the MS and the MSC. It comprises of procedures to establish, control, and terminate calls. Several parallel CC entities are provided, such that several parallel calls on different MM connections can be processed. For CC, finite state models are defined both on the mobile side as well as on the network side. The two entities at the MS and MSC sites each instantiate a protocol.
automaton, and these communicate with each other based on standard Layer 3 signalling message format.

### 2.2 Signalling Transport Protocols In GSM

The functional entities of the NSS use CCITT Signalling System No.7 (SS7) signalling. SS7 is used to implement both the data link and layer 3 transport functions for carrying CM and MM signalling messages over the MSC-BSS link (A-interface), as well as B,C,D and G interfaces. Call-related signalling between the MSCs or external networks use Telephony User Part (TUP) or ISDN User Part (ISUP), which may be modified according to national requirements. For all non-call-related functions the MAP protocol is used which makes use of services provided by the Transaction Capabilities Application Part (TCAP) and Signalling Connection Control Part (SCCP).

SS7 overcomes the limitations of conventional inband signalling by transmitting signalling information over a dedicated digital signalling network, separate from the voice channels. Additionally, SS7 enables the separation of switching intelligence from voice and data routing in telecommunications networks. SS7 also offers improvements in the speed and flexibility of call set-up, fast exchange of information between processors for a call requiring special routing or handling and access to customer information stored in network databases. Functions provided by the various layers in the SS7 protocol hierarchy in relation to GSM, are briefly described below.

#### 2.2.1 Message Transfer Part

The Message Transfer Part (MTP) provides reliable transfer and delivery of signalling information within the signalling network, and is responsible for maintaining the links between signalling connections in the presence of failures. A limited set of MTP protocol messages are used on the A-interface to transfer messages between the BSS and MSC. MTP 3 also provides for the support and management of multiple links (link-set) and messaging on fall-back link for maintenance purposes, on the A-interface. As shown in Figure 2.1 above, MTP consists of 3 layers. **MTP layer 1** corresponds to OSI physical layer. Typically 56 and 64 kb/s links are used to convey signalling information. **MTP layer 2** ensures reliable information transfer using features such as CRC and forward/backward sequence numbers and flow control. It is based on HDLC protocol with some enhancements e.g. MTP 2 uses FISUs (Fill-In Signalling Unit) which are sent when there are no message traffic, this allows for quick fault detection. Higher
layer information is encapsulated within the signalling Information field (SIO) of the layer 2 Message Signalling Unit or MSU. MTP layer 3 functions correspond to the lower half of OSI network layer, and provides for the routing of messages between signalling network nodes on a link-by-link basis. MTP 3 also provides for the support and management of multiple links (link-set) and messaging on fall-back link for maintenance purposes, on the A-interface. Routing of messages between different signalling points inside an SS7 network is specified in MTP layer 3. It defines network addressing capabilities using Point Code (PC) addressing where a single PC address identifies a signalling point exclusively in the SS7 network. Originating Point Code (OPC) identifies the sender signalling point of the message and Destination Point Code (DPC) identifies the receiving signalling point. PC is comparable to the OSI NSAP (Network Service Access Point) and the internet IP-address. Furthermore, MTP3 is responsible for signalling management and congestion control. The objective of management is to overcome link failures or link congestion in an SS7 network.

2.2.2 Signalling Connection Control Part

SCCP (Signalling Connection Control Part) provides the means to control logical signalling connections in the SS7 network, and transfer signalling data units across the SS7 network. SCCP allows different users within a node to be addressed separately. Another addressing enhancement to MTP, provided by SCCP is the ability to address messages to the end destination rather than only to the next link (hop). SCCP does this using either an internal network address, known as Signalling Point Code (SPC), or a Global Title (GT) which is an address such as dialed digits and does not explicitly contain information usable for routing by MTP3. (an analogy: SCCP provides the functionality to send someone a mail message using an e-mail address, while MTP can only route messages to the next link using a port number). Therefore, SCCP defines network addressing capabilities based on Global Title (GT), and Destination Point Code (DPC) and Subsystem Number (SSN). A global title is an address such as a set of dialed digits which does not explicitly contain routing information but can be used, for instance, to identify the caller and the party being called. The actual message routing between connected signalling points is done using DPC and the required service indication using SSN. DPC defines the SS7 network address for the signalling point (a service control point, for example) whereas the SSN addressing identifies the logical address for the desired user or application part. In this sense, SSN can be compared to the OSI TSAP (Transport Service Access Point) addressing, which is defined at the OSI transport layer. SCCP can support four distinct classes of operation:

- class 0: basic connectionless class
Chapter 2. Signalling Traffic in GSM Networks

- class 1: sequenced (MTP) connectionless class
- class 2: basic connection-oriented class
- class 3: flow controlled connection-oriented class

Thus, two service classes - connection oriented and connectionless service classes - are provided by SCCP, and as such SCCP services can be compared to the OSI TP (Transport Protocol) classes and the internet connection oriented TCP and connectionless UDP services. SCCP enhances the services provided by MTP3 layer and in conjunction with MTP3 layer provides full functional equivalent of OSI network layer. The GSM A-interface is used for messaging between BSC and MSC as well as for transfer of messages to/from the mobile station (using the CC or MM protocol discriminators). As shown in Figure 2.1, these two flows are managed by BSSMAP and DTAP protocol layers. Additionally, BSSMAP messages intended for BSC must be distinguished from messages specific to a connection to a mobile station. This latter functionality is provided in GSM based on a virtual circuit approach and makes use of services provided by SCCP. The distinction between the BSSMAP (messaging between BSC and MSC) and DTAP (messaging between MS and MSC) flows is achieved by introduction of a “distribution” protocol on top of SCCP. GSM uses SCCP for management of MS references on the BSS to MSC (A-interface) interface and also to provide addressing facilities between GSM network entities, where only SCCP class 0 and class 2 services are used. The class 0 mode is used on the A-interface for messages not directly related to a single mobile station, such as reset or overload indication. The class 2 mode enables separate independent connections to be setup, a function used on the A-interface for distinguishing transaction for different mobile stations.

2.2.3 Transaction Capabilities Application Part

The SS7 application part consists of two subparts:
- TCAP (Transaction Capabilities Application Part),
- TCAP-user ASE (Application Service Element) such as INAP (Intelligent Network Application Part) and MAP (Mobile Application Part)

TCAP is responsible for the management of dialogues between distributed signalling points in the SS7 network. TCAP is application layer signalling used for non-circuit related, or by direct signalling applications such as database inquiry/update/response. TCAP operation is such that the originating party may send a message to the remote site to perform a task, and after
performing the task, the remote site returns the result to the originating party. In GSM, TCAP is used for queries to the HLR e.g., number translation, database updates and retrieval of information needed for authentication and location updating. TCAP uses the services of SCCP directly. TCAP is divided into two sublayers:

- The Transaction SubLayer (TSL) and the Component SubLayer (CSL). The TSL is used to set-up a dialogue (session) between two remote users and manage exchange of the Protocol Data Units or PDUs between peer TCAP entities, using the main TSL primitives: BEGIN, CONTINUE and END. Within each TSL message, there may be several components each of which corresponds to an inquiry or response.

- The Component sub-layer provides the capability of invoking remote operations and receiving replies. The CSL manages the remote call functions for the services as well as coding and decoding call parameters. The commonly used CSL primitives are: INVOKE and RETURN RESULT. Four classes of operation within CSL allow 4 different types of responses to be made to an inquiry. In class 1, both success and failure in performing the operation are reported; In class 2 only failure is reported; In class 3 only success is reported; In class 4 neither success nor failure is reported.

### 2.2.4 Integrated Services Digital Network User Part

The Integrated Services Digital Network User Part (ISDN_UP or ISUP) is an application layer which provides the circuit-related signalling required for setting up of voice circuits for basic ISDN calls and for supplementary services. For basic ISDN and GSM calls, there are 5 messages needed to set up a call. In response to a set up message from terminal, the switch sends an *Initial Address Message* (IAM) into the network. The IAM causes trunks to be set up towards the destination switch. After the called party has been alerted, an *Address Complete Message* (ACM) is sent by the destination exchange towards the originating exchange. When the called party answers, an *Answer* (ANS) message is sent to the originating exchange. Call release is accomplished by using messages *Release* (REL) and *Release Complete* (RLC).

### 2.2.5 Mobile Application Part

Messages for different applications are defined in the SS7 application layer, residing above TCAP. The Mobile Application Part (MAP) defines the signalling messages required in GSM. MAP resides above TCAP and is an application layer protocol. MAP is a TCAP user and makes use of component and dialogue handling facilities provided by TCAP, for peer-to-peer communication. It makes use of connectionless services of SCCP. MAP is divided into five
Chapter 2. Signalling Traffic in GSM Networks

Application Entities or ASEs. These are MAP-MSC, MAP-VLR, MAP-HLR, MAP-EIR and MAP-AUC. These entities are each identifiable through individual Sub-System Numbers (SSN) assigned to each one. The SSNs are used by SCCP for addressing each ASE individually. From the application process point, MAP defines a number of procedures each consisting of one or several operations. Detailed descriptions of MAP messages for all operations/transactions between GSM entities may be found in [GSMr9.02].

2.3 Mobility Management and SS7 Traffic in GSM Networks

Mobility management is one of the most important and challenging problems for wireless communication and computing. It enables telecommunication networks to locate roaming terminals for call delivery and to maintain connections as the terminal is moving into a new service area. Mobility management contains two components: Location Management and Handoff Management.

Location management: a two-stage process required to locate the current attachment point of the mobile user for call delivery. The first stage is location update (or location registration). In this stage, the mobile terminal periodically notifies the network of its new access point, allowing the network to authenticate the user and revise the user's location profile. The second stage is call delivery, where the network is queried for the user location profile and the current position of the mobile host is found.

Location Update Procedure - In order to correctly deliver calls, the mobile network must keep track of the location of each MS. As described previously, location information is stored in two types of database, VLR and HLR. As the subscribers move around the network coverage area, the data stored in these databases may no longer be accurate. To ensure that calls can be delivered successfully, a mechanism is needed to update the databases with up-to-date location information. This updating process is known as location update. Current systems adopt an approach that the network coverage area is partitioned into a number of service areas (SA) i.e. area of coverage of each MSC, and each SA comprises of location areas (LAs) (note that IS-41 refers to this as the registration area). An MS performs a location update whenever it enters a new LA. Each LA consists of a number of cells and in general all base stations belonging to the same LA are connected to the same MSC. When an MS enters an LA, if the new LA belongs to the same VLR as the old LA, only the record at the VLR is updated to enter the ID of the new LA. Otherwise, if the new LA belongs to a different VLR, a number of extra steps are required to:
• Register the MS at the new serving VLR
• Update the HLR to record the ID of the new serving VLR
• De-register the MS at the old serving VLR

Figure 2-2 shows the location update procedure when an MS moves to a new LA. The following is the ordered list of tasks that are performed during location update/registration:

1. The MS enters a new LA and transmits a location update request message to the new base station.
2. The base station forwards the location update message to the MSC, which launches a registration query to its associated VLR.
3. The VLR updates its record on the location of the MS. If the new LA belongs to a different VLR, the new VLR determines the address of the HLR of the MS from its IMSI. This is achieved by a table lookup procedure called global title translation. The new VLR requests subscriber Id (IMSI) and authentication parameters from old VLR in the send parameters message.
4. The old VLR returns requested parameters in the IMSI response message, and authentication procedure by the new VLR begins.
5. Assuming a successful authentication is performed, the new VLR then sends a location registration message to the HLR; otherwise, location update is complete.
6. The HLR performs the required procedures to authenticate the MS and records the ID of the new serving VLR of the MS. The HLR then sends a registration acknowledgement message to the new VLR.
7. The HLR sends a registration cancellation message to the old VLR.
8. The old VLR removes the record of the MS and returns a cancellation acknowledgement message to the HLR.

Depending on the distance between the current and home locations of the MS, in steps 3, 4, 5, and 6 the signalling messages may have to go through several intermediate STPs before reaching their destinations. Therefore, the location registration may generate a significant traffic load to the SS7 network. As the number of mobile subscribers keeps increasing, the delay for completing a location registration may increase. More detailed message sequence charts for location update procedures can be found in Appendix A.
Call Delivery Procedure - Two major steps are involved in call delivery: determining the serving VLR of the called MS, and locating the visiting cell of the called MS. Locating the serving VLR of the MS involves the following database lookup procedure (Figure 2-3):

1. The incoming call is routed from the calling exchange in PSTN/ISDN to the mobile network GMSC.
2. The GMSC determines the address of the HLR of the called MS by global title translation (from provided MSISDN) and requests routing information from the HLR.
3. The HLR determines the serving VLR of the called MS and sends a provide roaming number message to this VLR.
4. The VLR returns the roaming number (MSRN) to the HLR.
5. The HLR informs GMSC by forwarding the MSRN. The gateway MSC uses the MSRN to set up call towards the serving MSC/VLR of the called subscriber.
6. The serving MSC initiates the paging process for the called MS in all the cells within the indicated LA and upon successful page response, initiates authentication and ciphering procedures on the radio path. ISDN Setup and Call Confirmed messages are then exchanged between the MS and the MSC.
7. Following the transmission of Alert and Connect messages by the MS, end-to-end speech path between the caller and the MS is setup.
The procedure described above allows the network to set up a connection from the caller to the serving MSC/VLR of the called MS. Since each MSC is associated with an SA, and there are more than one LA in each SA, a mechanism is therefore necessary to determine the cell location of the called MS. In current networks this is achieved by a paging (or alerting) procedure such that polling signals are broadcast to all cells within the LA of the called MS. On receiving the polling signal, the MS sends a reply which allows the MSC to determine its current residing cell. Appendix A, also provides a detailed depiction of the message sequences for Mobile-Originated (MO) and Mobile-Terminated (MT) calls in GSM.

_Handoff Management_ (or Handover) refers to the transfer of an existing connection to a new base station. The three-stage process for handoff first involves initiation, where either the user, a network agent, or changing network conditions identify the need for handoff. The second stage is new connection generation, where the network must find new resources for the handoff connection and perform any additional routing operations. The final stage is data flow control, where the delivery of the data from the old connection path to the new connection path is maintained according to agreed-upon service guarantees. In GSM handover decision is made by the network, and it is based on BSS criteria (received signal level, channel quality, MS-BTS distance) and on network operational criteria e.g. current traffic load of the cell. Periodically, the MS gathers signal field strength measurements of its current downlink and those of the neighboring base stations (quality monitoring). On the network side, the signal quality of the uplink is monitored, the measurement reports from MS are evaluated and handover decisions are made. Principally, two kinds of handovers can be distinguished:
Chapter 2. Signalling Traffic in GSM Networks

- **Intra-Cell Handover**: for administrative reasons or because of channel quality the MS may be assigned a new channel within the same cell. This decision is made by the RR management entity of the BSS and also executed within BSS.

- **Inter-Cell Handover**: the connection to an MS is transferred over the cell boundary, to a new BTS. The decision re. time of handover, is made by the RR protocol module of the network based on measurement data from MS and BSS.

As far as Inter-Cell handovers are concerned, two cases are distinguished. **Internal Handover** which refers to handovers between cells belonging to the same BSC. This does not require the involvement of controlling MSC except for notification of MSC re. successful execution of handover by the BSC). **External Handover** refers to handovers where participation of at least one MSC is required. Handovers between cells belonging to different BSCs (under control of same MSC – i.e. Intra-MSC HO) and those between cells belonging to different BSCs (under control of different MSCs – i.e. Inter-MSC HO) are classified as external. Figures A1.6 and 7 in Appendix A, present the message sequences for **Internal** (Intra-BSC) and **external** (Inter-BSC/Intra-MSC) handovers in GSM.

**MAP and Inter-MSC Handover**

Inter-MSC handovers require communication between the involved MSCs, which makes use of SS7 transport for MAP transactions. The serving MSC\textsubscript{A}, sends a handover request to the target MSC\textsubscript{B}, and once the link is setup between the two, MSC\textsubscript{A} will issue the handover command to the MS. There are two cases to consider:

1. **Basic Handover procedure** – the call is handed over from MSC\textsubscript{A} to MSC\textsubscript{B} (1).

2. **Subsequent Handover procedure** – (2a) the call is handed over from MSC\textsubscript{B} back to MSC\textsubscript{A} OR (2b) for handover to a third MSC, the call is first handed over from MSC\textsubscript{B} back to MSC\textsubscript{A} (known as Handback) and then from MSC\textsubscript{A} to MSC\textsubscript{C}.

Thus, **Basic** handover procedures are used for the first handover between MSCs, whilst **Subsequent** handover procedures are used for handovers to subsequent MSCs. The MAP message sequence charts for the Basic and Subsequent HO are provided in Appendix A. Unlike the location management protocols, the reliance of handoff protocols on routing, resource management and packet delivery make these algorithms very network protocol dependent. An additional dependency to be considered for interoperability is the dependency of the user on the local wireless network interfaces and infrastructure.
2.4 Analysis of SS7 Signalling Load in Cellular Networks

In this section a methodology for evaluation of the impacts of signalling traffic generated by both call-related and non-call related i.e. mobility management signalling, on the SS7 signalling network, based on simple analytical expressions, is presented. For the purpose of comparisons, GSM and PCS networks is presented. The traffic parameters in Table 2-1 assume higher penetration and smaller cell sizes for PCS, compared to GSM.

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>PCS</th>
<th>GSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population density (people/km²)</td>
<td>24000</td>
<td>24000</td>
</tr>
<tr>
<td>( \rho ) (subsc. density) i.e. subsc. per km²</td>
<td>15360</td>
<td>1920</td>
</tr>
<tr>
<td>penetration PCS/GSM : %64 / %8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Call per Hour/subsc. (BHCA/subsc.)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Traffic per Subscriber (ave. Erlang/terminal)</td>
<td>0.033</td>
<td>0.033</td>
</tr>
<tr>
<td>( T ) - Mean call holding time (s)</td>
<td>120</td>
<td>120</td>
</tr>
<tr>
<td>( S ) - Hexagonal Cell Area (km²) ( \approx 2.6r^2 )</td>
<td>0.026</td>
<td>0.036</td>
</tr>
<tr>
<td>( r ) - Hexagonal Cell radius (km)</td>
<td>0.1</td>
<td>0.8</td>
</tr>
<tr>
<td>Subsc. per cell = ( S x \rho )</td>
<td>400</td>
<td>3187</td>
</tr>
<tr>
<td>( SA_{MSC} ) - MSC service area size (km²)</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>% of MO-MT calls</td>
<td>45-55</td>
<td>45-55</td>
</tr>
<tr>
<td>( V_{km} ) - Ave. MS velocity (km/hr)</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2-1: GSM and PCS traffic parameters

In order to relate the information in Table 2-1 to the total traffic offered to the SS7 network, the following need to be determined:

- SS7 traffic (no. of bytes) generated by each key activity/transaction.
- Frequency/rate of occurrence of the activity – requires mobility model and determination of optimum size of LA in terms of number of cells

The GSM procedures that generate SS7 message traffic are the following transactions: mobile call originations, mobile call terminations, mobile-to-mobile calls, registration and deregistrations, location updating, inter-MSC handovers and short message services. A detailed description of message sequences and exchanges involved for each transaction type as well as message sizes can be found in Appendix A. SS7 message lengths are variable and the estimates include only the mandatory parameters with maximum length. Table 2-2 shows the SS7 traffic by network element.
Chapter 2. Signalling Traffic in GSM Networks

Table 2-2: SS7 traffic by network element

<table>
<thead>
<tr>
<th>TRANSACTION</th>
<th>MO Calls</th>
<th>MT Calls</th>
<th>Inter_MSC HO</th>
<th>Loc.Upd. same VLR</th>
<th>Loc.Upd. VLRold</th>
<th>Loc.Upd. VLRnew</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orig_MSC &lt;-&gt; Term_MSC</td>
<td>120</td>
<td>120</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MSC &lt;-&gt; VLR</td>
<td>550</td>
<td>612</td>
<td>-</td>
<td>148</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OriginatingMSC &lt;-&gt; HLR</td>
<td>-</td>
<td>126</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OldMSC &lt;-&gt; NewMSC</td>
<td>-</td>
<td>-</td>
<td>383</td>
<td>383</td>
<td>213</td>
<td>213</td>
</tr>
<tr>
<td>NewVLR &lt;-&gt; OldVLR VLR</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>95</td>
<td>-</td>
</tr>
<tr>
<td>OldVLR &lt;-&gt; HLR</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NewVLR &lt;-&gt; HLR</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>402</td>
</tr>
</tbody>
</table>

Table 2-3: aggregate network message traffic (SS7 bytes per transaction type)

The first row of Table 2-3 shows the number of bytes (sent to and from) the switch for standard ISDN call with minimum of optional parameters. The second and third rows show the number of bytes sent to and from the MSC assuming that VLR is either inside or outside the MSC. The fourth and fifth rows show the number of bytes sent to and from VLR and HLR, assuming both external to the MSC. The last two rows present the number of bytes that appear internal to the SS7 network. Figure 2-3 displays SS7 traffic between network entities for location updates. The dashed lines between the MSCs and their associated VLRs indicate that the VLR can be inside or outside the MSC. Consider the last two rows of Table 2-3 which give the internal SS7 traffic. For example, considering location update to a new VLR (last column) when VLR is inside the MSC, the number of required SS7 bytes is 490 ( = [(395+182) + (308+95)]/2 ), and when the VLR is outside the MSC, the number of bytes is 896 ( = [(406+801+182)+(308+95)]/2 ).
In order to use the information in Table 2-3 to calculate the total traffic offered to the SS7 network, the average rate of four main signalling transactions has to be known.

### 2.4.1 Mobility Model

Mobility models are used to estimate the rate of cell, location area and switch service area crossings, from which handover and location update rates can be determined. In the following it is assumed that the population distribution is uniform and that the direction of travel of user relative to the border is uniformly distributed on (0,2\(\pi\)). The fluid-flow mobility model as described in [Mora87],[Gil88] is used to calculate the average number of users crossing/leaving the area \(A\) per unit time, and is given by

\[
\rho \cdot v_{ave} \cdot \frac{L}{\pi}
\]

(2-1)

where \(L\) denotes the length of the exposed perimeter of cell/LA, \(\rho\) denotes subscriber density per unit area and \(v_{ave}\) is the average subscriber velocity. By conservation of flow, (2-1) is also the rate of crossings into \(A\). Although appealing in its simplicity, in applying (2-1) it is important to relate \(\rho\) and \(v_{ave}\) correctly. For example, we know that during traffic jams, traffic can come to a halt. Relationships between density and velocity can be found in chapter 10 of [May90].

### 2.4.2 Determination Of Optimum Location Area Size

In (2-1) expressions are required for \(L_f\) as a function of \(A\) for particular cell geometry. These are derived in [Alo92]. The expressions for \(L_f\) and traffic parameters listed in Table 2.1 are now used to calculate the optimum LA size in terms of no. of cells. The optimal size for LA is determined by trading off radio bandwidth required for paging and location updates. Paging normally takes place in all the cells in a location area, while registrations are only performed at the LA border cells. In [Gil88],[Min87] optimisation is achieved by attempting to reduce the
sum of the required bandwidth for paging and location updates. The paging bandwidth is given by
\[
\rho \cdot \Pr(\text{ON}) \cdot N_{\text{opt}} \cdot S \cdot [\lambda_1 + \lambda_2 \cdot PG_A] \cdot B_{PG}
\]
(2-2)

where
\[
\lambda_1 : \text{no. of MT calls during busy hour}
\lambda_2 : \text{no. of times an incoming call can not be connected during busy hour}
N_{\text{opt}} : \text{optimum no. of cells per location area}
PG_A : \text{page attempts per MS in busy hour}
B_{PG} : \text{paging bandwidth per attempt}
\]

and location update bandwidth
\[
[ L_f(N_{\text{opt}}) \cdot \rho \cdot v_{\text{ave}} \cdot \Pr(\text{ON}) \cdot B_{LU} ] / \pi
\]
(2-3)

where
\[
v_{\text{ave}} : \text{ave. subscriber speed}
B_{LU} : \text{bandwidth for one location registration}
L_f(N_{\text{opt}}) : \text{average fraction of length of perimeter of exposed cells, which is given in [Alo92] as}
\]
\[
L_f(N) = 6C_{rad} \cdot \left[ \frac{1}{3} + \frac{1}{2\sqrt{3}N - 3} \right]
\]
(2-4)

To evaluate the use of bandwidth for each procedure, the same approach as in [Gi188] is followed, which gives
\[
B_{PG} = 1 \text{ and } B_{LU} = \left[ 1 + 4T_{\text{loc.update}} \right] \text{ where } T_{\text{loc.update}} = 4 \text{ sec.}
\]

Minimizing the sum of (2-1) and (2-2), \(N_{\text{opt}}\) for different values of speed, cell size and call rate can be obtained. Let
\[
v_{\text{ave}} \cdot x \cdot [1 + 4T_{\text{loc.update}}] \cdot L'(N_{\text{opt}}) + S \cdot [\lambda_1 + \lambda_2 PG_A] \cdot \pi = 0
\]
(2-5)

where \(L'(N_{\text{opt}}) = -C_{rad} \sqrt{3}/2 \cdot N_{\text{opt}}^{-3/2}\) and \(L'(N_{\text{opt}}) = \text{derivative of } L(N_{\text{opt}}) \text{ w.r.t. } N\).

Thus with \(\lambda_1 = 0.55\) i.e. 55% & \(\lambda_2 = 0\) and \(v_{\text{ave}} = 10 \text{ km/hr}\) we obtain: \(N_{\text{opt}} = 12\) cells.

Table 2-4 below presents a summary of calculated parameters.
2.4.3 Transaction rates

The required quantities (as a function of switch service area) may now be calculated as follows:

- **MO(\(/sa\))** and **MT(\(/sa\))** are the number of mobile call origination and terminations in calls/hr., generated in the switch service area:

**GSM:**

\[
MO(\text{}/sa) = \frac{BHCA}{MSC} \cdot Pr\text{ (mobile origination)} = 51,848 /hr \tag{2-6}
\]
\[
MT(\text{}/sa) = \frac{BHCA}{MSC} \cdot Pr\text{ (mobile termination)} = 63,360 /hr \tag{2-7}
\]

**PCS:**

\[
MO(\text{}/sa) = \frac{BHCA}{MSC} \cdot Pr\text{ (mobile origination)} = 414,720 /hr \tag{2-8}
\]
\[
MT(\text{}/sa) = \frac{BHCA}{MSC} \cdot Pr\text{ (mobile termination)} = 506,880 /hr \tag{2-9}
\]

- **Handovers \((/sa)\)** refer to the total number of cell departures (intra plus inter-switch) made by calls in progress in the service area, in units of crossings/hr.:

**GSM:** We assume that only about one-third of cell crossings from the cells on border of SA i.e. \(N_{PSA}\), will in fact result in inter-MSC handovers (i.e. \(N_{PSA} = 18 - \text{exact} = 21\)): 

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>PCS</th>
<th>GSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N_{opt}) - optimum no. of cells per location area</td>
<td>48</td>
<td>12</td>
</tr>
<tr>
<td>(L_{A_{area}}\text{ (km)} = N_{opt} \times 6)</td>
<td>1.248</td>
<td>19.92</td>
</tr>
<tr>
<td>Subsc. Population per LA (= L_{A_{area}} \times \rho)</td>
<td>19169</td>
<td>38246</td>
</tr>
<tr>
<td>(N_{LA}) - no. of LAs per switch service area (= SA_{area} / L_{A_{area}})</td>
<td>48</td>
<td>3</td>
</tr>
<tr>
<td>Subsc. Population per SA (= SA_{area} \times \rho)</td>
<td>920112</td>
<td>114741</td>
</tr>
<tr>
<td>(N_{SA}) - no. of cells per SA (= N_{LA} \times N_{opt})</td>
<td>2304</td>
<td>36</td>
</tr>
<tr>
<td>(A_{cell}) - BH load/cell (Erlang) (= (S \times \rho) \times T \times (BHCA/subsc.) / 3600 \times Pr(ON))</td>
<td>13.3</td>
<td>106.2</td>
</tr>
<tr>
<td>(L_{cell}) - Length of cell perimeter (km) = (6r)</td>
<td>4.8</td>
<td>19.2</td>
</tr>
<tr>
<td>(N_{PLA}) - no. cells on perimeter of LA (= 6/(N_{opt}/3) - 3)</td>
<td>21</td>
<td>9</td>
</tr>
<tr>
<td>(L_{LA}) - Length of exposed LA perimeter (km) (= 6r[1/3 + 1/(2 \sqrt {3N_{opt}} \cdot \rho) - 3] \times N_{PLA})</td>
<td>4.8</td>
<td>19.2</td>
</tr>
<tr>
<td>(N_{PSA}) - no. cells on perimeter of SA (= (163.3))</td>
<td>163</td>
<td>(17.8) 18</td>
</tr>
<tr>
<td>(L_{SA}) - Length of SA perimeter (km) (= 6r[1/3 + 1/(2 \sqrt {3N_{SA}} \cdot \rho)] \times N_{PLA})</td>
<td>33.3</td>
<td>33.2</td>
</tr>
<tr>
<td>(X_{cell}) - Cell crossing rate (= \rho \cdot V_{ave} \cdot L_{cell} / \pi)</td>
<td>8.12/s</td>
<td>8.12/s</td>
</tr>
<tr>
<td>(X_{LA}) - LA crossing rate</td>
<td>65/s</td>
<td>32.5/s</td>
</tr>
<tr>
<td>(X_{SA}) - SA crossing rate</td>
<td>451/s</td>
<td>55.36/s</td>
</tr>
<tr>
<td>BHCA/MSC = (SA_{area} \times \rho \times (BHCA/subsc.))</td>
<td>921600</td>
<td>115200</td>
</tr>
<tr>
<td>MSC Traffic load (Erlang)</td>
<td>30412.8</td>
<td>3802</td>
</tr>
</tbody>
</table>

Table 2-4: List of calculated parameters
Chapter 2. Signalling Traffic in GSM Networks

\[ Handovers (/sa) = N_{SA} \times X_{cell} \times (\text{Erlang/terminal}) = 9.65 /s \]  
\[ Inter\_MSC\_H.O. (/sa) = 1/3 \times X_{cell} \times N_{PSA} \times (\text{Erlang/terminal}) = 1.87 /s \]  
\[ Intra\_MSC\_H.O. (/sa) = Handovers (/sa) - Inter\_MSC\_H.O. (/sa) = 7.78 /s \]  

**PCS:** once again, it is assumed that only about one-third of cell crossings from the cells on border of SA i.e. \( N_{PSA} \), will result in inter-MSC handovers (i.e. \( N_{PSA} = 163 - \text{exact} = 215 \)):

\[ Handovers (/sa) = N_{SA} \times X_{cell} \times (\text{Erlang/terminal}) = 545 /s \]  
\[ Inter\_MSC\_H.O. (/sa) = 1/3 \times X_{cell} \times N_{PSA} \times (\text{Erlang/terminal}) = 19.2 /s \]  
\[ Intra\_MSC\_H.O. (/sa) = Handovers (/sa) - Inter\_MSC\_H.O. (/sa) = 525.8 /s \]  

- \( LocUpdates (/sa) \) is the total number of location updates (intra- plus inter- VLR) in the switch service area, in crossings/hr.: 

**GSM:** The service area consists of \( N_{LA} = 3 \) LAs. We assume that no LA is located at the center of the service area, and that only half of cells located at LA border, are involved in intra-VLR location updates. Thus, the total number of cells involved in generating inter-VLR updates is given by

\[ N_{intra-vlr} = N_{LA} \times [1/2 \times N_{PLA}] = 3 \times [1/2 \times 6\sqrt{N_{on}/3} - 3] = 12 \text{ cells} \]  

We further assume that only about one-third of cell crossings from the involved cells i.e. \( N_{intra-vlr} \), will in fact result in intra-VLR location updates (i.e. \( N_{intra-vlr} = 12 - \text{exact} : 12 \)).

\[ LocUpdSameVLR (/sa) = 1/3 \times X_{cell} \times N_{intra-vlr} \times [Pr(ON) - (\text{Erlang/terminal})] = 24.9 /s \]  

The term \([Pr(ON) - (\text{Erlang/terminal})]\) arises since only terminals that have powered-on and do not have a call-in-progress, may generate location updates. The total number of location updates (intra- plus inter- VLR) in the switch service area i.e. \( LocUpdates (/sa) \) is

\[ LocUpdates (/sa) = N_{LA} \times X_{LA} \times [Pr(ON) - (\text{Erlang/terminal})] = 74.7 /s \]  

The rate of inter-VLR location updates can be calculated using (2-17) and (2-18) as:

\[ LocUpdNewVLR (/sa) = LocUpdates (/sa) - LocUpdSameVLR (/sa) = 49.8 /s \]  

**PCS:** The service area now consists of \( N_{LA} = 48 \) LAs, out of which 23 are located on the border of SA. We assume that only half of cells of border-LAs, are involved in intra-VLR location updates. Thus, the total number of cells involved in generating intra-VLR updates is given by

\[ N_{intra-vlr} = [N_{LA} - nLA_{on-SA\_border}] \times N_{PLA} + [nLA_{on-SA\_border} \times (1/2 \times N_{PLA})] \]
\[
= \left[ \frac{48 - 23}{3} \right] \times \left[ \frac{6\sqrt{N_{\text{opt}}/3}}{2} \right] = 766 \text{ cells}
\]

We further assume that only about one-third of cell crossings from the involved cells i.e. \(N_{\text{intra-vlr}}\), will in fact result in intra-VLR location updates i.e. (i.e. \(N_{\text{intra-vlr}} = 766\)).

\[\text{LocUpdSameVLR} (\text{sa}) = \frac{1}{3} X_{\text{cell}} \times N_{\text{intra-vlr}} \times [ \text{Pr (ON)} - (\text{Erlang/terminal}) ] = 1585.5 \text{ /s} \tag{2-21}\]

The total number of location updates (intra- plus inter- VLR) in the switch service area is

\[\text{LocUpdates} (\text{sa}) = N_{\text{LA}} \times X_{\text{LA}} \times [ \text{Pr (ON)} - (\text{Erlang/terminal}) ] = 2392 \text{ /s} \tag{2-22}\]

Using (2-21) and (2-22), the rate of inter-VLR location updates can be now be calculated as

\[\text{LocUpdNewVLR} (\text{sa}) = \text{LocUpdates} (\text{sa}) - \text{LocUpdSameVLR} (\text{sa}) = 806.5 \text{ /s} \tag{2-23}\]

### 2.4.4 Total Signalling Load

Equations (2-6) .. (2-23) may now be combined with Table 2-3 to calculate the total SS7 signalling load (in bytes per call) as a function of the switch service area:

\[
\text{SS7 bytes/call} (\text{sa}) = \left[ k_1 \times \text{MO} (\text{sa}) + k_2 \times \text{MT} (\text{sa}) + \left( k_3 + k_4 \right) \times \text{Inter MSC H.O.} (\text{sa}) + k_5 \times \text{LocUpdSameVLR} (\text{sa}) + \left( k_6 + k_7 \right) \times \text{LocUpdNewVLR} (\text{sa}) \right] / \left[ \text{MO} (\text{sa}) + \text{MT} (\text{sa}) \right] \tag{2-24}
\]

The quantities \(k_i\) are the bytes per transaction taken from the relevant row of Table 2-3. It is implicitly assumed that the number of mobile-to-mobile calls is negligible. If required, this could be easily incorporated through addition of an extra term in (2-24). Using (2-24) and applying the results in Table 2-3, the number of SS7 bytes per call, to and from key network elements is calculated and shown in Table 2-5. It is evident that the signalling load in the SS7 network can be 8-20 times greater for GSM than for ISDN, and 2-3 times greater for PCS than for GSM, depending on whether VLR is located inside or outside the MSC.

<table>
<thead>
<tr>
<th>Total SS7 bytes (normalised to BHCA/SA)</th>
<th>ISDN</th>
<th>GSM</th>
<th>PCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Call</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TO/FROM MSC</td>
<td>120</td>
<td>1301.6</td>
<td>1705.3</td>
</tr>
<tr>
<td>TO/FROM VLR</td>
<td>-</td>
<td>-</td>
<td>2877.4</td>
</tr>
<tr>
<td>TO/FROM HLR</td>
<td>-</td>
<td>500.4</td>
<td>500.4</td>
</tr>
<tr>
<td>INTERNAL SS7 Bytes</td>
<td>120</td>
<td>1017</td>
<td>2514.8</td>
</tr>
</tbody>
</table>

Table 2-5: SS7 bytes per call

Table 2-6 (row4) indicates that the signalling load due only to call-related transactions i.e. call origination/terminations and handovers, is 1.8-6.8 times greater for GSM or PCS than for ISDN.
Chapter 2. Signalling Traffic in GSM Networks

Based on the last rows of Table 2-4 and Table 2-5 it can be concluded that VLR should be placed inside the MSC to minimise SS7 traffic. However, when databases are placed inside the switch, the MSC must devote resources to database transaction processing, which could otherwise be used for call processing/switching. This cost must be traded off against the SS7 savings in order to determine appropriate network configuration.

### Table 2-6: Call related SS7 traffic

<table>
<thead>
<tr>
<th></th>
<th>ISDN Call</th>
<th>GSM or PCS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VLR in MSC</td>
<td>VLR out MSC</td>
</tr>
<tr>
<td>TO/FROM MSC</td>
<td>120</td>
<td>177.5</td>
</tr>
<tr>
<td>TO/FROM VLR</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TO/FROM HLR</td>
<td>-</td>
<td>69.3</td>
</tr>
<tr>
<td>INTERNAL SS7 Bytes</td>
<td>120</td>
<td>218</td>
</tr>
</tbody>
</table>

### Table 2-7: relative (PCS-to-GSM) signalling costs

Although the selected metric for the above comparisons i.e. SS7 bytes/call, is a particularly useful metric in transmission planning / network dimensioning, a different view of the impact of signalling traffic could be had based on the rate of transactions i.e. in terms of required updates and queries, processed by the network databases. Table 2-8 shows that the HLR update and query rates in PCS can be as much 16(806.5/49.8) and 8(140.8/17.6) times higher than in GSM, respectively.

### Table 2-8: PCS/GSM signalling transaction rates

<table>
<thead>
<tr>
<th>Rate of Transactions</th>
<th>GSM</th>
<th>PCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CALL-related (MO &amp; MT)</td>
<td>51,848 - 63,360</td>
<td>414,720 - 506,880</td>
</tr>
<tr>
<td>HLR query rate:</td>
<td>17.5/s</td>
<td>140.8/s</td>
</tr>
<tr>
<td>Handover (InterMSC &amp; IntraMSC)</td>
<td>6,732 - 28,008</td>
<td>69,120 - 1,922,990</td>
</tr>
<tr>
<td>Location Update (IntraVLR &amp; InterVLR)</td>
<td>89,840 - 179,280</td>
<td>5,707,800 - 2,903,400</td>
</tr>
<tr>
<td>HLR update rate:</td>
<td>49.8/s</td>
<td>806.5/s</td>
</tr>
</tbody>
</table>

Note that the HLR update/query rates presented in Table 2-8, are per-MSC values. Total network-wide rate of HLR related transactions will require multiplication of the above rates by
the number of MSCs within PLMN (assuming one HLR per PLMN). Thus, assuming 10 MSCs per PLMN, the total rate of update and queries processed by the HLR, will be as shown in Table 2-9. We can also observer that the ratio of updates-to-queries increases from 2.8 in GSM, to 5.7 for PCS, but more important is the rate of updates that need to be supported by the HLR in PCS, which represents a significant challenge to today’s capabilities in terms of database system performance requirements (i.e. real-time response times and transaction processing capability).

<table>
<thead>
<tr>
<th>Total Rate of DB-Transactions</th>
<th>GSM</th>
<th>PCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(per PLMN (10 MSC in PLMN) busy hour rates)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CALL-related (MT)</td>
<td>63,380 x 10</td>
<td>506,880 x 10</td>
</tr>
<tr>
<td>HLR query rate:</td>
<td>176 /s</td>
<td>1408 /s</td>
</tr>
<tr>
<td>Location Update (InterVLR)</td>
<td>179,260 x 10</td>
<td>2,903,400 x 10</td>
</tr>
<tr>
<td>HLR update rate:</td>
<td>498 /s</td>
<td>8065 /s</td>
</tr>
</tbody>
</table>

Table 2-9: per PLMN HLR signalling transaction rates

The results from the 2nd row of Table 2-10 below indicate that in terms of percentage of total SS7 cost to MSC, it may be advantageous in PCS, to place VLR outside the MSC. Although a near 10% saving can be made by placing the VLR outside MSC as far as location update cost is concerned, doing so will result in an increase of equal amount in respect of the call cost (from %6.5 to %16.9). This clearly shows the trade off involved (as mentioned earlier) but the advantages to be gained obviously depend on the relative weighting associated with the call and location update costs.

<table>
<thead>
<tr>
<th>() as % of total SS7 load</th>
<th>CALL</th>
<th>Location Update</th>
</tr>
</thead>
<tbody>
<tr>
<td>VLR in MSC</td>
<td>VLR out MSC</td>
<td>VLR in MSC</td>
</tr>
<tr>
<td>GSM (MSC load)</td>
<td>% 44.4</td>
<td>% 12.6</td>
</tr>
<tr>
<td>PCS</td>
<td>% 6.5</td>
<td>% 16.9</td>
</tr>
<tr>
<td>GSM (VLR load)</td>
<td>-</td>
<td>% 22.2</td>
</tr>
<tr>
<td>PCS</td>
<td>-</td>
<td>% 8.6</td>
</tr>
<tr>
<td>GSM (HLR load)</td>
<td>% 13.8</td>
<td>% 7.4</td>
</tr>
</tbody>
</table>

Table 2-10: Call and LU related signalling load as percentage of total SS7 load

The results presented in this section, highlight the impact of the SS7 signalling traffic on network databases and switching equipment, as subscriber density is increased. It is shown that whereas with GSM traffic parameters there is a near balance between the call and location update related signalling loads (1st row Table 2-9), the asymmetry of signalling load with PCS parameters becomes significant, indicating that more of the MSC resources will need to be dedicated to LU processing (given that more than %90 of total signalling load is due to LUs –
2nd row Table 2-9) instead of call processing or pure switching. Also the high rate of HLR-related updates in the case of PCS, indicate a 16-fold increase compared to the updates in GSM. Therefore, it is clear that methods and techniques are required to minimise the impacts of database-related transactions on the SS7 network, by reducing LU signalling and processing on both the radio-interface and network equipment, as the network expands or new services are introduced.

2.5 Radio Resource Occupancy due to Mobility

In order to quantify the impact of location updates (LU) on RF usage on cells located at the border of a LA and on transactions processed at the MSC/VLR managing a set of LAs, the following notation is first introduced:

- \( TLU_i \): average duration of one location update in case \( i \) (\( i=1: \) intra-VLR LU ; \( i=2: \) inter-VLR with TMSI; \( i=3: \) inter-VLR with IMSI) equal to the time occupancy of one SDCCH/SACCH
- \( NLU_i \): no. of transactions processed by the MSC/VLR for one LU in case \( i \).
- \( PLU_{ij} \): percentage of LUs in case \( i \) in cell number \( j \) (each cell is identified by a number)
- \( TTN_{LU} \): no. of transactions due to LUs generated in LA perimeter cells
- \( \lambda_j \): no. of location updates experienced by cell \( j \) – per hour
- \( N \): number of cells per LA
- \( N_p \): number of cells located on the perimeter of one LA – this is given by [Alo92]

\[
N_{PLA} = N_p = \left[6\sqrt{N/3}\right] - 3 = 21
\] (2-25)

MSs are assumed to be uniformly distributed, cells are hexagonal and the maximum blocking probability for the allocation of an SDCCH is set to 1 percent. The number of location updates in an LA perimeter cell \( j \) per hour is calculated using (2-1), which is then used to obtain the SDCCH/SACCH occupancy in cell \( j \), due to MS location updates from

\[
T_{LU_j} = \lambda_j \cdot \left[ \sum_{i=1}^{3} PLU_{ij} \cdot TLU_i \right]
\] (2-26)

The number of transactions due to LUs generated in the \( N_{LA} \cdot N_p \) LA perimeter cells (which are numbered 1 to \( N_{LA} \cdot N_p \)) and processed per hour by the serving MSC/VLR is given by the formula
Chapter 2. Signalling Traffic in GSM Networks

\[ TTN_{LU} = \lambda_j \cdot \left[ \sum_{i=1}^{N_{LU}} \left( \sum_{j=1}^{\lambda_i} \cdot NLU_j \right) \right] \]  

(2-27)

For numerical calculations, we consider the values in Table 2-4 for PCS. From Table 2-4 we have:

\[ \rho = 15,360, \ r = 100m, \ \nu_{ave} = 10 \text{ km/hr}, \ N_{LA} = 48, \ N = 48 \text{ and } N_p = 21. \]

The number of accesses to the MSC/VLR is \( N_{LU_1} = 2, \ N_{LU_2} = 14 \text{ and } N_{LU_3} = 16 \)

Also based on practical measures done on a GSM network [X95]

\( T_{LU_1} = 600 \text{ ms}, \ T_{LU_2} = 3.5 \text{ s}, \ T_{LU_3} = 4.0 \text{ s}. \)

2.5.1 MSC/VLR Mobility-related Transaction Load

Two cases are considered: case 1 - a cell where only intra-VLR LUs are generated, and case 2 - a cell where only inter-VLR LUs are generated.

**Case 1** - The considered cell-j is located at the border of two LAs related to the same VLR. Only intra-VLR location updates will be processed in the cell (IMSI-Attach procedure is ignored). Equation (2-26) yields \( T_{LU_j} = 39 \text{ Erl.} \), which based on 2% blocking probability, requires \( 49/8 = 6.1 \) channels (i.e. three quarter of channels available per carrier).

**Case 2** - The considered cell-j is located at the border of two LAs related to the different VLRs. Only inter-VLR location updates will thus be processed in the cell. It is assumed that 80% of LUs use TMSI and only 20% use IMSI. Equation (2-26) then yields \( T_{LU_j} = 234 \text{ Erl.} \), and based on 2% blocking probability, requires \( 244/8 = 30 \) channels (nearly 4 RF carriers).

To evaluate the load on the MSC/VLR, calculated parameters of Table 2-4 are used. Thus, the service area consists of \( N_{LA} = 48 \text{ LAs.} \) There are 23 LAs located on the border of service area. It is also assumed that the center LAs (25 LAs) and half the perimeter of border LAs will process only the intra-VLR location updates. For the other half of the perimeter cells of the border LAs, only inter-VLR LUs are generated. Thus the number of cells where intra-VLR LUs are generated is given by equation (2-28), and is equal to

\[ 25N_p + 23(N_p/2) = 25 \cdot 21 + 23 \cdot 21/2 = 766 \]  

(2-28)

and the number of cells where only inter-VLR LUs are processed is equal to

\[ 23(N_p/2) = 69 \cdot \left[ \sqrt{(N/3)} - (1/2) \right] = 242 \]  

(2-29)
It is further assumed that among the generated inter-VLR LUs, 80% use TMSI and remaining 20% use IMSI. Based on the above simple scenario, the number of transactions to be processed by the MSC/VLR due to LUs in its LAs is computed using (2.27) and is finally equal to

$$TTN_{LU} = 1.2 \times 10^9$$ transactions at peak hour.

The results presented in this section show that under heavy traffic conditions, the impact of LUs can be significant. In terms of radio channel usage, it was shown that between three-quarter to 4 RF carriers could be required to support location area crossings. Although this burden may not directly lead to call blocking on the radio interface (in a DCS 1800 network using a 12 reuse cluster, the average number of available RF carriers per cell, with three operators, is about 10), it has nevertheless a visible impact on the traffic channel consumption. Also in terms of processing at the MSC/VLR, requiring the processing of $1.2 \times 10^9$ transactions per hour, can quickly lead to blocking, as MSC/VLR resources dedicated to servicing LU processing, can not be used for providing call-related services.

### 2.6 Concluding Remarks

The above examples illustrate clearly that the SS7 signalling load due particularly to location management of subscribers constitutes a significant proportion of total signalling traffic, and support of mobility will substantially increase the rate of database-related transactions i.e. update and queries. Although the examples presented concerned specifically the case where subscriber penetration level was increased, it is easy to extend the argument to other cases where one could (given that all other parameters remain unchanged), assess the impact of cell size reduction, increase in mobility due to higher average velocity or increase in the number of service areas/switching equipment, under similar conditions. All other traffic parameters being equal, it was shown that for an eight-fold increase in subscriber density,

- The additional burden on the SS7 network can be 8-20 times greater for GSM than for ISDN, and 2-3 times greater for PCS than for GSM, depending on whether VLR is located inside or outside the MSC.
- HLR update/query rates in PCS can be as much 16/8 times higher, respectively,
- HLR update-to-query ratio increases from 2.8 in GSM to 5.7 for PCS,
- The asymmetry of the total signalling load becomes significant, indicating that more than 90% of the MSC resources could be required to be dedicated to location management instead of call processing,
Chapter 2. Signalling Traffic in GSM Networks

- Considering intra-VLR location updates, radio resource requirement per cell on the perimeter of a LA, would be 6 channels i.e. three quarter of channels available per carrier,
- Inter-VLR location updates in the cells on the border of SA, will require 30 channels i.e. nearly 4 RF carriers.

The rate of transactions to be processed by the MSC/VLR is high and may lead to call blocking as significant amount of resources will have to be dedicated to location management signalling processing.

It is therefore evident that methods and techniques are required to minimise the impacts location management on the switching equipment and network databases within the SS7 network, by reducing location management signalling via more efficient techniques or alternative methods, as the network expands or new services are introduced.
Chapter 3

THE ANCHORING TECHNIQUE FOR REDUCING SIGNALLING COSTS IN GSM NETWORKS

3 Introduction

Management of subscriber mobility in the next generation of mobile networks is expected to result in significant network signalling due to mobile terminal locating and tracking operations, and therefore more efficient mobility management techniques are required to reduce the network signalling traffic and the database-related signalling load. Based on the results presented in chapter II, the impact of mobility related signalling within the second generation cellular systems, can be such that call processing and location management of subscribers in PCS, generate two to three times the signalling load compared to GSM. Also, the support of new services will require changes in mobile switch software, at the same time that the new service logic is introduced into the HLR. Therefore introduction of new services and features will place additional signalling load of the network databases. As the number of mobile subscribers increases, new or improved location management schemes are needed to effectively support a continuously increasing subscriber population. The classical user location management strategies are based on a two-tier hierarchy of HLR and VLR databases. The user performs a registration at the HLR and the new serving VLR, every time the user changes the registration area, and de-registers at the previous VLR. When a call arrives, the network simply routes the call to the last known/reported location of the mobile terminal. However, if for instance, the incoming call arrival rate is low compared to the mobility rate of the terminal the classical schemes are not cost effective and in many cases it should be possible to avoid these registrations at the HLR. More is discussed in the literature review section (section 3.1).

In this chapter the Anchoring technique is introduced and analysed. The technique is intended to minimise the expensive HLR access, thereby reducing the overall network signalling load and preventing the HLR from becoming a bottleneck. The basic operation of the anchoring scheme is as follows. Under the scheme, the signalling traffic due to location updates is reduced by the requirement to report location changes (location updates alluded to here refer in fact to VLR
updates, which would normally be reported to the HLR) to another VLR, instead of the HLR. The serving VLR during the last call delivery, is designated as the Anchor VLR (AVLR) of the subscriber. As the subscriber moves to a new LA (assuming that each MSC/VLE service area consists of one LA), the location update messages are reported to the subscriber AVLR, instead of the HLR. Since the HLR keeps a pointer to the AVLR and not the serving VLR, when a call arrives, the HLR queries the AVLR of the called subscriber, to obtain a routing address to the subscriber's serving/current VLR. Also, when the call arrives, the serving VLR becomes the new AVLR of the subscriber. This change of AVLR is however reported to the HLR. Later in the chapter, optimisation of the basic anchoring technique is presented and analysed. The basic anchoring technique is shown to successfully reduce the location updating and the total network signalling costs compared to classical strategy of GSM. Although by distributing the signalling load over a number of VLRs, the HLR is prevented from becoming a bottleneck in the signalling network, the basic anchoring scheme can result in rather high call routing costs compared to GSM at low Call-to-Mobility Ratio (CMR). To alleviate this problem, an optimisation of the basic technique is proposed. The optimisation procedure is based on identification of the optimal trade-off point between the call routing and location updating costs, and analytical results indicate that the proposed optimisation strategy can significantly decrease the call routing and locating costs whilst maintaining the cost of HLR access well below that of GSM compared to the original anchor scheme. The cost evaluations take into account the signalling loads generated due to registration and de-registration procedures as well as call routing and location updating thus providing a more complete picture of the total network signalling cost. Then the proposed optimisation procedure is applied and results are compared with the basic scheme. Also more importantly, the effects of the proposed optimisation procedure on database access times are examined and compared to the original scheme. Call routing cost for the fixed-to-mobile and mobile-to-mobile calls are also evaluated separately.

The outline of this chapter is as follows. Section 3.1 presents an overview of recent research on location management and call delivery techniques. This is followed by the reference network architecture in section 3.2. In section 3.3 the Anchoring technique and its application in GSM networks is described. In section 3.4 cost evaluation methods are shown followed by presentation of the analytical model of the anchoring technique in section 3.5. The proposed optimisation procedure of the anchoring technique is introduced in section 3.6. The results based on the basic scheme are described and discussed in section 3.7. This is followed by the results of the optimised scheme and database impacts which are presented and discussed in section 3.8. Finally, the conclusions are presented in section 3.9.
3.1 Literature Survey

Recent Research On Location Registration and Call Delivery

Location registration involves updating location databases when current location information is available. On the other hand, call delivery involves the querying of location databases to determine the current location of a called MS. These can be costly processes, especially when the MS is located far from its assigned HLR. As the number of mobile subscribers keeps increasing, the volume of signalling traffic generated by location management is becoming extremely high [Lo92][Wol92]. Methods for reducing the signalling traffic are therefore needed. Research in this area generally falls into two categories. First, extensions to the current IS-41/GSM location management strategies are developed which aim to improve the existing schemes while keeping the basic database network architecture unchanged. This type of solution has the advantage of easy adaptation to the current PCS networks without major modification. These schemes are based on the centralized database architecture inherited from the GSM standard. Another category of research results in completely new database architectures which require a new set of schemes for location registration and call delivery. Most of these schemes are based on distributed database architectures. Some additional research efforts include: the reverse virtual call setup, a new scheme for delivering mobile-terminated calls [Pol97], an optimal routing scheme based on the ratio of source messaging to location update rates [Yate96], and a single registration strategy for multi-tier PCS systems [Li96]. In the following, centralised vs distributed architectures are discussed.

Centralized Database Architectures - In [Jain94] and [Lin94] a per-user location caching strategy is introduced. This location management scheme aims to reduces the cost of call delivery i.e. volume of signalling and database access traffic for locating an MS, by using the cached location information (stored at a nearby STP) obtained during a previous call, instead of expensive HLR access. Whenever the MS is accessed through the STP, an entry is added to the cache which contains a mapping from the ID of the MS to that of its serving VLR. When another call is initiated for an MS, the STP first checks if a cache entry exists for the MS. If no cache entry for the MS exists, the IS-41 call delivery scheme described earlier is used to locate the MS. If a cache entry exists, the STP will query the VLR as specified by the cache. If the MS is still residing under the same VLR, a hit occurs, and the MS is found. If the MS has already moved to another location which is not associated with the same VLR, a miss occurs, and the IS-41 call delivery scheme is used to locate the MS. Figure 3-1 demonstrates the operation of per-user location caching. When a call is initiated from MS 1 to MS 2, as indicated in Figure 3.1, the system can locate MS 2 by using the cached information at STP1. As a result,
MS 2 is successfully located without querying the HLR of MS2. The per-user location caching allows the STP to locate the VLR of the called MS after only one cache database lookup. This is true, however, only when the cached location information of the called MS is valid (a hit). The cost of per-user location caching is higher than the IS-41 scheme when a miss occurs. Based on the system parameters, the minimum hit ratio required to produce a performance gain using per-user location caching needs to be determined.

In [Jain94], the authors define the local call-to-mobility ratio (LCMR) as the average number of calls to an MS from a given originating STP divided by the average number of times the user changes VLR per unit time. The minimum LCMR necessary to attain the minimum hit ratio is obtained. In order to reduce the number of misses, it is suggested in [Lin94] that cache entries should be invalidated after a certain time interval. Based on the mobility and call arrival parameters, a T-threshold scheme is introduced in [Lin94] which determines the time when a particular cached location information should be cleared such that the cost for call delivery can be reduced.

The cache scheme as proposed in [Jain94][Lin94] were based on the assumption that HLR access time is constant and MS residence times have an exponential distribution. In comparing the PCS location tracking strategies [Lin96] the previous cache-based model is extended by considering the queuing effect at HLR (which is modelled as an M/G/1 queue) and an arbitrary distribution i.e. Gamma, for MS residence time. It is shown that the cache scheme is likely to outperform the IS-41 scheme when the hit-ratio in the cache scheme is larger than zero, when the MS mobility is low, and for a fixed mean HLR service time, the variance of the HLR service time distribution is large. Also, for a fixed mean MS residence time, it is shown that a higher cache hit ratio is possible for residence distributions with large variance. Although it is shown that the cost of location update can be lower than basic IS-41 scheme, this improvement over the basic scheme, is only significant for large call-arrival rates and low mobility subscribers i.e. high CMR region, since when the call frequency is low with respect to the mobility, the number
Chapter 3. The Anchoring Technique for Reducing the Signalling Costs

of location updates/registrations increases, significantly degrading the performance of the cache-based schemes.

The dynamic hierarchical database architecture [Ho97] proposal is based on the IS-41 standard, with the addition of a new level of databases, called directory registers (DR). Each DR covers the service area of a number of MSCs. The primary function of the DRs is to periodically compute and store the location pointer configuration for the MSs in its coverage area. Each MS has its unique pointer configuration and three types of location pointers are available at the DR:

- A local pointer is stored at an MS's serving DR which indicates the current serving MSC.
- A direct remote pointer is stored at a remote DR which indicates the current serving MSC.
- An indirect remote pointer is stored at a remote DR which indicates the current serving DR of the MS.

In addition, the HLR may be configured to store a pointer to either the serving DR or the serving MSC of the MS. In some cases, it may be more cost-effective not to set-up any pointers and the original IS-41 scheme will be used. In the proposed scheme, if there are significant number of incoming calls for MS, a direct or indirect remote pointer can be set up for the MS in the DR covering the region where calls are originated. Thus, when the next call is initiated for this MS from the same region, the calling MSC first queries the DR and the call can be immediately forwarded to serving MSC without requiring a query at the HLR which may be far from the current location of subscriber. This reduces the signalling overhead for call delivery. On the other hand, the HLR can be set up to record the ID of the serving DR (instead of the serving MSC) of the MS. When the MS moves to another MSC within the same LA in the current service area, only the local pointer at the serving DR of the MS has to be updated. Again it is not necessary to access the HLR. This reduces the signalling overhead for location registration. The advantage of this scheme is that it can reduce the overhead for both location registration and call delivery, however this is achieved at the cost of introducing an additional layer of databases (DR layer) and signalling between the DRs and the HLR, however the scheme is effective at reducing the signalling cost, particularly at low CMR values.

A user-profile replication scheme is proposed in [Shiv95]. Based on this scheme, user profiles are replicated at selected local databases. When a call is initiated for a remote MS, the network first determines if a replication of the called MS's user profile is available locally. If the user profile is found no HLR query is necessary, and the network can locate the called MS based on the location information available at the local database. Otherwise, the network locates the
called MS following the GSM procedures. When the MS moves to another location, the network updates all replications of the MS’s user profile. This results in higher signalling overhead for location registration. Depending on the mobility rate of the MS and the call arrival rate from each location, this method may significantly reduce the signalling and database access overhead for local management. A scheme is also introduced in [Shiv95] which determines the replication for each MS. Based on their scheme, the replication decision is made by a centralized system which must collect the mobility and calling parameters of the whole user population from time to time. This may not be feasible in current PCS networks because of the large number of PCS network providers involved. Besides, generating and distributing the replication decision for a large user population is a computation-intensive and time-consuming process which may incur a significant amount of network bandwidth. Future research should focus on the development of distributed user profile replication mechanisms.

A pointer forwarding strategy is introduced in [Jain95]. The proposed location tracking method based on forwarding strategy is used to reduce the cost of location registration. The basic idea is that instead of reporting a location change to the HLR every time the MS moves to an area belonging to a different VLR, the reporting can be eliminated by simply setting up a forwarding pointer from the old VLR to the new VLR. The technique is based on maintaining a pointer (resulting in a succession of pointers) to the new VLR, by the old VLR when a terminal crosses a service area boundary. For incoming calls, the HLR must then determine the current serving VLR by following the pointer chain to the current serving VLR. Location registration is performed when the length of the forwarding chain exceeds a particular length \( k \). A maximum of \( k+1 \) VLRs need to be interrogated in order to locate a terminal. This technique incurs additional delays at call set-up.

To minimize the delay in locating an MS, the length of the pointer chain is limited to a predefined maximum value, \( K \). When the length of the pointer chain reaches \( K \), additional forwarding is not allowed, and location change must be reported to the HLR when the next movement occurs. Figure 3-2 demonstrates the operation of pointer forwarding. Pointers are...
set up from VLR 1 to VLR 2 and from VLR 2 to VLR 3 as the MS moves from MSC 1 to MSC 2 and from MSC 2 to MSC 3, respectively. For \( K = 2 \), the pointer chain cannot be extended any further. An additional movement from MSC 3 to MSC 4 will result in a location registration at the HLR. The original pointers are deleted, and the HLR records the ID of the current serving VLR of the MS. It is demonstrated that, depending on the mobility and call arrival parameters and the value of \( K \), this scheme may not always result in a reduction in cost from the original IS-41 scheme. The authors determine the conditions under which the pointer forwarding scheme should be used based on the system parameters.

**Distributed Database Architectures** - A fully distributed database architecture for location registration is proposed in [Wan93]. The two-level HLR/VLR database architecture as described in the IS-41 standard is replaced by a large number of location databases. These location databases are organized as a tree with the root at the top and the leaves at the bottom. The MSs are associated with the leaf (lowest-level) location databases, and each location database contains location information for the MSs residing in its subtree. Figure 3-3 demonstrates the operation of the proposed scheme. Given that an MS, MS 1, is located at RA 1, an entry exists for MS 1 in each database along the path from its current location to the root of the tree.

The entries for MS 1 at these databases are as shown in Figure 3-3. When a call is initiated, the network locates the called MS by following its database entries; for example, if a call for MS 1 is initiated by MS 2, as shown in Figure 3.3. The call request is received by node A. Since the database of node A does not have an entry for MS 1, the call request is forwarded to node B and so on. When the request finally reaches node D, an entry for MS 1 is found and the location of MS 1 determined after another three database lookups, as demonstrated in Figure 3-3. When an MS moves to an RA that belongs to a different leaf database, the corresponding databases are updated to indicate the correct location of the MS. When compared to schemes based on a centralized database architecture, such as the IS-41 scheme, the proposed scheme reduces the distance travelled by signalling messages, through localisation of updates – the databases in the nodes at layer \( i \) and above need not be updates for all the movements of MS. However, this scheme increases the number of database updates and queries, and thus increases the delay in location registration and call delivery.
Chapter 3. The Anchoring Technique for Reducing the Signalling Costs

In [Kim95] a new distributed location management algorithm is introduced that attempts to address some of the problems associated with the centralised, hierarchical distributed databases, identified in the following. The centralised hierarchically organised databases though simple conceptually, would be impractical to implement and maintain due to large number of databases involved, and the hierarchical distribution requires complex signalling procedures and will incur long call set up delays as it has to process update and location query requests at multiple databases. For example, in [Wan93] the cost of updates and queries is reduced by localisation of updates and a tracing process in a layered distributed database, and though it is possible to reduce the signalling traffic and distance through localisation, the cost in terms of increased multiple database access and complexity of management procedure is significant. The paper proposes a new database architecture, based on eliminating VLRs and maintaining only the home databases for location information storage. When an MS moves into a different location area, a location updating message is delivered only to its home database to indicate the current location address. At call set up, location information is retrieved once from the called user's home database, to locate the MS. The technique has significant overheads when the MS is roaming far from home, as signalling traffic has to be delivered over long distances, but the authors maintain that this overhead from the signalling transport network is likely to be reduced in the future optical-fiber based networks and by introduction of high-speed switching. The registration scheme proposed in [Kim95] is based on the structure divided into \((n+1)\) hierarchical layer nodes with the lowest (layer 0) representing the smallest independent LA comprising several cells as shown in Figure 3-4.
Figure 3-4: Registration in the distributed hierarchical database architecture

Location update messages/requests to each layer node, is delivered using source routing whose destination address is the home database address of MS. In comparison with the distributed algorithm of [Wan93], the proposed method results in 10-50% reduction in the rate of database updates as the number of levels of hierarchy increases. Also, the number of database queries in the proposed scheme remains constant whereas this cost will increase as the switching hierarchy grows and more databases have to be enquired to locate the called MS, as would be the case with [Wan93]. The results and improvements however, are sensitive to the location area size assumed, and thus to the size of paging area and location update rate of subscribers.

A partitioning scheme for the fully distributed database hierarchy is introduced in [Bad92]. Since the mobility pattern of MSs varies among locations, partitions can be generated by grouping location servers among which the MS moves frequently. Based on the scheme introduced in [Bad92], location registration is performed only when the MS enters a partition. Figure 3-5 shows the partitions for a particular PCS network. Partition P2 consists of 5 location servers that have a least common ancestor location server LS2. When an MS moves into partition P2, LS2 is updated to indicate that the MS is residing in its subtree. No location registration is performed when the MS moves to another location server within the same partition. The scheme minimizes the number of location registrations in areas where the mobility rate of MSs is high. Simulation results demonstrate that the partitioning scheme is effective in reducing the signalling message cost. However, the cost reduction depends on the mobility and call arrival patterns as well as the method used for searching the subtree. Further study is needed to determine the effectiveness of this scheme under various parameters.
Chapter 3. The Anchoring Technique for Reducing the Signalling Costs

Based on the pointer forwarding technique of [12], the one-step pointer forwarding strategy for location tracking in distributed HLR environment is proposed in [Sue97]. Pointer forwarding was used to reduce the expensive HLR access and a distribute HLR scheme in IS-41 and GSM would require multiple HLR updates to maintain consistency of valid location information and re-synchronisation in case of a HLR failure, in [Sue97] the pointer forwarding technique is integrated into a distributed HLR scheme to prevent the HLR from becoming a bottleneck in the signalling network. However, since, the length of the forwarding pointer chain may be lengthened in a distributed HLR environment, the authors propose migration of the locating chains when an MS issues a location registration operation. Consequently, the length of any forwarding pointer chain does not exceed one. However, since migration of forwarding chains requires additional signalling between VLRs, for an MS which often moves across SA boundaries and does not receive calls frequently, i.e. low CMR region, a considerable signalling capacity will be wasted. Nevertheless the paper attempts to strengthen the feasibility and attractiveness of the concept of distributed HLRs.

In [Ana94], another distributed database architecture similar to that discussed in [Wan93] is introduced. Here, MSs may be located at any node of the tree hierarchy (not limited to the leaf nodes). The root of the tree contains a database, but it is not necessary for other nodes to have databases installed. These databases store pointers for MSs. If an MS is residing at the subtree of a database, a pointer is set up in this database pointing to the next database along the path to the MS. If there is no other database along this path, the pointer points to the residing node of the MS. When a call for an MS is initiated at a node on the tree, the called MS can be located by following the pointers of the MS. Figure 3-6 shows the operation of this scheme. We assume that a call is initiated at node A and the called MS is located at node B. The path for searching the called MS is given in Figure 3-6. If a database that does not contain a pointer for the called MS is reached, the next database along the path to the root is queried. Given the system parameters, such as the rate of movement between boundary location areas, the authors
introduce a method for determining the database placement that reduces the number of database accesses and updates.

![Figure 3-6: Distributed database architecture.](image)

### 3.2 Reference Network Architecture

The two level hierarchy of GSM network architecture is depicted in Figure 3-7. Signalling transfers between the nodes pass through signalling transfer points (STP) and for reliability and security reasons, the STPs are usually installed in so called “Mated Pair” configuration. Access links (A-link) provide access into the network and to the network databases through the STP. There are always at least two A-links, one to each of the home STP pairs. Cross links (C-link) connect an STP to its mate STP. To provide redundancy, STPs are deployed in pairs. Normal SS7 traffic is not routed over these links except in cases of congestion.

![Figure 3-7: Reference network architecture](image)
level. D-links are only found in networks based on a hierarchical structure, therefore not all networks deploy D-links. MSC/VLRs and the HLR, are inter-connected through the STPs via local and remote A-links.

### 3.3 Anchoring Technique in GSM

#### 3.3.1 Location Registration and Call delivery mechanisms

The anchoring technique relies on the observation that due to either the small size of the switch service areas - as might be the case in high-density urban areas - or simply the high mobility of the subscribers, many LA boundaries may be crossed before a call delivery and routing by the network is required. Note that we are assuming that a MSC/VLR service area contains a single LA. Such high rate of LA crossings will result in significant signalling traffic towards the HLR whilst it generates no revenues for the operator.

![Figure 3-8: (a) Location registration and (b) Mobile Terminated (MS) call in the proposed scheme](image)

Therefore by distributing the location updating signalling traffic to another VLR i.e. the current Anchor VLR (AVLR), instead of the HLR, the signalling traffic towards a single central node in the system i.e. the HLR, can be substantially reduced. Each mobile terminal may have a different AVLR and the AVLR for a mobile will change from time to time. When a VLR is selected as the Anchor for a particular MS, an extra entry indicating the current (which will not be the same as the Anchor, once the subscriber has left the Anchor) VLR is set-up. The address available to the HLR for call routing, is that of the AVLR. Updating of the HLR i.e. reporting an AVLR change to the HLR, is only carried out when a call is delivered to MS. At all other times HLR will contain the VLR address where the MS last received a call i.e. that of the current AVLR. Therefore, in the proposed architecture, the frequency of HLR updates is reduced. With reference to Figure 3-8a we outline the steps involved in the location updating procedure based on the anchoring technique. We assume that the mobile subscriber was initially
registered in MSC/VLR_{Anchor}, has then moved to MSC/VLR_{old} but has just crossed into MSC/VLR_{new}.

1. Subscriber has moved to a new LA. The mobile sends a *Location Update Request* message to MSC_{new}.
2. The MSC_{new} updates its associated VLR via an *Update Location Area* message.
3. Since MS has initiated location updating using LAI&TMSI, VLR_{new} requests IMSI and authentication parameters from VLR_{old} using the *Send Parameters (IMSI)* message.
4. The VLR_{old} then sends back the requested parameters in the *IMSI response* message which includes the identity of the VLR_{Anchor}. Note that the MS has identified itself using LAI and TMSI assigned originally by MSC/VLR_{old}. The VLR_{new} then initiates authentication and ciphering and then allocates a new TMSI (TMSI_{new}) to MS. TMSI_{new} is sent to MS in ciphered mode.
5. Assuming that authentication has succeeded, VLR_{new} informs VLR_{Anchor} of the new location using the *Update Location* message.
6. VLR_{Anchor} then responds with a *Location Updating Accepted* message.
7. VLR_{Anchor} then directs VLR_{old} to delete the entry for the mobile using the *Cancel Location* message.
8. This message is acknowledged with a *Location Cancel Accepted* message.

With reference to Figure 3-8b, the steps involved in call delivery procedure based on the proposed technique, are as follows:

1. The incoming call is routed from the ISDN/PSTN to the mobile network GMSC.
2. The Gateway MSC using called subscriber's MSISDN number will request routing information from HLR in order to route the call.
3. The HLR will then request AVLR for a roaming number.
4. AVLR will in turn request SVLR for a Mobile Subscriber Roaming Number (MSRN).
5. The MSRN provided by SVLR is passed onto AVLR.
6. MSRN is passed onto the HLR.
7. The HLR then sends this routing information back to the GMSC. The GMSC uses MSRN to set-up call towards serving MSC.
8. As SVLR is now the new AVLR for the called MS, the HLR must be informed of the change in AVLR. Now the SVLR sends an *Update Location* message to the HLR.
9. HLR responds with a *Location Updating Accepted* message, after the authentication has succeeded. HLR can then cancel the old AVLR using the *Cancel Location* message. This is acknowledged by the old AVLR with a *Location Cancel Accepted* message.

10. The serving MSC then pages the called MS in all the cells within the indicated LA, and upon successful page response, initiates authentication and ciphering of the radio path. ISDN SETUP and its acknowledgement CALL CONFIRMED messages are then exchanged between the serving MSC and the called MS and then following the transmission of ALERT and CONNECT messages by MS, End-to-End speech path between the caller and the MS is set-up.

Note that, instead of the HLR, the new AVLR can also initiate cancellation of the old AVLR. Also, the change to the new AVLR (and informing of the HLR of the change) takes place as part of the call delivery process. This does not affect the MS call set-up as the SVLR - shortly to become the new AVLR - can set-up signalling towards the HLR, independently and with no interference to the ongoing dialogue re. call delivery to the MS.

### 3.3.2 Detach / Attach Procedures

The signalling exchanges during the Detach procedure, for both GSM and the anchor technique are shown in Figure 3-9 and Figure 3-10 respectively.

![Figure 3-9: signalling exchange during DETACH in GSM](image)

In GSM phase 2, a GSM *DETACH* procedure results in the permanent removal of subscriber record from the SVLR and the setting of the *DETACH* flag within the HLR. Also, if the HLR receives a message from a VLR other than the one listed in the HLR, then the HLR automatically sends a *Cancel Location* message towards that VLR. This particular feature of GSM ensures consistency of data in the HLR. Figure 3-10 below shows MAP signalling exchanges during the *DETACH* procedure in the proposed scheme.

![Figure 3-10: signalling exchange during DETACH in the anchoring technique](image)
Chapter 3. The Anchoring Technique for Reducing the Signalling Costs

Note that since ATTACH (power-on/SIM insertion) is performed in the SVLR, the procedure only involves updating of the AVL and the SVLR. GSM ATTACH procedure is similar to the location updating message exchanges of Figure 3-8, except that during ATTACH, message exchanges re. cancellation of old VLR is not required, as this is done during the preceding DETACH procedure.

3.4 Cost Evaluation Procedure

In this section an analytical model for evaluation of the cost of the proposed anchoring scheme is presented. Location management cost based on different mobility and call arrival rates is evaluated. The location tracking cost is divided into Call Routing and Location Updating cost, however the total signalling cost includes ATTACH/DETACH cost as well as location tracking cost.

- **Call Routing Cost** ($C_R$): With reference to the call delivery mechanism of Figure 3-8b, the routing cost is the cost of delivering a call including informing the HLR of the change in AVL i.e. HLR update and the removal of the subscriber record from the previous AVL.
- **Location Updating Cost** ($C_U$): With reference to the location updating procedure of Figure 3.8a, the updating cost is the cost of informing the AVL of the current location of the terminal, and removal of the subscriber record from the old VLR by AVL.
- **ATTACH/DETACH Cost** ($C_{AD}$): This is the cost of informing the AVL of the current status of the terminal i.e. whether terminal is switched-on or switched-off.

The following parameters are used in the calculation of cost functions.

- $LSTP_c$: cost of message routing through the Local STP
- $RSTP_c$: cost of message routing through the Remote STP
- $LA_c$: cost of signalling over the Local A-Link
- $RA_c$: cost of signalling over the Remote A-Link
- $DL_c$: cost of signalling over the D-Link
- $H_c$: cost of an HLR access - an update or interrogation
- $V_c$: cost of an VLR access - an update or interrogation

Since both the call delivery and location updating procedures as well as ATTACH/DETACH operations involve exchange of signalling messages between two MSCs, a number of possible
paths between two MSCs can be identified. The cost of signalling through each path is different and these costs may be classified as:

- The signalling path from one MSC to another goes through the Remote STP and the HLR. Both the HLR and VLR may be interrogated. This cost is represented by:
  \[ S_1 = 2 \text{LSTP}_c + 2 \text{RSTP}_c + 2 \text{LA}_c + 2 \text{RA}_c + 2 \text{DL}_e \]

- The signalling path from one MSC to another goes through the Remote STP. An example would be reporting of the new LA to the AVLR - connected to the Remote STP, by the serving VLR which is geographically distant from AVLR and is served by a Local STP. This cost is represented by:
  \[ S_{II} = 2 \text{LSTP}_c + \text{RSTP}_c + 2 \text{LA}_c + 2 \text{DL}_e \]

- The signalling path from one MSC to another goes through the Local STP. An example would be reporting of the new LA to the AVLR, by the serving VLR. This cost is represented by:
  \[ S_{III} = \text{LSTP}_c + 2 \text{LA}_c \]

- The signalling path from GMSC to another MSC goes through the HLR. Both the HLR and VLR may be interrogated. For instance, in the case of a MS call, where the GMSC must interrogate the HLR in order to route the call to the MSC of the called subscriber. This cost is represented by:
  \[ S_{IV} = \text{LSTP}_c + \text{RSTP}_c + \text{LA}_c + 3 \text{RA}_c + \text{DL}_e \]

### 3.4.1 Determination of the Call Routing cost

Now, since the called mobile could be in any location in the network, the cost of searching for the mobile in order to route an incoming call is location dependent. Cost of locating a mobile for two types of mobile-terminated calls i.e. mobile-to-mobile and fixed-to-mobile are discussed separately.

**Mobile-to-Mobile Calls**

Note that the following categorisations assume that the calling and the called terminals are residing in different LAs. The following three cases can be identified:

- The called MS is residing in the AVLR. The cost in this case is represented by
Chapter 3. The Anchoring Technique for Reducing the Signalling Costs

CR₁ = 2 S₁ + Hₑ + 3Vₑ

- The called MS is in a LA different from its AVLR. The Serving MSC/VLR (SVLR) and AVLR are within the same local STP region. This cost is represented by
  CR₂ = 3 S₁ + 4 Sₘ + 2Hₑ + 5Vₑ

- The called MS is in a LA different from its AVLR. The SVLR and AVLR are within different local STP regions. This cost can be represented by
  CR₃ = 3 S₁ + 4 Sₘ + 2Hₑ + 5Vₑ

**Fixed-to-Mobile Calls**

The following three cases can be identified:

- The called MS is residing in the AVLR. The cost in this case is represented by
  CR₄ = 2 Sᵣᵥ + Hₑ + 2Vₑ

- The called MS is in a LA different from the AVLR. The SVLR and AVLR are within the same local STP region. GMSC is considered to be located in the remote STP region. The cost in this case is represented by
  CR₅ = 2 Sᵣᵥ + S₁ + 4 Sₘ + 2Hₑ + 4Vₑ

- The called MS is in a LA different from the AVLR. The SVLR and AVLR are within different local STP regions. GMSC is considered to be located in the remote STP region. The cost in this case is represented by
  CR₆ = 2 Sᵣᵥ + S₁ + 4 Sₘ + 2Hₑ + 4Vₑ

<table>
<thead>
<tr>
<th>MS Call Types</th>
<th>Case</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile-to-Mobile Calls</td>
<td>1</td>
<td>CR₁ = 2 S₁ + Hₑ + 3Vₑ</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>CR₂ = 3 S₁ + 4 Sₘ + 2Hₑ + 5Vₑ</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>CR₃ = 3 S₁ + 4 Sₘ + 2Hₑ + 5Vₑ</td>
</tr>
<tr>
<td>Fixed-to-Mobile Calls</td>
<td>4</td>
<td>CR₄ = 2 Sᵣᵥ + Hₑ + 2Vₑ</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>CR₅ = 2 Sᵣᵥ + S₁ + 4 Sₘ + 2Hₑ + 4Vₑ</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>CR₆ = 2 Sᵣᵥ + S₁ + 4 Sₘ + 2Hₑ + 4Vₑ</td>
</tr>
</tbody>
</table>

### 3.4.2 Determination of the Location Updating cost

Evaluation of the cost of location updating, requires information about the location of mobile terminal before and after location updating (reporting of the new LA to the AVLR + AVLR cancellation of old VLR) operation. Assuming that the mobile terminal has performed n LA
border crossings since the AVLR was last changed, Table 3-2 shows all possible types of crossings, when the \((n+1)\)th crossing is performed. AVLR is assumed to be in the LSTP.

<table>
<thead>
<tr>
<th>Present Location after (n) crossings</th>
<th>Location after the ((n +1))th crossing</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVLR</td>
<td>local STP</td>
<td>(LU_1 = 6 S_m + 4 V_e)</td>
</tr>
<tr>
<td>AVLR</td>
<td>remote STP</td>
<td>(LU_2 = 6 S_n + 4 V_e)</td>
</tr>
<tr>
<td>local STP</td>
<td>AVLR</td>
<td>(LU_3 = 2 S_m + 2 V_e)</td>
</tr>
<tr>
<td>local STP</td>
<td>local STP</td>
<td>(LU_4 = 8 S_m + 5 V_e)</td>
</tr>
<tr>
<td>local STP</td>
<td>remote STP</td>
<td>(LU_5 = 2 S_m + 6 S_n + 5 V_e)</td>
</tr>
<tr>
<td>remote STP</td>
<td>AVLR</td>
<td>(LU_6 = 2 S_n + 2 V_e)</td>
</tr>
<tr>
<td>remote STP</td>
<td>local STP</td>
<td>(LU_7 = 4 S_m + 4 S_n + 5 V_e)</td>
</tr>
<tr>
<td>remote STP</td>
<td>remote STP (same LSTP)</td>
<td>(LU_8 = 2 S_m + 6 S_n + 5 V_e)</td>
</tr>
<tr>
<td>remote STP</td>
<td>remote STP (different LSTP)</td>
<td>(LU_9 = 8 S_n + 5 V_e)</td>
</tr>
</tbody>
</table>

**3.4.3 Determination of the ATTACH/DETACH cost**

ATTACH/DETACH (A/D) operations are assumed to be independent processes, and so if a mobile is not location updating due to movement, or is not engaged in a MS call, then it contributes to the total network signalling load, through the A/D processes. Therefore, the signalling load due to these operations contributes to the total network signalling load and must therefore be included in the calculation of total signalling cost. The signalling exchange between the serving and Anchor VLRs re. A/D procedures can involve different paths. As the cost of signalling through each path is different, these costs are classified as shown in Tables 3.3 and 3.4 below.

<table>
<thead>
<tr>
<th>Present Location after (n) crossings</th>
<th>Cost of DETACH (Figure 3-10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVLR</td>
<td>(D_1 = S_1 + 2V_e + H_e)</td>
</tr>
<tr>
<td>LSTP</td>
<td>(D_2 = S_1 + S_m + 2V_e + H_e)</td>
</tr>
<tr>
<td>RSTP</td>
<td>(D_3 = S_1 + S_n + 2V_e + H_e)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Present Location after (n) crossings</th>
<th>Cost of ATTACH</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVLR</td>
<td>(A_1 = S_1 + V_e + H_e)</td>
</tr>
<tr>
<td>LSTP</td>
<td>(A_2 = 4S_m + S_1 + H_e + 3V_e)</td>
</tr>
<tr>
<td>RSTP</td>
<td>(A_3 = 4S_n + S_1 + H_e + 3V_e)</td>
</tr>
</tbody>
</table>
3.5 Analytical Model of the Anchoring Technique

Figure 3-11 shows the embedded Markov chain model which captures the mobility and call arrival patterns of a mobile terminal. The state of the chain, \( n \), is defined as the number of LA crossings since the AVLR was last changed. State transition occurs immediately before a mobile terminal’s departure from a LA. Since a movement will occur right after a state transition, the number of movements since AVLR was last updated, is \( n+1 \). A transition from state \( n \) to state \( n+1 \) (denoted by \( \mu \)) occurs when there are no call arrivals in the transition interval. Similarly, a transition from state \( n \) to zero (denoted by \( \gamma \)) occurs when at least one call arrives between the \((n+1)\)th and \((n+2)\)th movements, i.e. a new AVLR is selected only after a call arrival.

Let \( \omega_c \) and \( \omega_x \) be random variables representing call inter-arrival time and LA residence time, respectively. Assuming poisson call arrivals, the call inter-arrival times \( \omega_c \) are then exponentially distributed with rate \( \lambda_c \). It is also assumed that \( \omega_x \) has probability density function denoted by \( f_x(t) \) and mean \( 1/\lambda_x \) representing mean LA residence time. Then state transition probability from state \( n \) to state \( n+1 \) denoted by \( \mu_{n,n+1} \), can be obtained as:

\[
\mu_{n,n+1} = \int_0^\infty e^{-\lambda_c t} f_x(t) \, dt = \int_0^\infty e^{-\lambda_x t} \lambda_x e^{-\lambda_c t} \, dt
\]

and therefore, \( \mu_{n,n+1} = \lambda_x / (\lambda_x + \lambda_c) \) \hspace{1cm} (3-2)

The probability that one or more calls arrive between two LA crossings, denoted by \( \gamma_{n,0} \), can now be obtained as

\[
\gamma_{n,0} = \Pr \{ \omega_c \leq t : \forall n \} = \lambda_c / (\lambda_x + \lambda_c) \] \hspace{1cm} (3-3)

The model of Figure 3-11 can be used to obtain the relative costs of location update and call routing as well as total network signalling cost i.e. the sum of location update and call routing...
Chapter 3. The Anchoring Technique for Reducing the Signalling Costs

costs, of the anchor technique. The total cost however, does not take account of the signalling load due to the Attach/Detach operations. Therefore, in order to obtain a more realistic indication of the total network signalling cost, the markov model of Figure 3-11 has been modified - shown in Figure 3-12 - to include state transitions due to Attach/Detach operations. The modified markov model of Figure 3-12 ensures that in the calculations of the total network signalling cost, contributions due to attach/detach procedures are included.

### 3.6 Analytical Model of The Modified Anchoring Technique

Figure 3-12 shows the modified markov chain model that is used for the calculation of the total signalling cost incorporating mobility, call arrivals and user attach/detach. Parameters \( \lambda_A \) and \( \lambda_D \) represent the Attach and Detach probabilities respectively. The state transition probability from state \( n \) to state \( n+1 \) denoted by \( \mu_{n,n+1} \), can now be obtained as:

\[
\mu_{n,n+1} = \int_0^\infty e^{-\lambda_x t} \cdot e^{-\lambda_c t} \cdot e^{-\lambda_A t} \cdot f_X(t) \, dt = \int_0^\infty e^{-[\lambda_x + \lambda_c + \lambda_A + \lambda_D] t} \cdot \lambda_x \cdot e^{-\lambda_A t}
\]

(3-4)

The probability that one or more calls arrive between two LA crossings, denoted by \( \gamma_{n,0} \), can now be obtained as:

\[
\gamma_{n,0} = \Pr \{ \omega_c < t : \forall n \} = \frac{\lambda_c}{\lambda_x + \lambda_c + \lambda_A + \lambda_D}
\]

(3-5)

![Figure 3-12: The modified markov chain model](image)

The state transition probabilities for ATTACH and DETACH operations are denoted by \( \delta_{n,n} \) and \( \eta_{n,n} \) respectively. We assume \( \eta_n \) to be the equilibrium state probability of state \( n \). The expression for \( \eta_n \) \( n \geq 0 \) is \( \eta_n = \eta \cdot P_n + \mu \cdot P_{n+1} \). Hence \( \eta_n \) can be expressed as
\[
P_n = \mu / (\gamma + \delta + \mu) \cdot P_{n-1} = (\mu / (\gamma + \delta + \mu))^n \cdot P_0
\]

where \( P_0 \) is the equilibrium state probability of state 0. Since \( \sum_n P_n = 1 \) \( P_0 \) can be expressed as

\[
P_0 = (\gamma + \delta + \mu) \cdot P_0 + \sum_{n=1}^\infty P_n \cdot (\gamma + \delta) = (\gamma + \delta) / \mu \cdot (1 - P_0)
\]

\[
P_0 = (\gamma + \delta) / (\gamma + \delta + \mu).
\]

Let \( Q[n]_{AVLR} \), \( Q[n]_{LSTP} \), \( Q[n]_{RSTP} \) to represent respectively the probability that a mobile is located at the AVLR, in local STP (LSTP) and in remote STP (RSTP) regions, \( n \) crossings after the AVLR was last changed. Also let \( R_i[n+1] \) to represent the probability that the \( n \)th and \( (n+1) \)th crossings after AVLR was changed, are movements belonging to the set \{ 1 \leq i \leq 9 \} as shown in Table 3-2. Derivation of probabilities \( Q[n] \) and \( R_i[n+1] \) are given in [Ho96]. Let the expected LX and CR costs during the mobile's stay in state \( i \) (between the instants that the transition into and the transition out of the state \( i \) occur) of the modified markov chain, be \( C_{lx}[i] \) and \( C_{ci}[i] \) respectively. The expression for \( C_{lx}[i] \) is

\[
C_{lx}[i] = \sum_{k=1}^9 R_k (i+1).
\]

The average updating cost per state transition (\( P_x \)) is

\[
\overline{\text{cost}}_{lx} = \sum_{n=0}^\infty P_n \cdot C_{lx}[n] = \sum_{n=0}^\infty [\mu / (1 - \delta)]^nP_0 \cdot C_{lx}[n]
\]

\[
= [\gamma / (1 - \delta)] \cdot \sum_{n=0}^\infty [\mu / (1 - \delta)]^n \cdot C_{lx}[n]
\]

The average location updating cost per unit time is \( \text{cost}_{lx} = \lambda_x \cdot \overline{\text{cost}}_{lx} \)

Assuming that \( \rho_1 \) is the average number of call arrivals between two LA crossing, then

\[
\rho_1 = \lambda_c / \lambda_x \text{ where } \lambda_c \text{ is the call arrival rate. The expressions for } C_{cr}[i] \text{ can be obtained as}
\]

\[
C_{cr}[i] = \{ (\rho_1 - \gamma) \cdot CR_1 + \gamma \cdot Q[i+1]_{AVLR} \cdot CR_1 + Q[i+1]_{LSTP} \cdot CR_2 + Q[i+1]_{RSTP} \cdot CR_4 \} \cdot \alpha
\]

for mobile-to-mobile calls, where \( \alpha \) is the percentage of mobile-to-mobile calls, and

68
Chapter 3. The Anchoring Technique for Reducing the Signalling Costs

\[ C_{CR}[i] = \left\{ (P_1 - \gamma) \cdot CR_4 + \gamma \cdot Q[i+1]_{AVLR} \cdot CR_4 + Q[i+1]_{LSTP} \cdot CR_5 
+ Q[i+1]_{RSTP} \cdot CR_6 \right\} \cdot \beta \]  

(3-11)

for fixed-to-mobile calls, where \( \beta \) is the percentage of fixed-to-mobile calls. Note that the sum \((\alpha + \beta)\) represents the total number of mobile terminated (MS) calls irrespective of the call origin. So,

\[ \text{cost}_{cr}[i] = C_{cr}[i] + C_{cr}[i]. \]  

(3-12)

The average call routing cost per state transition is

\[ \overline{\text{cost}_{CR}} = \sum_{n=0}^{\infty} P_n \cdot C_{CR}[n] = \sum_{n=0}^{\infty} [\mu / (1 - \eta)]^n P_0 \cdot C_{CR}[n] \]

\[ = [\gamma / (1 - \eta)] \cdot \sum_{n=0}^{\infty} [\mu / (1 - \eta)]^n \cdot C_{CR}[n] \]  

(3-13)

Hence the average call routing cost per unit time is

\[ \text{cost}_{cr} = \lambda_x \cdot \overline{\text{cost}_{cr}}. \]  

(3-14)

Assuming that \( \rho_2 \) is the average number of ATTACH arrivals between two LA crossing, then

\[ \rho_2 = \lambda_A / \lambda_x \]  

where \( \lambda_A \) is the attach rate. Let \( C_A[i] \) represent the expected cost of the attach procedure. The expression for \( C_A[i] \) is:

\[ C_A[i] = (\rho_2 - \delta) \cdot A_1 + \delta \cdot \left\{ Q[i+1]_{LSTP} \cdot (A_1) + Q[i+1]_{RSTP} \cdot (A_2) 
+ Q[i+1]_{AVLR} \cdot (A_3) \right\} \]  

(3-14)

Similarly, the expected cost of the detach procedure during a mobile's stay in state \( i \), is:

\[ C_D[i] = (\rho_3 - \delta) \cdot D_1 + \delta \cdot \left\{ Q[i]_{LSTP} \cdot (D_1) + Q[i]_{RSTP} \cdot (D_2) 
+ Q[i]_{AVLR} \cdot (D_3) \right\} \]  

(3-15)

The average ATTACH and DETACH costs per state transition are

\[ \overline{\text{cost}_{A}} = \sum_{n=0}^{\infty} P_n \cdot C_A[n] = \sum_{n=0}^{\infty} [\mu / (1 - \eta)]^n P_0 \cdot C_A[n] \]

\[ = [\gamma / (1 - \eta)] \cdot \sum_{n=0}^{\infty} [\mu / (1 - \eta)]^n \cdot C_A[n] \]  

(3-16)
Chapter 3. The Anchoring Technique for Reducing the Signalling Costs

\[ \text{cost}_D = \sum_{n=0}^{\infty} P_n \cdot C_D[n] = \sum_{n=0}^{\infty} \left( \frac{\mu}{(1 - \eta)} \right)^n P_0 \cdot C_D[n] = \left[ \frac{\gamma}{(1 - \eta)} \right] \cdot \sum_{n=0}^{\infty} \left( \frac{\mu}{(1 - \eta)} \right)^n \cdot C_D[n] \quad (3-17) \]

The average ATTACH cost per unit time is \( \text{cost}_A = \lambda_A \cdot \text{cost}_A \) and the corresponding DETACH cost per unit time is

\[ \text{cost}_D = \lambda_D \cdot \text{cost}_D. \quad (3-18) \]

The total cost per unit time for the Anchoring technique is therefore:

\[ \text{Cost}_{\text{Total}} = \text{cost}_{\text{CR} \cdot} + \text{cost}_{\text{LX}} + \text{cost}_A + \text{cost}_D. \quad (3-19) \]

### 3.6.1 Determination of signalling costs in GSM

In GSM, the signalling exchange between the SVLR and the HLR re. A/D procedures involve different paths. As the cost of signalling through each path is different, these costs are classified as follows:

- **GSM DETACH** - The signalling path from the SVLR to the HLR passes through both the RSTP and LSTP regions. Both the serving VLR and the HLR are updated. This cost is represented by:
  \[ D_A = S_1 + H_e + V_e \]

- **GSM ATTACH** - The signalling path from the SVLR to the HLR passes through both the RSTP and LSTP regions. The HLR is accessed twice and both the SVLR and the HLR are updated. This cost is represented by:
  \[ A_A = 3S_1 + 2H_e + 2V_e \]

- **GSM Location Update** - The signalling path from the SVLR to the HLR passes through both the RSTP and LSTP regions. Both the SVLR and the HLR will be updated. The cost also includes cancellation and interrogation of old VLR. This cost is represented by:
  \[ \text{LU}_{16} = 3S_1 + S_m + H_e + 4V_e \]

- **GSM Call Routing** - The signalling cost for mobile-to-mobile and fixed-to-mobile calls are presented separately. These costs are represented by:

  - **Mobile-to-Mobile Calls**:
    \[ \text{CR}_7 = 2S_1 + H_e + 3V_e \]
  
  - **Fixed-to-Mobile Calls**:
    \[ \text{CR}_8 = 2S_{1V} + H_e + 2V_e \]
3.6.2 Total cost of the GSM scheme

Let \( G_R \), \( G_X \) and \( G_A \), \( G_D \) be per unit time routing, updating and A/D costs in GSM. The relevant expressions are:

\[
G_{CR} = (\alpha CR + \beta CR_0) \cdot \lambda_c
\]

\[
G_{LX} = LU_{10} \cdot \lambda_X
\]

\[
G_A = A_4 \cdot \lambda_A
\]

\[
G_D = D_4 \cdot \lambda_D
\]

Where \( \lambda_c \) is the call arrival rate and \( \lambda_X \) is the LA crossing rate (\( 1/\lambda_X \) denotes the mean LA residence time). The total cost per unit time, is therefore \( G_{\text{total}} = G_{CR} + G_{LX} + G_A + G_D \).

3.7 Analytical Results of Basic Scheme

Now, we will define the Call-to-Mobility Ratio (CMR) such that \( \text{CMR} = \lambda_c / \lambda_X \). A high CMR indicates a large number of calls arrive between two consecutive location updates and vice versa. Since a particular value of CMR can not be representative of the call/mobility patterns of all subscribers, we will consider a wide range of CMR values from 0.01 to 100, and determine the performance of the proposed scheme as CMR varies.

<table>
<thead>
<tr>
<th>Set</th>
<th>( S_1 )</th>
<th>( S_{II} )</th>
<th>( S_{III} )</th>
<th>( S_{IV} )</th>
<th>( H_0 )</th>
<th>( V_c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>4</td>
<td>1</td>
<td>10</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>8</td>
<td>1</td>
<td>10</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>23</td>
<td>1</td>
<td>10</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

As all costs are calculated based on \( S \) parameters, by assigning normalised values to these parameters, we can then calculate values for routing and updating cost formulae, based on these parameters. Table 3-5 represents six sets of normalised values selected for \( S_n \) parameters. The value of \( S_{III} \) is normalised to 1 as it represents the lowest cost parameter. Parameter sets 1 to 3 represent cases when it is significantly more expensive to send a message through the HLR than sending a message through the LSTP. This is true when the HLR is far away from the MSC and communication cost is high, or when the cost for accessing the HLR is high. Parameter sets 4 to
Chapter 3. The Anchoring Technique for Reducing the Signalling Costs

6 capture the cases when the communication cost between the MSC and HLR is low and HLR access is inexpensive. Note that set 6 represents an extreme situation and it is expected that the anchoring technique will perform the worst under this data set, as no cost savings can be obtained by reducing the access to HLR. Although this set may not correspond to a realistic situation, it can be used as the upper bound and for comparison purpose. In order to show the cost reduction obtained by the proposed scheme, compared to GSM, plots of the relative costs $C_{CR}/G_{CR}$, $C_{LX}/G_{LX}$, $C_{A}/G_{A}$, $C_{D}/G_{D}$, $C_{Total}/G_{Total}$ are derived and presented in Figure 3-13. The size of a LSTP region is set to 64 and CMR is varied from 0.01 to 100. User attach/detach rates are set to 10% of the call rate i.e. $\lambda_{A} = 0.1 \cdot \lambda_{C}$, in the analysis for both the anchor technique and GSM. The Relative cost (Y-axis) represents the cost of Anchor scheme, relative to the cost of GSM scheme.

Figure 3-13: (a) Call routing cost (MM: 20% - FM: 80%) (b) Call routing cost (MM: 80% - FM: 20%) (c) Location updating cost and (d) Total network signalling costs. ($\lambda_{A} = 0.1 \cdot \lambda_{C}$)
As can be seen in Figure 3-13(d), for low CMR the reduction in total cost is very significant when the cost for sending a message through the HLR, \( S_1 + H_c + V_c \), is relatively high (parameter sets 1 to 3 and 5). However, when \( S_1 + H_c + V_c \) is relatively low (parameter sets 4 and 6) the total cost can be higher than that of GSM. These results are expected as because Anchoring reduces the number of messages going through the HLR at the expense of increasing the number of messages exchanged between two MSCs. If this assumption does not hold, (as in cases 4 and 6) there may not be a saving in total cost. Figure 3-13 (a) and (c) show the call routing cost and location updating cost of Anchoring scheme as compared to GSM. It can be seen that Anchoring always results in higher call set-up cost although this cost approaches that of GSM as call rate increases i.e. at high CMR, while the location updating cost is significantly lower in most cases.

It can be seen from Figure 3-13 (d) that the total cost increases as the CMR decreases below 0.1 for all the data sets other than data sets 1 and 5. When the CMR is low a large number of movements occur before the arrival of a call and there is a high probability that the mobile terminal is residing outside the Anchor LSTP region. As a result signalling messages between the AVLR and the SVLR have to go through the RSTP instead of LSTP, resulting in higher location update and call routing costs. When the CMR is high, the call arrival rate is high relative to mobility rate. As the AVLR is changed after each call arrival, the AVLR is the same as the SVLR most of the time. In this situation, anchoring is similar to GSM scheme, and the total cost approaches that of GSM regardless of the cost parameters selected. It is important that in the calculation of the total network signalling cost, all contributions are taken into account. Failure to do so can result in misleading estimates of the total signalling cost. In the calculation of the total signalling cost, we use the ratio of the sum of costs for the proposed technique, to the corresponding sum of costs for GSM Figure 3-13 (b) shows call routing cost of the anchor scheme with mobile-to-mobile calls making up 80% and fixed-to-mobile calls forming the remaining 20% of the mobile terminated calls as CMR varies.

When CMR is high, large number of call arrivals occur between two location area crossing. Then, the current SVLR is the same as the AVLR most of the time. In this case, Anchoring is similar to GSM where location updating cost is independent of the call arrival pattern while the call routing cost is independent of the mobility pattern of the mobile terminal. As long as the mobility and call arrival rates are not changed, the total cost is not affected. When CMR is low, a large number of movements occur between two call arrivals and the AVLR is seldom changed between two movements. In this case, the effect of call arrivals on the total cost is small. The total cost is therefore, almost equal to the movement cost. When the CMR is close to one, the
call arrival and mobility rates are similar. If a mobile terminal stays at a AVLR for longer than the mean LA residence time, a number of calls may arrive before the next movement. Since the AVLR is changed whenever a call arrives, the AVLR is the same as SVLR after the first call arrival. Setting up a connection for subsequent call arrivals involves only two VLR interrogations, thus resulting in lower overall cost. If a mobile terminal stays at a AVLR for a period shorter than the LA residence time, there is a low probability that more than one call will arrive. As a result, setting up of a connection involves an intermediate AVLR which results in higher search costs.

3.7.1 Database impacts

Figure 3-14(a) shows that as CMR increases, HLR access cost i.e. sum of interrogations and updates, reaches that of GSM. However, within the region of interest i.e. CMR between 0.1 and 10, a cost reduction of between 80 - 20% compared to GSM can be achieved.

![Figure 3-14: (a) HLR access cost, (b) VLR access cost (λ_A = 0.1 • λ_C)](image)

The reductions in HLR access cost are achieved at the expense of increased VLR related signalling. This is evident from Figure 3-14(b), where VLR access cost shows an increase of about 22% for CMR values in the range 0.1 to 10. Note that this increase is bounded and is a small price to pay for significant reductions in the HLR access cost, as it is often the HLR which as a single network node can become a bottleneck and source of significant call set-up delay particularly during congestion periods.
3.8 Optimisation of the Anchoring Technique

It has been shown that the basic anchoring scheme can reduce the cost of location updating compared to GSM. However in most cases, the cost of call routing is 2 to 2.5 times that of GSM within CMR range 0.01 - 1.0. We now propose an optimisation of basic anchoring scheme (BAS) is based on the optimisation procedure described below, and is denoted as the optimised anchoring scheme (OAS). This technique is only applied to the CMR region of interest i.e. 0.01 to 1.0 to reduce the cost of call routing. The procedure is in three parts:

1. based on the trade-off between the (raw) call routing and location update costs of the modified anchor scheme, obtain the optimal number of crossings (denoted as optimal_n, at which costs of call routing and location updating both, are minimum, as depicted in Figure 3-15) for a given CMR, and
2. then use the optimal_n values to re-calculate the relative costs, for each CMR value.
3. repeat procedure parts 1 and 2 for different data sets.

Given any CMR, it is postulated that an optimal point exists at which the costs of call routing and location updating can both be minimised. By plotting the raw (normalised) costs of call routing and location updating of the anchor scheme against the number of crossings, as shown below, the optimal trade-off point i.e. optimal_n, is determined for each CMR and under different network signalling loads i.e. for different data sets. For example, in Figure 3-15(a) we observe that at CMR of 0.01 the optimal_n is 7 i.e. for a CMR of 0.01, it is optimal (in terms of balancing both the call routing and location updating costs of the network) to report location changes to the HLR, every 7 location crossing. The results depicted in Figure 3-15 below show how the optimal_n values have been derived. A value of 500 crossings (denoted as the range value) have been used in the simulations, but the x-axis values do not extend to the range value of 500 to ease viewing and reading of optimal_n values from the graphs. The range values indicate our assumption for the total number of crossings that a mobile user may experience, in 24 hour period for example. Note that since in the calculation of the relative costs, presented later, the same range value is used for both GSM and the anchor schemes, the particular choice of the range value has no impacts on the final results of the optimisation procedure (the results provided in Figure 3.17).
The observed behaviour in the diagrams of Figure 3-15 can be explained as follows. In the anchoring scheme, as the number of LA crossings by the user becomes more frequent, the call routing cost shows an initial increase, up to about 15-20 crossings, but levels off and remains almost constant thereafter. This is because as the number of crossings extends beyond a certain point, the mobile will be away from the anchor VLR or the LSTP region where AVLR is situated, and may be located with a high probability at either a neighbouring LSTP or the RSTP regions, because of which the call routing cost will remain at almost constant levels. The cost of reporting location updates to the AVLR however will increase with the number of crossings, as expected. The optimal trade-off points, optimal \( n \) values, can now be read from the graphs as points of intersection between call routing and location update costs for each CMR value, as depicted in the diagrams of Figure 3-15. Figure 3-16 shows the results of the simulations with a
range value of 100 crossings. The CMR and corresponding optimal \( n \) values are shown on the diagrams.

![Diagram](image)

Figure 3-16: Call routing vs location update cost: (a) and (b) set 4 , (c) and (d) set 6 results.

Having obtained the optimal \( n \) (at integer multiples of which, the serving MSC will report the location change to the HLR instead of AVLR), we return to the modified anchoring scheme and re-calculate the relative costs, whilst taking into account the cost of reporting updates to the HLR (every optimal \( n \)), cost of de-registration of old AVLR and registration of serving VLR as the new AVLR. Prior to re-calculation of the relative costs however, the cost of next movement that requires information about the type of the movement, needs to be determined. Table 3-6 presents expressions for each possible location type, when the \((n+1)th\) crossing is performed. AVLR is assumed to be in the LSTP.
Table 3-6. Classification of Location Updating cost based on movement

<table>
<thead>
<tr>
<th>Present Location after n crossings</th>
<th>Location after the (n+1)th crossing</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVLR</td>
<td>local STP</td>
<td>LU_{10} = 6 S_i + 2 S_m + 4 V_e + H_e</td>
</tr>
<tr>
<td>AVLR</td>
<td>remote STP</td>
<td>LU_{11} = 6 S_i + 2 S_m + 4 V_e + H_e</td>
</tr>
<tr>
<td>local STP</td>
<td>AVLR</td>
<td>LU_{12} = 2 S_m + 2 V_e</td>
</tr>
<tr>
<td>local STP</td>
<td>remote STP</td>
<td>LU_{13} = 6 S_i + 2 S_m + 4 V_e + H_e</td>
</tr>
<tr>
<td>local STP</td>
<td>remote STP (same LSTP)</td>
<td>LU_{14} = 6 S_i + 2 S_m + 4 V_e + H_e</td>
</tr>
<tr>
<td>remote STP</td>
<td>AVLR</td>
<td>LU_{15} = 2 S_m + 2 V_e</td>
</tr>
<tr>
<td>remote STP</td>
<td>local STP</td>
<td>LU_{16} = 6 S_i + 2 S_m + 4 V_e + H_e</td>
</tr>
<tr>
<td>remote STP</td>
<td>remote STP (different LSTP)</td>
<td>LU_{17} = 6 S_i + 2 S_m + 4 V_e + H_e</td>
</tr>
<tr>
<td>remote STP</td>
<td>remote STP</td>
<td>LU_{18} = 6 S_i + 2 S_m + 4 V_e + H_e</td>
</tr>
</tbody>
</table>

Using the optimal \( n \) values for different data sets, we can now re-calculate the expressions required for the calculation of the relative signalling costs. Let the expected location update and call routing costs during the mobile’s stay in state \( i \) of the modified Markov chain, be denoted by \( C_{LX}[i] \) and \( C_{CR}[i] \) respectively. The expression for new \( C_{LX}[i] \) (denoted by \( NC_{LX}[i] \)), applied only to CMR range 0.01 - 0.5, is then determined as follows:

\[
C_{LX}[i] = \sum_{k=1}^{9} R_i(i+1) LU_k \quad \text{for} \quad 0 \leq i < \text{optimal}_n - 1 \tag{3-20}
\]

\[
C_{LX}[^{	ext{optimal}_n - 1}] = \sum_{k=10}^{18} R_i(\text{optimal}_n) LU_k \tag{3-21}
\]

Therefore, \( NC_{LX}[i] = C_{LX}[i] + C_{LX}[^{	ext{optimal}_n - 1}] \). The average updating cost per state transition (\( P_i \)) is then

\[
\bar{\text{cost}}_{LX} = \sum_{n=0}^{\text{optimal}_n} P_i \cdot NC_{LX}[n] = \sum_{n=0}^{\text{optimal}_n} \left[ \frac{\mu}{(1 - \delta)} \right]^n P_0 \cdot NC_{LX}[n]
\]

\[
= \left[ \frac{\gamma}{(1 - \delta)} \right] \cdot \sum_{n=0}^{\text{optimal}_n} \left[ \frac{\mu}{(1 - \delta)} \right]^n \cdot NC_{LX}[n] \tag{3-22}
\]

The average location updating cost per unit time is \( \text{cost}_{LX} = \lambda_X \cdot \bar{\text{cost}}_{LX} \).

Assuming that \( \rho_1 \) is the average number of call arrivals between two LA crossing, then \( \rho_1 = \frac{\lambda_c}{\lambda_X} \) where \( \lambda_c \) is the call arrival rate. The new expressions for \( C_{CR}[i] \) is
Chapter 3. The Anchoring Technique for Reducing the Signalling Costs

\[ C_{CR}[i] = \left( (\rho_1 - \gamma) \cdot CR_1 + \gamma \cdot (Q[i+1]_{AVLR} \cdot CR_1 + Q[i+1]_{LSTP} \cdot CR_2 + Q[i+1]_{RSTP} \cdot CR_3) \right) \cdot \alpha \quad \text{for } 0 \leq i < \text{optimal}_n - 1 \] (3-23)

\[ C_{CR}[\text{optimal}_n - 1] = \left( (\rho_1 - \gamma) \cdot CR_1 + \gamma \cdot (Q[\text{optimal}_n]_{AVLR} \cdot CR_1) \right) \cdot \alpha \] (3-24)

Therefore, \[ ^1NC_{CR}[i] = C_{CR}[i] + C_{CR}[\text{optimal}_n - 1] \], for mobile-to-mobile calls, and

\[ C_{CR}[i] = \left( (\rho_1 - \gamma) \cdot CR_4 + \gamma \cdot (Q[i+1]_{AVLR} \cdot CR_4 + Q[i+1]_{LSTP} \cdot CR_5 + Q[i+1]_{RSTP} \cdot CR_6) \right) \cdot \beta \quad \text{for } 0 \leq i < \text{optimal}_n - 1 \] (3-25)

\[ C_{CR}[\text{optimal}_n - 1] = \left( (\rho_1 - \gamma) \cdot CR_4 + \gamma \cdot (Q[\text{optimal}_n]_{AVLR} \cdot CR_4) \right) \cdot \beta \] (3-26)

Therefore, \[ ^2NC_{CR}[i] = C_{CR}[i] + C_{CR}[\text{optimal}_n - 1] \], for fixed-to-mobile calls. So

\[ C_{CR}[i] = ^1NC_{CR}[i] + ^2NC_{CR}[i] \] (3-27)

The average call routing cost per state transition is

\[ \text{cost}_{CR} = \sum_{n=0}^{\text{optimal}_n} P_n \cdot C_{CR}[n] = \sum_{n=0}^{\text{optimal}_n} \left[ \mu / (1 - \eta) \right]^n \cdot P_n \cdot C_{CR}[n] \]

\[ = \left[ \gamma / (1 - \eta) \right] \cdot \sum_{n=0}^{\text{optimal}_n} \left[ \mu / (1 - \eta) \right]^n \cdot C_{CR}[n] \] (3-28)

Hence the average call routing cost per unit time is \( \text{cost}_{CR} = \lambda_x \cdot \text{cost}_{CR} \).

### 3.8.1 Results of the proposed optimisation procedure

The results of the optimisation procedure are shown in Figure 3-17. For each CMR value (range 0.01 to 1), the corresponding \text{optimal}_n value is used to evaluate the costs in both GSM and the anchor schemes. This process is then repeated for each data set independently.
Chapter 3. The Anchoring Technique for Reducing the Signalling Costs

The improvements in call routing cost depicted in Figure 3-17(a) are two fold. The relative cost of call routing is now between 1.5 - 1.7 times that of GSM as CMR varies in the range 0.01 to 1, which compares favourably with the non-optimised anchor the cost of call routing is 2 - 3.5 times that of GSM. Moreover, not only the costs for all data sets i.e. sets 1 to 6, have been reduced by between 50 to 70% compared to SAS, the application of optimisation procedure to all data sets, has resulted in a relative cost of about 1.6 for all data sets. Thus, irrespective of network loading, the OAS results in a call routing cost which is 1.6 times that of GSM within CMR range of 0.01 to 1, for all data sets.

The effect of the proposed optimisation procedure on the location updating cost is shown in Figure 3-17(b). Although in response to lower call routing cost, the raw cost of location
updating has increased, the relative cost of location updating as shown in Figure 3-17 (b), is less than that of SAS shown Figure 3-13(c). The gain obtained in applying OAS is due to the fact that the optimisation procedure, results in more frequent reporting to the HLR and hence the corresponding increase in the number of new AVLRs. Since the relative cost is calculated based on the number of transitions/movements from the anchor VLR, the ratio of the location update costs (anchor to GSM) decreases, resulting in the observed behaviour. Also, comparison of Figure 3-13(c) and Figure 3-17(b) reveal that except for data set 6 which as mentioned previously represents an extreme case, the location update cost under OAS is now less than half that of GSM. Finally, the total signalling cost under OAS is now below that of GSM for all data sets except set 6.

3.8.2 Database Implications

Further insight into the impacts of OAS may be gained by observing the effect of OAS on total HLR and VLR access costs.

![Graph](image.png)
The results depicted in Figure 3-18(a) to (f) below, clearly show that the application of the optimisation procedure results in increased total HLR access cost while at the same time, the total cost of VLR access is reduced, which has been expected, as there are now more frequent accesses to the HLR than before (and less frequent VLR access).

Comparison of Figure 3-14(a) and (b) with Figure 3-18(a) and (b) for instance shows how the increase in the relative (total) HLR access cost form 0.2 to 0.4 is accompanied by a corresponding decrease in the relative (total) VLR access cost from 1.25 to 1.1, within CMR region of 0.01 to 0.1. Comparison of Figure 3-14(a) and (b) with Figure 3-18(c) and (d) for data set 4 and with Figure 3-18(e) and (f) for data set 6, all show that the OAS procedure can
maintain the total HLR access cost of the anchor technique always below that of GSM, while managing to reduce the cost associated with the VLR access compared to BAS. From an operational point of view, the implementation of the proposed optimisation procedure can be based on either of the following methods:

1. The AVLR collects statistics about the number and frequency of movements or LA crossings. This information is used to estimate/calculate the appropriate range value. The VLR also needs to collect statistics about incoming call rates in order to estimate subscriber CMR. Once CMR and range values are determined, based on the optimisation procedure of section 3.8, the optimal \( n \) value is derived, and passed onto the terminal. The terminal then reports movement updates to the HLR, instead of AVLR, at integer multiples of \( n \). The terminal will need to maintain a movement counter however.

2. The terminal is assumed to maintain counters for the number of movements and call arrival and a timer to calculate the average the time between movements, over a given period. Estimates for the range value can be calculated using the number of movements and time between movements. The required CMR value can then be estimated based on the number of call arrivals and the movements, over the same period. Then, the optimal \( n \) value may be: (1) either fetched from an internal table or (2) obtained by the terminal over the air-interface if such information is broadcast by the system. The \( n \) value thus derived, can be used by the terminal to report movement updates to the HLR, instead of AVLR at integer multiples of \( n \).

3.9 Conclusions

Each of the proposed schemes for PLMN mobility management can improve the GSM/IS-41 strategies to a certain extent. However, it is difficult to select a scheme that clearly outperforms the others under all system parameters. In most cases, the performance of the proposed schemes exceeds that of the GSM/IS-41 only under certain mobility and call arrival parameters. When a different set of parameters is used, the performance may be changed significantly. It is, however, possible to make several general observations. As described in section 3.1, recent research efforts in location registration and call delivery are based on either the centralized or the distributed database architectures. The centralized approach records the location information of all MTs in the centralized HLR. Signaling messages are exchanged between the current location of an MT and the HLR during location registration and call delivery. As the number of MTs increases, the signaling traffic may significantly degrade the performance of the PLMN network. One undesirable consequence is that the connection setup delay may become
very high. On the other hand, an advantage of the centralized approach is that the number of database updates and queries for location registration and call delivery is relatively small. This minimizes the delay due to database accesses. The distributed database approach has the advantage that database accesses are localized. An update or query to a far away database is executed only when necessary. However, the number of database accesses required for location registration and call delivery is significantly increased from that of the centralized approach. Careful design is needed to ensure that database accesses will not significantly increase the signaling delay. Based on these observations, it is likely that the ideal architecture should lie between the centralized and the fully distributed approach. In fact, in order to attain better cost effectiveness, most of the on-going research efforts either try to i) increase the distribution of location information under a centralized database architecture or ii) limit the distribution of location information in a distributed database architecture. Besides, mobility and call arrival patterns vary among users, it is highly desirable that the location registration and call delivery procedures can be adjusted dynamically on a per-user basis. Dynamic schemes usually require the on-line collection and processing of data. This may consume significant computing power and careful design is necessary so that the computation can be effectively supported by the network.

This chapter has described the main features, operation and implementation of the basic anchoring technique and an optimisation procedure for application in GSM networks. Distribution of the HLR related signalling load is achieved through replacing of the expensive location registration messages between the serving VLR and the HLR by message exchanges between two MSC/VLRs. In response to subscriber call and mobility patterns, the optimised anchoring scheme removes the need to always report location changes to the AVLR whilst balancing the costs of both location update and call routing. As the benchmark for the comparison of the results, GSM standard signalling transactions and database traffic under the same traffic model and user mobility assumptions, are used. An analytical model for the proposed optimisation procedure is introduced. It is shown that within the call-to-mobility (CMR) range considered i.e. 0.01 to 100, the location update signalling cost of the optimised scheme can be reduced to less than half that of GSM. It is also shown that HLR related signalling costs can be reduced by as much as 25-65% subject to variations in network signalling load, compared to GSM. Also under typical operating conditions, reductions in the total network signalling cost of about 60-25% at low to high CMR respectively, can be achieved. Where as the cost of call routing in the original anchor scheme is about 2 - 3.5 times that of GSM (at low CMR values), the optimisation procedure manages to reduce this cost to within 1.6 times that of GSM under a variety of network loading conditions.
In the proposed scheme no modifications to call set-up messages are required, and the cost of location updating as well as the number of HLR updates and interrogations have been significantly reduced. Although not evaluated, the cost of HLR updates and interrogations can be further reduced if registration / de-registration notifications were not relayed to the HLR. This further reduction however would be at the expense of nominal increase in the VLR signalling. Call routing cost is shown to approach that of GSM when the rate of call arrivals is high. The results show that a distributed mobility management and appropriate exchange of messages between databases, could result in better distribution and balancing of the signalling traffic in GSM. Simulation results demonstrated that the optimisation procedure can result in lower total network signalling and location updating while connection set-up delay is also reduced.
Chapter 4

PERFORMANCE EVALUATION OF LOCATION UPDATE AND PAGING TECHNIQUES

4 Introduction

Location management schemes as discussed in chapter 3, are essentially based on users' mobility and incoming call rate characteristics. Whereas location registration and call delivery procedures primarily concern the fixed network (network signalling traffic, database query/update), the location updating and paging mechanisms referred to in this chapter, concern signalling traffic management over the air-interface. The location update and paging procedures are responsible for generating a significant proportion of the signalling traffic in cellular networks. The location update procedure allows the VLR/MSC to keep the knowledge of user's location, more or less accurately, in order to locate the subscriber, in case of an incoming call for example. The paging process achieved by the system consists of sending paging messages in all cells where the mobile terminal could be located. Therefore, if the locating cost is high (and thus the user location knowledge is accurate), the paging cost will be low (paging messages will only be transmitted over a small area). Conversely, if the location cost is low (and thus the user location knowledge is fuzzy), the paging cost will be high (paging messages will have to be transmitted over a wide area).

In first-generation cellular mobile systems, traffic was highly unbalanced. Less than one third of calls were incoming calls; the remaining were outgoing. Therefore, the signalling traffic due to paging process was not significant, and the location updating procedure also had little impact due to the large cells. However in the current generation of mobile systems there is a near balance between incoming and outgoing call rates. The paging process is therefore more important, due to smaller cell sizes the location update traffic has become more significant, and thus the location management has become more important. For instance, recent statistics from GSM operators show that, in the Paris dense urban environment, the location updating rate can be 10 times that of the call rate at peak traffic hours [X95]. Presently, the location method most widely implemented in first- and second-generation cellular systems (GSM, IS95, etc.) makes
Chapter 4. Performance Evaluation of Location Management Techniques

use of fixed-size location areas (LAs). In these wide-area radio networks, location management is done automatically. Location areas allow the system to track the mobiles during their roaming in the network(s); subscriber location is known if the system knows the LA in which the subscriber is located. When the system must establish a communication with the mobile (to route an incoming call, typically), the paging only occurs in the current user LA. Thus, resource consumption is limited to this LA; paging messages are only transmitted in the cells of this particular LA. Implementing LA-based methods requires the use of databases. Two basic location updating methods implemented based on LA structuring are:

**Periodic Location Updating** - This method just requires the mobile to periodically (period set by the operator) transmit its identity to the network. Its drawback is its resource consumption, which is user-independent and can be unnecessary if the user does not move from an LA for extended periods of time. Generally, this method is combined with the next one.

**Location Updating on LA Crossing** - This method first requires each BS to periodically broadcast the identity of the LA it belongs to. The mobile is required to permanently listen to network broadcast information (on the broadcast channel) and to store the current LA identity. If the received LA identifier differs from the stored one, a location update (LU) procedure is automatically triggered by the mobile. The advantage of this method is that it only requires LUs when the mobile actually moves. A highly mobile user will generate a lot of LUs; a low mobility user will only trigger a few.

A hybrid method which combines the two previous ones can also be implemented. The mobile generates its LUs each time it detects an LA crossing. Nevertheless, if no communication (related to an LU or a call) has occurred between the mobile (in idle mode, i.e., powered on but not communicating) and the network for a fixed period (e.g., three hours), the mobile generates an LU (i.e., a periodic LU). This periodic LU typically allows the system to recover user location data in case of a database failure.

**Optimal Design of LAs**

In order to minimize total location management cost (i.e., location update + paging traffic and processing), the LA must be designed carefully. Two types of methods can be used to design optimum LAs: analytic and heuristic. Analytic approaches are based on assumptions of homogeneous cell shape, LA structure, and user movements. One interesting problem is to determine a subscriber mobility model that can capture as much as possible the real subscribers movements. Common approaches for analytical modelling of movements include:
The Markovian model, describes individual movement behaviour and captures direction of movement by assigning different probabilities to different neighbor cells. This is a discrete stochastic process, where the position of particle at time $t = nT$, i.e. $x(nT)$ are independent random variables (RVs) taking values $\pm s$, with equal probability. The other main sub-classification i.e. the Generalized Random Walk, is a random walk model where RVs take values $\pm s$, with different probabilities. Random walk models can be of two types: 1-D Random Walk when movement in only 1 direction e.g. up-down or left-right, is considered, and 2-D Random Walk where movement in 2 directions e.g. X and Y directions, is assumed. The cell residence time follows a geometric distribution, and there is no concept of movement history for a user.

The Brownian Motion model, used to describe movement of particles in a liquid, subject to collision and other forces. Macroscopically, the position $x(t)$ can be modelled as a stochastic process satisfying a 2nd order differential equation. Two main sub-classifications are: Bound Motion if restoring force is not equal to 0, and Free Motion if restoring force is 0). Brownian Motion can also be modelled as 1- or 2-dimensional. The model is used to describe probability distribution of user location. It uses the concept of motion drift (defined as mean velocity in a given direction). 1D Brownian motion has been used to model directed traffic e.g. along highways, and it shows that the variance of distribution, as a measure of location uncertainty, does not depend on mean drift velocity.

The Random Walk model, is used to describe individual user movement behaviour. Upon leaving a cell, equal probabilities are assigned for movement into different neighbor cells, e.g. 1D linear model $Pr = 1/2$ and for 2D hexagonal model $Pr = 1/6$. Although the model simplifies the analysis, one of its limitations is that the direction of movement of user is not taken into account.

The Fluid Flow model considers traffic flow as the flow of a fluid, modelling macroscopic movement behaviour. The model assumes average constant velocity, direction of movement is uniformly distributed. Users are also uniformly distributed. The model is only useful to calculate boundary crossing rates.

The fixed size LA location management methods are the most widely used in current cellular (e.g. GSM, IS-41, etc.) systems. Nevertheless, the traffic and processing generated may lead to congestion problems in high-density systems. One of the main concerns of system designers is therefore to define methods allowing the system to reduce the overhead traffic.
Chapter 4. Performance Evaluation of Location Management Techniques

The objective of the work presented in this chapter is to present evaluation results of the proposed adaptive multiplayer location management scheme. We also examine dynamic and distance-based location management methods proposed within the last few years which attempt to reduce the signalling traffic over the air-interface. The fixed size LA scheme of GSM is used as the baseline for comparisons.

The outline of this chapter is as follows. Section 4.1 presents an overview of recent research on location update and paging techniques. This is followed by Performance Comparison of Location Area Optimization Techniques in section 4.2. In section 4.3 the simulation parameters are outlined. Section 4.4 presents numerical results, followed by a summary in section 4.5. Finally, the conclusions are presented in section 4.6.

4.1 Literature Survey

Recent Research On Location Update and Paging

As discussed above, the current cellular networks partition their coverage areas into a number of location areas (LAs). Each LA consists of a group of cells and each MS performs a location update when it enters an LA. When an incoming call arrives, the network locates the MS by simultaneously paging all cells within the LA. There are a number of inefficiencies associated with this location update and paging scheme:

- Excessive location updates may be performed by MTs that are located around LA boundaries and are making frequent movements back and forth between two LAs.
- Requiring the network to poll all cells within the LA each time a call arrives may result in excessive volume of wireless broadcast traffic.
- The mobility and call arrival patterns of MTs vary and it is generally difficult to select an LA size that is optimal for all users. An ideal location update and paging mechanism should be able to adjust on a per-user basis.
- In addition, the fixed-size LA location update and paging scheme is a static scheme as it cannot be adjusted based on the parameters of an MT from time to time.

Recent research efforts attempt to reduce the effects of these inefficiencies. Most efforts focus primarily on dynamic location update mechanisms which perform location update based on the mobility of the mobile terminals and the frequency of incoming calls.
4.1.1 Location Update and Paging Schemes

Dynamic Location Update Schemes - The first Dynamic LA Management scheme due to [Xie93], proposes a method for calculating the optimal LA size given the respective costs for location update and paging. The authors consider a mesh cell configuration with square shaped cells. Each LA consists of $k \times k$ cells arranged in a square and the value of $k$ is selected on a per-user basis according to the mobility and call arrival patterns and the cost parameters. Each mobile terminal is assigned a different size LA, reflecting different mobility and call arrival patterns. The mechanism introduced in [Xie93], performs better than the static scheme in which LA size is fixed. However, it is not generally easy to use different LA sizes for different MTs as the MTs must be able to identify the boundaries of LAs which are continuously changing. The implementation of this scheme is complicated when cells are hexagonal shaped, or in the worst case, when irregular cells are used. In [Xie93] a mobility cost function is minimized so that $k$ is permanently adjusted. Each user is therefore related to a unique LA for which size $k$ is adjusted according to his/her particular mobility and incoming call rate characteristics.

In [Bar95] the authors examine the performance of Three Dynamic Update Schemes:

- **Time-based**: The mobile terminal (MT) performs location updates periodically at a constant time interval $VT$. Figure 4-1(a) shows the path of an MT. If a location update occurred at location $A$ at time 0, subsequent location updates will occur at locations $B$, $C$ and $D$ if the MT moves to these locations at times $VT$, $2VT$ and $3VT$, respectively.

  ![Figure 4-1: (a) Time-Based Location Update Scheme.](image)

- **Movement-based**: An MT performs a location update whenever it completes a predefined number of movements across cell boundaries (this number is referred to as the movement threshold). Assuming a movement threshold of 3 is used, the MT performs location updates at locations $B$ and $C$ as shown in Figure 4-1 (b).

  ![Figure 4-1: (b) Movement-Based Update Scheme.](image)
• **Distance-based**: An MT performs a location update when its distance from the cell where it performed the last location update exceeds a predefined value i.e. the distance threshold. A location update is performed at location B where the distance of the MS from location A exceeds the threshold distance (the distance from location A to the thick solid line as shown in Figure 4-2 is equal to the threshold distance).

![Figure 4-2: Distance-Based Update Scheme.](image)

The authors evaluated the performance of the above schemes based on one-dimensional random walk movement models. Results demonstrated that, the distance-based scheme produces the best performance but its implementation incurs the highest overhead, and time-based method is the worst. For the time-based and the movement-based schemes, the MT has to keep track of the time elapsed and the number movements performed, respectively, since the last location update. This can be achieved simply by implementing a timer or a movement counter at the MT. The distance-based scheme, however, assumes that the MTs have knowledge of the distance relationship among all cells. The network must be able to provide this information to each MT in an efficient manner.

**Distance-based location Update** - Another *distance-based location update scheme* is considered in [Mad94]. The authors introduce an iterative algorithm that can generate the optimal threshold distance that results in the minimum cost. When an incoming call arrives, cells are paged in a shortest-distance-first order such that cells closest to the cell where the last location update occurred are polled first. The delay in locating an MT is, therefore, proportional to the distance travelled since the last location update. Results demonstrated that, depending on the mobility and call arrival parameters, the optimal movement threshold varies widely. This demonstrates that location update schemes should be per-users based and should be dynamically adjusted according to the current mobility and call arrival pattern of the user. However, the number of iterations required for this algorithm to converge varies depending on the mobility...
and call arrival parameters considered. Determining the optimal threshold distance may require significant computation at the MT.

**Time-based location Update** - A dynamic time-based location update scheme is introduced in [Aky95]. The location update time interval is determined after each movement based on the probability distribution of the call inter-arrival time. This scheme does not make any specific assumptions on the mobility pattern of the mobile subscribers and the shortest-distance-first paging scheme as described in [Mad94] is used. It is demonstrated that the results obtained are close to the optimal results given in [Ho95]. Computation required by this scheme is low and similar to the scheme described in [Mad94], the drawback of this scheme is that paging delay is not constrained. The time required to locate an MT is directly proportional to the distance travelled since the last location update. Another timer based location update scheme with single-step dynamic paging is introduced in [Lei97]. Under this scheme, each user-network interaction resets a timer and modifies location record. The user registers when the timer expires. Users are paged over personally constructed paging areas whose size is an increasing function of the time elapsed since the last contact. The authors investigate the influence of user mobility and call statistics on the optimum timer value and on the optimum size of paging areas. The results indicate a cost reduction in the signalling load of up to 50% compared with fixed periodic time-based scheme. The results are however sensitive to the level of average location uncertainty and it is assumed that distributions of user location are available or can be estimated.

**Movement-based location Update** - Another movement-based location update (with selective paging) scheme is reported in [Aky96]. Each MT keeps a counter of the number of cells visited, and a location update is performed when the number of moves since the last location registration equals to a pre-defined threshold. The scheme also allows for the dynamic selection of movement threshold on a per-user basis. Similar to [Ho95], paging delay is confined to a maximum value. Movement-based location update schemes have the advantage that implementation is simple. The MTs do not have to know the cell configuration of the network. Although the selective paging reduces the cost of locating a mobile terminal, this is achieved at the expense of increased paging delay.

**Paging Under Delay Constraints** - Paging subject to delay constraints is considered in [Yat95]. The authors assume that the network coverage area is divided into location areas and the probability that an MT is residing in a location area is given. It is demonstrated that when delay is unconstrained, the polling cost is minimized by sequentially searching the location
areas in decreasing order of probability of containing the MT. For constrained delay, the authors obtain the optimal polling sequence that results in the minimum polling cost. The authors, however, assume that the probability distribution of user location is provided. This probability distribution may be user dependent. A location update and paging scheme that facilitates derivation of this probability distribution is needed in order to apply this paging scheme. Besides, the trade off between the costs of location update and paging is not considered in [Yat95].

**Update and Paging Under Delay Constraints** - Location update and paging subject to delay constraints is considered in [Ho95]. Similar to [Mad94], the authors consider the distance based-location update scheme. However, paging delay is constrained such that the time required to locate an MT is smaller than or equal to a predefined maximum value. When an incoming call arrives, the residing area of the MT is partitioned into a number of subareas. These subareas are then polled sequentially to locate the MT. By limiting the number of polling areas to a given value such as N, the time required to locate a mobile is smaller than or equal to the time required for N polling operations. Given the mobility and call arrival parameters, the threshold distance and the maximum delay, an analytical model is introduced that generates the expected cost of the proposed scheme. A random walk mobility model based on a geometrically distributed cell residence time is introduced. An iterative algorithm is then used to locate the optimal threshold distance that results in the lowest cost. It is demonstrated that the cost is the lowest when the maximum delay is unconstrained. However, by slightly increasing the maximum delay from its minimum value of 1, the cost is significantly lowered. The implementation however requires availability of information on distance relationships of all cells within the network, in the mobile terminal.

In [Bar93] the idea of reporting cells is introduced. Under this scheme, a subset of cells is selected as reporting cells (reporting centres). A mobile terminal reports its location only when entering one of these reporting cells. When an incoming call arrives, the search starts from the cell where the mobile last reported its location. As described in [Bar93], the selection of the set of reporting cells such that the total cost (update + paging) is minimized is an NP-complete problem. The scheme is static in the sense that the reporting cells are fixed and can not be easily adjusted dynamically in response to subscriber call and/or mobility characteristics. Excessive location updates are addressed by [Ros95] and [Bar96]. A timer-based strategy that uses a universal timeout parameter is presented in [Ros95], while a tracking strategy for mobile users in PCS networks based on cell topology is explored and compared with the time-based strategy.
in [Bar96]. For excessive polling, a one-way paging network architecture and the interfaces among paging network elements are examined in [Lin97]. Additional schemes attempt to reduce the cost of finding a user when the MT moves during the paging process [Ros99], [Yen98].

**Combining Location Areas and Paging Areas** - LA size optimization can be achieved by minimization of total cost of locating and paging. Based on this observation, several proposals have defined location management procedures that make use of LAs and paging areas (PAs) of different sizes [Pla94]. One method often considered consists of splitting an LA into several PAs. A MT registers only once i.e. when it enters the LA. It does not register when moving between the different PAs of the same LA. For an incoming call, paging messages will be broadcast in the PAs according to a sequence determined by different strategies. For example, the first PA of the sequence can be the one where the MS was last detected by the network. The drawback of this method is the possible delay increase due to large LAs. Another method based on the same partitioning principle is investigated in [Goo96], in which each cell of an LA can be considered one individual PA. The interesting result obtained shows that partitioning the LA into several PAs paged sequentially allows first a reduction of the expected number of messages, and second a significant decrease of the paging delay at high loads.

**Multilayer LAs** - In present location management methods, LU traffic is mainly concentrated in the cells of the LA border. Based on this observation and to overcome this problem, the multilayer location updating method is introduced in [Oka91]. In this method, each MT is assigned to a given group, and each group is assigned one or several layers of LAs. According to Figure 4-3, it is clear that group 1 and group 2 MTs will not generate LUs in the same cells, thus allowing LU traffic load to be distributed over all the cells.

![Multilayer LAs Diagram](image)
The method proposes deployment of multiple LA layers, such that the layers are staggered and overlay each other. Two classifications are introduced:

- **Multi-Grouping (MULTI_G)** – The MSs are divided into several groups and each group is assigned to one or more layers.
- **Multi-Switching layer (MULTI_S)** – MSs of each group have several LA layers. When an MS updates, it switches layers (updates to a different layer).

With MULTI_G, the cells which the MSs access for location update, are different from group to group, thus location update signalling traffic is distributed over all cells and not concentrated in the location area border cells. MULTI_S in addition provides location update activity hysteresis so that short-term switching instability (due to users moving frequently back and forth) can be avoided. The technique is aimed at reducing the location update traffic alone. No results regarding the performance and impacts on the total system signalling cost are provided in either [Oka91] or [Bra96]. In [Bra96] performance of the multilayer location update method, initially proposed in [Oka91] is analysed under Markovian user mobility assumption. In [Bra96], partial overlapping of adjacent location areas is allowed, introducing a hysteresis effect, in order to overcome the potential problem of back-and-forth movement of users near LA boundary as indicated in [Oka91]. MT initiates a location update procedure only when the location area ID it has memorised, is no longer in the set of IDs broadcast in the cell it has just entered. Different degrees of overlap and different location area sizes are considered. The improvements (reduction in LUs) reported strongly depend on the number of cells per LA, with the highest improvement reported when location areas overlap by half, which incidentally increases the unnecessary paging traffic in the overlap area. The improvement factor also depends on the degree of randomness of the movement of subscribers. Similar to [Oka91], the proposal in [Mar95] is also based on Multilayer LAs to reduce location update cost.

**Intelligent Paging** - The intelligent paging strategies in [Lyb95] are aimed at reducing the signalling overhead over the radio interface, due to paging procedure. The main idea is the application of a multi-step paging strategy. At the instance of call arrival, the user is paged initially, in the paging area indicated by the "paging related information". On no page response, the mobile is paged sequentially in the remaining portion of the location area. *paging related information* elements comprise of recent interaction information, mobility flag etc. Based on assumptions derived from classification of subscribers into ordinary, low and high mobility classes and with similar categorisation of call arrival characteristics, the simulation results indicate up to 70% savings in the paging signalling cost can be achieved, thus allowing for
Chapter 4. Performance Evaluation of Location Management Techniques

definition of larger location areas (leading to reduction of location update signalling costs). The significant portion of cost savings are due to low mobility users and gathering the so called paging related information requires additional processing functionality and storage space at the network databases. Also evident, is the fact that multi-step paging strategies can result in unpredictable call-connection times.

4.1.2 Memory-Based Methods

The design of memory-based location management methods has been motivated by the fact that systems do a lot of repetitive actions, which can be avoided if predicted. This is particularly the case for LUs. Indeed, present cellular systems achieve every day, at the same peak hours, almost the same LU processing. Short-term and long-term memory processes can help the system avoid these repetitive actions. Some methods have thus been proposed that are based on user and system behaviour observation and statistics.

Assignment/Adjustment – In current systems, the size of LAs is optimized according to mean parameter values, which in practical situations can vary over a wide range during the day and from one user to another. Based on this observation, it is proposed to manage user location by defining multilevel LAs in a hierarchical cellular structure [Hu95], as depicted in Figure 4-4. At each level the LA size is different, and a cell belongs to different LAs of different sizes. According to past and present MS mobility behavior, the scheme dynamically changes the hierarchical level of the LA to which the MS registers. LU savings can thus be obtained, however, no analytical/simulation results have been reported.

A variant of this strategy evaluated in [Pas96] consists in requiring from mobiles to register in the cells where they are camped on. Registrations involve a periodic timer which value has to be optimized. Thus, rather than paging a mobile in all the cells of a LA, the mobile will be paged only in the cells visited during the last period: these are the cells where the mobile camped on during its traversal of the LA. High (incoming) call rate and low-mobility users are directed to small LAs, medium-mobility users are directed to medium-sized LAs, and high-velocity and low incoming call rate users are directed to large LAs.
Dynamic/Adaptive Individual LAs - An opposite approach considers that instead of defining LA sizes a priori, these can be adjusted dynamically for every user according to incoming call rate and LU rate.

In [Lei98] the LA size (in terms of number of cells) is chosen/calculated on a per-user basis. Upon each location registration at the boundary crossing of the personal LA, a new LA (new set of cells) is assigned to the user, according to mobility and call characteristics in its previous LA (Figure 4-5 and Figure 4-6). The goal is to minimize the combined (LU and paging) average signalling cost function for each individual mobile subscriber, such that the overall system-wide signalling cost for location management is minimized. The one-dimensional Brownian motion with drift process [Kar75] is used for mobility modelling. The cost function is defined as the sum of paging and location registration costs, similar to [Xie93][Ros96]. The derivative of the cost function is taken and set equal to zero, in order to obtain optimum LA length and position. Effects of motion drift velocity and diffusion constant (representing motion uncertainty region around initial position) have been examined. The main conclusions are:

- increasing velocity results in larger optimum LA sizes,
- minimum cost increase with speed and
- LA shape, besides LA size plays important role in signalling reduction,
- The scheme can offer a cost reduction of up to 30% compared to dynamic LA techniques of [Xie93] and [Goo93].
- The upper bound for cost reduction is given at 50%.

The results presented in [Lei98] are however based on a simplified one-dimensional random walk model and thus present an over simplification of user movement model. As authors admit,
it is extremely difficult to solve the same formulation with a two-dimension Brownian movement model, as analytical solutions are hard to find. For the two-dimensional mobility tracking problem, the authors recommend the probability-based approach of [Lei97]. Adapting the LA size to each user’s parameters values may be difficult to manage in practical situations. This has led, in [Rac92], to the definition of a method where the LAs sizes are dynamically adjusted for the whole population, not per user as in the two previous methods. Statistical information about users and mobility in the network is collected in databases and computed. Network characteristics in function of time, place, density, and so on are thus evaluated. Results of this computation allow the network to dynamically (daily, weekly, monthly, yearly, etc.) adjust the sizes of location areas. For instance, during the day, when call rates are high, it is preferable to deal with small LAs. Conversely, at night the call rate is much lower, larger LAs are preferred.

In [Kim95] a dynamic decision scheme is introduced which takes into account regional information as well as mobility and call characteristics of users. The determination of the dynamic LA is formulated as integer programming problem, and LA size and shape can be adjusted dynamically for every user according to incoming call rate and mobility characteristic.

Another adaptive scheme is proposed in [Yeu95]. Based on extensions of formulations in [Xie93], the authors compare the performance of fixed LA and distance-based schemes, where it is shown that contrary to expectations, the fixed scheme outperforms the distance-based strategy in many cases. This being due to the large areas of overlap associated with the latter strategy. A composite adaptive location tracking strategy based on the multilayer approach of [Xie93] is then introduced. It is shown that the proposed new scheme has the ability to
dynamically change/adjust the location area size, and it results in a lower total signalling cost, compared to both fixed and distance-based techniques. The proposed scheme is not however capable of adjusting to different shape LAs (although LA size is adaptive, LA shape remains constant).

**Subscriber Profile-based Strategies** - Observing that users show repetitive mobility patterns, the *alternative strategy* (AS) is defined in [Tab92][Tab95]: its main goal is to reduce the traffic related to mobility management - and thus reduce the LUs by taking advantage of users' highly predictable patterns. In AS, the system handles a *profile* recording the most probable mobility patterns of each user. The profile of the user can be provided and updated manually by the subscriber or determined automatically by monitoring the subscriber's movements over a period of time. When the user receives a call the system pages him sequentially over the LA until getting an acknowledgement from the mobile. When the subscriber moves away from the recorded zone the terminal processes a voluntary registration by pointing out its new LA to the network. The main savings achieved by this method are due to the non-triggered LUs when the user keeps moving inside his profile LAs. So, the more predictable the user's mobility, the lower the mobility management cost. A variant of this method, called the Two Location Algorithm (TLA), is proposed and studied in [Lin97]. In this strategy, a mobile stores the two most recently visited LA addresses. The same is done at the HLR level. Obviously, the main advantage of this method relies on the reduction of LUs when a mobile goes back and forth between two LAs. Another profile-based location strategy is introduced in [Pol97]. The aim is to reduce the signalling traffic on the radio link by increasing the intelligence within the fixed network. The system is required to maintain a sequential list of most likely places that a subscriber could be located. The list is ranked, and upon call arrival, MT is paged sequentially in each location within the list. When user moves between location areas within the list, no update is required. The list could be provided by the user or compiled by the system based on past calling history. The paper does not address methods for list generation and maintenance. The focus is on the potential performance improvements that can be gained from maintaining such a list. The proposed scheme is based on that of [Tab95] and assumes that the probability distribution of user location is known and the information required to update the list can be obtained from billing records. Based on key performance metrics (fixed network SS7 traffic, radio bandwidth and call setup delay) the performance of the proposed scheme is compared to that of classic schemes. In terms of radio link performance, three representative profile distributions - uniform, linear and exponential - are considered. Profile distribution \( \alpha_i \) refers to the probability of the user being in location area \( A_i \). Based on minimization of the total cost
formulation with respect to $K$, optimal $K$-size i.e. $K_{opt}$ is obtained (where $K_{opt}$ indicates optimum number of location areas). It is shown that:

- as in [Tab95], the optimum $K$ and cost saving depend on probability distribution of $a_i$.
- as the number of location areas in a list increases, the upper bound on the paging delay increase (with exponential dist.) can be limited to 50% though this delay increase much more in the case of uniform and linear distributions of $a_i$.
- Savings on the radio link traffic when $K$ is less than a threshold value,
- Savings on the fixed network signalling traffic when $K$ is greater than a threshold value and list maintenance cost is significantly lower than location update cost,
- Stationary users will contribute towards increased fixed network signalling cost (due to list maintenance cost), so the technique should be applied to mobile users only,
- Call setup delay for mobile-terminated calls increases with $K$,
- For practical purposes, $K$ values of 3-5 are recommended.

Predicting Short-Term Movements of Subscribers - The method proposed in [Rok92] uses a process which predicts the movements of the MT according to its direction of movement, velocity, and so on. Processing and prediction are made at both the MT and the HLR. When actual movements of the MT do not fit with those predicted, a registration is triggered by the mobile to inform the network of its actual location. Otherwise, no exchange is required, which allows savings in LU processing and signalling. In [Shi94], a mobility management method similar to AS is defined. It is called Statistical Paging Area Selection (SPAS) and is based on location statistics collected by each MT which periodically reports them to the network. These statistics consist of a list of the average duration the MT has been located in each LA. A priority rule is determined to settle the sequence of LAs visited by the mobile. If this sequence is different from the last one reported to the network, the MT transmits it; otherwise, nothing is done. The paging process is achieved in the same way as in AS. When the MT moves to an area that is not on the reported list, it has to process a temporary location registration to the network.

In [Liu95], the proposed method provides a means of allowing pre-connection and pre-assignment of data or services at the location before the user moves into it, so user can immediately receive service or data. This method clearly applies to location management, particularly handoffs. Just as are the previous two methods, it is based on users’ movement history patterns. Called Mobile Motion Prediction (MMP), it allows the system to predict the
future location of the user. Schematically, the MMP combines two movement models: Movement Circle (MC), based on a closed-circuit model of user movement behaviour, and Movement Track (MT), used to predict routine movements. MC is used to predict long-term regular movements. Based on the proposed MC/MT detection algorithm, the accuracy of prediction subject to randomness factor i.e. location uncertainty, is evaluated. Prediction accuracy ratios of up to 50% is reported assuming 70% randomness in user movement behaviour, and for users which have 30% or less randomness in their movement, the prediction accuracy is more than 70%.

The reverse virtual call (RVC) setup is a new scheme for delivering mobile terminal calls [Git97]. It allows, under the constraint that the LA is not smaller than the VLR area, a reduction in the number of signalling messages exchanged between the called and calling databases and switches. The call setup delay is shown to always be reduced by about 50% when using the RVC scheme.

**Meta-Heuristics for LA Partitioning** - The LA planning/partitioning requires optimal grouping of cells into LAs, such that the location update traffic is minimized without violating the paging bound. This can be mapped into a graph partitioning problem, which is more complex and appears as an NP-complete combinatorial problem [Gar79]. Only empirical methods can thus be used to approach the optimal solution due to exponential growth in execution times. In [Gam91] graph modelling is used to represent the LA planning problem and greedy approaches are used to obtain solutions. In [Mar93][Mar94] two heuristic algorithms are used to obtain better results than in [Gam91], by adjusting the border weights. In [Ple95] a weighted greedy algorithm consisting of two phases – the merge and exchange phases – is introduced. It is shown that the proposed greedy technique can outperform previous methods. In [Wan98] the methodology for application of GAs to LA planning problem is outlined, and performance comparison results, between the GA and Hill-Climbing (HC) methods are reported. It is shown that GAs consistently outperform HC however the improvement percentage gradually diminishes as the number of cells are increased. In [Car95], for example, the method proposed proceeds in an iterative way by reducing a mobility cost function and starts by a configuration of LAs using an analytical method. Another approach, proposed in [Gon96], makes use of genetic algorithms. Genetic algorithms are used to efficiently group the cells under a mobility cost function constraint. They use several interconnected processes, such as elitism and edge-based crossover. Other empirical methods can be used, such as simulated
annealing [Beh92] - which is currently used by some GSM operators. Single move heuristics, Hill-Climbing and steepest descent optimization are also used.

4.2 Performance Comparison of LA Optimization Techniques

In this section 4 different adaptive techniques are studied and compared to the fixed-LA strategy of GSM (Alg.1). These are:

- Alg.2 - Dynamic Location Management of [Xie93]
- Alg.3 - Dynamic Location Registration and Paging of [Kim96]
- Alg.4 - Adaptive Multilayer Scheme (proposed)
- Alg.5 - Distance-based Strategy of [Bar95][Yeu95] (same as Adaptive Location Management of [Hu95])

Furthermore, the impact of Recent Interaction Paging (RIP) - a two-step paging technique - on all the above is also investigated. The analytical results and improvements reported in the literature and all of the above references, are either based on simple/unrepresentative mobility assumptions e.g. 1D random walk models or high-way models (constant velocity with semi-directed trajectories), etc., or else different techniques are often compared against different baseline schemes. Simulation studies are therefore a better approach for conducting performance evaluations under a common set of assumptions, and range of mobility and call related parameters.

Dynamic Location Management [Xie93] - In [Xie93] location area covers a square area of \( K \times K \) cells, referred to as size-\( k \) location area. The total bandwidth cost for location update and paging per terminal per unit time is given as

\[
\Omega(K, \alpha, U_k) = K^2 \alpha \Omega_{pg} + U_k \Omega_{lu}
\]

(4-1)

where

- \( \alpha \): call arrival rate (calls/hr./terminal)
- \( U_k \): location update rate per terminal assuming size-\( k \) location areas (updates/hr./terminal)
- \( \Omega_{pg} \): paging cost for a call (bytes/page/cell)
- \( \Omega_{lu} \): location update cost for a call (bytes/update)
From [Tho88], the mean number of outgoing (=incoming) terminals from a cell per unit time is given by

\[ E[m] = \rho \cdot 4K\delta \cdot E[v] / \pi \]  

(4-2)

Thus,

\[ U_k = E[m] / \rho (K\delta)^2 = 4E[v]/K\delta \pi = U_1 / K \]  

(4-3)

Where

- m: number of outgoing terminals from LA per unit time
- \( \rho \): subscriber density per unit area
- v: subscriber speed (km/hr.)
- \( \delta \): length of cell side (square cells assumed.)

Normalising \( \Omega_{lu} \) to 1, the normalized cost function becomes

\[ \Omega[K, \alpha, U_1] = K^2 \alpha \eta + U_1 / K \]  

(4-4)

where \( \eta = \Omega_{ps} / \Omega_{lu} \).

To obtain the optimum size location area, a cost difference equation between the system with size-\( k \) and one with size \( k-1 \) (\( k \geq 2 \)) is defined, i.e.

\[ V[K, \alpha, U_1] = \Omega[K, \alpha, U_1] - \Omega[K-1, \alpha, U_1] = [K \cdot (2K-1) \cdot (K-1) \cdot \alpha \eta - U_1] / K(K-1) \]  

(4-5)

Given \( \alpha, \eta \) and \( U_1 \), the optimum \( K (\geq 1) \) by which the total cost is minimized is

\[ K_{opt}(\alpha, U_1) = \max \{ K: V[K, \alpha, U_1] \leq 0 \} \]  

(4-6)

\[ 1 \quad \text{if} \quad V[2, \alpha, U_1] > 0. \]

Under the classical scheme of GSM, the LAs are fixed and whose optimum size can be given by \( K_{opt}(\bar{a}, \bar{U}_1) \) where \( \bar{a} \) is the average call arrival rate and \( \bar{U}_1 \) is the average \( U_1 \), both over all system users. In this system the cost per user is given by

\[ \Omega[K_{opt}(\bar{a}, \bar{U}_1), \alpha, U_1] = K_{opt}^2(\bar{a}, \bar{U}_1) \cdot \alpha \eta + U_1 / K_{opt}(\bar{a}, \bar{U}_1) \]  

(4-7)

In the proposed algorithm, the optimum size of location area is obtained and adapted to each terminal according to \( \alpha \) and \( U_1 \) of terminal.
Dynamic Location Registration and Paging [Kim96] – Here a dynamic decision scheme is introduced which takes into account regional information as well as mobility and call characteristics of users. The determination of the dynamic LA is formulated as integer programming problem i.e.

\[
\text{Min. } a L \sum x_i + \beta \sum \lambda_i(v) \cdot (1 - x_i) \tag{4-8}
\]

Subject to

\[
\begin{align*}
\rho_i \cdot \tau_i / (1 + \rho_i \tau_i) &< 10^{-9} \\
\nu_i \cdot \tau_i / (1 + \nu_i \tau_i) &< 10^{-9} \\
x_i &= 0 \text{ or } 1 
\end{align*} \tag{4-9, 4-10}
\]

where the following notation is used:

- \(x_i = 1\) if cell \(i\) is in the LA of the subscriber,
- \(x_i = 0\) otherwise.
- \(a\) call arrival rate to subscriber
- \(\lambda_i(v)\) visiting rate of cell \(i\), when average subscriber velocity is \(v\).
- \(\alpha\) weight associated with the paging signal (bits/signal)
- \(\beta\) weight associated with the update signal (bits/signal)
- \(\rho_i\) mean number of paging signals to cell \(i\).
- \(\nu_i\) mean number of updating signals from cell \(i\).
- \(\tau_1\) mean service time of paging signal
- \(\tau_2\) mean service time of updating signal
- \(\theta_1\) forward signalling ch. Blocking probability requirement
- \(\theta_2\) reverse signalling ch. Blocking probability requirement
- TC set of all cells
- BC set of boundary cells not included in the LA
- \(N\) number of cells
- \(d\) driving distance
- \(v\) mean subscriber velocity

In the formulation of (4-8), the first term in the cost function represents the paging cost for each user. If cell \(i\) belongs to the location area of the subscriber, the paging signal occurs \(a\) times, for
a unit time period. The second term represents the LA updating cost. Since updates occur when
the subscriber moves out of its current LA, the cost at cell i can be computed by examining the
visiting rate to cell i. The visiting rate to cell i is computed by including regional information at
each cell using
\[ \lambda_i(v) = \nu \cdot k \cdot \mathbb{E} \text{[no. of visits to cell i with free flow]} / d. \]  
\[ (4-11) \]

where k is a constant.

The visiting rate is proportional to the velocity and the expected number of visits to cell i, when
no blocking is assumed in the flow. The rate is inversely proportional to the driving distance
from the current cell. Thus, in the cost function, the paging cost increases in proportion to the
number of cells in LA whilst the update cost is inversely proportional to LA size.

Constraints (4-9) and (4-10) represent the blocking probability of paging and location updating
signals respectively. It is assumed that arrival of signalling messages is poisson, and blocking
probabilities are defined by Erlang-B formula. The formulation of (4-8) can not be solved
without relaxation of one of the constraints. The boundary cells to which the second constraint
(4-10) are to be applied, can be identified before the problem is solved. Thus the problem is first
solved by relaxing the second constraint, and after obtaining the initial LA, the second
constraint can be applied (at the boundary cells). The objective/cost function of (4-8) is therefore
transformed as follows:

\[ \text{Min. } \alpha \sum_i \alpha \chi_i + \beta \sum_i \lambda_i(v) \cdot (1 - \chi_i) = \alpha \sum_i \alpha \chi_i + \beta \sum_i \lambda_i(v) - \beta \sum_i \lambda_i(v) \cdot \chi_i \]  
\[ (4-12) \]

Thus, the problem to solve initially can be formulated as

\[ \text{Min. } (\alpha a - \beta \lambda_i(v)) \cdot \chi_i \]  
\[ (4-13) \]

Subject to \[ \rho_i \cdot \tau_i/(1 + \rho_i \tau_i) < 10^{-6} \]  
\[ (4-14) \]

\[ \chi = 0 \text{ or } 1. \]

In (4-13) if \( \chi_i = 1 \) satisfies the constraint (4-14) and the coefficient is negative or equal to zero,
then cell I is included in the new LA of subscriber. Otherwise, if \( \chi_i = 1 \) does not satisfies the
constraint or the coefficient is positive, then cell i, is excluded from the new LA.
After solving the problem, it may be that the location area obtained is not a contiguous, and fully connected. In this case, the area to which the subscriber belongs i.e. the serving cell and those adjacent are kept, and other cells discarded. Having determined the initial LA of subscriber, the boundary cells not contained in the LA are now considered for possible inclusion. If any of cells on the boundary satisfy constraint 2 (4.11) then the cell is included in the LA. Otherwise if constraint 2 is not satisfied, then constraint 1 (4.10) is checked. If the cell satisfies this constraint, then the cell is included in the new LA. This process has the effect of avoiding location updates at cells where the reverse control channel is extremely busy. The decision scheme above, updates the LA of subscriber each time the mobile moves out of current LA. The size as well as shape of the new LA is dynamically adjusted by reflecting the call arrival rate, velocity and regional flow rates. The structure of the dynamic LA decision is depicted in Figure 4-7 below.

The service times of paging and updating are assumed to have exponential distributions with means of 0.03 and 0.3 seconds. The blocking probability requirement is assumed to be 0.001 i.e. \( \theta_1, \theta_2 = 3 \).

**Distance-based Strategy** [Yeu95] [Hu95] [Bar95] – In the distance-based strategy, mobile starts movement from the centre cell (cell A in Figure 4-8) of its current LA. After travelling through D cells (assuming no change of direction) the mobile performs a location update at cell B, and takes the current cell B as the centre cell of the new LA. Thus the number of cells traversed between two consecutive location updates is D. As shown in Figure 4-8, the

```plaintext
procedure Dynamic LA search
Begin dynamic decision scheme
For i = 1 to the last cell number
\[ \text{solve } (ca - \beta(x,v)).X_i \text{ s.t. constraint (4.15)} \]
End (for);
IF LA disconnected Then
IF cell i is in the area to which mobile belongs Then
\[ X_i = 1; \]
ELSE \[ X_i = 0; \]
End (if);
End (if);
While ( all boundary cells are considered ) do
IF ( 2nd constraint not satisfied ) Then
IF ( 1st constraint is satisfied by setting \( X_i = 1 \) ) Then
\[ X_i = 1; \]
ELSE \[ X_i = 0; \]
End (if);
End (if);
Until (termination-condition i.e. all boundary cells are checked )
End (dynamic decision scheme).
```

Figure 4-7: The structure of decision technique
consecutive values taken by D are 3, 2, 3, 3, 4, and 2. Note that there exists a large overlap area between location areas.

Figure 4-8: Trajectory of a mobile using the distance-based strategy

Simple counting shows that the number of cells in the $i$th tier of a cell is $6i$. Thus, for any given value of D, the corresponding location area size in number of cells $k$, is given by

$$K = 1 + \sum_{i=0}^{D-1} 6i = 1 + 3D(D-1). \quad \text{- Hexagonal cells.} \quad (4-15)$$

$$K = 1 + \sum_{i=0}^{D-1} 8i = 1 + 4D(D-1) \quad \text{- square cells.} \quad (4-16)$$

If the location area can be represented by a hexagonal region, then the radius of this hexagonal area is

$$R_h = (2D - 1)r \cdot \cos 30 = \frac{3(2D-1)}{2}r. \quad (4-17)$$

The radius of equivalent circular LA with the same area is given by

$$R = \sqrt{\frac{3}{\pi}} \cdot R_h = 0.788(2D - 1)r. \quad (4-18)$$

This is also the distance travelled by a mobile in a location area. Therefore the location update rate per mobile is given by

$$U_k = \frac{E[v]}{R} = \frac{E[v]}{0.788(2D - 1)r} \quad \text{- Hexagonal LAs} \quad (4-19)$$

$$U_k = U_1/K = 4E[v]/K \delta \pi \quad \text{- Square LAs (from 4.4)} \quad (4-20)$$

Based on approach similar to [Xie93], to obtain the optimum D, a cost difference equation between the system with size-D and one with size D-1 ($D \geq 2$) is defined, i.e.
\[ V[D, a, U_1] = \Omega[D, a, U_1] - \Omega[D-1, a, U_1] \]  

where

\[ \Omega[D, a, U_1] = (1 + 3D^2 - 3D) \cdot a \cdot \Omega_{pg} + E[v] / 0.788 \cdot (2D - 1) \cdot r \]  

\[ \Omega[D-1, a, U_1] = (1 + 3(D-1)^2 - 3(D-1)) \cdot a \cdot \Omega_{pg} + E[v] / 0.788 \cdot (2(D-1) - 1) \cdot r \]

Putting (4-22) and (4-23) into (4-21) yields

\[ V[D, a, U_1] = 6(D-1) \cdot a \cdot \Omega_{pg} - \left[ 2.538v/ (2D-1) \cdot (2D-3) \cdot r \right] \]

Given \( a \) and \( U_1 \), the optimum \( D \) (\( \geq 1 \)) by which the total cost is minimized is

\[ D_{opt}(a,U_1) = \max \{ D: V[D, a, U_1] \leq 0 \} \]

1 if \( V[2, a, U_1] > 0 \).

Adaptive Multilayer Scheme (proposed) - In this scheme, the service area structure is divided into \((n+1)\) hierarchical layer nodes, with the lowest layer representing the smallest independent LA, comprising a single cell. For instance, in the simulation experiments of section 4.3, the service area considered, consists of 100 cells, and total/max. of 8 hierarchical layers are considered, where location area size \( k \), can assume values \( k = 1, 2, 4, 5, 10, 20, 25, \) and \( 50 \) cells. Selection of appropriate LA layer is based on Call-to-Mobility Ratio (CMR) metric. The MT keeps counters of the number of call arrivals as well as LA updates (at any given layer) to calculate a running count of its CMR, and if and when the current CMR value changes from the old CMR (kept in memory), a location update to another layer is initiated. Thus, the MT may perform intra-layer updates whilst user moves across different LAs of a given layer, and may also perform inter-layer updates when the current CMR value changes from the old CMR by a Reference_CM value, scaled by variable threshold \( T_r \). The \( T_r \) which can assume values in the range \((0,1)\) is used to scale the old CMR value - to reduce call-related signalling cost \( T_r \) can be set to high values e.g. > 0.5, whilst if the location update costs are high, the \( T_r \) value should be set low e.g. < 0.1. This scaled, old CMR value is then used for comparison with the current CMR.

The operation of the techniques is described in the pseudo-code of Figure 4-9 below:
procedure Adaptive Multilayer
Begin adaptive decision scheme
For $i = 1$ to the last user
   Reference_CMR = CMR_Old[i] x T,
   IF (((CMR_New[i] > (CMR_Old[i] - Reference_CMR))
      OR
      (CMR_New[i] > (CMR_Old[i] + Reference_CMR)))
      AND
      MS[i].LA_LAYER > 0) Then
      MT[i].LA_LAYER -- ; i.e. move to layer n-1;
   ELSE
      IF (((CMR_New[i] < (CMR_Old[i] + Reference_CMR))
         OR
         (CMR_New[i] < (CMR_Old[i] - Reference_CMR)))
         AND
         MT[i].LA_LAYER < No_Layers-1) Then
         MT[i].LA_LAYER ++ ; i.e. move to layer n+1.
      ELSE
         remain at the same layer;
         perform location updates,
         update CMR[i],
      End (for);
End (adaptive decision scheme).

Figure 4-9: The structure (pseudo code) of the adaptive multilayer technique

Note that the reference used i.e. scaled old CMR, is in fact based on the CMR value in memory
and as such is reflective of the call and mobility characteristics of individual subscriber.

Paging and Location Update Procedures – Each base station broadcasts its beacon signal
containing its LA_id associated with layer 0, and the number of layers. As the mobile user
moves from cell to cell, the MT may detect change in the stored LA-id (at a given layer) or
receive incoming call(s). Upon either of these events, the MT must calculate the new CMR
value to replace the old value. Thus the MT has to store the most recent CMR, LA_id and the
Layer_id. The decision as to perform an inter-layer or an intra-layer update is based on the
procedure in Figure 4-9. Upon initiation of location update messaging, the MT sends a location
update request containing (among other parameters) its identity, the new LA_id and if a layer
change is decided (based on current CMR value and procedure of Figure 4-9), the Layer_id is
also sent to the network. The network databases are updated accordingly, and following the
authentication, ciphering and TMSI re-allocation, a location update acknowledgement is
returned to the MT. Upon receiving an incoming call, all the cells in the last reported LA of the
subscriber are paged. If RIP is enabled however, paging is performed through a 2-step process
described previously.
Chapter 4. Performance Evaluation of Location Management Techniques

4.3 Simulation Parameters and Assumptions

All simulations were carried out on 100 cell structure. The effect of boundary cells are excluded by duplicating the cell numbers. All simulations are performed with 10 subscriber per cell, with cell $r = 500$ m. It is assumed that call arrivals are poisson and that velocity has normal distribution. The results presented cover two categories:

- the Busy Hour (BH) period, at different call rates and subscriber velocities
- over a 24 hr. period.

Two basic types of movements are considered. The movements for both the vehicular and pedestrian subscriber categories in the Manhattan Grid (MG) type of environment are characterised by horizontal and vertical only movements where as in environment with Free Flow (FF), movement is characterised by the range trajectories that an MT can assume i.e. the new direction is drawn from a uniform distribution over $(0, 2\pi)$. Also, two classes of subscriber population are considered. First a class of All Vehicular (AV) with velocities in the range 5-50 km/hr. (in FF environment), or 5-30 km/hr. range (in MG environment). Secondly, a class of mixed types – pedestrians at 70%, vehicular 30% -, with pedestrians having an average velocity of 4 km/hr. and vehicular users with velocities in the range 5-50 km/hr. (in FF environment) or 5-30 km/hr. range (in MG environment). Table 4-1, shows the list of mobility-related parameters used in the simulations:

<table>
<thead>
<tr>
<th>Movement Type</th>
<th>Vehicular subscriber</th>
<th>Pedestrian subscriber</th>
<th>Vehicular subscriber</th>
<th>Pedestrian subscriber</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Velocity (km/hr.) - range:</td>
<td>5 - 30</td>
<td>4</td>
<td>5 - 50</td>
<td>4</td>
</tr>
<tr>
<td>standard deviation for speed (Normal dist.)</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Position update every (m)</td>
<td>10</td>
<td>5</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Prob. to change speed at position update</td>
<td>0.3</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>Prob. to change direction at position update</td>
<td>n/a</td>
<td>n/a</td>
<td>0.2</td>
<td>n/a</td>
</tr>
<tr>
<td>Prob. to turn at junctions</td>
<td>0.5</td>
<td>0.5</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>max. angle for direction update</td>
<td>n/a</td>
<td>n/a</td>
<td>2\pi</td>
<td>2\pi</td>
</tr>
<tr>
<td>cell rad.(m)</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>no. of users/cell</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Total no. of cells</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Subscriber Density (users/km$^2$)</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Block size (m x m) - no. Blocks 400</td>
<td>225x225</td>
<td>225x225</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Street width (m)</td>
<td>25</td>
<td>25</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Area (km$^2$)</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 4-1: Mobility-related simulation parameters

Furthermore, performance of selected algorithms with and without Recent Interaction Paging (RIP) is also considered. With RIP enabled, paging for the called subscriber is done initially in the last known subscriber location/cell, and following a time-out or no response from the paged cell, the whole LA (minus the paged cell) is then paged. This simple 2-step paging process can
be beneficial particularly when rate of updates and call arrival rates is high. In summary the performance results presented in the next section are categorised as follows:

- All subscribers Vehicular, movement type Free-Flow (AVFF)
- All subscribers Vehicular, movement type Free-Flow, RIP enabled (AVFF w/RIP)
- Pedestrians & Vehicular subscribers, movement type Free-Flow (PVFF)
- Pedestrians & Vehicular subscribers, movement type Free-Flow, RIP enabled (PVFF w/RIP)
- All subscribers Vehicular, movement type Manhattan Grid (AVMG)
- All subscribers Vehicular, movement type Manhattan Grid, RIP enabled (AVMG w/RIP)
- Pedestrians & Vehicular subscribers, movement type Manhattan Grid (PVMG)
- Pedestrians & Vehicular subscribers, movement type Manhattan Grid, RIP enabled (PVMG w/RIP)

4.4 Numerical Results

4.4.1 Results with FF Movement (All Vehicular) – AVFF
Chapter 4. Performance Evaluation of Location Management Techniques

Figure 4-10: BH - Signalling cost vs. velocity (km/hr.) (a) 2 calls/hr. (b) 6 calls/hr.  
% vehicular: 100, % pedestrian: 0 – Movement type: Free flow

Figure 4-11: BH - Signalling cost vs. call rate (a) 10 km/hr. (b) 30 km/hr.  
% vehicular: 100, % pedestrian: 0 – Movement type: Free flow

Figure 4-12: Signalling cost per user vs. time of day (a) call cost (b) total cost  
% vehicular: 100, % pedestrian: 0 – Movement type: Free flow
Figure 4-10 shows the total signalling cost (per subscriber) with different velocities, while Figure 4-11 presents those with different call arrival rates. Dynamic schemes of [Xie93] and [Kim96] outperform the fixed LA scheme (i.e. GSM) by about 5-20% after 20 km/hr at low call rates (Figure 4-10a) while a 5-30% improvement can be observed at velocities below 20 km/hr, at high call rates (Figure 4-10b). Thus as the incoming call rate increases, dynamic schemes tend to offer better performance at velocities below 20 km/hr. The adaptive multilayer scheme (Alg.4) however maintains higher cost savings after 10 and 20 km/hr. Comparison of the two dynamic schemes indicates a small advantage of the scheme of [Xie93] over [Kim96] (Figure 4-10 and Figure 4-11b) as indicated in Table 4-3. Figure 4-11a also clearly indicates the better performance of the dynamic schemes at higher call arrival rates but at low velocity, while the scheme of Alg.4, continues to outperform all other schemes (at high velocity) as the call rate increases (Figure 4-11b). A better indication of the performance of all 5 schemes is depicted in Figure 4-12b, where the total signalling cost i.e. location update + paging, during a day is shown. The variant velocity values and call arrival rates assumed at different times of the day are given in Table 4-2. The simulation results indicate that the paging cost between the hours of 11-14 and 5-7 in the afternoon are high (Figure 4-12a) the cost savings due to location updates, by the dynamic schemes results in a combined lower cost (26.5 and 19.9% average reduction in total cost) as shown in Table 4-3, whilst the scheme of Alg.4 manages to outperform all other schemes, by a significant 32% cost reduction compared with GSM. Also note the minimum total cost in the case of Alg.4 scheme at periods 9-10, 10-11 and 17-18. The minimum cost at these particular time periods correspond to minimum velocities, clearly showing the superior capability of the adaptive multilayer scheme (Alg.4) in adjusting to changes in the speed, by proper selection of LA size. The significant cost increase of about 330% in the case of the distance-based scheme (Alg.5) – despite comparable call related signalling cost (Figure 4-12a) – is mainly due to its inability to adjust properly to the changes in velocity, thereby incurring substantially higher location update costs (4-5 times higher than GSM).

Table 4-2: Total signalling cost, velocity range and call rates (AVFF)

Table 4-3: Performance comparison of various schemes , % improvement over GSM (AVFF)
4.4.2 Results with FF Movement (All Vehicular) - AVFF w/RIP

Figure 4-13: BH - Signalling cost vs. velocity (km/hr.) (a) 2 calls/hr. (b) 6 calls/hr.
% vehicular: 100, % pedestrian: 0 - Movement type: Free flow

Figure 4-14: BH - Signalling cost vs. call rate (a) 10 km/hr. (b) 30 km/hr.
% vehicular: 100, % pedestrian: 0 - Movement type: Free flow

Figure 4-15: Signalling cost per user vs. time of day (a) call cost (b) total cost
% vehicular: 100, % pedestrian: 0 - Movement type: Free flow
Chapter 4. Performance Evaluation of Location Management Techniques

Comparing Figure 4-10, Figure 4-11 and Figure 4-12 with those of Figure 4-13, Figure 4-14 and Figure 4-15 respectively, indicate a similar trend – higher cost savings due to RIP, but similar performance improvements over the GSM scheme - as indicated by results in Table 4-5. Table 4-6 above presents the results of comparisons and the level of improvements with RIP. For the case of all vehicular subscribers with Free flow type of movement, the average performance gain (reduction in total signalling cost) is between 5-9%. Despite significant call-related signalling cost reduction with RIP, observed in Figure 4-15a for the distance-based scheme (Alg.5), the total signalling cost remains much higher than GSM (and all other schemes) due to the high levels of the location update signalling load. Thus it may be concluded that when all subscribers are highly mobile, dynamic schemes perform better at velocities below 20 km/hr. (Figure 4-13b) or at call rates below 3 calls/hr. (Figure 4-14b), whilst much better performance gains can be obtained through the adaptive scheme of Alg.4 right across the whole range of velocities (Figure 4-13a, Figure 4-13b) and call rates (Figure 4-14b) considered.

Table 4-4: Total signalling cost, velocity range and call rates (AVFF w/RIP)

Table 4-5: Performance comparison of various schemes, % improvement over GSM (AVFF w/RIP)

Table 4-6: Performance comparison of ALL schemes, % improvement with RIP (AVFF)
4.4.3 Results with FF Movement (Pedestrians & Vehicular) – PVFF

Figure 4-16: BH - Signalling cost vs. velocity (km/hr.) (a) 2 calls/hr. (b) 6 calls/hr.
% vehicular: 30, % pedestrian: 70 – Movement type: Free flow

Figure 4-17: BH - Signalling cost vs. call rate (a) 10 km/hr. (b) 30 km/hr.
% vehicular: 30, % pedestrian: 70 – Movement type: Free flow

Figure 4-18: Signalling cost per user vs. time of day (a) call cost (b) total cost
% vehicular: 30, % pedestrian: 70 – Movement type: Free flow
Figure 4-16 shows the total signalling cost (per subscriber) with different velocities, while Figure 4-17 presents those with different call arrival rates. With a mix of subscriber types (vehicular: 30%, pedestrians: 70%) improvements of 10-50% by the dynamic schemes of [Xie93] and [Kim96] can be observed, as velocity and/or call rates increase. The adaptive multilayer scheme (Alg.4) begins to offer cost savings after 30 (at 2 calls/hr) and 40 km/hr (at 6 calls/hr), thus indicating that it is more suited to high-mobility environments.

In terms of performance improvement over the GSM scheme, the dynamic schemes and the adaptive multilayer schemes show 16, 11 and 18% improvement, as indicated in Table 4.7. Comparing the results in Table 4-8 and Table 4-3 shows a clear 10% difference (reduction) in the level of improvements. This is due to the fact that with a mix type of subscriber, whose majority are slow-moving pedestrians, the amount of location update traffic generated will be reduced, hence the total signalling cost per subscriber is reduced (compare Table 4-2 and Table 4-7). Given that there is now less location update traffic to minimize, the levels of overall reductions obtainable with mix type of subscribers, will also be reduced.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM</td>
<td>78.1</td>
<td>73.1</td>
<td>85.3</td>
<td>75.6</td>
<td>85.3</td>
<td>83.2</td>
<td>114.2</td>
<td>127.8</td>
<td>182.0</td>
<td>142.4</td>
<td>85.0</td>
<td>74.1</td>
<td>77.5</td>
<td>84.3</td>
<td>80.7</td>
<td>81.1</td>
<td>79.9</td>
<td>92.0</td>
</tr>
<tr>
<td>Alg.2</td>
<td>44.7</td>
<td>41.8</td>
<td>76.1</td>
<td>72.8</td>
<td>77.5</td>
<td>69.8</td>
<td>110.4</td>
<td>97.9</td>
<td>104.8</td>
<td>84.8</td>
<td>69.9</td>
<td>73.7</td>
<td>80.8</td>
<td>78.9</td>
<td>77.5</td>
<td>67.1</td>
<td>78.0</td>
<td>68.4</td>
</tr>
<tr>
<td>Alg.3</td>
<td>57.1</td>
<td>61.3</td>
<td>76.2</td>
<td>77.2</td>
<td>80.9</td>
<td>68.2</td>
<td>102.0</td>
<td>107.4</td>
<td>117.0</td>
<td>94.8</td>
<td>76.9</td>
<td>70.1</td>
<td>90.0</td>
<td>86.5</td>
<td>74.1</td>
<td>75.6</td>
<td>68.4</td>
<td>68.9</td>
</tr>
<tr>
<td>Alg.4</td>
<td>59.0</td>
<td>48.1</td>
<td>89.4</td>
<td>88.3</td>
<td>75.7</td>
<td>67.8</td>
<td>101.7</td>
<td>105.7</td>
<td>106.5</td>
<td>87.4</td>
<td>70.3</td>
<td>64.2</td>
<td>82.0</td>
<td>74.8</td>
<td>71.2</td>
<td>64.0</td>
<td>71.2</td>
<td>70.0</td>
</tr>
<tr>
<td>Alg.5</td>
<td>342.2</td>
<td>314.4</td>
<td>386.3</td>
<td>258.4</td>
<td>221.8</td>
<td>392.0</td>
<td>400.0</td>
<td>358.4</td>
<td>390.1</td>
<td>397.2</td>
<td>296.8</td>
<td>272.2</td>
<td>257.7</td>
<td>325.8</td>
<td>321.8</td>
<td>374.0</td>
<td>394.7</td>
<td></td>
</tr>
</tbody>
</table>

Table 4-7: Total signalling cost, velocity range and call rates (PVFF)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Alg.2</td>
<td>42.7%</td>
<td>42.0%</td>
<td>7.2%</td>
<td>3.7%</td>
<td>-2.4%</td>
<td>11.1%</td>
<td>5.3%</td>
<td>23.1%</td>
<td>26.4%</td>
<td>31.8%</td>
<td>17.5%</td>
<td>4.9%</td>
<td>-5.6%</td>
<td>-2.5%</td>
<td>4.4%</td>
<td>18.0%</td>
<td>21.9%</td>
<td>29.2%</td>
<td>15.9%</td>
</tr>
<tr>
<td>Alg.3</td>
<td>25.0%</td>
<td>29.7%</td>
<td>9.9%</td>
<td>-3.3%</td>
<td>-5.1%</td>
<td>29.8%</td>
<td>18.4%</td>
<td>10.8%</td>
<td>11.8%</td>
<td>18.5%</td>
<td>-2%</td>
<td>9.3%</td>
<td>-7.7%</td>
<td>-12.7%</td>
<td>-9.6%</td>
<td>-7.2%</td>
<td>10.7%</td>
<td>20.8%</td>
<td>10.8%</td>
</tr>
<tr>
<td>Alg.4</td>
<td>31.5%</td>
<td>31.0%</td>
<td>16.5%</td>
<td>9.8%</td>
<td>3.8%</td>
<td>18.6%</td>
<td>13.2%</td>
<td>16.6%</td>
<td>19.1%</td>
<td>21.5%</td>
<td>17.1%</td>
<td>4.8%</td>
<td>7.1%</td>
<td>12.2%</td>
<td>19.0%</td>
<td>20.9%</td>
<td>28.5%</td>
<td>18.0%</td>
<td></td>
</tr>
<tr>
<td>Alg.5</td>
<td>30.6%</td>
<td>30.3%</td>
<td>-31.7%</td>
<td>-1.4%</td>
<td>-5.2%</td>
<td>36.4%</td>
<td>-78.1%</td>
<td>-167.5%</td>
<td>-318.5%</td>
<td>-300.3%</td>
<td>-257.6%</td>
<td>-205.7%</td>
<td>-194.5%</td>
<td>-577.3%</td>
<td>-390.7%</td>
<td>-306.6%</td>
<td>-303.7%</td>
<td>-286.6%</td>
<td></td>
</tr>
</tbody>
</table>
4.4.4 Results with FF Movement (Pedestrians & Vehicular) – PVFF w/RIP

Figure 4-19: BH - Signalling cost vs. velocity (km/Hr.) (a) 2 calls/hr. (b) 6 calls/hr. 
% vehicular: 30, % pedestrian: 70 - Movement type: Free flow

Figure 4-20: BH - Signalling cost vs. call rate (a) 50 km/hr. (b) 10 km/hr. 
% vehicular: 30, % pedestrian: 70 - Movement type: Free flow

Figure 4-21: Signalling cost per user vs. time of day (a) call cost (b) total cost 
% vehicular: 30, % pedestrian: 70 - Movement type: Free flow
In Figure 4-19a performance improvements of the dynamic schemes begin after 30 km/hr, whilst Alg.4 indicates improvements of 10-40% at 2 calls/hr. The situation changes rather dramatically, as depicted in Figure 4-19b, where the dynamic schemes show an average of 20-25% improvement over the GSM scheme, at 6 calls/hr, whilst improvements due to Alg.4 begin to set in after about 40 km/hr. Thus it may be concluded that when the majority of subscribers are slow-moving pedestrians, dynamic schemes perform better as the call rate increases (i.e. with higher paging load, or lower location update load) whilst the reverse is true of the scheme of Alg.4, where better gains can be obtained when the call rates are relatively low (Figure 4-19a) and location update costs are high (Figure 4-20a vs. Figure 4-20b).

### Table 4-9: Total signalling cost, velocity range and call rates (PVFF w/RIP)

<table>
<thead>
<tr>
<th>Time</th>
<th>Speed (km/hr)</th>
<th>Call Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1</td>
<td>68.5</td>
<td>70.5</td>
</tr>
<tr>
<td>1-2</td>
<td>72.2</td>
<td>72.6</td>
</tr>
<tr>
<td>2-3</td>
<td>78.4</td>
<td>80.6</td>
</tr>
<tr>
<td>3-4</td>
<td>80.0</td>
<td>82.8</td>
</tr>
</tbody>
</table>

### Table 4-10: Performance comparison of various schemes, % improvement over GSM (PVFF w/RIP)

<table>
<thead>
<tr>
<th>Time</th>
<th>Speed (km/hr)</th>
<th>Call Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1</td>
<td>68.5</td>
<td>70.5</td>
</tr>
<tr>
<td>1-2</td>
<td>72.2</td>
<td>72.6</td>
</tr>
<tr>
<td>2-3</td>
<td>78.4</td>
<td>80.6</td>
</tr>
<tr>
<td>3-4</td>
<td>80.0</td>
<td>82.8</td>
</tr>
</tbody>
</table>

### Table 4-11: Performance comparison of ALL schemes, % improvement with RIP (PVFF)

In terms of improvement gains due to RIP, all schemes manage a modest -9% improvement, except for the dynamic scheme of Alg.3 where the improvement with RIP is almost 16%, as shown in Table 4-11.
4.4.5 Results with MG Movement (All Vehicular) – AVMG

Figure 4-22: BH - Signalling cost vs. velocity (km/hr.) (a) 2 calls/hr. (b) 6 calls/hr.
% vehicular: 100 , % pedestrian: 0 – Movement type: Manhattan Grid

Figure 4-23: BH - Signalling cost vs. call rate (a) 30 km/hr. (b) 10 km/hr.
% vehicular: 100 , % pedestrian: 0 – Movement type: Manhattan Grid

Figure 4-24: Signalling cost per user vs. time of day (a) call cost (b) total cost
% vehicular: 100 , % pedestrian: 0 – Movement type: Manhattan Grid
Figure 4-22 shows the total signalling cost (per subscriber) with different velocities, while Figure 4-23 presents those with different call arrival rates, within a Manhattan grid (MG) environment. Figure 4-22a and Figure 4-22b clearly show that the dynamic scheme of [Xie93] (Alg.2) only begins to show marginal gains over the GSM scheme after about 15-20 km/hr., the dynamic scheme of [Kim96] (Alg.3) manages an average of 25% gain across the range of velocities considered. Adaptive scheme of Alg.4 also produces improvements after velocities of 10-15 km/hr., with higher gains at lower call rates (Figure 4-22a). A better indication of the performance of all 5 schemes is depicted in Figure 4-24b, where the total signalling cost i.e. location update + paging, during a day is shown. Adaptive scheme of Alg.4 and dynamic scheme of Alg.3 clearly out perform the dynamic scheme of Alg.2 and GSM, also confirmed by the performance comparison results in Table 4-13.

Comparing results of Table 4-13 and Table 4-3, indicate lower level of gains in the Manhattan grid environment, which is due to the lower range of velocities considered in that case (Table 4-12 vs Table 4-2), however, the dynamic scheme of Alg.3 seems more suited to the MG environment than the dynamic scheme of Alg.2, as it ranks higher in the table of gains i.e. Table 4-13 and Table 4-3.

Table 4-12: Total signalling cost, velocity range and call rates (AVMG)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Call</td>
<td>0.5</td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
<td>2.5</td>
<td>3.0</td>
<td>3.5</td>
<td>4.0</td>
<td>4.5</td>
<td>5.0</td>
<td>5.5</td>
<td>6.0</td>
<td>6.5</td>
<td>7.0</td>
</tr>
<tr>
<td>9-10</td>
<td>24.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
</tr>
<tr>
<td>10-11</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
</tr>
<tr>
<td>11-12</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
</tr>
<tr>
<td>12-13</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
</tr>
<tr>
<td>13-14</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
</tr>
<tr>
<td>14-15</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
</tr>
<tr>
<td>15-16</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
</tr>
<tr>
<td>16-17</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
</tr>
<tr>
<td>17-18</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
</tr>
<tr>
<td>18-19</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
</tr>
<tr>
<td>20-21</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
</tr>
<tr>
<td>21-22</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
</tr>
<tr>
<td>22-23</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
</tr>
<tr>
<td>23-24</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
</tr>
</tbody>
</table>

Table 4-13: Performance comparison of various schemes, % improvement over GSM (AVMG)
4.4.6 Results with MG Movement (All Vehicular) – AVMG w/RIP

Figure 4-25: BH - Signalling cost vs. velocity (km/hr.) (a) 2 calls/hr. (b) 6 calls/hr.
% vehicular: 100, % pedestrian: 0 – Movement type: Manhattan Grid

Figure 4-26: BH - Signalling cost vs. call rate - (a) 30 km/hr. (b) 10 km/hr.
% vehicular: 100, % pedestrian: 0 – Movement type: Manhattan Grid

Figure 4-27: Signalling cost per user vs. time of day (a) call cost (b) total cost
% vehicular: 100, % pedestrian: 0 – Movement type: Manhattan Grid
Results depicted in Figure 4-25 to Figure 4-27 show a similar trend to those of Figure 4-22 to Figure 4-24 respectively. In Table 4-15 higher cost savings due to RIP can be observed, and the gains with RIP over GSM are slightly higher than those reported in Table 4-13. The results of performance comparisons with and without RIP, are depicted in Table 4-16. Thus in the case of all fast-moving subscribers within an MG environment all schemes (except distance-based scheme of Alg.5) show reasonable gains of about 5-13% when RIP is enabled.

<table>
<thead>
<tr>
<th>time</th>
<th>6-7</th>
<th>7-8</th>
<th>8-9</th>
<th>9-10</th>
<th>10-11</th>
<th>11-12</th>
<th>12-13</th>
<th>13-14</th>
<th>14-15</th>
<th>15-16</th>
<th>16-17</th>
<th>17-18</th>
<th>18-19</th>
<th>19-20</th>
<th>20-21</th>
<th>21-22</th>
<th>22-23</th>
<th>23-24</th>
</tr>
</thead>
<tbody>
<tr>
<td>speed</td>
<td>24.0</td>
<td>17.0</td>
<td>16.0</td>
<td>7.0</td>
<td>21.0</td>
<td>21.0</td>
<td>17.0</td>
<td>21.0</td>
<td>24.0</td>
<td>24.0</td>
<td>14.0</td>
<td>10.0</td>
<td>14.0</td>
<td>17.0</td>
<td>21.0</td>
<td>24.0</td>
<td>28.0</td>
<td>32.0</td>
</tr>
<tr>
<td>ute</td>
<td>0.5</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>2.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

Total signalling cost per subscriber

<table>
<thead>
<tr>
<th></th>
<th>GSM</th>
<th>Alg.2</th>
<th>Alg.3</th>
<th>Alg.4</th>
<th>Alg.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
<td></td>
</tr>
</tbody>
</table>

Table 4-14: Total signalling cost, velocity range and call rates (AVMG w/RIP)

Table 4-15: Performance comparison of various schemes, % improvement over GSM (AVMG w/RIP)

Results without RIP:

<table>
<thead>
<tr>
<th></th>
<th>GSM</th>
<th>Alg.2</th>
<th>Alg.3</th>
<th>Alg.4</th>
<th>Alg.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
<td>10-11</td>
</tr>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
<td>10-11</td>
</tr>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
<td>10-11</td>
</tr>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
<td>10-11</td>
</tr>
</tbody>
</table>

Percentage performance improvement over GSM

<table>
<thead>
<tr>
<th></th>
<th>Alg.2</th>
<th>Alg.3</th>
<th>Alg.4</th>
<th>Alg.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
</tr>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
</tr>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
</tr>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
</tr>
</tbody>
</table>

Results with RIP:

<table>
<thead>
<tr>
<th></th>
<th>GSM</th>
<th>Alg.2</th>
<th>Alg.3</th>
<th>Alg.4</th>
<th>Alg.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
<td>10-11</td>
</tr>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
<td>10-11</td>
</tr>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
<td>10-11</td>
</tr>
<tr>
<td>time</td>
<td>6-7</td>
<td>7-8</td>
<td>8-9</td>
<td>9-10</td>
<td>10-11</td>
</tr>
</tbody>
</table>

Table 4-16: Performance comparison of ALL schemes, % improvement with RIP (AVMG)
4.4.7 Results with MG Movement (Pedestrians & Vehicular) – PVMG

![Diagram](image.png)

Figure 4-28: BH - Signalling cost vs. velocity (km/hr.) (a) 2 calls/hr. (b) 6 calls/hr. 
% vehicular: 30, % pedestrian: 70 - Movement type: Manhattan Grid

![Diagram](image.png)

Figure 4-29: BH - Signalling cost vs. call rate (a) 30 km/hr. (b) 10 km/hr. 
% vehicular: 30, % pedestrian: 70 - Movement type: Manhattan Grid

![Diagram](image.png)

Figure 4-30: Signalling cost per user vs. time of day (a) call cost (b) total cost 
% vehicular: 30, % pedestrian: 70 - Movement type: Manhattan Grid
Figure 4-28 shows the total signalling cost (per subscriber) with different velocities, while Figure 4-29 presents those with different call arrival rates. With a mix of subscriber types (vehicular: 30%, pedestrians: 70%) improvements of 5-50% by the dynamic schemes of [Xie93] and [Kim96] can be observed, as velocity and/or call rates increase (Figure 4-28a and b), although dynamic scheme of Alg.3 shows marginal superiority over the dynamic scheme of Alg.2. Similar to the results in Figure 4-16 and Figure 4-17, the adaptive multilayer scheme (Alg.4) begins to offer cost savings after 30 km/hr (at 6 calls/hr), once again indicating that it is more suited to environments where the majority of subscribers are vehicular/high speed. In terms of performance improvement over the GSM scheme, the dynamic and the adaptive multilayer schemes show 9, 15 and 16 % improvement, as indicated in Table 4-18.

Table 4-17: Total signalling cost, velocity range and call rates (PVMG)

<table>
<thead>
<tr>
<th>Time</th>
<th>Vehicular</th>
<th>Pedestrian</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-7</td>
<td>145.6</td>
<td>53.8</td>
<td>94.4</td>
</tr>
<tr>
<td>7-8</td>
<td>140.0</td>
<td>59.0</td>
<td>97.5</td>
</tr>
<tr>
<td>8-9</td>
<td>130.0</td>
<td>63.3</td>
<td>95.2</td>
</tr>
<tr>
<td>9-10</td>
<td>120.0</td>
<td>66.0</td>
<td>96.5</td>
</tr>
<tr>
<td>10-11</td>
<td>110.0</td>
<td>69.0</td>
<td>98.0</td>
</tr>
<tr>
<td>11-12</td>
<td>100.0</td>
<td>72.0</td>
<td>99.5</td>
</tr>
<tr>
<td>12-13</td>
<td>90.0</td>
<td>75.0</td>
<td>101.0</td>
</tr>
<tr>
<td>13-14</td>
<td>80.0</td>
<td>78.0</td>
<td>102.0</td>
</tr>
<tr>
<td>14-15</td>
<td>70.0</td>
<td>81.0</td>
<td>103.0</td>
</tr>
<tr>
<td>15-16</td>
<td>60.0</td>
<td>84.0</td>
<td>104.0</td>
</tr>
<tr>
<td>16-17</td>
<td>50.0</td>
<td>87.0</td>
<td>105.0</td>
</tr>
<tr>
<td>17-18</td>
<td>40.0</td>
<td>90.0</td>
<td>106.0</td>
</tr>
<tr>
<td>18-19</td>
<td>30.0</td>
<td>93.0</td>
<td>107.0</td>
</tr>
<tr>
<td>19-20</td>
<td>20.0</td>
<td>96.0</td>
<td>108.0</td>
</tr>
<tr>
<td>20-21</td>
<td>10.0</td>
<td>99.0</td>
<td>109.0</td>
</tr>
</tbody>
</table>

Table 4-18: Performance comparison of various schemes, % improvement over GSM (PVMG)

4.4.8 Results with MG Movement (Pedestrians & Vehicular) – PVMG w/RIP

Figure 4-31: BH - Signalling cost vs. velocity (km/hr.) (a) 2 calls/hr. (b) 6 calls/hr. % vehicular: 30 , % pedestrian: 70 - Movement type: Manhattan Grid
Chapter 4. Performance Evaluation of Location Management Techniques

Figure 4-32: BEI - Signalling cost vs. call rate (a) 30 km/hr. (b) 10 km/hr.  
% vehicular: 30, % pedestrian: 70 - Movement type: Manhattan Grid

(a)  
(b)  
Figure 4-33: Signalling cost per user vs. time of day (a) call cost (b) total cost  
% vehicular: 30, % pedestrian: 70 - Movement type: Manhattan Grid

Figure 4-34: Signalling cost vs. call rate (a) 30 km/hr. (b) 10 km/hr.  
% vehicular: 30 - Movement type: Manhattan Grid

Table 4-19: Total signalling cost, velocity range and call rates (PVMG w/RIP)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>24.0</td>
<td>17.0</td>
<td>10.0</td>
<td>7.0</td>
<td>21.0</td>
<td>21.0</td>
<td>17.0</td>
<td>21.0</td>
<td>24.0</td>
<td>24.0</td>
<td>14.0</td>
<td>14.0</td>
<td>19.0</td>
<td>21.0</td>
<td>20.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
</tr>
<tr>
<td>Rate</td>
<td>2.0</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Call</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 4-20: Performance comparison of various schemes, % improvement over GSM (PVMG w/RIP)
Chapter 4. Performance Evaluation of Location Management Techniques

Table 4-21: Performance comparison of ALL schemes, % improvement with RIP (PVMG)

Results depicted in Figure 4-31 to Figure 4-33 show a similar trend to those of Figure 4-28 to Figure 4-30 respectively. The results of performance comparisons with and without RIP, are depicted in Table 4-22 thus in the case of a mix of fast and slow-moving subscribers within an MG environment all schemes show reasonable gains of about 15-24% when RIP is enabled.

4.4.9 Number of layers and threshold in the Adaptive Multi-Layer scheme

Figure 4-34 depicts variations in the call and total signalling cost with different number of layers. The higher the number of layers, the more pronounced is the reduction in the total signalling cost. However, call-related signalling cost tend to increase as the number of layers increase, indicating the ability of the scheme to effectively reduce the location update costs by selecting appropriate size LAs.
For example when the location update rates are high, larger LAs are selected, resulting in lower update costs at the expense of higher paging costs. Figure 4-35 shows the effect of varying the CMR threshold $T_r$ from 0.01 through to 0.9. Variations in $T_r$ from 0.01 to 0.5 results in about 2% change in the total signalling cost, however, as shown in Figure 4-35a, setting $T_r$ to values greater than 0.5 helps significantly reduce the call-related signalling load, whilst lower $T_r$ values can result in savings in location update costs. The value of $T_r$ can be set by the terminal and adjusted dynamically in response to variations/swings in the rate of incoming calls or frequency of location updates. Also since the network will have a better overall view of the signalling traffic, it is feasible that the appropriate $T_r$ value is decided by the network and forwarded to the MT as a parameter in the location update acknowledgement message from the network.

![Figure 4-35: Signalling cost per user vs. time of day (a) call cost (b) total cost](image)

% vehicular: 30, % pedestrian: 70 - Movement type: Manhattan Grid

![Figure 4-36: Signalling cost per user vs. time of day (a) call cost (b) total cost](image)

% vehicular: 100, % pedestrian: 0 - Movement type: Manhattan Grid
Therefore dynamic adjustments to the threshold value by the MT or from the network are both possible, and for example correct settings can help minimize paging load in cells which would otherwise carry a heavy call-related signalling traffic. Figure 4-36 depicts variations in the call and total signalling cost with different number of layers, when all subscribers are fast-moving. Once again the higher the number of layers, the more pronounced is the reduction in the total signalling cost.

Figure 4-37: Signalling cost per user vs. time of day (a) call cost (b) total cost
% vehicular: 100, % pedestrian: 0 - Movement type: Manhattan Grid

4.5 Summary of Results
The summary presented in this section covers the results drawn from simulations based on BH period (AVFF, AVMG, PVFF and PVMG), and the 24 hr. period.

4.5.1 Over the BH period

4.5.1.1 The AVFF case
- Of the two dynamic schemes, Alg. 2 is the better performer in FF environments
- Both dynamic schemes, show gains/improvements in FF environments
  - At high call-arrival rates, at velocities < 20 km/hr.
  - At low call-arrival rates, at velocities > 20 km/hr.
  - At high velocities, with call-arrival rates < 3 calls/hr.
  - At low velocities, with call rates > 3 calls/hr.
- Alg. 4 adaptive multilayer scheme, shows gains/improvements in MG environments
  - At high call-arrival rates, at velocities > 20 km/hr.
  - At low call-arrival rates, at velocities > 10 km/hr.
  - At high velocities, with call-arrival rates <= 8 calls/hr.
Chapter 4. Performance Evaluation of Location Management Techniques

4.5.1.2 The AVMG case

- Of the two dynamic schemes, Alg. 3 is the better performer in MG environments.
- Dynamic scheme of Alg. 2, shows gains/improvements in MG environments:
  - At high call-arrival rates, at velocities > 20 km/hr.
  - At low call-arrival rates, at velocities < 20 km/hr.
  - At high velocities, with call-arrival rates < 2 calls/hr.
  - At low velocities, with call-arrival rates > 2 calls/hr.
- Dynamic scheme of Alg. 3, shows gains/improvements in MG environments:
  - At high call-arrival rates, at velocities > 5 km/hr.
  - At high/low velocities, with call-arrival rates > 1 calls/hr.
- Alg. 4 adaptive multilayer scheme, shows gains/improvements in MG environments:
  - At high call-arrival rates, at velocities > 20 km/hr.
  - At low call-arrival rates, at velocities > 10 km/hr.
  - At high velocities, with call-arrival rates <= 8 calls/hr.
  - At low velocities, with call rates <= 2 calls/hr.

4.5.1.3 The PVFF case

- Of the two dynamic schemes, Alg. 2 is the better performer in FF environments.
- Both dynamic schemes, show gains/improvements in FF environments:
  - At high/low call-arrival rates, at velocities > 5 km/hr.
  - At high/low velocities, with call-arrival rates >= 2 calls/hr.
- Alg. 4 adaptive multilayer scheme, shows gains/improvements in FF environments:
  - At high/low call-arrival rates, at velocities > 30 km/hr.
  - At high/low velocities, with call-arrival rates in range 1-10 calls/hr.

4.5.1.4 The PVMG case

- Of the two dynamic schemes, Alg. 3 is the better performer in MG environments.
- Both dynamic schemes, show gains/improvements in MG environments:
  - At high/low call-arrival rates, at velocities > 5 km/hr.
  - At high/low velocities, with call-arrival rates >= 1 calls/hr.
- Alg. 4 adaptive multilayer scheme, shows gains/improvements in MG environments:
  - At high/low call-arrival rates, at velocities > 30 km/hr.
  - At high/low velocities, with call-arrival rates >= 6 calls/hr.
It is therefore evident that the type of environment, mix of subscribers and variations in velocity/call arrival rates can have a wide-ranging impact on the performance of various algorithms. The conclusions in section 4.5.2, are based on the summary of the results in Table 4-22.

4.5.2 Over the 24 hr. period - irrespective of subscriber classifications (AV/PV).

- Of the two dynamic schemes, Alg.2 is the better performer in the FF environments.
- Of the two dynamic schemes, Alg.3 is the better performer in the MG environments.
- Alg.4 outperforms all other schemes considered, in both environments.
- In terms of RIP impact, Alg.3 benefits the most.

<table>
<thead>
<tr>
<th></th>
<th>AVFF (Max. velocity 50 km/hr)</th>
<th>AVMG (Max. velocity 50 km/hr)</th>
<th>PVFF (Max. velocity 30 km/hr)</th>
<th>PVMG (Max. velocity 30 km/hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>24 Hr. period</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ave. improvement % over GSM</td>
<td>Alg.2 26.5%</td>
<td>Alg.3 19.95%</td>
<td>Alg.4 31.8%</td>
<td>Alg.5 -334.3%</td>
</tr>
<tr>
<td>Ranking</td>
<td>Alg.4</td>
<td>Alg.2</td>
<td>Alg.3</td>
<td>Alg.4</td>
</tr>
<tr>
<td>(highest gain 1st)</td>
<td>Alg.4</td>
<td>Alg.2</td>
<td>Alg.3</td>
<td>Alg.4</td>
</tr>
<tr>
<td>RIP impact - Improvement % with RIP</td>
<td>GSM 6.1%</td>
<td>Alg.2 9.0%</td>
<td>Alg.3 6.3%</td>
<td>Alg.4 4.6%</td>
</tr>
<tr>
<td>Total Improvement % inc. RIP over GSM</td>
<td>Alg.2 26.5% + 9.0% = 27.4%</td>
<td>Alg.3 19.9% + 6.3% = 26.2%</td>
<td>Alg.4 31.8% + 4.6% = 36.4%</td>
<td>Alg.5 31.8% + 4.6% = 36.4%</td>
</tr>
</tbody>
</table>

Table 4-22: Summary of the results over the 24 hr. period

4.6 Conclusions

As discussed in Section 4.1, there are two basic types of location update and paging schemes: static and dynamic. Static schemes have the disadvantage that they cannot be adjusted according to the parameter of individual users. For example, under the fixed LA-based location update scheme, the LA size most suitable for one user may be ineffective for another user. Most of the recent research efforts focus on the development of dynamic location update and paging schemes. Dynamic schemes allow online adjustments based on the characteristics of each individual MT. Some of these schemes require information, such as the distance between
cells, that is not generally available to the MTs. Besides, the operation of dynamic schemes may require significant computing power. Most of the proposed schemes for PLMN location management – as indicated in the section on literature survey - can improve the GSM strategy to a certain extent. However, it is difficult to select a scheme that clearly outperforms the others under all system parameters. In most cases, the performance of the schemes exceeds that of the GSM only under certain mobility and call arrival parameters. When a different set of parameters is used, the performance may be changed significantly. The work presented in this chapter has focused on performance evaluation and comparisons (through simulations) of 4 representative dynamic schemes, with the baseline scheme of GSM. The simulations have been conducted under a variety of common call and mobility profiles. As mentioned previously, the results available in the literature, are either based on simple/unrepresentative mobility assumptions or else different techniques are often compared against different baseline schemes. Simulation studies are therefore a better approach for conducting performance evaluations under a common set of assumptions. The simulation results presented have covered two categories:

- the Busy Hour (BH) period, at different call rates and subscriber velocities
- over a 24 hr. period.

Two basic types of movements were considered. The movements for both the vehicular and pedestrian subscribers in the Manhattan Grid (MG) environment are characterised by horizontal and vertical only movements where as in Free Flow (FF) environment, movement is characterised by the range trajectories an MT can assume i.e. the new direction is drawn from a uniform distribution over \((0, 2\pi]\). Also, the performance of selected algorithms with and without Recent Interaction Paging (RIP) is also considered. With RIP enabled, paging for the called subscriber is done initially in the last known subscriber location/cell, and following a time-out or no response from the paged cell, the whole LA (minus the paged cell) is then paged. This simple 2-step paging process was shown to be beneficial particularly when rate of updates and/or call arrival rates are high.

**Comparison of Fixed and the Distance-based scheme**

Although the distance-based strategy can dynamically adjust location area sizes of individual subscribers in response to their mobility and incoming call patterns, based on the computational results obtained, it suffers from a number of shortcomings, which render it not as efficient as the fixed LA strategy. These are:
Higher combined cost of location update and paging – Generally, when the LA size of the distance-based and the fixed-LA strategies are the same i.e. both have same paging cost, the distance-based strategy incurs higher location update cost, and on the other hand, when the location update costs are the same, the location area size for the distance-based scheme is greater than that of the fixed scheme, resulting in higher paging costs.

Location area size $k$ can not be scaled down when user is stationary – In the distance-based scheme, in response to changes in mobile call and mobility characteristics, a new location area size is calculated and applied after the next location update. Thus, $K$, should be chosen carefully, as it will not change until the next location update becomes necessary. For example, in the case of a user driving home, and staying at home at night, $k$ would be set to a large value, due to high mobility when driving home, but the user's mobility has dropped to 0 after a while. For calls arriving during this period of time, the system still has to page the original $K$ cells.

LA size $k$ can only assume specific values – For the cellular system shown in Figure 4-9 the possible values of $K$ are 1, 7, 19 ... as given by (4-16) for hexagonal cells, and 1, 6, 18 ... as given by (4-17) for square cells. But mobility and call arrival rates usually assume real values, and therefore, most values of $K$ that are used, are in fact sub-optimal values.

The main advantage of distance-based schemes is LA overlapping that can be used to provide a certain level of location area hysterisis, and whereas in the fixed scheme, the base stations on the border of location areas carry all the location update signalling traffic, in the distance-based scheme, this signalling load is distributed over many base stations.

Given the scope and the extent of the simulations carried out, and based on the results from simulations over the 24hr. period, the following conclusions are drawn:

- Of the two dynamic schemes, Alg.2 is the better performer with FF movement, whilst Alg.3 is the better performer with MG movement. The proposed Adaptive Multilayer scheme out performs all other schemes, irrespective of the type of movement, and it is the only scheme where LA size can be scaled down to 1 cell, when user is stationary.

- Recent Interaction Paging (RIP) has shown the most significant impact in mixed subscriber environments where majority of subscribers are slow-moving i.e. the PVMG case, and then PVFF, followed by AVFF and AVMG cases. Except for the PVMG case where the average overall improvement with RIP is around 18%, in all other cases enabling RIP results in an average (over all schemes considered) improvement of 8%.
Of all the schemes considered, RIP produces most gain when applied in conjunction with schemes of Alg.3 and Alg.4, and both schemes with RIP enabled, provide highest gains/ cost savings in terms of total signalling load (ave. 31%, max. 38%) in almost all environments when RIP is enabled compared with the baseline fixed-LA scheme of GSM.

High performance gains by the scheme of Alg.3, is due to the ability of the scheme to take advantage of available regional information, implicit in the calculation of the cell visiting rate and implied within the formulation of the cost function. Nevertheless, improvements only begin to set in after about 20 km/hr. and at call rates below 3 calls/hr. Despite the observed significant performance gains/savings by the schemes of Alg.3 and Alg.4, it should be noted that implementation of the scheme of Alg.3 requires the MTs to forward the identities of all the cells within the personal LA of MT to the network, during location update procedure. Inevitably, this operation results in higher resource consumption i.e. more signalling bandwidth will be required, and also the network databases will be required to maintain and update the list of cells comprising the personal LA of subscribers, on a per-user basis. On the other hand, the scheme of Alg.4, only requires broadcast of a single additional parameter i.e. the no. of layers, and since the identities of all LAs are predetermined at the network side, and there are no additional requirements, implementation of the scheme of Alg.4 is less complex than the scheme of Alg.3.

Based on the computational results obtained, the distance-based strategy of Alg.5 suffers from a number of shortcomings, which render it not as efficient as the fixed LA strategy. A higher combined cost of location update and paging is associated with the scheme of Alg.5, since when the LA size of the distance-based and the fixed-LA strategies are the same (both have same paging cost), the distance-based strategy incurs higher location update cost, and on the other hand, when the location update costs are the same, the location area size for the distance-based scheme is greater than that of the fixed scheme, resulting in higher paging costs. Also Location area size can not be scaled down when user is stationary , and the LA size can only assume specific values and most values that are used, are in fact sub-optimal values.

A final notable observation relates to relatively high levels of location update signalling associated with the Distance-based Scheme, where the high levels of LU traffic observed, are compensated by extremely low levels of call-related signalling traffic compared to other schemes. It seems that the distance-based strategy is able to take full advantage of RIP, and this opens up the possibility that combinations of techniques would perhaps result in even greater savings in the signalling load. For example, the dynamic scheme of Alg.3 and the Adaptive Multilayer scheme (Alg.4) have shown particularly high aptitude in
adapting/tracking local mobility profiles and appear rather successful in reducing the location update related signalling traffic. A combination of Alg.3/Alg.4 and the distance-based strategy (with RIP enabled) could in principle be used such that during periods of time when the call-to-mobility of subscriber is high, the system (network or terminal) initiates a distance-based LU mechanism, and when the mobility begins to or is estimated to result in high rate of location crossings i.e. at low call-to-mobility ratios, the system reverts back to either of the schemes of Alg.3 or Alg.4. Such a combination not withstanding the implementation issues, seems a suitable combination for further reducing the signalling across the air-interface, especially since simulation results on the whole, indicate that in fact apart from the distance-based scheme, all other schemes manage to offer reductions in the total signalling load mostly due to minimisation of the location update related signalling, with little impact on the call-related signalling.
Chapter 5

A HYBRID GA-HEURISTIC APPROACH FOR PACKET NETWORK DESIGN

5 Introduction

Telecommunications networks are complex systems. They can consist of mixes of many physical and logical layers, built using different technologies. A typical network can be viewed as a multi-layered structure, with each layer requiring different network design to account for different design objectives, constraints and technologies. Topological aspect of network design relates to service layers of a network, whilst routing and capacity design problems are network layer related.

The role of network design in providing Quality of Service (QoS) is often overlooked; a poorly designed network architecture will not - regardless of the sophistication of the bandwidth allocation strategy used - be able to match the performance of a well designed network. Regular and random partially meshed network designs can overcome many of the problems associated with conventional network designs. These partially meshed networks have bounded hop counts, relatively few links (~N^2) and an even distribution of routes across the network. The even route distribution allows path selection i.e. routing, algorithms to load traffic evenly, thus preventing network hotspots that degrade performance. This also improves networks response to node/link failures. Mesh and ring network topologies are good starting points for network design, as they embody various trade-off involved in almost all network design problems. Supposing that a network is required to carry a constant level of traffic C between all pairs N of nodes. In a fully-meshed configuration, there are links on the order of N^2 i.e. (N(N-1)/2) links. Each N(N-1) streams of traffic is switched twice, once at the originating node, and once at the destination node. Each stream has one-hop between source and destination, where a hop is defined as the number of links traversed. Alternatively, all nodes can be connected alternatively into a ring using N links. Then, the traffic between non-adjacent nodes has to be switched by the intervening nodes and the average number of hops per traffic stream is proportional to N. The total transit traffic in a ring network scales as N^3, and thus more traffic carrying capacity has to be provided by the links and nodes to carry this transit traffic. In fully-meshed configurations,
the network makes optimum use of its nodes i.e. no capacity is used for switching transit traffic. However, too much redundancy (too many links) are used to achieve this objective, and the number of required links grows so fast with $N$ that full-meshes are only practical for small networks. In ring networks on the other hand, more and more capacity has to be devoted to carrying transit traffic, as the size of the network increases. In both cases the designs do not scale well—a ring has too many hops and a fully-meshed network requires too many links.

Networks with better scaling properties can be designed by selectively adding links between nodes to form a random partial-mesh. Typically, each node will be connected to at least two other nodes to ensure that the network survives a link failure, and that no single node failure can split the network into two pieces. Therefore, a random partial-mesh represents a good compromise between ring and full-mesh configurations.

**Some QoS Support Requirements in GPRS**

The eminent introduction of General Packet Radio Service (GPRS) will greatly improve and simplify wireless access to packet data networks e.g. internet, and whilst it supports interworking with IP and X25 based networks, it will provide data services based on a packet-switched core network. GPRS users will benefit from shorter access times and high data transfer rates.

In order to integrate GPRS into the existing GSM network architecture, a new class of network nodes, called GPRS Support Nodes (GSN) have been introduced [GSMr3.02]. GSN are essentially responsible for the delivery and routing of data packets between the mobile station and the external packet data networks (PDNs). Figure 5-1 illustrates the system architecture. A serving GSN (SGSN) is responsible for delivery of packets to and from mobile stations within its service area. SGSNs are also in charge of packet routing and transfer, mobility management (attach/detach, location management), logical link management and authentication and charging functions. The GPRS location register stores location information and user service profiles of all subscribers registered with this SGSN. A gateway GSN (GGSN) acts as the interface between the GPRS backbone and external packet data networks. It is responsible for converting GPRS packets coming from the SGSN into appropriate packet data protocol (PDP) format e.g. IP or X25, and sending them out to the corresponding external PDN. PDP addresses of incoming data packets are converted to GSM address of the called user. The re-addressed packets are the forwarded to the appropriate SGSN.
All GSNs are interconnected via an IP-based backbone network. Within this backbone, the GSNs encapsulate the PDN packets and transmit (tunnel) them using GPRS tunnelling protocol (GTP). GPRS will also support different QoS classes, which can be specified for each individual session. GPRS supports definition of QoS profiles using service precedence, reliability, delay and throughput parameters [GSMr2.60]. For example, reliability indicates the transmission characteristics required by an application.

As shown in Table 5-1, three reliability classes are defined which guarantee certain maximum values for the probability of loss, duplication, mis-sequencing and corruption of packets. The delay parameters define maximum values for the mean delay and 95-percentile delay (Table 5-2 [GSMr2.60]). The latter being the maximum delay guaranteed in 95% of all transfers. The delay is defined as the end-to-end transfer time between two communicating mobile stations or between a mobile and an external PDN. This includes all the delays within the GPRS network i.e. delay for request and assignment of radio channel and the transit delay in the GPRS backbone. Using these QoS classes, QoS profiles can be negotiated between the mobile subscriber and the network for each session, depending on the network QoS demand and available resources. The charging for the service is then based on the transmitted data volume, type of service and the chosen QoS profile. Although a QoS profile is defined, this profile does not differentiate treatment of data flows within the core GPRS network i.e. between GSNs. All flows within the network are treated uniformly since the present use of IP tunnels (based on GTP protocol) creates difficulties with the applicability of IP QoS schemes (e.g. IntServ, DiffServ) Thus all flows within the network are treated uniformly (IP best effort delivery) and therefore the aforementioned QoS profiles/parameters can only be used to allocate resources outside the core network. So after activation of a certain QoS profile, the MT is responsible for shaping its traffic to the negotiated QoS, whilst GGSN is responsible for restricting dataflows to
the MS based on its QoS profile. This implementation can be considered a network with guaranteed bandwidth on the access interface, but with no provision for resource allocation/reservation within the core GPRS network.

<table>
<thead>
<tr>
<th>Class</th>
<th>Probability for</th>
<th>Lost packet</th>
<th>duplicated packet</th>
<th>out of sequence packet</th>
<th>Corrupted packet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>$10^{-6}$</td>
<td>$10^{-6}$</td>
<td>$10^{-6}$</td>
<td>$10^{-6}$</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>$10^{-6}$</td>
<td>$10^{-6}$</td>
<td>$10^{-6}$</td>
<td>$10^{-6}$</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>$10^{-2}$</td>
<td>$10^{-5}$</td>
<td>$10^{-5}$</td>
<td>$10^{-2}$</td>
</tr>
</tbody>
</table>

Table 5-1: Reliability classes

<table>
<thead>
<tr>
<th>Class</th>
<th>128 byte packet</th>
<th>Mean delay</th>
<th>95% delay</th>
<th>1024 byte packet</th>
<th>Mean delay</th>
<th>95% delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$&lt;0.5s$</td>
<td>$&lt;1.5s$</td>
<td>$&lt;2s$</td>
<td>$&lt;7s$</td>
<td>$&lt;15s$</td>
<td>$&lt;75s$</td>
</tr>
<tr>
<td>2</td>
<td>$&lt;5s$</td>
<td>$&lt;25s$</td>
<td>$&lt;15s$</td>
<td>$&lt;75s$</td>
<td>$&lt;250s$</td>
<td>$&lt;75s$</td>
</tr>
<tr>
<td>3</td>
<td>$&lt;50s$</td>
<td>$&lt;250s$</td>
<td>$&lt;75s$</td>
<td>$&lt;375s$</td>
<td>$&lt;750s$</td>
<td>$&lt;375s$</td>
</tr>
<tr>
<td>4</td>
<td>best effort</td>
<td>best effort</td>
<td>best effort</td>
<td>best effort</td>
<td>best effort</td>
<td>best effort</td>
</tr>
</tbody>
</table>

Table 5-2: Delay classes

It is therefore evident that in order to meet the required quality of service within the core network, particularly in terms of the maximum allowable delays and resource allocation requirements, some form of IP QoS protocol will need to be incorporated into GPRS. For proposals on how QoS enabling schemes may be utilized to enhance GPRS architecture the reader is referred to [Mic98][Puu99][Hof98][Pri00]. However, the role of network design is important as a poorly designed architecture, regardless of the sophistication of the resource allocation strategy used, will not perform as efficiently as a well designed network. Therefore cost-efficient design of the network with respect to capacity, link flows and reliability constraints will play an increasingly important role in the provision of network level QoS, prior to deployment of any network-wide QoS schemes.

**Network design**

Network design invariably involves a compromise between cost and performance. Designing a communication network often can be viewed as a complex multi-constraint optimisation problem, where the design must meet a set of possibly conflicting, interdependent requirements and invariably, the design must provide optimal assignments for link capacity, routing and topological connectivity, such that costs are minimised and yet satisfy the traffic requirements and keep network delays within permissible bounds. In telecommunication networks, the common optimisation problems of concern are capacity and delay constraints, routing assignment, topology and cost. Addressing these issues in isolation or as a multi-objective problem, often involves non-linear and discrete function formulations for optimisation, which may not yield adequate solutions using classical gradient-based type of search algorithms. The problems can be more severe when multiobjective functions are encountered.
Chapter 5. A Hybrid GA-Heuristic Approach for Packet Network Design

The basic principles of network design do not change significantly as the technology develops, although the nature of the tradeoffs may change if for example, one type of network components become much less expensive. In fact the problem is NP-complete (i.e. no algorithm exists that can find the optimum solution in polynomial time, OR one that has a solution in polynomial time, but can only be solved through non-deterministic algorithms) and for practical problems with modest number of nodes, only approximate solutions can be obtained through heuristic algorithms. Techniques such as simulated annealing and genetic algorithms can be used find near-optimal, but useful solutions and some design tools incorporate these algorithms and other heuristics. Heuristic methods for communication network design include branch exchange [Ber87], cut-saturation, and the more efficient MENTOR algorithm [Ker93]. These can be applied (with certain modifications), more or less to any network design problem. However, there is a great demand for more efficient design techniques since quite often, the above methods can yield satisfactory yet far from optimal solutions. Thus, the subject remains open to further study. In the design of mesh packet-switched networks, the component subproblems of routing and capacity assignment are difficult problems. In particular the relation between the cost and capacity is usually assumed to be linear. For the more realistic discrete cost-capacity functions, many of the current design algorithms can only provide approximate solutions with a bound to the optimum. Methods which guarantees an optimal solution, with a discrete cost-capacity function, require very long search times to provide solutions for realistic network sizes (chapter 7 of [Ker93]). Another heuristic technique, Genetic Algorithms (GAs) effectively conduct a parallel search of the solution space, starting from a number of points in the search domain and maintaining this multidimensionality during most of the time that the process lasts. GAs appear ideal for the design of mesh networks, with the capability for handling discrete values, multiobjective functions and multi-constraint problems [Ma97] and there exists a number of applications of GAs to the network design problem [Man97][Elb95][Pie96].

The objective of the work presented in this chapter is to develop a novel and efficient GA-based network design technique for partially-meshed communication networks. The proposed method for the solution of the complete network design problem is a hybrid technique in that it has borrowed useful component algorithms from conventional network design techniques and incorporated them into the GA optimisation process. The proposed hybrid approach can be applied to the design of packet-based communications networks such as the GPRS backbone i.e. between SGSNs and GGSNs. The technique developed is used to solve the class of TCFA problem and is shown to be able to satisfy the traffic requirements as well as capacity, delay and reliability constraints.
Chapter 5. A Hybrid GA-Heuristic Approach for Packet Network Design

The organisation of this chapter is as follows. In section 5.1 a number of existing techniques in the related literature is described. Section 5.2 introduces the design parameters of concern and the methodology for formulation of the network design problem. The 10-node network design problem is then described in some detail in section 5.3. Section 5.3 also contains descriptions of the proposed hybrid GA-based method and two other techniques used in comparisons. In section 5.4, the numerical results including GA parameter settings, process performance and comparisons with two other methods (as applied to a specific network design problem), are presented. Finally, in section 5.5 a summary of results and conclusions are presented.

5.1 Literature Review

On Existing Methods for the Telecommunication Network Design

Previous work in this area are few and often restrictive in their formulation. As mentioned previously, heuristic methods for communication network design include Branch-exchange [Ber87], the Concave Branch Elimination (CBE) methods [Ger76], cut-saturation and the Mentor algorithm [Ker93]. Starting with a complete graph, Branch-exchange optimises connections by dropping links with the highest cost per bit. This process repeats until there is no further improvement. Branch-exchange optimises routing based on finding the minimum-distance routes. It assigns capacity with the delay constraint fulfilled (based on a heuristic approach). The Mentor approach though more efficient, can not be directly applied to the mesh network design problem, as it would require modifications to satisfy biconnectivity requirements. Similar to the Branch-exchange, the Concave Branch Elimination methods start from a given topology (a randomly generated one, or a fully connected one) and drop or add links while solving a CFA problem, each time the network topology is transformed. The optimisation process suggests a walk in the huge domain of the problem’s solution space and focuses on to the first local minima it approaches, with the hope that even if this is not the globally optimal one (this is unlikely to happen but not impossible), it is not a mediocre local one. The main disadvantage of the above heuristics, concerning efficiency, is that they start from a single point in the search domain. Thus, they don’t have the ability to reach any solutions that are not located sufficiently close to the starting point. In contrast, GAs conduct a parallel search of the solution space. GAs have also been applied to several aspects of network design problem: the unconstrained Minimum Spanning Tree [Mic91], a solution for local area networks [Elb95], concentrator assignment [Rou94] and routing [Sin93]. However, they have only optimised network topology and ignore the more difficult subproblems of routing and
Chapter 5. A Hybrid GA-Heuristic Approach for Packet Network Design

capacity assignment, crucial metrics of network service quality and cost. In [Web98] GAs and random search techniques are used to obtain optimal backbone topologies. It has been shown that GAs can be applied to produce feasible network designs that minimise overall cost and how tariffing (distance- or usage-based) can affect optimal component placement. [Sin98] reports on the results of minimum cost routing and wavelength allocation of multiwavelength all optical transport networks using a GA/Heuristic approach. Since hybrid approaches are often able to incorporate problem specific heuristics, such combined algorithms generally perform better than pure GA or the heuristic alone techniques [Dav91]. Based on the survey of work on the application of GAs to network design in [Pro96] which shows that even without inclusion of heuristic-based operators, GAs are often able to match heuristics alone in terms of results, if not speed, the author in [Sin98] advocates a hybrid approach.

5.1.1 Decomposition of the Network Design Problem

The network design problem as stated in the previous section is a complex problem, in that there exists no efficient technique for the exact optimal solution of the complete network design problem. The starting point for formulation of a network design methodology, is the definition of the network architecture and the design parameters. The work presented in this chapter, is based on the well-known problem formulation for packet-switched communication network design in [Ger76]. Therefore the network design problem is formulated according to specifications in [Ger76], thus the complexity of the problem is broken down and seen to be driven only by the constraints and cost requirements. Based on the methods proposed and implemented ([Ber87], [Ger76], [Ker93]) the network design problem can be decomposed into the following subproblems:

**The Link Capacity Assignment (CA) Problem:** Given a topology and a routing policy (and thus flow vector), optimise link capacity assignment so as to achieve a packet delay value below a given value (delay constraint), with minimum cost. The optimal assignment of capacities to a distributed network with arbitrarily fixed routes is not very interesting as a stand-alone problem, since routing plays an essential role in the optimisation of network performance. Rather the CA is of practical interest as a subproblem of more generalised problems. The technique used for the solution of the CA problem depends on the nature of the cost-capacity function. For the case of continuous costs, optimal methods are available [Ger76]. For the case of discrete costs, optimal methods (based on dynamic programming techniques) still exist, but sometimes suboptimal methods, with lower order of complexity, are preferred.
The Routing or Flow Assignment (FA) Problem: Given a topology, the capacity of the established links and the traffic requirements, optimise the routing policy (i.e. find a fixed policy) so as to achieve minimum packet delay. There exists a number of heuristic and Calculus Based (down hill search) methods that can be employed [Ger76]. Some of them provide an optimal solution (Flow Deviation method and the superior Bertsekas-Gallager algorithm).

The Capacity and Flow Assignment (CFA) Problem: Given a topology and the traffic requirements, optimise routing policy and capacity assignment so as to achieve satisfaction of the delay constraint with minimum cost. For this problem, no technique that guarantees the optimal solution exists. However very efficient heuristics have been employed which can provide suboptimal solutions, anticipated not to be far from the optimal. These methods are based on a succession of CA and FA problem settings, to be solved alternately [Ger76].

The Topology, Capacity and Flow Assignment (TCFA) Problem: Given a topology and the traffic requirements, optimised routing policy and capacity assignment, minimize network delay (or cost in the dual-form of the problem) so as to achieve satisfaction of the cost (or delay in dual form) constraint. TCFA represents the most demanding of the four subproblems and for this problem also, no technique that guarantees the optimal solution exists. However efficient heuristics have been employed such as CBE which are believed to come within 5-10% of optimum at reduced computation costs. These methods are based on a succession of CA and FA problem settings, to be solved alternately [Ger76][Ger73].

5.2 Design of the Mesh Communication Networks

5.2.1 Requirements, costs and constraints

In a packet-switched network, packets are transmitted through the network using a store and forward mechanism. That is, a packet travelling from source node $s$ to destination node $d$ is received and “stored” in queue at any intermediate node $k$, while awaiting transmission, and is then “forwarded” to node $p$, the next node from $s$ to $d$, when channel (or link) $(k, p)$ permits. Even when this channel is free, the packet must first be received fully by node $k$ before transmission to node $p$ may be started. Given the destination and the present node $k$, the selection of the next node $p$ is made by a well defined decision rule referred to as the routing policy. A routing policy is said to be fixed (or static) if a predetermined fraction of the packets arriving at $k$ and directed to $d$ is sent to each output queue; it is said to be adaptive if the selection of the output channel at each node depends on some estimate of current network traffic. For more detailed discussion of routing mechanisms in telecommunications networks the reader is referred to [Ash95][Kha94] and references contained therein.
Traffic requirements between nodes arise at random times and the size of the requirement is also a random variable. Consequently, queues of packets build up at the nodes and the system behaves as a stochastic network of queues [Kle76]. For routing purposes, packets are distinguished only on the basis of their destination. Thus, messages having a common destination can be considered as forming a “class of customers” [Ger76]. The data network, therefore, can be modelled as a network of queues with N classes of customers where N is the number of different destinations (number of the nodes of a network).

5.2.1.1 Traffic Requirements

Average (busy-hour) traffic requirements between nodes can be represented by a requirement matrix \( R = \{ r_{jk} \} \), where \( r_{jk} \) is the average transmission rate from source j to destination k. In general \( R \) cannot be estimated accurately apriori because of its dependence upon the network parameters (e.g. allocation of resources to servers, demand for resources etc.) which are difficult to forecast and subject to fluctuations with time. Fortunately, the analysis of several different traffic situations has shown that the optimal design is rather insensitive to traffic pattern variations. This insensitivity property, which seems typical of distributed networks, justifies the use of traffic averages for network design.

5.2.1.2 Routing Policy

The routing procedure is one of the most important functions a network performs. Routing can be studied from the perspective of designing a network or from the perspective of managing a network in real time adjusting to momentary changes in load (adaptive routing case). The algorithms used in both cases are similar. The difference is that during the design process we identify and define the optimal paths for each requirement based on the average/peak anticipated traffic on each link of the network at hand, but once the network is on-line (the network has been established and is in operation) optimal routes are sought with respect to the current state of the network (routes are updated with a frequency defined by the network operator). There are a number of different routing procedures that are employed in various networks ([Ker93], [Tan96]). For example, in fixed routing policies a single path or a number of paths (with given fraction of traffic to serve: this is known as alternate fixed routing policy), are defined during the design process. More popular in the current networks, appear to be the adaptive routing policies that employ collection of the information about the state of the network (load, link state or node failures etc.). Any node in such a network, routes the packets it receives with respect, to the current state of the network. Shortest paths (in terms of delay or cost - it is up to the operator to choose the criteria) through the mesh topology are sought, and
routing decisions are taken in real time. Flooding is another early routing procedure. In designing network topologies, one generally assumes fixed routing, since it is easy to describe (by means of routing tables) and allows the direct evaluation of channel flows and average delays as a function of routing tables and traffic requirements [Ger76]. Adaptive routing, on the other hand, is complex to describe, and requires simulations to evaluate channel flows and delay.

5.2.1.3 Link Flows
The routing policy and the traffic requirements uniquely determine the vector $f=(f_1, f_2, \ldots, f_b)$ where $f_i$ is the average data flow on link $i$ and $b$ is the number of links in the network. The evaluation of $f$ is straightforward in the case of fixed routing. For the case of adaptive routing, it can be obtained by simulation [Ger76].

5.2.1.4 Communication Cost
This can include the monthly lease cost, installation cost, cost per unit traffic, and maintenance costs. Costs vary depending on whether the facility is owned or leased. Usually, cost is modelled simply by a cost figure, or if cost is usage sensitive, by a cost per unit of usage (e.g., cost per hour or cost per bit). Both the methodology used and the topologies selected in the design of a network are strongly influenced by the variations in the cost structure, and it is important for the designer to take this into account. Thus, if costs vary roughly linearly with distance we choose a very different topology than if cost is relatively invariant. Similarly, if there is a strong economy of scale with respect to capacity, that is if cost goes up much more slowly than capacity, it has a profound effect on the topology selected [Ker93]. There is also a significant difference between fixed monthly costs and usage sensitive costs. Often, a network incurs both types of cost and one can be traded off against the other. The typical situation is that fixed costs are used for large volumes of traffic (where a fixed cost can be justified by spreading it out over a high volume of traffic) and then the remaining traffic flows over usage sensitive facilities.

With $C_i$ the capacity of link $i$, we let $d_i(C_i)$ be the cost of leasing capacity value $C_i$ for link $i$. The communication cost $D$ is then defined as:

$$D = \sum_{i=1}^{b} d_i(C_i)$$

(5-1)

The capacity cost function $d_i(C_i)$ can be either a continuous or a discrete function. In most real applications, however, capacities are specified as discrete values.
5.2.1.5 Delay Analysis

A vital performance measure of a data network is the average source-to-destination packet delay $T$, defined as follows

$$T = \sum_{j,k} \gamma_{jk} Z_{jk}$$  \hspace{1cm} (5-2)

where

$\gamma_{jk}$ average packet rate flowing from source j to destination k

$Z_{jk}$ average packet delay (queue and transmission) from j to k

$$\gamma = \sum_{j,k} \gamma_{jk}$$  \hspace{1cm} (5-3)

A straightforward application of Little's result [Kle76] to the network of queues model leads to the following very useful expression for $T$:

$$T = \sum_{i=1}^{b} \frac{\lambda_i}{\gamma} T_i$$  \hspace{1cm} (5-4)

where $b$ is the number of links (arcs), $\lambda_i$ is the average traffic rate, and $T_i$ is the average queuing plus transmission delay on link $i$. This expression is very general and extremely simple. Unfortunately, we are not able in general to evaluate $\lambda_i$ and $T_i$ analytically. However with the following assumptions [Kle76]: 1) Poisson arrivals (exponential packet length distribution), 2) fixed routing, 3) error free channels, and 4) independence between interarrival times and transmission times on each channel, the evaluation of the above expression can be carried out analytically. In fact the network of queues reduces to a model in which each queue behaves as an independent M/M/1 queue [Kle76]. Thus, the average delay $T_i$ on link $i$ is given by

$$T_i = \frac{1}{\mu C_i - \lambda_i}$$  \hspace{1cm} (5-5)

where

$1/\mu$ average packet length (bits/packet)

$C_i$ capacity of channel $i$ (bits/sec)

$\lambda_i$ average packet rate on link $i$ (packets/sec)

The average rates $\lambda_i$ are easily computed from the routing tables and the traffic requirement matrix. By substituting (5-5) into (5-4) and letting $f_i$ be the average bit rate on channel $i$ (bits/sec), we obtain the following expression for $T$:
where $T$ expresses the average source to destination delay a packet will incur (sec/packet). The above formula is known as “Kleinrock independence approximation”.

5.2.1.6 Capacity Constraint

Capacity refers to the amount of traffic a channel can carry. A channel may be full duplex, permitting simultaneous communication in both directions, half duplex, permitting communication in one direction at a time, or simplex permitting communication in only one direction. It is essential that we distinguish among these cases in modelling a network. The presence of capacity constraints $f \leq C$ (where $C=(C_1, C_2, ..., C_b)$) makes the design problem a constrained problem. From the delay expressions we notice that if the link flow approaches the link capacity, then the delay approaches infinity, thus violating the delay constraint. Therefore, if both capacity and delay constraints must be satisfied the capacity constraint is implied by the delay constraint and can be disregarded.

5.2.1.7 Reliability Considerations

Links and nodes in a real network can fail with nonzero probability, thus interrupting some communication paths. It is important to evaluate the overall network reliability in the presence of such failure probabilities. A proposed reliability measure is the two-connectivity (biconnectivity) of the network (i.e. two node-disjoint paths available between each node pair). This measure is easy to include as a constraint in the topological design. Furthermore, it is adequate for networks with a relatively small number of nodes (on the order of 20-40) and relatively small component failure probability (on the order of 0.01). For larger networks (or higher failure rates), stronger constraints must be applied to the network topology (e.g. three-connectivity, no long chains, etc.) in order to obtain adequate reliability. It is the need for such reliability guarantees that gives rise to a mesh topology for the network to be designed.

5.3 A 10-node Network Design Problem

The proposed GA-based network design method, is applied to the design of a 10-node mesh network. The mesh network in question, is aimed at interconnecting ten major Chinese cities as depicted in Figure 5-2 [Man97]. This reasonable size network problem, with realistic topology and traffic requirements, is suitable for evaluating the proposed technique. Besides, in [Man97]
two other solutions based on different techniques have been provided; so it will be possible to make comparisons and evaluate the quality of solutions provided by the proposed technique.

![Figure 5-2: The ten nodes (cities) to interconnect [Man97]](image)

The traffic requirements (average peak hour) for each pair of nodes are given in Table 5-3 (from [Man97]). The average packet length is 1Mbits. Traffic requirements are symmetric implying that the amount of data that node \( s \) has to send to node \( d \) (data originating \( s \)) is equal to the amount, \( d \) has to send to \( s \).

<table>
<thead>
<tr>
<th>Nodes</th>
<th>B</th>
<th>S</th>
<th>G</th>
<th>H</th>
<th>W</th>
<th>C</th>
<th>X</th>
<th>K</th>
<th>Ha</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>S</td>
<td>20</td>
<td>0</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>G</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>20</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>H</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>W</td>
<td>20</td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>X</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>K</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ha</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>T</td>
<td>20</td>
<td>20</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5-4: Distances between the 10 cities (nodes) in Km [Man97]

<table>
<thead>
<tr>
<th>Nodes</th>
<th>B</th>
<th>S</th>
<th>G</th>
<th>H</th>
<th>W</th>
<th>C</th>
<th>X</th>
<th>K</th>
<th>Ha</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>1200</td>
<td>2000</td>
<td>2100</td>
<td>1120</td>
<td>1600</td>
<td>960</td>
<td>2160</td>
<td>1120</td>
<td>160</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>1200</td>
<td>1280</td>
<td>1250</td>
<td>720</td>
<td>1680</td>
<td>1240</td>
<td>2000</td>
<td>2240</td>
<td>1120</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>2000</td>
<td>1280</td>
<td>0</td>
<td>240</td>
<td>840</td>
<td>1240</td>
<td>1360</td>
<td>1160</td>
<td>3120</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>2100</td>
<td>1250</td>
<td>240</td>
<td>0</td>
<td>960</td>
<td>1480</td>
<td>1440</td>
<td>1400</td>
<td>3220</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>1120</td>
<td>720</td>
<td>840</td>
<td>960</td>
<td>0</td>
<td>1000</td>
<td>680</td>
<td>1320</td>
<td>2190</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>1600</td>
<td>1680</td>
<td>1240</td>
<td>1480</td>
<td>1000</td>
<td>640</td>
<td>680</td>
<td>2720</td>
<td>1600</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>960</td>
<td>1240</td>
<td>1360</td>
<td>1440</td>
<td>680</td>
<td>640</td>
<td>0</td>
<td>1240</td>
<td>2080</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>2160</td>
<td>2000</td>
<td>1160</td>
<td>1400</td>
<td>1320</td>
<td>680</td>
<td>1240</td>
<td>0</td>
<td>3280</td>
<td></td>
</tr>
<tr>
<td>Ha</td>
<td>1120</td>
<td>2240</td>
<td>3120</td>
<td>3220</td>
<td>2190</td>
<td>2720</td>
<td>2080</td>
<td>3280</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>160</td>
<td>1120</td>
<td>2400</td>
<td>2400</td>
<td>1040</td>
<td>1600</td>
<td>960</td>
<td>2200</td>
<td>1150</td>
<td>0</td>
</tr>
</tbody>
</table>
The distances between each pair of nodes are given in Table 5-4. The available channels to be deployed are bidirectional ones. So links are full duplex, allowing simultaneous data flow in both directions. Thus, in this design problem, the network can be modelled as an undirected graph. The solution(s) produced must comply with the following two constraints:

- **Delay Constraint**: The average source to destination packet delay in the network, must be below $T_{max} = 0.1$ seconds. Average packet delay values are computed using (5-6) (Kleinrock independence approximation) adopting all the relevant assumptions stated in section 5.2.1.

- **Reliability Constraint**: The network has to be biconnected. Thus, at least two disjoint (with no common node and therefore no common link) paths will exist between each pair of nodes. Biconnectivity guarantees that in the case of failure of any node (but one at a time) the network still remains connected, and thus maintains acceptable performance.

A compact statement of the design problem is in Table 5-5. This is simplified into three sets of design variables as indicated, corresponding to the number of optimization levels depicted in Figure 5-3.

<table>
<thead>
<tr>
<th>Given</th>
<th>Node locations</th>
<th>Topology Optimization</th>
</tr>
</thead>
<tbody>
<tr>
<td>minimise</td>
<td>Peak-hour traffic</td>
<td>Routing Optimization</td>
</tr>
<tr>
<td>Over design variables</td>
<td>requirements between</td>
<td>Capacity Optimization</td>
</tr>
<tr>
<td>Subject to</td>
<td>node pairs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total network cost</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Topology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Routing policy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Channel capacities</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Satisfaction of traffic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>requirements</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Link capacity constraint</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average packet delay</td>
<td></td>
</tr>
<tr>
<td></td>
<td>constraint</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reliability constraint</td>
<td></td>
</tr>
</tbody>
</table>

Table 5-5: Design problem formulation

The optimal solution must meet the two constraints of capacity, delay and reliability while serving the traffic requirements and all that with minimum cost. The cost referred to has exclusively to do with the cost of the links establishment. The evaluation of the cost of the equipment required in each node is beyond the design process objectives since it does not govern the network design tradeoffs. Thus, the cost of the network is simply computed as the summation of the costs of the established links.

### 5.3.1 Design Parameters

The network design parameters subject to optimisation are:
• **Network Topology**

• **Routing Policy**: During the design process fixed routing policy is assumed.

• **Capacity Assignment**: Available link capacities, have discrete costs (this is the real case). The different capacity types are the line rates [Man97].

### 5.3.2 The MAN-GA Approach [Man97]

In this section the GA approach in [Man97] that has been applied to the specific 10-node network design problem (section 5.2.2) is described. The overall strategy in [Man97] is governed by the 3-level optimisation process indicated in Figure 5-3. A detailed account of the optimisation procedures adopted in [Man97] is provided in the following subsections.

**Topology Optimisation** – In this process the topology of the network need not be specified beforehand, as the set of links connecting the nodes is used to define a network topology. The process starts with fixed node locations and mapping of node names to node number (x) (where x is an integer between 1 and n). The optimisation process selects a set of links connecting the nodes. A topology is then defined as a set of nodes and edges connecting the nodes. The so-called adjacency matrix representation is used to represent topology. Using this representation, an $n \times n$ matrix of Boolean values is developed, where entry $a_{ij}$ is set to 1 if there is an edge connecting from node $i$ to $j$, and 0 otherwise. It is assumed that the link is bidirectional, resulting in a symmetric matrix. Hence only $n(n-1)/2$ binary numbers are required to specify a topology. For simplicity, the two-dimensional matrix is transformed into a one-dimensional matrix $A$, suitable for bit-string representation in genetic algorithms:

$$A_k = a_{ij}$$

Where for $j > k, k = jx - [j(j+1)/2] + i - j - 1$.

For $i = j, a_{ij} = 0$ (no node sends messages to itself).

The topological chromosome can thus be formulated as a binary string with $n(n-1)/2$ elements. To manipulate the chromosomes, one-point crossover and bit mutation are applied. Although this first-step of optimisation process yields a working topology, the design is not complete without the involvement of routing and capacity assignments. The results of these – link flow and optimal link capacities - optimisation procedures, become available at the end of their respective GA cycles.
Routing Optimisation – for a given topology, the link capacity assignment and routing require simultaneous optimisation of both the flow and link capacities. In this second-step of the optimisation process, the routing scheme is considered as a GA chromosome. The routing chromosome is then represented as

\[ R = \{ R_1, R_2, R_3, \ldots, R_{n(n-1)/2} \} \]

Where \( R_k = \text{Path}(i,j) = [i,x_1,x_2,\ldots,x_n,j] \) is the gene that represents the routing path from node \( i \) to \( j \); where \( i = 1, \ldots, (n-1) \) and \( j = (i+1), \ldots, n \).

With a particular routing scheme \( R \), the flow of the link can be assigned according to the traffic requirement in Table 5-3. The fitness of the routing scheme is evaluated using the optimal solution obtained from optimising capacity assignment. With parental routing chromosomes \( P \) and \( Q \), where \( P = \{ p_1, p_2, p_3, \ldots, p_{n(n-1)/2} \} \) and \( Q = \{ q_1, q_2, q_3, \ldots, q_{n(n-1)/2} \} \), the offspring due to crossover operation is represented as

\[ R = \{ s(p_1,q_1), s(p_2,q_2), \ldots, s(p_{n(n-1)/2},q_{n(n-1)/2}) \} \]

Where \( s(p_i,q_j) = \begin{cases} p_i & \text{if } p_i \text{ has shorter distance than } q_j \\ q_j & \text{otherwise} \end{cases} \)

The routing path \( P_k \) will be randomly re-routed if the probability test is passed. The mutation rate is set at 0.05.

Capacity Assignment Optimisation – the routing optimisation process yields a flow vector that specifies the link flow. The capacity assignment problem refers to the choice of capacity value \( C_i \) for link \( i \), such that network cost is minimised. This in turn is subject to the maximum average delay constraint and the assigned flow requirement (specified by the flow vector). Generally, the smaller the average delay, the higher will be the required link capacities. This is therefore a multiobjective optimisation problem for which a pareto-optimal set exists between the average delay and the required capacity or connection cost. To identify the pareto-optimal set, a multiobjective approach [Fon95] is adopted, where delay is not a constraint but rather another objective function for minimisation. The linkage chromosome structure is \( C = \{ C_1, C_2, \ldots, C_b \} \) where \( b \) is the number of links. The GA operators for capacity optimisation are one-point crossover and random mutation. For continuous capacity assignment problem, \( C_k \) is simply the capacity assigned to link \( k \). For a discrete capacity assignment case, each element of \( C_k \) is an integer array of dimension \( m \), which is the number of capacity types, i.e. \( C_k = \{ C_{k1}, C_{k2}, \ldots, C_{ki}, \ldots, C_{km} \} \) where \( C_{ki} \) is the number of capacity of type \( i \) used by link \( k \). combinations
of different capacity types are allowed on each link. The different capacity types used are [Man97]:

- 6 Mbps costs 1 unit per km
- 45 Mbps costs 4 units per km
- 150 Mbps costs 9 units per km

The above channels are full duplex. Thus, for example, at a cost of 1 unit per km, 6Mbps capacity is available in each direction. By observing the price values more carefully, the strong economy of scale with respect to capacity becomes evident. The cost goes up much more slowly than capacity (this is the usual real case), a fact that has a profound effect on the network design problem. Therefore, it is more cost-effective to:

- use a single 45-Mbps line rather than four 6-Mbps lines
- use a single 150-Mbps line rather than three 45-Mbps lines
- use a single 150-Mbps line rather than two 45-Mbps lines plus one 6-Mbps line

Similar relationships appear to be in effect in most network design problems. The above conclusion are very useful in the capacity assignments process since it facilitates the construction of a table of available cost effective capacity choices. This table is given below Table 5-6. The last element of this table yields a capacity value which is greater than the sum of all the requirements. Mathematically, the above conditions translate into

- \( 0 \leq C_{kl} \leq 3 \)
- \( 0 \leq C_{k2} \leq 2 \)
- if \( C_{k2} = 2 \) then \( C_{kl} = 0 \).

Where \( C_{kl} \), \( C_{k2} \) and \( C_{k3} \) are the number of 6-, 45- and 150- Mbps links. Using these conditions reduces the size of the search domain for \( C_{kl} \). In [Man97], fitness evaluation of each individual in each generation of the main GA (CFA problem) is conducted using GA techniques as well. In each generation of the main GA, for each individual, a second level GA is set to optimise routing for the corresponding topology. The individuals of the population of each second level GA represent potential routing policies. Finally, a third level GA is used for each individual of the latter population in each generation, to optimise capacity assignment for a given routing policy for a given topology, in terms of minimum total cost and packet delay values (multiobjective approach). The second and third level GAs yield an optimal capacity assignment.
for an optimal routing policy which implies a minimum cost and packet delay for a given topology of the network. Thus, the fitness value for each individual of the main GA can be assigned. The Genetic Algorithm strategy of this scheme uses a second level GA as a routine within the main GA, to be executed in each generation for each individual. Similarly, each second level GA uses a third level GA as a routine. The above imply a huge computational complexity of the whole optimisation process. In the following sections the GA strategy of [Man97] is referred to as the “Man-GA” method.

5.3.3 The Proposed Approach

One of the main disadvantages of the existing methods for the solution of the complete network design problem (topology, routing and capacity assignment optimisation) is the fact that they start from a single arbitrary topology. Then, they try to optimise the properties of the network of this topology, allowing for gradual changes (e.g. links are added or dropped), so that the network becomes sometime one which relaxes the constraints of the problem with the hope that the topology that will arise is a cost effective one. The GA-based technique proposed in this section, offers the advantage that the optimisation process is not restricted to a single starting topology and the transformations that can be produced from it. Instead during the optimisation process, a population of candidate topologies are processed. Thus, it is very probable that a solution can be reached which apart from satisfying the requirements will constitute if not the global optimal point of the search domain, a high quality local optimal one, in terms of cost. The strategy employed is described in the remaining part of this section. The main structure of the proposed optimisation process is based on a main GA engine which processes a population of candidate topologies i.e. individuals, for the network design problem. The initial population is randomly generated. Afterwards, in each generation, the individuals are reproduced according to the genetic operations (selection, crossover, mutation) and high-fitness ones are expected to
Randomly generated initial population

Two-connectivity testing / corrections applied

Main GA population: candidate topologies

Fitness evaluation via conventional CFA technique / routing and capacity assignment optimisation for each candidate topology / cost and delay assessment

Termination criteria met?

Yes

Final product (minimum cost Network)

No

Selection

Crossover

Mutation

New population

Table 5-6: Table of available cost effective capacity choices

<table>
<thead>
<tr>
<th>Set Number</th>
<th>Capacity (Mbps)</th>
<th>Cost (units)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>12.0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>18.0</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>45.0</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>51.0</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>57.0</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>63.0</td>
<td>7.0</td>
</tr>
<tr>
<td>8</td>
<td>90.0</td>
<td>8.0</td>
</tr>
<tr>
<td>9</td>
<td>150.0</td>
<td>9.0</td>
</tr>
<tr>
<td>10</td>
<td>155.0</td>
<td>10.0</td>
</tr>
<tr>
<td>11</td>
<td>162.0</td>
<td>11.0</td>
</tr>
<tr>
<td>12</td>
<td>168.0</td>
<td>12.0</td>
</tr>
<tr>
<td>13</td>
<td>195.0</td>
<td>13.0</td>
</tr>
<tr>
<td>14</td>
<td>201.0</td>
<td>14.0</td>
</tr>
<tr>
<td>15</td>
<td>207.0</td>
<td>15.0</td>
</tr>
<tr>
<td>16</td>
<td>213.0</td>
<td>15.0</td>
</tr>
<tr>
<td>17</td>
<td>240.0</td>
<td>17.0</td>
</tr>
<tr>
<td>18</td>
<td>300.0</td>
<td>18.0</td>
</tr>
<tr>
<td>19</td>
<td>305.0</td>
<td>19.0</td>
</tr>
<tr>
<td>20</td>
<td>312.0</td>
<td>20.0</td>
</tr>
<tr>
<td>21</td>
<td>318.0</td>
<td>21.0</td>
</tr>
<tr>
<td>22</td>
<td>345.0</td>
<td>22.0</td>
</tr>
</tbody>
</table>

Figure 5-4: The optimisation strategy of the proposed technique

appear among the population in a later generation. The fittest individual that finally emerges constitutes a proposed solution to the problem. The key point in this GA approach is the way that the fitness of a topology is assessed. The assessment is based on the cost and delay values that the network topology exhibits. In order to evaluate an individual, it is judged based on "the best it can give" in terms of delay and cost. Thus, both the routing policy and capacity assignment have to be optimised in advance. Therefore, in each generation, for each individual in the fitness evaluation cycle, we have to solve an individual flow and capacity assignment (CFA) network design optimisation problem.

To achieve this, an efficient method based on the Bertsekas-Gallager (BG) Algorithm ([Ber87], [Ker93]), which is fully detailed in a later section is employed. What remains is the reliability constraint, which is catered for by employing a procedure that check for biconnectivity of each individual in each generation and applying the necessary transformations to any topology found to violate it. The proposed network optimisation approach can be viewed as a hybrid technique that combines GA and conventional techniques, the latter as a fitness evaluation routine incorporated within the former. The main idea is that in order to judge the quality of a topology in the main GA population, routing and capacity need to be optimised in advance. Figure 5-4
demonstrates the logic diagram of the proposed algorithm used for the complete network design problem.

### 5.3.3.1 Main GA Process - Topology optimisation

The main GA processes a population of candidate topologies in each generation. The size of the population remains constant from generation to generation. The proposed scheme uses non-overlapping populations. The characteristics of this GA are presented in this section.

#### String Representation of Individuals (Topologies)

One of the most important features of GA applications, is the way in which the potential solutions to the problem are encoded into chromosome structures (strings). In the case of the main GA process the adjacency matrix representation of [Ker93] and [Man97] is used to code the topology of the network into a binary string. Thus, the topological chromosome becomes a binary string with N(N-1)/2 elements (where N is the number of nodes). The information contained in such a chromosome defines which nodes of the network are straight connected, which defines a specific topology. In order to demonstrate the simplicity of the above topology representation technique, a simple example is presented. Suppose that there are 5 nodes to connect. The topology shown in Figure 5-5 is represented as a symmetric matrix as shown in Table 5-7.

![Topology Diagram](image)

**Figure 5-5: Example of topology representation**

<table>
<thead>
<tr>
<th>Nodes</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>X</td>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

The equivalent bit string is: 1110100011.

For a network of 10 nodes, the strings representing candidate topologies have a length of 10(10-1)/2 = 45bits.
Chapter 5. A Hybrid GA-Heuristic Approach for Packet Network Design

Fitness Evaluation of Individuals

As mentioned in section 5.4.1, the fitness evaluation of each individual is conducted with respect to cost and delay parameters. These parameters are received from the fitness evaluation routine which is described in detail in the next section. Thus, the optimisation problem appears as a multiobjective one, since there are two network properties to optimise (i.e. minimise). These objectives are conflicting since in general, decrease in packet delays implies an increase in network’s cost (greater capacity values) and vice versa. This is the case where a Pareto-Optimal set of solutions exists, where none is dominated by another solution. The Pareto-Optimal set shape (as in baseline appears with respect to each solution’s properties) for the problem is shown in Figure 5-6. It represents a set of infinite solutions. If the proposed algorithm were to search for the solution to this multiobjective problem, then the proposed GA approach would be a multiobjective one. However in the case of the proposed technique, things are more simple: delay is considered a constraint. The fitness evaluation routine is set to return for each topology a minimum cost required to achieve delay below the stated constraint value \( T_{\text{max}} \). During this routine if a topology appears to exhibit smaller packet delay than \( T_{\text{max}} \), it is possible to trade-off cost for delay, and that’s what has been done. The result is that all individuals in the main GA appear to have a packet delay value just below \( T_{\text{max}} \). Thus, the focus returns to a single objective (i.e. cost minimisation) during the GA process. Going back to the Pareto-Optimal set concept, it is clear that the region of interest defines the border between the constraint violating and non-violating solutions. (shaded region in Figure 5-6). Optimal solution(s) sought, reside within this region of the set. However, sub-optimal solutions in terms of cost and delay close to the optimal points of the search domain, are also acceptable.

![Figure 5-6: Network Optimisation Problem; Pareto-Optimal set of solutions](image)

5.3.3.2 Mapping Objective Values to Fitness Form

Sigma Truncation, Fitness Scaling
Let the cost and delay values for each individual \( i \) be denoted by \( c[i] \) and \( d[i] \), respectively (\( i = 1 \ldots P_{\text{size}} \) where \( P_{\text{size}} \) denotes the population size). First the cost is processed, and then the delay property, to evaluate the number of trials (the number of times selected to mate) each individual should expect to be assigned for the formation of the next generation, with respect to its cost and delay performance.

**Cost Performance evaluation**

Let \( c_{\text{max}} \) and \( c_{\text{min}} \) denote the maximum and minimum cost value encountered across the current population respectively. As the first step the fitness of each individual is computed as

\[
\text{fit}1[i] = c_{\text{max}} + (c_{\text{min}} - c[i])
\]  

(5-7)

Using this transformation a positive figure of merit for each individual (the higher its value, the fitter the individual) is obtained, as the GA concept dictates. Afterwards, sigma-truncation and linear scaling are applied successively for the reasons stated in Appendix D:

**Sigma truncation:**

\[
\text{fit}2[i] = \text{fit}1[i] - (\text{fit}1_{\text{avg}} - \Lambda_{\text{sig}} \cdot \sigma)
\]  

(5-8)

\[
= 0 \quad \text{if } \text{fit}2[i] < 0 ; \quad i = 1 \ldots P_{\text{size}}
\]

where:

- \( \text{fit}1_{\text{avg}} \) is the mean value of \( \text{fit}1 \) in the current population
- \( \sigma \) is the standard deviation of \( \text{fit}1 \) in the current population
- \( \Lambda_{\text{sig}} \) is a small integer (we set \( \Lambda_{\text{sig}} = 2 \))

**Linear Scaling:**

\[
\text{fit}3[i] = a \cdot \text{fit}2[i] + b \quad ; \quad i = 1 \ldots P_{\text{size}}
\]  

(5-9)

where the scaling coefficients are:

\[
a_{\text{scaling}} = \frac{\text{fit}2_{\text{avg}} \cdot (\Lambda_{\text{scaling}} - 1)}{(\text{fit}2_{\text{max}} - \text{fit}2_{\text{avg}})} \quad (5-10)
\]

\[
b_{\text{scaling}} = \text{fit}2_{\text{avg}} \cdot (1 - a_{\text{scaling}})
\]

or if \((a_{\text{scaling}} \cdot \text{fit}2_{\text{min}} + b_{\text{scaling}} < 0)\)

\[
a_{\text{scaling}} = \frac{\text{fit}2_{\text{avg}}}{(\text{fit}2_{\text{avg}} - \text{fit}2_{\text{min}})} \quad (5-11)
\]

\[
b_{\text{scaling}} = \text{fit}2_{\text{avg}} \cdot (1 - a_{\text{scaling}})
\]
where $\text{fit2}_{\text{avg}}$, $\text{fit2}_{\text{min}}$ and $\text{fit2}_{\text{max}}$ are the mean, minimum and maximum fit2 values respectively, in the current population.

- $\Lambda_{\text{scaling}}$ is an integer defining the maximum number of offsprings that the fittest individuals will receive (we set $\Lambda_{\text{scaling}}=2$).

The expected number of trials (i.e. $\text{cost}\_\text{exp}\_\text{trials}[i]$) according to the cost performance for an individual is computed as

$$\text{cost}\_\text{exp}\_\text{trials}[i]=\frac{\text{fit3}[i]}{\text{fit3}_{\text{avg}}}; \quad i = 1 \ldots P_{\text{size}}$$  \hfill (5-12)

where $\text{fit3}_{\text{avg}}$ is the mean fit3 value in the current population, and the $\text{cost}\_\text{exp}\_\text{trials}[i]$ is for each individual a decimal number: $0 \leq \text{cost}\_\text{exp}\_\text{trials}[i] \leq \Lambda_{\text{scaling}}$

**Delay Performance evaluation**

Repeating the above procedure for processing the $\partial[i]$ values, yields the expected number of trials ($\text{delay}\_\text{exp}\_\text{trials}[i]$) an individual should receive according to its delay performance.

**Overall Performance evaluation**

The overall (cost and delay performances taken in account) expected number of trials ($\text{exp}\_\text{trials}[i]$), an individual $i$ receives, is computed as the weighted sum of the values $\text{cost}\_\text{exp}\_\text{trials}[i]$ and $\text{delay}\_\text{exp}\_\text{trials}[i]$. Specifically:

$$\text{exp}\_\text{trials}[i] = 0.9 \cdot \text{cost}\_\text{exp}\_\text{trials}[i] + 0.1 \cdot \text{delay}\_\text{exp}\_\text{trials}[i]; \quad i = 1 \ldots P_{\text{size}}$$  \hfill (5-13)

It should be noted that the delay performance of individuals is not completely ignored, as would be the case in an exclusively single objective (cost) optimisation process. As a bonus opportunity to individuals with delay values lower than the average (close to/and or below the constraint value), such individuals are permitted to survive and produce successors. However, the process is mainly cost minimisation oriented since the weighting factor of cost (0.9) is significantly greater than that for the delay (0.1). By changing these values, the orientation of the optimisation procedure can be changed. The computed values i.e. $\text{exp}\_\text{trials}[i]$, are then assigned to the selection scheme so as to start reproduction.

### 5.3.3.3 Selection Scheme

The Stochastic Remainder Sampling Without Replacement selection scheme described in Appendix D is used for the reasons stated.
5.3.3.4 Crossover
Both single and multiple crossover operations, with various values of the crossover probability $p_{crossover}$ have been applied.

5.3.3.5 Mutation
Single bit mutation is employed and the optimisation process with various values of the mutation probability $p_{mutation}$ is tested.

5.3.3.6 Fitness Evaluation
Routing Policy and Capacity Assignment Optimisation
As mentioned previously, the proposed GA-based process employs an efficient heuristic deterministic technique to solve the network flow and capacity assignment (CFA) problem that implies the fitness evaluation of each individual topology of the population of the main GA. The methodology employed is inspired by the Bottom Up Algorithm in [Ger76].

Having chosen as a starting point, a set of capacities assigned to the existing links that the topology at hand defines, the routing problem (FA: flow assignment problem) that arises needs to be solved. For this problem, as mentioned in section 5.1, a number of optimal methods exists, but not all are effective in practice. We choose to employ the most effective one: the Bertsekas-Gallager Algorithm ([Ber87], [Ker93]), which is detailed in Appendix C. To do this we have to start from a feasible flow vector (one where in any link, the flow is not exceeding link capacity). The way that we choose both the starting capacity assignment policy and the starting feasible flow vector, will be explain later. The main concern now, is to present how the process “moves” so as to achieve relaxation of the delay constraint, with minimum network cost. When the above referred routing optimisation process has converged, and thus, packet delays cannot be furthered decreased for the specific network structure, the delay value achieved are checked to ensure that it is below the constraint value. If not, the most utilised link is identified and its capacity is increased (by assigning to it the next permissible capacity value in order of ascending cost, according to Table 5-6. Routing is then re-optimised using the Bertsekas-Gallager algorithm, starting from the current flow vector. If the packet delay, at the end of a routing optimisation cycle, is found to be below the constraint value (sooner or later this happens) we attempt to reduce the cost of the network by applying the most cost effective link capacity reductions. To do so, the link whose capacity decrease to the next lower value (in order of descending cost according to Table 5-6 leading to maximum cost savings, is identified. This maximum cost
saving link capacity reduction is applied only if it does not incur a delay constraint violation. Otherwise, the next (in order of descending cost efficiency) link capacity reduction is identified and applied, and so on, until a state is reached where no link capacity reduction is allowed. The values of cost and packet delay at this state, are the return values of the CFA optimisation procedure. What is achieved by employing this strategy is relaxation of the delay constraint for any network topology, while taking care not to waste capacity and thus, increase cost.

We now refer to the mechanism by which the flow and capacity assignment starting points are chosen. This is a crucial point of the algorithm since this starting point will orient the searching process which follows. We follow a simple and common sense deployment strategy which is described next. At first the shortest path for each pair of nodes (using the Dijkstra's algorithm or any equivalent) in the mesh topology is identified - shortest in terms of number of hops or physical length to break ties - where necessary. Requirements are loaded on the shortest paths computed. The starting flow vector is now defined. Then to each link, \( i \), is allocated the least expensive capacity value available, \( C_i \) (according to Table 5-6) which renders the flow vector feasible; \( f_i \), the flow in link \( i \), has to be: \( f_i < C_i \). Thus, we obtain a feasible flow vector and a non-expensive starting capacity assignment policy. The optimisation process is now ready to begin.

The CFA problem solution algorithm that is described, is not optimal, in the sense that it does not guarantee finding the globally optimal solution (no method guaranteeing the globally optimal for this problem exists) but it always obtains a high quality local optimal solution, which can be even the global optimal one. Moreover, having been incorporated into the proposed algorithm as a fitness evaluation routine, it offers a very credible metric for the fitness of a topology. Even if the best cost and delay performance of the topology in question are not always reached, the algorithm definitely identifies whether a particular topology deserves to receive the interest of the GA process and spread its genes among the population of the next generations.

**5.3.3.7 Reliability Considerations: Biconnectivity**

The reliability constraint in the network design problem, is interpreted as a requirement for biconnectivity of any candidate topology in each generation of the main GA. This means that any topology of the starting randomly generated population has to be filtered and refined by a procedure which either verifies inherent biconnectivity, or applies transformations so as to obtain a biconnected topology. Similarly, this check has to be applied to each individual of the new population in each generation, since the offspring of two biconnected networks are not
necessarily biconnected ones. Besides, the mutation operation can cause a violation of the biconnectivity constraint for a “healthy” topology. This procedure appears as a routine in the network design optimisation algorithm (prior to the fitness evaluation cycle: Figure 5.5), which is executed for each individual in each generation. This increases the order of complexity of the whole algorithm but it is necessary, since it guarantees that any solution that will arise at the end, will satisfy the reliability requirement.

The refining procedure is presented in detail in Appendix B. In this section only the main features are described. The routine operates as a filter (Figure 5-7) with an input and an output. It receives as its input an unrefined topology (string of bits using adjacency matrix representation). This topology might not satisfy even the connectivity constraint, let alone biconnectivity. Initially, the procedure checks whether connectivity condition is met. If not, links are added so as the graph gets connected. Afterwards, a similar check for biconnectivity is conducted and links are added when articulation points are discovered, to connect the separate biconnected components of the network. Whenever in the whole process it is found that links have to be added, this is done in a cost effective way. This means only a minimum number of additional links are established each time, so as to disturb the genetic process as little as possible. The procedure produces as output, a definitely biconnected topology. The operation of this refinement procedure is depicted in Figure 5-7. This biconnectivity testing procedure removes any concerns regarding the minimum number of links a topology should include. All relevant decisions are taken by the optimisation process automatically. Highly connected topologies are definitely not the most cost effective solutions and it would be beneficial for the GA process, if they were not allowed to evolve significantly. Such topologies slow down the algorithm without any profit, and degrade performance. The connectivity testing procedure is based on a Depth First Search (DFS) and the formation of DF (Depth First) trees among the given graph. The biconnectivity testing procedure employed is a modification of the Lowpoint Algorithm [Aho74]. If k-connectivity with k>2 were the reliability constraint, a similar test procedure could be applied, by implementing algorithms proposed in [Ber87] and [Ker93] for example.
5.3.3.8 Implementation of the Proposed Technique

The technique presented in the section 5.4 was implemented and set to solve the stated network design problem. The results are very encouraging. The network solutions obtained show significantly lower cost than the solutions provided by other algorithms in [Man97] (i.e. in Figure 5-14 and Figure 5-15), while the packet delay still remains at the same levels and definitely below the constraint value. The key element for the success of the proposed method, was the fitness evaluation concept (as described in section 5.3.3.6) which not only provided a credible metric of the fitness of each candidate topology, but also produced each time, a very good estimate of how routing and capacity should be optimised so as to achieve relaxation of the delay constraint with minimum cost. The main GA process, proved to be very effective in constructing high-quality solutions by bringing together meaningful “building blocks” collected from individuals that although far from a good solution by themselves, possessed some good features. The proposed GA-based approach as any other GA-based method, is to a great extent a stochastic procedure. Also, since the search domain of the problem is huge and an infinite number of local optimal solutions exist, one of which is the global optimum point but no one can tell for sure which one, one should not expect from the GA technique to propose exactly the same solution whenever it is executed for the same problem. Indeed, a number of very satisfactory solutions were derived executing the algorithm several times. Sometimes the solution proposed was identical with a previous one, and generally all solutions had some common features concerning their topology (i.e. some links appeared to be always present in the GA solution and some others never). All final solutions had more or less in common the most important feature of a topology from our point of view: cost and delay performance. In the following sections some representative network solutions (including the best found), the solutions obtained using other methods (comparison) and plots demonstrating quantitatively the process of the GA from generation to generation, are presented.
5.4 Numerical Results

In this section a collection of results obtained for the network design problem using the proposed GA technique are presented and discussed. Network topologies obtained based on five different solutions are shown in the figures below. The link capacity assignment and flow information is given for each case in corresponding tables (Table 5-8 and Table 5-12). The best solution (lowest cost) is shown in Solution 1 (Figure 5-8). In all five solutions we observe significant similarities. Some links appear to be established/common in all solutions, while some other linkages appear only once (per particular solution).

![Figure 5-8: Network Solution.1](image1)
![Figure 5-9: Network Solution.2](image2)
![Figure 5-10: Network Solution.3](image3)
![Figure 5-11: Network Solution.4](image4)
Chapter 5. A Hybrid GA-Heuristic Approach for Packet Network Design

In all cases, the average delay constraint is satisfied (packet delay ranges from 0.056 sec to 0.097 sec), while the cost ranges from 46,620 units to 48,450 units. Solution 5 (Figure 5-12) exhibits the lowest average packet delay value (0.056 sec) but is not considered the best one, since our objective was to minimise cost subject to the delay constraint $T < 0.1$ sec. The above solutions were obtained for the best GA parameters settings, given in the section 5.5.3. In Table 5-13, the routing policy for the best solution Solution 1 (Figure 5-8) is given.

![Network Solution 5](image)

Figure 5-12: Network Solution 5

In all cases, the average delay constraint is satisfied (packet delay ranges from 0.056 sec to 0.097 sec), while the cost ranges from 46,620 units to 48,450 units. Solution 5 (Figure 5-12) exhibits the lowest average packet delay value (0.056 sec) but is not considered the best one, since our objective was to minimise cost subject to the delay constraint $T < 0.1$ sec. The above solutions were obtained for the best GA parameters settings, given in the section 5.5.3. In Table 5-13, the routing policy for the best solution Solution 1 (Figure 5-8) is given.

**Table 5-8: Capacity Flow Assignment Information for Solution 1**

<table>
<thead>
<tr>
<th>Established links</th>
<th>Capacity (Mbps)</th>
<th>Flow (Mbps)</th>
<th>Utilisation</th>
<th>Cost (units)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B,X</td>
<td>45.000</td>
<td>37.331</td>
<td>0.830</td>
<td>3840.0</td>
</tr>
<tr>
<td>B,T</td>
<td>150.000</td>
<td>120.331</td>
<td>0.802</td>
<td>1440.0</td>
</tr>
<tr>
<td>S,W</td>
<td>150.000</td>
<td>113.737</td>
<td>0.758</td>
<td>6480.0</td>
</tr>
<tr>
<td>S,Ha</td>
<td>6.000</td>
<td>2.737</td>
<td>0.456</td>
<td>2240.0</td>
</tr>
<tr>
<td>G,H</td>
<td>150.000</td>
<td>101.994</td>
<td>0.680</td>
<td>2160.0</td>
</tr>
<tr>
<td>G,W</td>
<td>150.000</td>
<td>123.916</td>
<td>0.826</td>
<td>7560.0</td>
</tr>
<tr>
<td>H,K</td>
<td>45.000</td>
<td>34.669</td>
<td>0.770</td>
<td>5600.0</td>
</tr>
<tr>
<td>W,T</td>
<td>150.000</td>
<td>127.954</td>
<td>0.851</td>
<td>9360.0</td>
</tr>
<tr>
<td>C,X</td>
<td>45.000</td>
<td>31.585</td>
<td>0.702</td>
<td>2560.0</td>
</tr>
<tr>
<td>C,K</td>
<td>45.000</td>
<td>29.669</td>
<td>0.659</td>
<td>2720.0</td>
</tr>
<tr>
<td>Ha,T</td>
<td>12.000</td>
<td>11.263</td>
<td>0.939</td>
<td>2300.0</td>
</tr>
</tbody>
</table>

**Table 5-9: Capacity Flow Assignment Information for Solution 2**

<table>
<thead>
<tr>
<th>Established links</th>
<th>Capacity (Mbps)</th>
<th>Flow (Mbps)</th>
<th>Utilisation</th>
<th>Cost (units)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B,X</td>
<td>45.000</td>
<td>31.904</td>
<td>0.709</td>
<td>3840.0</td>
</tr>
<tr>
<td>B,Ha</td>
<td>6.000</td>
<td>4.508</td>
<td>0.751</td>
<td>1120.0</td>
</tr>
<tr>
<td>B,T</td>
<td>150.000</td>
<td>103.775</td>
<td>0.692</td>
<td>1440.0</td>
</tr>
<tr>
<td>S,W</td>
<td>90.000</td>
<td>74.246</td>
<td>0.825</td>
<td>5760.0</td>
</tr>
<tr>
<td>S,T</td>
<td>45.000</td>
<td>38.754</td>
<td>0.861</td>
<td>4480.0</td>
</tr>
<tr>
<td>G,H</td>
<td>90.000</td>
<td>84.102</td>
<td>0.934</td>
<td>1920.0</td>
</tr>
<tr>
<td>G,W</td>
<td>150.000</td>
<td>125.102</td>
<td>0.834</td>
<td>7560.0</td>
</tr>
<tr>
<td>H,K</td>
<td>18.000</td>
<td>13.192</td>
<td>0.733</td>
<td>4200.0</td>
</tr>
<tr>
<td>W,X</td>
<td>45.000</td>
<td>35.229</td>
<td>0.783</td>
<td>2720.0</td>
</tr>
<tr>
<td>W,T</td>
<td>90.000</td>
<td>78.519</td>
<td>0.872</td>
<td>8320.0</td>
</tr>
</tbody>
</table>
### Table 5-10: Capacity Flow Assignment Information for Solution 3

<table>
<thead>
<tr>
<th>Established links</th>
<th>Capacity (Mbps)</th>
<th>Flow (Mbps)</th>
<th>Utilisation</th>
<th>Cost (units)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B,S</td>
<td>150.000</td>
<td>125.270</td>
<td>0.842</td>
<td>10800.0</td>
</tr>
<tr>
<td>B,X</td>
<td>45.000</td>
<td>35.788</td>
<td>0.818</td>
<td>3840.0</td>
</tr>
<tr>
<td>B,Ha</td>
<td>6.000</td>
<td>4.635</td>
<td>0.773</td>
<td>1120.0</td>
</tr>
<tr>
<td>B,T</td>
<td>90.000</td>
<td>59.208</td>
<td>0.658</td>
<td>1280.0</td>
</tr>
<tr>
<td>S,W</td>
<td>150.000</td>
<td>139.212</td>
<td>0.928</td>
<td>6480.0</td>
</tr>
<tr>
<td>S,Ha</td>
<td>6.000</td>
<td>4.573</td>
<td>0.762</td>
<td>2240.0</td>
</tr>
<tr>
<td>G,H</td>
<td>150.000</td>
<td>102.212</td>
<td>0.681</td>
<td>2160.0</td>
</tr>
<tr>
<td>G,W</td>
<td>150.000</td>
<td>123.212</td>
<td>0.821</td>
<td>7560.0</td>
</tr>
<tr>
<td>H,K</td>
<td>45.000</td>
<td>35.302</td>
<td>0.784</td>
<td>5600.0</td>
</tr>
<tr>
<td>C,X</td>
<td>45.000</td>
<td>35.788</td>
<td>0.818</td>
<td>2560.0</td>
</tr>
<tr>
<td>C,K</td>
<td>45.000</td>
<td>31.788</td>
<td>0.706</td>
<td>2720.0</td>
</tr>
<tr>
<td>Ha,T</td>
<td>6.000</td>
<td>4.792</td>
<td>0.799</td>
<td>1150.0</td>
</tr>
</tbody>
</table>

### Table 5-11: Capacity Flow Assignment Information for Solution 4

<table>
<thead>
<tr>
<th>Established links</th>
<th>Capacity (Mbps)</th>
<th>Flow (Mbps)</th>
<th>Utilisation</th>
<th>Cost (units)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B,X</td>
<td>45.000</td>
<td>25.986</td>
<td>0.577</td>
<td>3840.0</td>
</tr>
<tr>
<td>B,Ha</td>
<td>6.000</td>
<td>4.457</td>
<td>0.743</td>
<td>1120.0</td>
</tr>
<tr>
<td>B,T</td>
<td>150.000</td>
<td>104.557</td>
<td>0.697</td>
<td>1440.0</td>
</tr>
<tr>
<td>S,H</td>
<td>45.000</td>
<td>39.901</td>
<td>0.887</td>
<td>5000.0</td>
</tr>
<tr>
<td>S,W</td>
<td>90.000</td>
<td>77.345</td>
<td>0.859</td>
<td>5760.0</td>
</tr>
<tr>
<td>G,H</td>
<td>90.000</td>
<td>80.655</td>
<td>0.896</td>
<td>1920.0</td>
</tr>
<tr>
<td>G,W</td>
<td>90.000</td>
<td>81.726</td>
<td>0.908</td>
<td>6720.0</td>
</tr>
<tr>
<td>G,C</td>
<td>12.000</td>
<td>8.994</td>
<td>0.750</td>
<td>2480.0</td>
</tr>
<tr>
<td>G,K</td>
<td>12.000</td>
<td>8.621</td>
<td>0.718</td>
<td>2320.0</td>
</tr>
<tr>
<td>W,X</td>
<td>45.000</td>
<td>29.441</td>
<td>0.654</td>
<td>2720.0</td>
</tr>
<tr>
<td>W,T</td>
<td>150.000</td>
<td>119.014</td>
<td>0.793</td>
<td>9360.0</td>
</tr>
<tr>
<td>C,X</td>
<td>45.000</td>
<td>31.427</td>
<td>0.698</td>
<td>2560.0</td>
</tr>
<tr>
<td>C,K</td>
<td>45.000</td>
<td>4.379</td>
<td>0.730</td>
<td>680.0</td>
</tr>
<tr>
<td>Ha,T</td>
<td>12.000</td>
<td>9.543</td>
<td>0.795</td>
<td>2300.0</td>
</tr>
</tbody>
</table>

### Table 5-12: Capacity Flow Assignment Information for Solution 5

<table>
<thead>
<tr>
<th>Established links</th>
<th>Capacity (Mbps)</th>
<th>Flow (Mbps)</th>
<th>Utilisation</th>
<th>Cost (units)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B,X</td>
<td>45.000</td>
<td>37.362</td>
<td>0.830</td>
<td>3840.0</td>
</tr>
<tr>
<td>B,T</td>
<td>150.000</td>
<td>120.362</td>
<td>0.802</td>
<td>1440.0</td>
</tr>
<tr>
<td>S,W</td>
<td>150.000</td>
<td>113.608</td>
<td>0.757</td>
<td>6480.0</td>
</tr>
<tr>
<td>S,Ha</td>
<td>6.000</td>
<td>2.608</td>
<td>0.435</td>
<td>2240.0</td>
</tr>
<tr>
<td>G,H</td>
<td>150.000</td>
<td>101.921</td>
<td>0.679</td>
<td>2160.0</td>
</tr>
<tr>
<td>G,W</td>
<td>150.000</td>
<td>123.710</td>
<td>0.825</td>
<td>7560.0</td>
</tr>
<tr>
<td>H,K</td>
<td>45.000</td>
<td>34.638</td>
<td>0.770</td>
<td>5600.0</td>
</tr>
<tr>
<td>W,Ha</td>
<td>6.000</td>
<td>2.917</td>
<td>0.486</td>
<td>2190.0</td>
</tr>
<tr>
<td>W,T</td>
<td>150.000</td>
<td>127.888</td>
<td>0.853</td>
<td>9360.0</td>
</tr>
<tr>
<td>C,X</td>
<td>45.000</td>
<td>31.731</td>
<td>0.705</td>
<td>2560.0</td>
</tr>
<tr>
<td>C,K</td>
<td>45.000</td>
<td>29.638</td>
<td>0.659</td>
<td>2720.0</td>
</tr>
<tr>
<td>Ha,T</td>
<td>12.000</td>
<td>8.474</td>
<td>0.706</td>
<td>2300.0</td>
</tr>
</tbody>
</table>
### Table 5-13: Routing information for the best solution (solution 1)

<table>
<thead>
<tr>
<th>Endpoints</th>
<th>Nodes in Route (path description)</th>
<th>Requirement portion served</th>
</tr>
</thead>
<tbody>
<tr>
<td>B,S</td>
<td>Path1:[B][T][W][S]</td>
<td>100%</td>
</tr>
<tr>
<td>B,G</td>
<td>Path1:[B][T][W][G]</td>
<td>100%</td>
</tr>
<tr>
<td>B,H</td>
<td>Path1:[B][T][W][G][H]</td>
<td>100%</td>
</tr>
<tr>
<td>B,W</td>
<td>Path1:[B][T][W]</td>
<td>100%</td>
</tr>
<tr>
<td>B,C</td>
<td>Path1:[B][X][C]</td>
<td>100%</td>
</tr>
<tr>
<td>B,X</td>
<td>Path1:[B][X]</td>
<td>100%</td>
</tr>
<tr>
<td>B,K</td>
<td>Path1:[B][X][C][K]</td>
<td>100%</td>
</tr>
<tr>
<td>B,Ha</td>
<td>Path1:[B][T][Ha]</td>
<td>100%</td>
</tr>
<tr>
<td>B,T</td>
<td>Path1:[B][T]</td>
<td>100%</td>
</tr>
<tr>
<td>S,G</td>
<td>Path1:[S][W][G]</td>
<td>100%</td>
</tr>
<tr>
<td>S,H</td>
<td>Path1:[S][W][G][H]</td>
<td>100%</td>
</tr>
<tr>
<td>S,W</td>
<td>Path1:[S][W]</td>
<td>100%</td>
</tr>
<tr>
<td>S,C</td>
<td>Path1:[S][W][T][B][X][C]</td>
<td>68.46%</td>
</tr>
<tr>
<td>S,X</td>
<td>Path1:[S][W][T][B][X]</td>
<td>100%</td>
</tr>
<tr>
<td>S,K</td>
<td>Path1:[S][W][G][H][K]</td>
<td>100%</td>
</tr>
<tr>
<td>S,Ha</td>
<td>Path1:[S][Ha]</td>
<td>100%</td>
</tr>
<tr>
<td>S,T</td>
<td>Path1:[S][W][T]</td>
<td>100%</td>
</tr>
<tr>
<td>G,H</td>
<td>Path1:[G][H]</td>
<td>100%</td>
</tr>
<tr>
<td>G,W</td>
<td>Path1:[G][W]</td>
<td>100%</td>
</tr>
<tr>
<td>G,C</td>
<td>Path1:[G][H][K][C]</td>
<td>100%</td>
</tr>
<tr>
<td>G,X</td>
<td>Path1:[G][H][K][C][X]</td>
<td>90.78%</td>
</tr>
<tr>
<td>G,K</td>
<td>Path1:[G][H][K]</td>
<td>100%</td>
</tr>
<tr>
<td>G,Ha</td>
<td>Path1:[G][W][T][Ha]</td>
<td>44.17%</td>
</tr>
<tr>
<td>H,Ha</td>
<td>Path1:[H][G][W][T][Ha]</td>
<td>55.83%</td>
</tr>
<tr>
<td>H,W</td>
<td>Path1:[H][G][W]</td>
<td>100%</td>
</tr>
<tr>
<td>H,C</td>
<td>Path1:[H][K][C]</td>
<td>100%</td>
</tr>
<tr>
<td>H,X</td>
<td>Path1:[H][K][C][X]</td>
<td>91.90%</td>
</tr>
<tr>
<td>H,K</td>
<td>Path1:[H][K]</td>
<td>100%</td>
</tr>
<tr>
<td>H,Ha</td>
<td>Path1:[H][G][W][T][Ha]</td>
<td>38.36%</td>
</tr>
<tr>
<td>H,T</td>
<td>Path1:[H][G][W][T][Ha]</td>
<td>61.64%</td>
</tr>
<tr>
<td>W,C</td>
<td>Path1:[W][T][B][X][C]</td>
<td>0.7%</td>
</tr>
<tr>
<td>W,X</td>
<td>Path1:[W][T][B][X][C]</td>
<td>99.3%</td>
</tr>
<tr>
<td>W,K</td>
<td>Path1:[W][G][H][K][C][X]</td>
<td>55.01%</td>
</tr>
<tr>
<td>W,Ha</td>
<td>Path1:[W][G][H][K]</td>
<td>100%</td>
</tr>
<tr>
<td>W,T</td>
<td>Path1:[W][T][Ha]</td>
<td>43.78%</td>
</tr>
<tr>
<td>C,X</td>
<td>Path1:[C][T][C][K]</td>
<td>100%</td>
</tr>
<tr>
<td>C,K</td>
<td>Path1:[C][T][C][K]</td>
<td>100%</td>
</tr>
<tr>
<td>C,Ha</td>
<td>Path1:[C][T][C][X][B][T][Ha]</td>
<td>100%</td>
</tr>
<tr>
<td>C,T</td>
<td>Path1:[C][T][C][X][B][T][Ha]</td>
<td>100%</td>
</tr>
<tr>
<td>X,K</td>
<td>Path1:[X][C][K]</td>
<td>100%</td>
</tr>
<tr>
<td>X,Ha</td>
<td>Path1:[X][C][K]</td>
<td>100%</td>
</tr>
<tr>
<td>X,T</td>
<td>Path1:[X][C][K]</td>
<td>100%</td>
</tr>
<tr>
<td>K,Ha</td>
<td>Path1:[K][C][X][B][T][Ha]</td>
<td>100%</td>
</tr>
<tr>
<td>K,T</td>
<td>Path1:[K][C][X][B][T][Ha]</td>
<td>100%</td>
</tr>
<tr>
<td>Ha,T</td>
<td>Path1:[Ha][T]</td>
<td>100%</td>
</tr>
</tbody>
</table>

### Table 5-14: Routing information for the MAN-GA solution

<table>
<thead>
<tr>
<th>Endpoints</th>
<th>Nodes in Route (path description)</th>
<th>Requirement portion served</th>
</tr>
</thead>
<tbody>
<tr>
<td>B,S</td>
<td>Path1:[B][T][S]</td>
<td></td>
</tr>
<tr>
<td>B,G</td>
<td>Path1:[B][T][S][G]</td>
<td></td>
</tr>
<tr>
<td>B,H</td>
<td>Path1:[B][T][S][G][H]</td>
<td></td>
</tr>
<tr>
<td>B,W</td>
<td>Path1:[B][T][S][G][H][W]</td>
<td></td>
</tr>
<tr>
<td>B,C</td>
<td>Path1:[B][T][S][G][H][W][C]</td>
<td></td>
</tr>
<tr>
<td>B,X</td>
<td>Path1:[B][T][S][G][H][W][X]</td>
<td></td>
</tr>
<tr>
<td>B,K</td>
<td>Path1:[B][T][S][G][H][W][X][K]</td>
<td></td>
</tr>
<tr>
<td>B,Ha</td>
<td>Path1:[B][T][S][G][H][W][X][K][Ha]</td>
<td></td>
</tr>
<tr>
<td>B,T</td>
<td>Path1:[B][T][S][G][H][W][X][K][Ha][T]</td>
<td></td>
</tr>
<tr>
<td>S,G</td>
<td>Path1:[S][H][G][W]</td>
<td></td>
</tr>
<tr>
<td>S,H</td>
<td>Path1:[S][H][G][W][H]</td>
<td></td>
</tr>
<tr>
<td>S,W</td>
<td>Path1:[S][H][G][W][H][W]</td>
<td></td>
</tr>
<tr>
<td>S,C</td>
<td>Path1:[S][H][G][W][H][W][C]</td>
<td></td>
</tr>
<tr>
<td>S,X</td>
<td>Path1:[S][H][G][W][H][W][X]</td>
<td></td>
</tr>
<tr>
<td>S,K</td>
<td>Path1:[S][H][G][W][H][W][X][K]</td>
<td></td>
</tr>
<tr>
<td>S,Ha</td>
<td>Path1:[S][H][G][W][H][W][X][K][Ha]</td>
<td></td>
</tr>
<tr>
<td>S,T</td>
<td>Path1:[S][H][G][W][H][W][X][K][Ha][T]</td>
<td></td>
</tr>
<tr>
<td>G,H</td>
<td>Path1:[G][H][W]</td>
<td></td>
</tr>
<tr>
<td>G,W</td>
<td>Path1:[G][H][W][H]</td>
<td></td>
</tr>
<tr>
<td>G,C</td>
<td>Path1:[G][H][W][H][C]</td>
<td></td>
</tr>
<tr>
<td>G,X</td>
<td>Path1:[G][H][W][H][C][X]</td>
<td></td>
</tr>
<tr>
<td>G,K</td>
<td>Path1:[G][H][W][H][C][X][K]</td>
<td></td>
</tr>
<tr>
<td>G,Ha</td>
<td>Path1:[G][H][W][H][C][X][K][Ha]</td>
<td></td>
</tr>
<tr>
<td>H,Ha</td>
<td>Path1:[H][G][W][H][C][X][K][Ha]</td>
<td></td>
</tr>
<tr>
<td>H,W</td>
<td>Path1:[H][G][W][H][C][X][K][Ha][T]</td>
<td></td>
</tr>
<tr>
<td>H,C</td>
<td>Path1:[H][G][W][H][C][X][K][Ha][T][C]</td>
<td></td>
</tr>
<tr>
<td>H,X</td>
<td>Path1:[H][G][W][H][C][X][K][Ha][T][C][X]</td>
<td></td>
</tr>
<tr>
<td>H,K</td>
<td>Path1:[H][G][W][H][C][X][K][Ha][T][C][X][K]</td>
<td></td>
</tr>
<tr>
<td>H,Ha</td>
<td>Path1:[H][G][W][H][C][X][K][Ha][T][C][X][K][Ha]</td>
<td></td>
</tr>
<tr>
<td>H,T</td>
<td>Path1:[H][G][W][H][C][X][K][Ha][T][C][X][K][Ha][T]</td>
<td></td>
</tr>
<tr>
<td>W,C</td>
<td>Path1:[W][B][X][C]</td>
<td></td>
</tr>
<tr>
<td>W,X</td>
<td>Path1:[W][B][X][C][K]</td>
<td></td>
</tr>
<tr>
<td>W,K</td>
<td>Path1:[W][B][X][C][K][X]</td>
<td></td>
</tr>
<tr>
<td>W,Ha</td>
<td>Path1:[W][B][X][C][K][X][K]</td>
<td></td>
</tr>
<tr>
<td>W,T</td>
<td>Path1:[W][B][X][C][K][X][K][Ha]</td>
<td></td>
</tr>
<tr>
<td>C,X</td>
<td>Path1:[C][X][C][K]</td>
<td></td>
</tr>
<tr>
<td>C,K</td>
<td>Path1:[C][X][C][K][C]</td>
<td></td>
</tr>
<tr>
<td>C,Ha</td>
<td>Path1:[C][X][C][K][C][X][B][T][Ha]</td>
<td></td>
</tr>
<tr>
<td>C,T</td>
<td>Path1:[C][X][C][K][C][X][B][T][Ha][T]</td>
<td></td>
</tr>
<tr>
<td>X,K</td>
<td>Path1:[X][C][K][C][X][B][T][Ha]</td>
<td></td>
</tr>
<tr>
<td>X,Ha</td>
<td>Path1:[X][C][K][C][X][B][T][Ha][T]</td>
<td></td>
</tr>
<tr>
<td>X,T</td>
<td>Path1:[X][C][K][C][X][B][T][Ha][T][C]</td>
<td></td>
</tr>
<tr>
<td>K,Ha</td>
<td>Path1:[K][C][X][B][T][Ha]</td>
<td></td>
</tr>
<tr>
<td>K,T</td>
<td>Path1:[K][C][X][B][T][Ha][T]</td>
<td></td>
</tr>
<tr>
<td>Ha,T</td>
<td>Path1:[K][C][X][B][T][Ha][T][C][X]</td>
<td></td>
</tr>
</tbody>
</table>
5.4.1 GA Parameter Settings and Process Performance

Population size
The GA process was tested with various settings of the population size. As shown in Figure 5-13, the process appeared to converge at about 100 generations in all cases. However, the best performance was achieved when the population size was set at 200 individuals. For smaller population sizes (100, 50) the process converged faster, but to higher network cost solutions. For larger population sizes (400 individuals) no significant improvement was observed and the process was noticeably slowed down.

![Figure 5-13: Convergence of the GA process for different sizes of the population](image)

Number of crossover points
The best performance was achieved when single point crossover was applied. With 2-point crossover, no better results were obtained in terms of cost of the proposed network solution. Moreover, the convergence of the process was slowed down without any gain. When the number of cross-over points were more than two, despite higher fitness individuals emerging relatively early, the final attained fitness was low and the solutions obtained were not satisfactory. The main reason for inefficiency of multi-point crossover seems to be disruption of useful blocks of bits within the chromosome structure, since greater number of string segments are exchanged between chromosome pairs. Thus GA couldn’t process efficiently the meaningful blocks and it’s random nature was intensified. Furthermore, ineffectiveness of multipoint crossover in large populations is confirmed. In [Sri94] it is suggested that multi-point crossover seems to operate more effectively in small populations where a few solutions could be simultaneously processed. Single point crossover turned out to be the best choice.
Crossover probability

Simulations were also carried out for identification of the impact of different crossover and mutation probabilities on the search process. Since the selection of crossover and mutation rates depend on encoding strategy and population size, simulations with different setting of crossover and mutation probabilities were conducted. It was observed that for this particular problem high settings of crossover probability provided better performance. The setting $P_{\text{crossover}} = 0.9$ turned out to be the best choice. For lower values the obtained solutions were not that satisfactory. Setting $P_{\text{crossover}} = 1$ speeded up convergence, although it resulted in low quality solutions.

Mutation

The best setting for mutation probability was determined to be $P_{\text{mutation}} = 0.001$. For lower values the obtained solutions were not satisfactory (any lost genetic information could not be recovered). For greater values the process tended to become a random walk.

In summary, the optimal performance in terms of quality of the solution and computational complexity, was achieved when the population size was set to 200 individuals and single point crossover was applied with crossover probability $P_{\text{crossover}} = 0.9$, while mutation rate was set to $P_{\text{mutation}} = 0.001$.

5.4.2 Comparison with Other Methods

The results (topologies, cost and average packet delay values) obtained using the Branch Exchange method and the Man-GA method referred to in section 5.3 are shown below (Figure 5-14, Figure 5-15). The Branch Exchange method yields a network solution which exhibits higher values for both cost and delay, when compared to all the solutions (1-5) produced based on the proposed technique. The Branch Exchange method is restricted to search for a network solution throughout the transformations that can be produced from a single arbitrarily initial topology, whilst the proposed method searches for a solution by exploring in parallel a wide region of the search domain which implies that the probability of obtaining a high performance network solution is much greater.
The MAN-GA solution is inferior to the solutions based on the proposed method. All the five produced solutions in the previous section outperform the solution MAN-GA proposes (Figure 5-15), in terms of cost performance. Concerning delay, Solution 5 (Figure 5-12) gives a lower value. Note however that the original design objective was to minimise cost subject to the delay constraint $T < 0.1$ sec. The cost and average packet delay values for all the network topologies obtained are summarised in Table 5-15. In terms of average link utilisation, solutions 1-5 exhibit utilisation at or close to 80%, which compare favourably with the 70% and 65% utilisation levels achieved by MAN-GA and the Branch exchange methods respectively. Solutions 1 and 5 also show higher overall network capacity compared to MAN-GA and Branch exchange solutions, whilst maintaining higher utilisation levels.

<table>
<thead>
<tr>
<th>Network solutions</th>
<th>Cost (units)</th>
<th>Average Packet Delay (sec)</th>
<th>Total capacity (Mbps)</th>
<th>Flow (Mbps)</th>
<th>Ave. Utilisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solution 1</td>
<td>46,620</td>
<td>0.074</td>
<td>948</td>
<td>735.2</td>
<td>77.2%</td>
</tr>
<tr>
<td>Solution 2</td>
<td>47,580</td>
<td>0.097</td>
<td>798</td>
<td>644.6</td>
<td>80%</td>
</tr>
<tr>
<td>Solution 3</td>
<td>47,510</td>
<td>0.080</td>
<td>888</td>
<td>704.8</td>
<td>79.4%</td>
</tr>
<tr>
<td>Solution 4</td>
<td>48,220</td>
<td>0.090</td>
<td>798</td>
<td>626</td>
<td>78.2%</td>
</tr>
<tr>
<td>Solution 5</td>
<td>48,450</td>
<td>0.056</td>
<td>954</td>
<td>734.8</td>
<td>77%</td>
</tr>
<tr>
<td>MAN-GA</td>
<td>50,590</td>
<td>0.064</td>
<td>894</td>
<td>622</td>
<td>69.4%</td>
</tr>
<tr>
<td>Branch X</td>
<td>55,310</td>
<td>0.099</td>
<td>834</td>
<td>547</td>
<td>65.4%</td>
</tr>
</tbody>
</table>

The superior performance of the proposed technique compared to Man-GA, in terms of the quality of solutions obtained, can be attributed to:
The Man-GA method views packet delay as a network property to minimise (two objectives: multiobjective GA approach), and thus becomes completely insensitive to the value of the delay constraint.

In the Man-GA method the excessive use of GAs throughout the optimisation process, seems to be depriving the evaluation function of the ability to assess individuals as they should be. The stochastic nature of the GA (employing evaluation function that it uses) leads to a frequent overestimation or underestimation of individuals (topologies) within the main GA population.

5.5 Conclusions

A total solution to mesh network design problem based on a genetic algorithm approach was presented. Combining the robustness of a GA with the efficiency of specialised network algorithms i.e. the BG algorithm used to solve CFA problem, a hybrid scheme was deployed. The key points for the success of the proposed method are:

- The parallel search that a GA employs throughout the solution space of the problem,
- The fitness evaluation function of individuals of the GA process achieved the fair evaluation of candidate topologies and
- an efficient management of the genetic information available in each generation – the choice of selection mechanism (Stochastic Remainder Without Sampling) and Linear scaling, both contributing to preservation of population diversity and avoidance of premature convergence.

The optimal performance in terms of quality of the solution and computational complexity, was achieved when the population size was set to 200 individuals and single point crossover was applied with crossover probability $P_{\text{crossover}} = 0.9$, while mutation rate was set to $P_{\text{mutation}} = 0.001$. Single and double point crossovers seem to preserve the "high fitness" segments of bits. In the final results, single-point crossover was applied as it was considered to be the best trade-off between maintaining useful string segments and preservation of population diversity. This has also confirmed the suggestion in [Gold89] that an important factor in convergence ability of GA is the preservation of chromosomes useful segments, which could lead to better solutions. The secondary nature of the role of mutation parameter was verified, as the final choice of very low mutation rate of 0.001, enables preservation of meaningful/useful string segments, whilst managing to introduce useful corrections in the string structure at the same time.
The objective function of a problem is the main source of and mechanism for evaluating the status of each chromosome. It takes the chromosome as input and produces an objective value as performance measure. However, since its range of values varies from problem to problem, principally to maintain uniformity over various problem domains, the objective value is rescaled to a fitness value. In the linear scaling that was implemented, the coefficients were chosen to enforce the quality of average objective value and the scaled average fitness values and caused maximum scaled fitness to be a specified multiple of the average fitness. Linear scaling allowed the best individuals to create higher number of offspring, by multiplying their fitness by a suitable coefficient. To generate good offspring, efficient parent selection mechanisms are also necessary. The Stochastic Remainder Sampling Without Replacement scheme used, is a technique that combines minimum spread beyond zero bias, with low computational complexity. Its effectiveness is related to its ability in preventing early dominance by fit individuals and the range of the possible number of trials that an individual may achieve i.e. the minimum spread.

Comparison of the results obtained by the proposed technique with other two methods reveal that solution 5 for example, results in the network costs that are 4.3% and 12.5% less than costs obtained by MAN-GA and Branch exchange solutions respectively. Note that using MAN-GA has produced a solution that costs 9% less and has two-thirds the delay of Branch exchange solution. In terms of delay, solution 5 results in 12.5% and 43.4% less delay, compared to MAN-GA and Branch exchange solutions respectively, whilst solution 1 results in 7.8% and 15.7% less cost than the above two solutions. In terms of average link utilisation, solutions 1-5 exhibit utilisation at or close to 80%, which compare favourably with the 70% and 65% utilisation levels achieved by MAN-GA and the Branch exchange methods respectively. Solutions 1 and 5 also show higher overall network capacity whilst maintaining higher utilisation levels, at lower costs than MAN-GA and Branch exchange solutions. The proposed GA-based method used for the design of communication network has thus produced solutions that cost less and have lower delay than the two other typical design methods.
Chapter 6

CONCLUSIONS AND FUTURE WORK

6 Research contributions

This thesis has addressed database signalling in GSM networks with particular emphasis on signalling costs and database related transactions, location management signalling traffic in the access network and the efficient design of packet communication networks.

In chapter 3, the operation and implementation of the basic anchoring technique and an optimisation of the basic scheme for application in GSM networks, was evaluated. Distribution of the HLR related signalling load was achieved through replacing of the expensive location registration messages between the serving VLR and the HLR by message exchanges between MSC/VLRs. In response to subscriber call and mobility patterns, the optimised anchoring scheme removes the need to always report location changes to the anchor VLR whilst balancing the costs of both location update and call routing. As the benchmark for the comparison of the results, GSM standard signalling transactions and database traffic under the same traffic model and user mobility assumptions, were used. An analytical model for the proposed optimisation procedure was introduced. It was shown that within the call-to-mobility (CMR) range considered i.e. 0.01 to 100, the location registration signalling cost of the optimised scheme can be reduced to less than half that of GSM. It was also shown that HLR related signalling costs can be reduced by as much as 25-65% in the CMR range considered, compared to GSM. Also under typical operating conditions, reductions in the total network signalling cost of about 60-25% at low to high CMR respectively, can be achieved. Where as the cost of call routing in the original anchor scheme is about 2 - 3.5 times that of GSM (at low CMR values), the optimisation procedure managed to reduce this cost to within 1.6 times that of GSM under a variety of network loading conditions.

In chapter 4, the performance of the proposed adaptive multiplayer location management technique was evaluated and compared with a number of dynamic schemes for reducing the
location update and paging control traffic, through extensive simulations. The simulations were conducted under a variety of common call and mobility profiles. The performance of selected algorithms with and without Recent Interaction Paging (RIP) was also evaluated. With RIP enabled, paging for the called subscriber is done initially in the last known subscriber location/cell, and following a time-out or no response from the paged cell, the whole LA (minus the paged cell) is then paged. This simple 2-step paging process was shown to be beneficial particularly when rate of updates and/or call arrival rates are high. Of the two dynamic schemes, \textit{Alg.2} showed better performer with FF movement, whilst \textit{Alg.3} was the better performer with MG movement. \textit{Alg.4} outperformed all schemes considered, irrespective of the type of movement considered, and it was the only scheme where LA size could be scaled down to 1 cell, when user is stationary. Recent Interaction Paging (RIP) had the most significant impact in mixed subscriber environments where majority of subscribers were slow-moving and the observed average (over all schemes considered) improvement with RIP was around 8-18%. Also, of all the schemes considered, RIP produced most gain when applied in conjunction with schemes of \textit{Alg.3} and \textit{Alg.4}, and both schemes with RIP enabled, resulted in about 30-38% signalling cost savings compared with the baseline fixed-LA scheme of GSM.

High performance gains by the scheme of \textit{Alg.3}, is attributed to the ability of the scheme to take advantage of available regional information, implicit in the calculation of the cell visiting rate and implied within the formulation of the cost function. Nevertheless, improvements only began to set in after about 20 km/hr. and at call rates below 3 calls/hr. Despite the observed significant performance gains/savings by the schemes of \textit{Alg.3} and \textit{Alg.4}, it should be noted that implementation of the scheme of \textit{Alg.3} requires the MTs to forward the identities of all the cells within the personal LA of MT to the network, during location update procedure. Inevitably, this operation results in higher resource consumption i.e. more signalling bandwidth will be required, and also the network databases will be required to maintain and update the list of cells comprising the personal LA of subscribers, on a per-user basis. On the other hand, the scheme of \textit{Alg.4}, only requires broadcast of a single additional parameter i.e. the no. of layers, and since the identities of all LAs are pre-determined at the network side, and there are no additional requirements, implementation of the scheme of \textit{Alg.4} is less complex than the scheme of \textit{Alg.3}.

Based on the computational results obtained, the distance-based strategy of \textit{Alg.5} was shown to be not as efficient as the fixed LA strategy. A higher combined cost of location update and paging is associated with the scheme of \textit{Alg.5}, since when the LA size of the distance-based and the fixed-LA strategies are the same (both have same paging cost), the distance-based strategy
incurs higher location update cost, and on the other hand, when the location update costs are the same, the location area size for the distance-based scheme is greater than that of the fixed scheme, resulting in higher paging costs. Also, Location area size could not be scaled down when the user is stationary, and the LA size could only assume specific (sub-optimal) values.

The Hybrid GA-Heuristic approach in chapter 5, was aimed at achieving a cost efficient design for partial mesh packet-switched communication networks with application to GPRS backbone architecture. Combining the robustness of a GA with the efficiency of specialised algorithms i.e. the BG algorithm used to solve CFA problem, a hybrid scheme was deployed. The key points for the success of the proposed method could be attributed to the efficient management of the genetic information available in each generation i.e. the choice of selection mechanism (Stochastic Remainder Sampling Without Replacement) and Linear scaling, both contributing to preservation of population diversity and avoidance of premature convergence, as well as the fitness evaluation mechanism used in the main GA process which was responsible for a fair evaluation of candidates in each generation. The optimal performance in terms of quality of the solution and computational complexity, was achieved when the population size was set to 200 individuals and single point crossover was applied with crossover probability $P_{\text{crossover}} = 0.9$, while mutation rate was set to $P_{\text{mutation}} = 0.001$. Single and double point crossovers seemed to preserve the “high fitness” segments of bits. In the final results, single-point crossover was applied as it was observed to be the best trade-off between maintaining useful string segments and preservation of population diversity. The secondary nature of the role of mutation parameter was also verified, as the final choice of very low mutation rate of 0.001, enabled preservation of meaningful/useful string segments, whilst managing to introduce useful corrections in the string structure at the same time. In the linear scaling that was implemented, the coefficients were chosen to enforce the quality of average objective value and the scaled average fitness values and caused maximum scaled fitness to be a specified multiple of the average fitness. Linear scaling allowed the best individuals to create higher number of offspring, by multiplying their fitness by a suitable coefficient.

Comparison of the the proposed hybrid technique, with two other methods revealed that the network costs of solution 5 were 4.3% and 12.5% less than costs obtained by pure GA based i.e. MAN-GA approach, and Branch exchange solutions respectively. Note that using MAN-GA has produced a solution that costs 9% less and has two-thirds the delay of Branch exchange solution. As far as the delay was concerned, solution 5 resulted in 12.5% and 43.4% less delay, compared to MAN-GA and Branch exchange solutions respectively, whilst solution 1 resulted
in 7.8% and 15.7% less cost than the above two solutions. In terms of average link utilisation, solutions 1-5 exhibited utilisation at or close to 80%, which compared favorably with the 70% by MAN-GA and 65% by the Branch exchange methods. Solutions 1 and 5 also showed higher overall network capacity compared to MAN-GA and Branch exchange solutions, whilst maintaining higher utilisation levels. The MAN-GA method views packet delay as a network property to minimise (i.e. multiobjective GA approach), and thus becomes completely insensitive to the value of the delay constraint. Furthermore, the excessive use of GAs throughout the optimisation process, seems to be depriving the evaluation function of the ability to assess individuals as they should be. The stochastic nature of the GA (employing evaluation function that it uses) lead can to a frequent overestimation or underestimation of individuals (topologies) within the main GA population.

6.1 Future Research Directions

As far as mobility management techniques are concerned, schemes based on user movement and call-arrival profiles have been shown in the existing literature to be able to reduce the signalling costs, however, they are sensitive to mobility and call patterns assumed. On the other hand, dynamic methods usually require on-line collection and processing of great deal of data which can result in significant processing overheads. As far as DB architectures are concerned, centralised database architectures of which GSM and IS-41 are the prime examples, are based on Fixed LA partitioning scheme and are therefore associated with high signalling loads due to location updating procedure. Besides, centralised architectures do not scale well. Distributed database architectures can lead to minimisation of the signalling costs (for location registration and updating) however, the gains achieved may be at the expense of increased database updates and interrogations. Also, deployment and network management of many databases is at best problematic and distributed architectures tend to increase call set up delays unless full or partial replication of user data is used. The main requirements for efficient location management are:

- Minimization of database access/transactions due to location update traffic - It is expected that in the near future, the actual overhead of transporting the signalling traffic by the signalling network will be much reduced, since optical fiber based transmission and high-speed switching capabilities will provide bandwidths on the order of 100 Gb/s. It will therefore be possible to provide signalling capacity on the order of several Mb/s in broadband cellular networks. Hence, the simplification of signalling procedures and minimization of database access for mobility management will be the key issues to address.
Chapter 6. Conclusions and Future work

- Minimization of the signalling traffic on the air-interface, through intelligent partitioning of LAs - The optimum solution would be allocation of “personal location areas” i.e. LAs on a per-subscriber basis, taking into account, user mobility and call patterns, as opposed to “fixed location areas”. However, since the processing of data on a per-user basis may consume significant computing resources, and taken into consideration the fact that high mobility subscribers tend to exhibit similar mobility characteristics, solutions based on group/volume mobility i.e. dynamic LA partitions based on volume mobility as opposed to individual subscriber mobility patterns, may provide more efficient mechanisms. Thus the trade-off between the location updating and paging signalling load remains the main performance optimization problem during LA planning.

- The observed high levels of LU traffic in the case of the distance-based scheme (with RIP enabled) is compensated by extremely low levels of call-related signalling traffic compared to other schemes. It seems that the distance-based strategy is able to take full advantage of RIP, and this opens up the possibility that combinations of techniques could perhaps result in even greater savings in the signalling load. For example, the schemes of Alg.3 and Alg.4 have shown particularly high aptitude in adapting/tracking local mobility profiles and appear rather more successful in reducing the location update related signalling traffic. A combination of Alg.3/Alg.4 and Alg.5 (with RIP enabled) could in principle be used such that during periods of time when the call-to-mobility of subscriber is high, the system (network or terminal) initiates a distance-based LU mechanism, and when the mobility begins to or is estimated to result in high rate of location crossings i.e. at low call-to-mobility ratios, the system reverts back to either of the schemes of Alg.3 or Alg.4. Such a combination not withstanding the implementation issues, seems a viable combination for further reducing the signalling across the air-interface. A combined scheme should therefore be implemented and fully investigated.

The proposed GA-based method used for the design of communication network produced solutions that cost less and have lower delay than the two other typical design methods. However, the method has not given all it has to offer. Further enhancements can be achieved, and in that direction the following might prove beneficial:

- A the end of the optimisation process apply a local search technique around the finally proposed GA solution, using a conventional method (e.g. Branch-exchange). This can help explore any possible improvements that the GA missed to locate (after all GA is a global search optimisation technique).
Incorporate fitness sharing and mate restriction schemes to promote the evolution of species among the GA populations. Thus, a number of optimal solutions will be yielded at the end of each execution (GA convergence to more than one points of the search domain).

- Explore possibilities of further performance improvement when other Genetic Operations are incorporated (e.g. reinsertion, inversion).

- The proposed technique should be tested for network problems with a greater number of nodes, to check its performance for more difficult problems than the 10-node one considered here. Also impact of different tariff mechanisms/cost models on optimal component placement or topologies, need to be further assessed.

- Incorporation of repair heuristics, to avoid search for infeasible solutions or false minima.
Chapter 7

BIBLIOGRAPHY & REFERENCES


Chapter 7. Bibliography


[GSMr2.60] ETSI, GSM rec. 02.60, GPRS, service description stage1.

[GSMr3.02] ETSI, GSM rec. 03.02, Network architecture.

[GSMr8.02] ETSI GSM Rec. 08.02 series.

[GSMr8.52] ETSI GSM Rec. 08.52 series.

[GSMr9.02] ETSI GSM Rec. 09.02 series.


Chapter 7. Bibliography


180


[UMTS30.03] Selection procedures for the choice of radio transmission technologies of the UMTs, TR 101 112 v.3.2.0, 1998.


[Yat95] C.Rose and R.Yates “Minimizing the Average Cost of Paging under Delay Constraints”,


Publications

1. Performance Evaluation of Signalling Protocols on Different Middleware Platforms for Service Provision in Distributed Mobile Communication Networks
Klaus Moessner, Seiamak Vahid, Rahim Tafazolli
Published :: WPMC'99 (VTC99), September 21-23, Amsterdam, Netherlands.

2. A Minimum Air Interface Implementation for Software Radio Based on Distributed Object Technology
Klaus Moessner, Seiamak Vahid, Rahim Tafazolli
Published :: ICPWC'99, February 17-19, Jaipur, India.

3. Extension of the PNNI to Support Mobility Management and Handoffs in WATM Networks Based on CORBA
Klaus Moessner, Seiamak Vahid, Rahim Tafazolli
Published :: VTC'99, May 16-20, Houston, Texas, USA.

Seiamak Vahid, Rahim Tafazolli

5. Performance Comparison of Dynamic Techniques for reduction of Location Management Signalling in Cellular Networks
Seiamak Vahid, Rahim Tafazolli
Submitted (Sept. 2000) :: IEEE Electronics Letters

6. Optimization of Anchoring Technique for reduction of Signalling Traffic in Cellular Networks
Seiamak Vahid, Rahim Tafazolli
Submitted (Sept. 2000) :: IEE Proceedings in Communications

Klaus Moessner, Seiamak Vahid, Rahim Tafazolli
(invited paper) IEEE Special workshop on SWR – ALBA, Italy, 11th September 2000

Seiamak Vahid, Klaus Moessner, Rahim Tafazolli
Submitted (Oct. 2000) :: IEEE OPENARCH

Seiamak Vahid, Klaus Moessner, Rahim Tafazolli
In preparation.
Patents

   K. Moessner, S. Vahid, R. Tafazolli. PAT. App. no.: 0011954.5

   S. Vahid, K. Moessner, R. Tafazolli. PAT. App. no.: 0011955.1
Appendix A - GSM Signalling Message Charts & SS7 Tables

Figure A1.1 - MAP signalling message sequence during first location update/registration

Figure A1.2 - MAP signalling message sequence during location update to a new VLR (neighbour)
Figure A1.3 - MAP signalling message sequence during location update to a new VLR (non-neighbour)

Figure A1.4 - MAP signalling message sequence for Mobile_Terminated call
A.1 Handover

Handover refers to the transfer of an existing connection to a new base station. In GSM handover decision is made by the network, and it is based on BSS criteria (received signal level, channel quality, MS-BTS distance) and on network operational criteria e.g. current traffic load of the cell. Periodically, the MS gathers signal field strength measurements of its current downlink and those of the neighboring base stations (quality monitoring). On the network side, the signal quality of the uplink is monitored, the measurement reports from MS are evaluated and handover decisions are made. Principally, two kinds of handovers can be distinguished:

- **Intra-Cell Handover**: for administrative Reasons or because of channel quality the MS may be assigned a new channel within the same cell. This decision is made by the RR management of the BSS and also executed within BSS.

- **Inter-Cell Handover**: the connection to an MS is transferred over the cell boundary, to a new BTS. The decision re. Time of handover, is made by the RR protocol module of the network based on measurement data from MS and BSS.

As far as Inter-Cell handovers are concerned, two cases are distinguished. **Internal Handover**: which refers to handovers between cells belonging to the same BSC. This does not require the involvement of controlling MSC except for notification of MSC re. successful execution of handover by the BSC. **External Handover** refers to handovers where participation of at least one MSC is required. Handovers between cells belonging to different BSCs (under control of same MSC – i.e. Intra-MSC HO) and those between cells belonging to different BSCs (under control of different MSCs – i.e. Inter-MSC HO) are classified as external. Figures A1.6 and 7 show the message sequences for Internal (Intra-BSC) and external (Inter-BSC/Inter-MSC) handovers.

A.1.1 MAP and Inter-MSC Handover

Inter-MSC handovers require communication between the involved MSCs, which makes use of SS7 transport for MAP transactions. The serving MSC\(_A\) sends a handover request to the target MSC\(_B\), and once the link is setup between the two, MSC\(_A\) will issue the handover command to the MS. There are cases to consider:
• **Basic Handover procedure** -- the call is handed over from MSC_A to MSC_B (1).
• **Subsequent Handover procedure** – (2) the call is handed over from MSC_B back to MSC_A OR (3) for handover to a third MSC, the call is first handed over from MSC_B back to MSC_A (known as Handback) and then from MSC_A to MSC_C.

Thus, Basic handover procedures are used for the first handover between MSCs, whilst Subsequent handover procedures are used for handovers to subsequent MSCs. The MAP message sequence charts for the Basic and Subsequent HO are shown in Figures A1.8,9 and 10.

---

**Figure A1.6** - MAP signalling message sequence for Internal (Intra-BSC) handover

**Figure A1.7** - MAP signalling message sequence for External (Inter-BSC/Intra-MSC) handover

**Figure A1.8** - MAP signalling message sequence for Basic (Inter-MSC) handover (MSC_A to MSC_B)

---

188
Note that subsequent handover from MSC_b to MSC_c consists of two parts:
- A Subsequent Handover from MSC_b to MSC_a
- A Basic Handover between MSC_a and MSC_c

Figure A1.9 - MAP signalling message sequence for Subsequent handover (MSC_b to MSC_a) - Handback

Figure A1.10 - MAP signalling message sequence for Subsequent handover (MSC_a to MSC_c)
Appendices

A.2 SS7 Tables

A.2.1 SS7 Message Lengths

Based on GSM Rec. 09.02 which defines the SS7 MAP protocol, estimates of MAP message lengths can be calculated. For each of the messages defined, the header sizes relating to lower protocol layers were calculated as follows:

- **Tag and Length**: 2 octets per parameter used.
- **MTP L2**: 7 octets (inc. SIO, and 1 flag)
- **MTP L3**: 4 octets
- **SCCP**: 14 octets (SCCP with PC. For SCCP with GT header would be 16-32 octets.)
- **TSL**: 10 octets for a BEGIN, or END and 16 octets for a CONTINUE
- **CSL**: 10 octets for an INVOKE, 12 octets for a RETURN RESULT, 13 octets for an INVOKE and 5 octets for a RETURN RESULT which is only an acknowledgement.
### Table A2.1 - Basic Inter-MSC Handover Message Lengths (octets)

<table>
<thead>
<tr>
<th>MESSAGE NAME</th>
<th>TSL</th>
<th>CSL</th>
<th>Parameter Length</th>
<th>TCAP Total</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEGIN [INV(Perform Handover)]</td>
<td>10</td>
<td>10</td>
<td>47</td>
<td>67</td>
<td>92</td>
</tr>
<tr>
<td>TargetCellID</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ServingCellID</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChansTyps</td>
<td></td>
<td></td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ClassMarkInfo</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOPriority</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KS</td>
<td></td>
<td></td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONT [RES(RadioChan.Ack.)]</td>
<td>16</td>
<td>12</td>
<td>27</td>
<td>55</td>
<td>82</td>
</tr>
<tr>
<td>HONumber</td>
<td></td>
<td></td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ProtocolID</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CellDescription</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChannelDescription</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PowerCommand</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IAM</td>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td>59</td>
</tr>
<tr>
<td>ACM</td>
<td></td>
<td></td>
<td>16</td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>CONT [INV(SendEndSignal)]</td>
<td>16</td>
<td>10</td>
<td>26</td>
<td>44</td>
<td>61</td>
</tr>
<tr>
<td>ANS REL</td>
<td>10</td>
<td>5</td>
<td>15</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>RLC</td>
<td>10</td>
<td>5</td>
<td>15</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>BEGIN [INV(ReslocateHONumber)]</td>
<td>10</td>
<td>10</td>
<td>12</td>
<td>30</td>
<td>63</td>
</tr>
<tr>
<td>CONT [INV(RetResHORep)]]</td>
<td>16</td>
<td>16</td>
<td>28</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>END [RES(HORep)]</td>
<td>10</td>
<td>5</td>
<td>15</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

### Table A2.2 - Location Update (New VLR) Message Lengths (octets)

<table>
<thead>
<tr>
<th>MESSAGE NAME</th>
<th>TSL</th>
<th>CSL</th>
<th>Parameter Length</th>
<th>TCAP Total</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEGIN [INV(UpdateLoc.Area)]</td>
<td>10</td>
<td>10</td>
<td>22</td>
<td>43</td>
<td>65</td>
</tr>
<tr>
<td>TMS</td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LocIDOld</td>
<td>10</td>
<td>5</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LocIDNew</td>
<td>10</td>
<td>5</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CKSN</td>
<td>10</td>
<td>5</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONT [INV(Authenticate.)]</td>
<td>16</td>
<td>10</td>
<td>36</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>Rand</td>
<td>10</td>
<td>5</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CKSN</td>
<td>10</td>
<td>5</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONT [INV(Authenticate Ack.)]</td>
<td>16</td>
<td>12</td>
<td>28</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>SRES</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONT [INV(SendCipher.)]</td>
<td>16</td>
<td>10</td>
<td>13</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>Cipher Moda</td>
<td>16</td>
<td>10</td>
<td>13</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>KS</td>
<td>16</td>
<td>10</td>
<td>13</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>CONT [RES(Loc.AreaUpd. Ack.)]</td>
<td>16</td>
<td>10</td>
<td>13</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>END [RES(TMS Ack.)]</td>
<td>16</td>
<td>10</td>
<td>13</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>BEGIN [INV(BandPare, VLR)]</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>TMS</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter IMSI</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>END [RES(IMSIResponse)]</td>
<td>16</td>
<td>10</td>
<td>13</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>IMSI</td>
<td>16</td>
<td>10</td>
<td>13</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>KS, R, 5 (assume 3 sets)</td>
<td>16</td>
<td>10</td>
<td>13</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>BEGIN [INV(Upd. Location)]</td>
<td>16</td>
<td>10</td>
<td>13</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>TMS</td>
<td>16</td>
<td>10</td>
<td>13</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>Loc. Info.</td>
<td>16</td>
<td>10</td>
<td>13</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>MSRN</td>
<td>16</td>
<td>10</td>
<td>13</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>END [RES(Upl. Loc., Acc.)]</td>
<td>16</td>
<td>10</td>
<td>13</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>IMSI</td>
<td>16</td>
<td>10</td>
<td>13</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>BEGIN [INV(BandPare, VLR)]</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>IMSI</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter [AUTH.]</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>END [RES(Rep authentic. Skew)]</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>IMSI</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kc, R, 5 (assume 5 sets)</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEGIN [INV(Reslocate)]</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>IMSI</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONT [INV(CancelLocation)]</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>IMSI</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>END [RES(Loc. Cancel Ack.)]</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>20</td>
<td>40</td>
</tr>
</tbody>
</table>

191
For call setups (both mobile-originated and mobile terminated call types) the number of bytes between SS7 nodes involved per transaction are:

- **MO_Calls** - Originating MSC and terminating switch: 120 bytes
- **MO_Calls** - Originating MSC and associated VLR: 550 bytes
- **MT_Calls** - Originating switch and terminating MSC: 120 bytes
- **MT_Calls** - Terminating MSC and associated VLR: 612 bytes
- **MT_Calls** - Originating switch and HLR: 126 bytes

For an inter_MSC handover the number of bytes between SS7 network entities is:

- **New_MSC** and associated VLR: 148 bytes
- **New_MSC** and **Old_MSC**: 383 bytes

For an intra_VLR location update, the number of bytes between SS7 network entities is:

- **MSC** and associated VLR: 406 bytes

For an inter_VLR location update, the number of bytes between SS7 network entities is:

- **New_VLR** and associated MSC: 406 bytes
- **New_VLR** and **Old_VLR**: 213 bytes
- **HLR** and **Old_VLR**: 95 bytes
- **HLR** and **New_VLR**: 182 bytes

The above results have been summarised in Table 2.2 chapter II.
<table>
<thead>
<tr>
<th>MESSAGE NAME</th>
<th>TSL</th>
<th>CSL</th>
<th>Parameter Length</th>
<th>TCAP Total</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEGIN [INV(SendRoutingInfo.)]</td>
<td>10</td>
<td>10</td>
<td>12</td>
<td>32</td>
<td>57</td>
</tr>
<tr>
<td>MSISDN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>END [RES(RoutingInfo. Ack.)]</td>
<td>10</td>
<td>12</td>
<td>22</td>
<td>44</td>
<td>69</td>
</tr>
<tr>
<td>IMSI RoamingNumber</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IAM</td>
<td></td>
<td></td>
<td></td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>BEGIN [INV(VC CallSetup)]</td>
<td>10</td>
<td>10</td>
<td>12</td>
<td>32</td>
<td>57</td>
</tr>
<tr>
<td>MSRN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONT [INV(Page MS)]</td>
<td>16</td>
<td>10</td>
<td>17</td>
<td>43</td>
<td>68</td>
</tr>
<tr>
<td>IMSI LocAreaID</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONT [INV(Packet Access Req.)]</td>
<td>16</td>
<td>10</td>
<td>15</td>
<td>41</td>
<td>66</td>
</tr>
<tr>
<td>TMSI CMServiceType</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access Conn.Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CKSN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONT [INV(Authenticate)]</td>
<td>16</td>
<td>10</td>
<td>21</td>
<td>47</td>
<td>72</td>
</tr>
<tr>
<td>RAND CKSN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONT [RES(Auth. Ack.)]</td>
<td>16</td>
<td>12</td>
<td>6</td>
<td>34</td>
<td>59</td>
</tr>
<tr>
<td>TMSI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONT [INV(Start Ciphering)]</td>
<td>16</td>
<td>10</td>
<td>13</td>
<td>30</td>
<td>64</td>
</tr>
<tr>
<td>Cipher Mode Kc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONT [RES(Packet Access Resp.)]</td>
<td>16</td>
<td>12</td>
<td>10</td>
<td>38</td>
<td>63</td>
</tr>
<tr>
<td>IMSI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONT [INV(Forward New TMSI)]</td>
<td>16</td>
<td>10</td>
<td>6</td>
<td>32</td>
<td>57</td>
</tr>
<tr>
<td>TMSI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONT [RES(Complete Call)]</td>
<td>16</td>
<td>13</td>
<td>12</td>
<td>41</td>
<td>66</td>
</tr>
<tr>
<td>MSISDN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>END [INV(TMSI Ack.)]</td>
<td>10</td>
<td>5</td>
<td>15</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>ACM</td>
<td></td>
<td></td>
<td></td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>ANSI</td>
<td></td>
<td></td>
<td></td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>REL</td>
<td></td>
<td></td>
<td></td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>RLC</td>
<td></td>
<td></td>
<td></td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>

Table A2.4 - MT Call Establishment Message Lengths (octets)
Appendix B – Modelling Communication Networks

Graph theory is the basic mathematical tool when addressing network design problems [Ker93]. Some fundamentals of this theory, necessary to be understood even for the simplest design problem to be solved, are briefly given in the following sections. In this section the basic terminology is introduced for describing graphs and their properties. Although graph theory is a well-established discipline, there are, unfortunately several different accepted terms for many of its basic concepts. The terminology given here is just one of a number of accepted sets of terms and is used consistently throughout. We also present the baseline up on which communication networks are modelled as graphs.

B.1 Graph Definition
A graph, G, is defined by its vertex set, V, and its edge set, E. The vertices are more commonly called nodes and these represent locations (e.g. sources of traffic or sites of communication equipment). The edges are also called links, and these represent communication facilities. This is expressed:

\[ G = (V, E) \]

While in theory, V can be empty or infinite, V is usually a non-empty set that is,
\[ V = \{ u_i \mid i = 1, 2, \ldots, N \} \]
where N is the number of nodes.
Similarly, E is denoted by
\[ E = \{ e_j \mid j = 1, 2, \ldots, M \} \]
where M is the number of edges.

A link, \( e_j \), corresponds to a connection between a pair of nodes. This is sometimes noted by a link \( e_j \) between nodes i and k as
\[ e_j = (u_i, u_k) \]
or more simply as
\[ e_j = (i,k) \]
A link is incident on a node if the node is one of its endpoints. Nodes i and k are adjacent if there exists a link \((i, k)\) between them. Such nodes are also referred to as neighbours. Sometimes the link is bidirectional. In this case the order of the nodes in the link does not matter. Other times, the order does matter. If the order of the links does matter, the link is referred to as an arc and denoted by
\[ a_j = (u_i, u_k) \]
Or more simply as
\[ a_j = (i,k) \]
k is said to be outwardly adjacent to i if an arc \([i,k]\) exists. In this case i is said to be inwardly adjacent to k.

A graph is called a network if the links and nodes associated with it have properties (e.g. length, capacity, type etc). Networks are used to model problems in communications, and the specific properties of the nodes and links relate to the specific problem at hand. In this report, since any graph we confront refers to the topology of a communication network, the terms graph and network are used interchangeably. The difference between links and arcs is important, both in terms of how they model networks and how they function within algorithms, so the difference will always be clear. Graphically, links are lines connecting pairs of nodes. Arcs are lines with an arrow at one end, indicating the direction of the arc.

B.2 Undirected and Directed Graphs
A graph with links is called an undirected graph, while one with arcs is called a directed graph. Some directed graphs also include undirected links. Usually, graphs are assumed to be undirected. Otherwise,
their directed nature is specified. In figure B.1 and figure B.2 we show examples of an undirected and a directed graph respectively.

![Figure B.1: An undirected graph](image1)

![Figure B.2: A directed graph](image2)

**B.2.1 Simple Graphs**

It is possible for there to be more than one links between the same pair of nodes. This can for example correspond to multiple communication channels between two switches. Such links are referred to as parallel links. A graph with parallel links is called a multigraph. It is also possible for there to be a link between a node and itself. These links are called self-loops. This is relatively rare, but might result from treating two nodes as one in modelling a network, or may arise during the course of an algorithm which merges nodes. A graph without parallel links or self-loops is called a simple graph. In the network design problem, usually one has to deal only with simple graphs. This is the case for our project work as well.

**B.2.2 Paths in Graphs**

A path in a network is a sequence of links which begins at some node, s, and ends at some node, t. Such a path is also referred to as an s,t-path. The order of the links in the path matters. A path can be directed or undirected depending on whether its components are links or arcs. A path is said to be simple if a node appears no more than once in the path. A simple path in a simple graph can be specified by the sequence of nodes it contains, since a unique sequence of links is specified by such a node sequence. If s is the same as t, the path is called a cycle, and if an intermediate node appears no more than once, the cycle is said to be a simple cycle. A simple cycle in a simple graph can also be specified by a node sequence.

**B.3 Modelling Communication Channels in Graphs**

We have already referred to the terms duplex, half duplex and simplex channel in section 2.1. Now we see, how these channel properties are modelled in graphs. Clearly if network channels have capacity in only one direction, they are modelled as arcs. Half duplex lines are usually modelled as pairs of arcs. Full duplex links can be modelled as links with two capacities (two separate properties), or a pair oppositely directed arcs between the same pair of nodes. This is often an important decision which can have an effect on both the simplicity and effectiveness of the algorithm used to solve the problem. It is also possible, of course, to convert from one form to the other if necessary, but it is best to try and avoid this as it complicates implementation and increases runtime and storage requirements.

**B.3.1 Connected Graphs and Components**

A graph is connected if there is at least one path between every pair of nodes. The sets of nodes with paths to one another are connected components, or more simply, components. The links between these nodes are also part of the components. A connected graph has a single component. Each node or link is a member of exactly one component. Therefore the component structure of a graph can be described as a partition on its node set or link set. A directed graph with a directed path from every node to every other node is called strongly connected. Connectivity in directed graphs is not symmetric. There may be a directed path from i to j without there being one from j to i. A set of nodes with directed paths from any one node to any other is called a strongly connected component. A node is part of exactly one strongly connected component but that an arc can be part of at most one. Specifically, some arcs may not be part of any strongly connected component.
B.3.2 Trees

A tree is a graph without cycles. A spanning tree is a connected graph without cycles. Such a graph is referred to more simply as a tree (Figure B.3). If the graph is not necessarily connected, it is referred to as a forest. We generally speak of trees in undirected graphs.

In directed graphs, there is an analogous structure, called an arborescence. An arborescence is a directed graph which forms paths from one node (called the root of the arborescence) to all other nodes; alternatively, the paths might be from all other nodes to the root. An arborescence would be a tree if it were undirected.

Spanning trees have many interesting properties which make them useful in designing communication networks. First, they are minimally connected, that is, they are connected graphs but no subset of the edges in a tree forms a connected graph. They provide connectivity without any unnecessary additional links. Thus, if the objective was to simply design a connected network of minimum cost, a tree would be the optimal solution. Closely related to this is the fact that there is exactly one path between every pair of nodes in a tree. This makes routing a trivial problem in trees and greatly simplifies the communications equipment involved. However, since trees are minimally connected they are also minimally reliable and robust. This is why actual networks are usually more highly connected. Nevertheless, tree network configurations play a major role in the design process.

B.3.3 Cutsets and Cuts

A set of edges whose removal disconnects a graph (or, more generically, increases the number of its components) is called a disconnecting set. A disconnecting set which partitions the set of nodes into two sets, X and Y, is called a cutset or sometimes an XY-cutset. The most concern is with minimal cutsets (i.e., cutsets which are not subsets of other cutsets). In a tree, any edge is a minimal cutset. A minimal set of nodes whose removal partitions the remaining nodes into two sets is called a cut. Again, the interest is usually with minimal cuts. A single node whose removal disconnects the graph is called an articulation point.

B.4 k-Connectivity: A Reliability Measure of a Network Topology

The notion of connectivity in a graph can be generalised. If a network contains k disjoint paths between every pair of nodes, the network is k-connected. Disjoint paths are paths that have no elements in common. Paths are said to be edge disjoint (node disjoint) if they have no edges (nodes) in common. The network is said to be k-connected or k-edge-connected depending on whether the paths are node disjoint or edge disjoint. Node disjoint paths are, of course, also edge disjoint and that is why we are mainly concerned with k-node connectivity; we imply this one by the term k-connectivity. k-connectivity implies that the removal of any (k-1) or less nodes in the network (the links incident on each of them are considered removed as well) does not split the remaining network into two or more connected components (we keep on referring to 1-connectivity when we use the term connectivity). Thus, the graph still remains connected. This a very important reliability measure in communication networks where an acceptable performance (connectivity) has to be guaranteed, even in the case that a number of nodes collapses. Usually the demand for a network and thus for the corresponding graph is bi-connectivity. Biconnectivity of a connected graph is relaxed when no articulation point exists in it (see Figure B.4 and Figure B.5).
B.4.1 Tree Traversals: Finding Connected Components in a Graph

Given a tree, we may wish to visit all its nodes. This is called a tree traversal. There are several ways of traversing a tree. We begin by identifying a node in the tree as the root. The traversal is carried out relative to this node. Sometimes there is a logical reason for selecting the root (e.g., it is the central computer site). Other times, the root is selected arbitrarily. There are two approaches that can be employed to traverse a tree: the Breadth First Search (BDS) and the Depth First Search (DFS).

In the BDS case we start from the root (the root is considered visited) and set in a list all its neighbours. Then we visit all the nodes of the list. Afterwards, we identify for each node of the current list, all its unvisited neighbours and set them in the list which in the meantime has been emptied. We proceed by visiting all the nodes of the updated list, and so on, until all the nodes of the tree are visited.

In the DFS case we start again from the root and proceed on the basis of visiting an unvisited neighbour of the node just visited. Whenever we encounter a situation where all neighbours of a node are visited, we back off until we reach a visited node that has an unvisited neighbour. The end of the procedure is signalled when all the neighbours of the root are found visited. This implies that all nodes of the tree have been visited. Applying a DFS to the tree of Figure B.3, and selecting node A to be the root, the nodes are visited in the following order: first A, then C and B, then D and E, then F, G and H. The exact order is not unique (e.g., node D can be visited before or after node E). Applying a DFS to the same tree and selecting node A to be the root, the nodes could be visited in the following order: first A, then C, D, E, F, G, H and finally B.

As with trees, we usually want to traverse a graph and visit its nodes. The algorithms used are the same as in the tree traversal case. Again we select a node as the root and apply a tree traversal procedure. If the graph is connected, all its nodes will be visited. Otherwise, only the nodes of the connected component that the root belongs to will be visited. Thus tree traversals can be applied to a graph to identify its connected components. Since during the process, each node of the graph is visited at most once, a tree is defined in the graph. In the case that a DFS approach is employed, we refer to such a tree as a Depth First spanning tree (DFSTree). Throughout this work, graph traversals appeared several times as a necessary step in a more complex algorithm to be implemented. Whenever this happened, a DFS approach was employed. A pseudo-code implementing a DFS in a graph, is given below. The algorithm employs a recursive function \texttt{Visit( )}.

**DFS Implementation Pseudocode**

\textbf{Step 1:} Set all nodes of the graph to state "unvisited".
\textbf{Step 2:} Select at random a node, \( v \), as the root.
\textbf{Step 3:} \texttt{Visit(\( v \))}

\texttt{Function Visit(\( v \))}

- Set node \( v \) to state visited.
- For each unvisited neighbour of \( v \), \( w \)
  \texttt{Visit(\( w \)).}
B.5 Finding Shortest Paths: Dijkstra’s Algorithm

Consider a mesh topology connected graph (such as the one in Figure B.5) where a length value has been assigned to each link. This length value in the communication networks case, may refer to distance, capacity, cost or a combinatorial description of the above. The length of a path connecting two nodes is simply the sum of the lengths of the links traversed along the path. The question that usually arises is which is the shortest path, in the mesh topology, that connects a pair of nodes. A number of algorithms for this problem exists among which the most well known are the Dijkstra’s, the Bellman’s and the Floyd’s algorithms ([Ber87], [Ker93], [Tan96]). These algorithms can be applied to either undirected or directed (connected) graphs. They operate based on the optimality principle which states that if node j is on the shortest path from node i to node k, then the optimal path from j to k also falls along the same route. Which algorithm has the lower order of complexity depends on the problem at hand. We have chosen to employ the Dijkstra’s algorithm whenever a shortest path problem arises. This algorithm is briefly described in the following of this section.

B.5.1 Dijkstra’s Algorithm

The algorithm finds the shortest paths from a target node X to all the other nodes of the connected graph, at once. The lengths of the established links in the topology are available and denoted by \( d_{v,w} \) where v and w are the nodes that a link connects. The algorithm makes use of a list (tentative list) where initially all nodes except from node X, are set to belong. A node belongs to the tentative list, as long as it remains labeled tentative. The label of each node denotes the length of the so far found shortest path to node X. Progressively, nodes are labeled permanently, one by one, whenever we find a shortest path that definitely cannot be outperformed. Therefore, nodes are removed from the tentative list, one by one. The algorithm is over when the tentative list remains empty. That means that shortest paths from all nodes to the target node X have been computed. For a detailed description of the algorithm, the interested reader may refer to [Ber87], [Tan96]. What we give below is a pseudo-code which implements the Dijkstra’s Algorithm.

\[\text{Dijkstra’s Algorithm Implementation Pseudo-Code}\]

\[\begin{align*}
\text{Step 1:} & \quad \text{Set all nodes except from node } X, \text{ to belong to the tentative list, mark them “tentatively labeled” and set their labels to infinity value. } X \text{ is the working node (set } v = X). \\
& \text{Mark } X \text{ “permanently labelled”. Set label}[X] = 0. \\
\text{Step 2:} & \quad \text{Updating labels: For each neighbour (w) of the current working node (v), which is marked “tentatively labelled”, check whether label}[v] + d_{v,w} < \text{label}[w]. \text{ If so:} \\
& \quad \text{set label}[w] = \text{label}[v] + d_{v,w} \\
& \quad \text{set previous}[w] = v \\
\text{Step 3:} & \quad \text{Finding the new working node: From all the nodes marked “tentatively labelled”, find the one (Y) which has the minimum label value. This will be the new working node (set } v = Y). \\
\text{Step 4:} & \quad \text{Mark node } Y \text{ “permanently labelled”. Remove node } Y \text{ from the tentative list. If now the list is not empty go to Step 2.} \\
\text{Step 5:} & \quad \text{The shortest paths from node } X \text{ to each node of the graph have been found. The length of each path is equal to the permanent label of the respective node.}
\end{align*}\]

Now, for each node, we can easily obtain the sequence of nodes that the shortest path found implies, by making use of the previous[ ] information that any node, except from X, possesses. For a given Y node (YεX) we set K=Y, then identify node Z=previous[K], then set K=Z and so on, until we encounter the target node X. The list of nodes identified defines the shortest path from X to Y.

Connectivity and Biconnectivity Testing of a Graph

B.6 Connectivity Testing Algorithm

This algorithm receives as an input a graph, not necessarily a connected one, and checks if it meets the connectivity requirement. If not, it adds links at random such that:
- the transformed graph is connected
- a minimum number of links are added (cost effective transformation)

Thus, the algorithm yields as an output, a definitely connected graph.

**B.6.1 The Strategy**

We are given a graph of $N$ nodes, not necessarily a connected one. We choose at random a node $A$ and applying DFS we find a DFtree rooted on $A$. We check if all the nodes of the graph have been visited. If not (in this case the graph is not connected) we choose at random a node $B$ not visited, and similarly we apply DFS to find a DFtree rooted on $B$. The procedure goes on like this until all nodes of the graph are eventually visited. Each DFtree we find, corresponds to a separate connected component of the graph. Assume that $k$ ($k \geq 1$ always) connected components are identified. If $k=1$ the graph is connected and no additional links are needed. If $k>1$ the graph is not connected and we have to add links so as to bring into a single body its separate connected components. To achieve this, we need to establish at least $k-1$ new links between nodes of separate components. In the example of the unconnected graph in Figure B.6, our algorithm identifies the three connected components $\{A,B,C,D\}$, $\{E\}$, $\{F,G,H\}$ and adds two links at random to bridge them. A possible choice (among several) is to establish links (CE) and (AF) as shown in Figure B.6.

**B.6.2 Pseudocode**

The pseudo-code given below implements the above technique.

![Figure B.6: Example of connectivity testing and necessary transformations](image)

Connectivity Testing Implementation Pseudo-Code

**Step1:** Set all nodes of the graph to state "unvisited". Set $k=0$.

**Step2:** Choose at random a node $X$ of the graph, not visited.

**Step3:** Find an FDtree rooted on $X$. The nodes of this tree (including $X$) are set to state "visited". They comprise a connected component. Increase $k$ ($k=k+1$)

**Step4:** If not all nodes of the network have been visited go to Step2.

**Step5:** A number of $k$ ($k \geq 1$) connected components $\{C_1,C_2,C_3...C_k\}$ have been identified. If $k=1$ the network (graph) is itself connected, go to Step9.

**Step6:** Set all connected components to state "not embodied". Choose at random a component and mark it "embodied".

**Step7:** Choose at random a not embodied connected component $Y$. Add at random one new link between component $Y$ and a randomly chosen, embodied component. Mark $Y$ "embodied".

**Step8:** If not all components have been embodied, go to Step7.

**Step9:** The network (graph) is now connected; $(k-1)$ new links were added
B.7 Bi-connectivity Testing Algorithm
This algorithm receives as an input a graph, not necessarily a biconnected one, and checks if it meets the biconnectivity requirement. If not, it adds links in a random such that:
- the transformed graph is biconnected
- a minimum number of links is added (cost effective transformation)

Thus, the algorithm yields as output a definitely biconnected graph.

B.7.1 Some Comments on a DFS applied to a Connected Graph
- A DFS traversal (as any traversal) applied to a connected graph, partitions its links into two sets: T and B. An edge \( (V,W) \) is placed in set T (tree edges) if vertex W has not been previously visited when we are at vertex V, considering edge \( (V,W) \). Otherwise edge \( (V,W) \) is placed in set B (back edges).
- A DFS traversal in a connected graph, visits its nodes in a certain order. The DFS can be set to assign a number (depth first number: DFN) to each node it visits, according to this order. Thus, the DFN number for node \( V \) on which the FDTree is rooted, will be \( DFN(V) = 1 \).
- Consider a visited node, \( V \), in a FDTree. Its neighbour, \( W \), that had been visited just before \( V \) was visited, is referred to, as its father. \( V \) is a son of \( W \).
- A node \( W \) in the FDtree is considered a descendant of node \( V \), if the path that connects them, does not include the father of \( V \). Then, node \( V \) is an ancestor of node \( W \). A node with no descendants is called a leaf. The root has no father or any ancestors. The root is an ancestor of all the other nodes of the FDtree.

B.7.2 Identifying Articulation Points in a Connected Graph
The Lowpoint algorithm [Aho74] suggests to employ a DFS traversal in the connected graph in question, in order to identify its articulation points (if any). Then, articulation points are identified using the lemma given below

Lemma B.1
A node (vertex) \( V \) is an articulation point if and only if either
- \( V \) is the root of the DFtree and has more than one sons
- \( V \) is not the root and there exists a son \( S \) of \( V \) such that there exists no back edge between \( S \) or any of its descendants, and an ancestor of \( V \) (i.e. there is no ancestor of \( V \) straight connected, via a back edge, to \( S \) or any of its descendants).

In order to express the above argument in terms of mathematical equations and inequalities, so as to be applicable in an algorithm, the notion of the Low property of each node in the traversed graph is introduced:

For a node \( V \) its Low value \( \text{Low}(V) \) is given by:
\[
\text{Low}(V) = \min \left( \{ \text{DFN}(V) \} \cup \{ \text{Low}(S) \mid S \text{ is a son of } V \} \cup \{ \text{DFN}(W) \mid (V,W) \text{ is a back edge } \} \right)
\]

It can be shown [Aho74] that Lemma B.1 can be interpreted as follows

A node (vertex) \( V \) is an articulation point if and only if either
- \( V \) is the root of the DFtree and has more than one sons
- \( V \) is not the root and there exists a son \( S \) of \( V \) such that \( \text{Low}(S) \geq \text{DFN}(V) \)
B.7.3 Implementing an Algorithm to Find Biconnected Components in a Connected Graph (Lowpoint Algorithm)

The algorithm given below (Lowpoint algorithm), in terms of a pseudocode, identifies the articulation points (if any) that may exist in a connected graph. Then it is simple to identify the bi-connected components. To achieve this, it employs a stack where edges are pushed when they are encountered for the first time during the DFS process which is conducted to locate articulation points. We refer to this stack as the stack of edges. The algorithm employs a recursive function Low_evaluation_of_node() which effectively conducts a DFS throughout the graph, while at the same time assigns “Low values” to its nodes in a recursive way as eq.B.1 implies. For a more detailed description of the algorithm the interested reader should refer to [Aho74].

Lowpoint Algorithm Pseudo-Code

Step1: Set all nodes of the graph to state “unvisited”. Set $k_b=0$. Set the stack of edges to state empty. Set counter=0.
Step2: Choose at random a node to be the root.
Step3: Low_evaluate_of_node(root).

Function Low_evaluation_of_node(V)
• Set V to state “visited”
• Set DFN[V]=counter
• Increase counter (counter =counter+1)
• Set Low[V]=DFN[V]
• For each node W, which is a neighbour of node V
  • if ( (W is unvisited) or ( (W is visited ) and ( DFN(W) < DFN(V)) and (W is not the father of V) ) )
    • push edge (V,W) into the stack of edges
  • if W is unvisited
    • mark edge (V,W) as basic edge
    • set that V is the father of W
    • Low_evaluation_of_node(W)
    • If $(Low[W] \geq DFN[V])$
      • If V is not the root
        • node V is an articulation point
        • The existence of a new biconnected component has been identified.
        • increase $k_b$: $k_b=k_b+1$
        • Remove edges from the stack of edges until you get edge (V,W). This is the last to remove. The removed edges comprise the $k_b^{th}$ biconnected component.
      • If $(Low[W] \geq Low[V])$
        • set Low[V]=Low[W]
    • else, if (W is not the father of V)
      • if (FDN<Low[V])
        • set Low[v]=FDN[W]
B.7.4 Restoring Biconnectivity in a non-Biconnected graph

The above algorithm yields the number of the biconnected components ($k_b : k_b \geq 1$) found in the graph. If $k_b = 1$ the graph is biconnected itself and there is no need for additional links. Otherwise, links must be added. To do this we have to identify the biconnected components that contain only one articulation point. Bringing these biconnected components (we call them single biconnected components) into a single body is the most cost effective way (in terms of number of links to be added) to achieve a biconnected topology. In the example of non-biconnected graph in Figure B.7 (topology a) we identify three biconnected components {A,C}, {B,C,D}, {E,D,F}. However only two of them are single ( {A,C} (E,D,F)). By adding one link to bring them together as shown in topology b (in the same figure ), biconnectivity is restored. On the other hand if we add link (A,B) to topology a, topology c is produced, which is still not biconnected.

Assume that $k_s$ single biconnected components have been identified ($B_1, B_2, B_3... B_{ks}$). The simple following pseudocode implements an algorithm that restores biconnectivity in a graph.

**Biconnectivity Restoration Pseudo-Code**

**Step 1:** Set all single biconnected components to state “not embodied”. Choose at random a single biconnected component and mark it “embodied”.

**Step 2:** Choose at random a not embodied single biconnected component Y. Add at random one new link between component Y and a randomly chosen, embodied single biconnected component. Mark Y “embodied”.

**Step 3:** If not all the components have been embodied, go to Step 2.

**Step 4:** The network (graph) is now biconnected; ($k_s - 1$) new links were added.
Appendix C - The Bertsekas-Gallager Algorithm

C.1 The Concept and the Strategy

The Bertsekas-Gallager (BG) [Ber87],[Ker93] algorithm is a very efficient method for finding the optimal solution for the routing problem in a communication network i.e. the FA problem. The routing problem is a convex one, which means that once a feasible solution has been produced, we can apply gradual transformations to it (e.g. using a “hill climbing” method) which will eventually lead to the optimal solution. However few algorithms can do this efficiently. The BG algorithm it is a numerical steepest descent method which achieves the above target.

The BG algorithm is an iterative process which starts from a feasible solution (a routing policy where in each established link, the flow $f_i$ does not exceed its capacity $C_i$) and applies gradual transformations to it, until the optimal routing policy is achieved in terms of minimisation of average packet delay ($T$). In chapter 5 we have shown a way to obtain a starting feasible routing policy. What is presented in this appendix is a modification of the original BG algorithm [KER.93]. This version of BG algorithm, is an accelerated one. In each iteration, the algorithm identifies for each requirement (pair of nodes) the least congested path, that connects its origin and destination points. Therefore, an amount of traffic of each requirement is to be moved to those paths, and this is expected to bring a decrease in the average packet delay. It is actually very simple to compute those least congested paths. We just need a metric of the congestion of each link. The first derivative $\frac{dT}{df_i}$, where $T$ is the average packet delay and $f_i$ the flow in link $i$, serves as such a metric and is termed incremental delay. Thus, when incremental delays have been computed for each link, applying a shortest path algorithm (e.g. Dijkstra's algorithm) where the length of each link is considered equal to its incremental delay value, the least congested paths are identified. The crucial point is how much traffic to move to the current identified shortest paths. In order the process not to oscillate and perform efficiently, the amount of traffic diverted must be a small one at a time.

At the end of each iteration the new packet delay value is computed. The process is considered to have converged when between successive iterations the difference between packet delay values is within a tolerance $T_{toler}$, which defines the accuracy to which we wish to minimise average packet delay. Usually a $T_{toler} = 0.01\%$ is set to judge whether the process has converged, and thus can be terminated since effectively, no further improvement can be obtained.

C.2 Basic Strategy Elements

C.2.1 Calculating Incremental Delay Values

For the case where we accept Kleinrock's Independence Approximation (5.9) to assess the packet average delay $T$:

$$T = \frac{1}{\gamma} \sum_{i} f_i \frac{C_i}{C_i - f_i}$$

C.1

the incremental delay value $l_i$ for each link $i$ is given by:

$$l_i = \frac{1}{\gamma} \sum_{i} \frac{C_i}{(C_i - f_i)^2}$$

C.2

C.2.2 Maintaining and Updating Path Lists for each Requirement

In each stage (iteration) of the progress for each requirement:

- a list of the identities of the paths, that have been so far deployed to serve it, is maintained
- the information of which is the requirement's traffic amount, loaded on each deployed path, is also stored.
Whenever a shortest path computation is over, for each requirement, the algorithm checks whether the current shortest path is in its path list. If so, it identifies it. If not, the list is updated and a new path is registered in it. Afterwards, the algorithm moves traffic from each non-shortest deployed path to the shortest.

### C.2.3 Computing the amount of traffic to move each time

Assume, that in the current iteration, for a requirement \( r \), a path \( S \) has been found to be the shortest one, and that we want to compute the amount of traffic \( \delta \), to move from a non-shortest path \( P \) of the corresponding requirement’s list to path \( S \). The BG algorithm computes this amount \( \delta \), with respect to the second derivative of delay with respect to flow. Thus \( \delta \) is given by the formula:

\[
\delta = a \cdot \frac{(L_P - L_S)}{H_{ps}}
\]

where
- \( a \) is the stepsize
- \( L_P, L_S \) are the lengths of paths \( P \) and \( S \) respectively
- \( H_{ps} \) is the second derivative path length

We will refer to the parameter step size \( a \) later. For now consider it a small positive value (0 < \( a < 1 \)). \( L_P, L_S \) are simply the sums of the incremental delays of the links participating in routes (paths) \( P \) and \( S \) respectively. \( H_{ps} \) is actually an approximation. For the case where we accept Kleinrock Independence Approximation (C.1) \( H_{ps} \) is given by the formula

\[
H_{ps} = \sum_{j \in U} B_j
\]

where \( U \) is the set of links which either belong to path \( S \) or to path \( P \), but not to both, and \( B_j \) is given by

\[
B_j = \frac{1}{\gamma} \cdot \frac{2 \cdot C_j}{(C_j - f_j)^3}
\]

However we cannot always apply the traffic shift that eq.C.3 suggests. We have to check whether this shift is feasible to take place and adjust (decrease) the proposed amount of traffic \( \delta \) to be diverted. This adjustment of \( \delta \) is done in two steps:
- Check whether \( \delta \) is less than the requirement traffic amount, path \( P \) currently serves. If not set \( \delta \) equal to this value (i.e. the flow from path \( P \) is to be removed completely).
- Check whether the spare capacity in each link of path \( S \) is greater than \( \delta \). If not, adjust \( \delta \) (decrease it), so as not to violate the capacity constraint of the link with the least spare capacity. This guarantees that no link in \( S \) path will appear overloaded if the traffic shift \( \delta \) is applied.

### C.2.3 The stepsize \( a \) parameter

The choice of the value that this parameter is assigned is very crucial for the performance of the process. Small values guarantee a shift towards the right direction, but they slow down the whole process, sometimes unacceptably. On the other hand, great values sometimes cause the algorithm to oscillate without being able to bring any average packet delay decrease at all. More than that, the process in this case can lead to an unbalanced state where packet delay appears to increase dramatically. An effective choice is the following strategy:

Set an initial value for \( a \) in the range

\[
\frac{1}{N} \leq a_{init} \leq \frac{L}{N^2}
\]
where \( N \) is the number of nodes and \( L \) is the number of directed links in the network. The value of \( a \) is allowed to vary (be reduced) during the process, among iterations, according to the following rules:

- Whenever packet delay appears to increase (instead of decreasing) we reduce the value of \( a \) (e.g. \( a \) is halved).
- Whenever the process seems to have converged we reduce the value of \( a \) (e.g. \( a \) is halved). Each time we are to apply such a reduction we check whether the stepsize value has already reached a minimum pre-specified value (e.g. \( a_{\text{init}}/10 \)). If so we terminate the process.

C.3 Bringing all together
To sum up, the whole BG routing optimisation process can be outlined in the following pseudo-code.

**BG Route-Optimisation Implementation Pseudo-Code**

**Step1**: Obtain a feasible routing policy. Set values for the acceptable packet delay tolerance \( T_{\text{toler}} \), the initial stepsize \( a_{\text{init}} \) and the minimum allowable stepsize \( a_{\min} \).

**Step2**: Compute the incremental delay for each established link.

**Step3**: Find shortest paths in the network according to the incremental delay values.

**Step4**: For each requirement, if the new shortest path is already in its list, identify it. Else, register it in the list.

**Step5**: For each requirement

- for each non-shortest path \( P \) in its path list
  - compute the amount of traffic \( \delta \) to remove from path \( P \) to the shortest path \( S \)
  - apply the traffic shift implied above
  - update the link flows state throughout the network (this step is necessary for valid further \( \delta \) computations in the current and future iterations)

**Step4**: Compute packet delay \( T \) for the new flow state.

**Step6**: Update if necessary (convergence within tolerance \( T_{\text{toler}} \) or packet delay increase cases), stepsize value \( a \). If \( a \) is not found below the pre-specified minimum value \( a_{\min} \), got to **Step2**

**Step7**: Terminate the process. Average Packet Delay has been minimised with respect to routing.
Appendix D – On Genetic Algorithms

D.1 Meta Heuristic Search Techniques

Genetic Algorithms (GA) are directed random search techniques, inspired by the mechanisms of natural selection. GA is not considered a mathematically guided algorithm, as the optima obtained is evolved from generation to generation, without rigorous mathematical formulation of say traditional gradient-type optimization procedures. GAs are merely stochastic, discrete event and non-linear processes where the obtained optima is an end product containing the best elements of previous generations and where the attributes of stronger individuals, tend to be carried forward into the following generations. GAs and hybrid techniques based on GAs, have been used as optimization tools in the fields of industrial design, signal processing, control engineering and communications. GA techniques are considered attractive, since they can handle problem constraints by embedding them into the chromosome coding, can address multiobjective problems, and since GAs are techniques that are independent of the error surface, they can be used to solve multimodal, non-continuous or NP-complete problems.

In general, any abstract task to be accomplished can be thought of as solving a problem, which, in turn can be perceived as a search through a space of potential solutions. Since we are after “the best” solution, we can view this task as an optimization process. For small spaces, classical exhaustive search methods usually suffice; for larger spaces special heuristic techniques must be employed. Genetic Algorithms are among such techniques. Genetic Algorithm search strategy - first introduced by Holland [Holl75] - is based on a theme borrowed from population genetics. Holland uses concepts from population genetics and evolution theory to construct algorithms that try to optimize the fitness of a population of elements through recombination and mutation of their genes. The individuals (or phenotypes, structures) in a population are quite often also structures or chromosomes. Chromosomes are made of units - genes (also features, characters, or decoders) - arranged in linear succession, with every gene controlling the inheritance of one or several characters. Genes of certain characters are located at certain places of the chromosome, which are called loci (string positions). Any character of individuals (such as hair color) can manifest itself differently; the gene is said to be in several states, called alleles (feature values). Each genotype (i.e. a single chromosome) thus represents a potential solution to a problem (the meaning of a particular chromosome, i.e., its phenotype, is defined externally by the user); an evolution process run on a population of chromosomes corresponds to a search through a space of potential solutions.

GAs have been quite successfully applied to optimization problems like routing, scheduling, adaptive control, game playing, cognitive modelling, transportation problems, travelling salesman problems, optimal control problems, database query optimization, etc. (see [Mic94]). Genetic algorithms aim at complex large-scale combinatorial optimization problems. They belong to the class of probabilistic algorithms, yet they are very different from random algorithms as they combine elements of directed and stochastic search. Because of this, GA are also more robust than existing directed search methods. For instance, the hill-climbing methods use the iterative improvement techniques; the technique is applied to a single point (the current point) in the search space. During a single iteration, a new point is selected from the neighbourhood of the current point (this is why this technique is known also as neighbourhood search or local search [Laa87]). If the new point provides a better value of the objective function, the new point becomes the current point. Otherwise, some other neighbour is selected and tested against the current point. The method terminates if no further improvement is possible. It is clear that the hill-climbing methods provide local optimum values only and these values depend on the selection of the starting point. Moreover, there is no information available on the relative error (with respect to the global optimum) of the solution found. To increase the chances to succeed, hill-climbing methods usually are executed for a number of starting points (these points need not be selected randomly - a selection of a starting point for a single execution may depend on the result of the previous runs). The simulated annealing technique [Aar89] eliminates most disadvantages of the hill-climbing methods in that solutions do not depend on the starting point any longer and are (usually) close to the optimum point.

D.2 GA Building Blocks

As mentioned earlier, a GA performs a multi-dimensional search by maintaining a population of potential solutions and encourages information formation and exchange between these dimensions. The population undergoes a simulated evolution: at each generation the relatively "good" solutions reproduce,
while the relatively "bad" solutions die. To distinguish between different solutions we use an objective (evaluation) function which plays the role of an environment. As depicted in Figure D.1, there are three common genetic operators: selection, crossover and mutation. An additional reproduction operator, inversion, is sometimes employed. The choice or design of these operators depend on the problem at hand and the representation scheme selected. For instance, operators designed for binary strings cannot be directly used on strings coded with integers or real numbers.

```plaintext
procedure genetic algorithm
  begin
    t ← 0
    initialize P(t)
    evaluate P(t)
    while (not termination-condition) do
      begin
        t ← t + 1
        select Pnew(t) from P(t-1)
        recombine Pnew(t)
        mutate Pnew(t)
        evaluate Pnew(t)
      end
    end
```

Figure D.1 – The structure of a GA programme

The structure of a simple genetic algorithm is the same as the structure of any evolution program (Figure D.1). During iteration \(t\), a genetic algorithm maintains a population of potential solutions (chromosomes, vectors), \(P(t) = \{x_1, \ldots, x_n\}\). Each solution \(x_i\) is evaluated to give some measure of its "fitness". Then, a new population - \(P_{\text{new}}(t)\) (iteration \(t + 1\)) - is formed by selecting the more fit individuals. Some members of this new population undergo alterations by means of crossover and mutation operators, to form new solutions. Crossover combines the features of two parent chromosomes to form two similar offspring by swapping corresponding segments of the parents. For example, if the parents are represented by five-dimensional vectors \((a_1, b_1, c_1, d_1, e_1)\) and \((a_2, b_2, c_2, d_2, e_2)\) then crossing the chromosomes after the second gene would produce the offspring \((a_1, b_1, c_2, d_2, e_2)\) and \((a_2, b_2, c_1, d_1, e_1)\). The intuition behind the applicability of the crossover operator is information exchange between different potential solutions. Mutation arbitrarily alters one or more genes of a selected chromosome, by a random change with a probability equal to the mutation rate. The intuition behind the mutation operator is the introduction of some extra variability into the population. A genetic algorithm (as any evolution program) for a particular problem must have the following five components:

- a genetic representation for potential solutions to the problem,
- a way to create an initial population of potential solutions,
- an evaluation function for rating solutions in terms of their "fitness",
- genetic operators that alter the composition of children,
- values for various GA parameters (pop-size, probabilities of applying genetic operators, etc.).

D.2.1 Chromosome Representations

Individuals or chromosomes, are encoded as strings, composed over some alphabet(s), so that the genotypes (chromosome values) are uniquely mapped onto the decision variable (phenotypic) domain. The method of representation can have major impacts (in terms of accuracy and computation time) on the performance of the GA. There are two common representation methods for numerical optimisation problems [Mic92][Dav91]. The classical approach is the method of bit string encoding over the binary alphabet \((0, 1)\), which offers the maximum number of schemata per bit compared to other coding techniques. Binary coding can be based on uniform or gray-scale coding schemes. The second representation method, is to use a vector of integers or real numbers, with each integer or real number
representing a single parameter [Zal97]. This direct manipulation of real-valued chromosomes [Jan91][Wri91] has been introduced relatively recently, for dealing with real-parameter problems. However, despite the results in [Jan91] indicating that floating-point representation would enable faster computations, the opinion expressed in [Gold90] suggests that real-coded GA would not necessarily yield better results in all circumstances, and therefore, so far there is not sufficient consensus to draw general conclusions on the superiority of real-valued chromosomes over bit-string encoding.

D.3 The Main Genetic Operators
In order the GA process to begin, a starting population of individuals is randomly generated. Afterwards, its individuals are evaluated by means of a fitness evaluation mechanism, and genetic operations are applied to them so as the next generation (new population) to be produced. This process is repeated at each generation and throughout the genetic evolution, a fitter chromosome has the tendency to yield good-quality offspring, which means a better solution to the problem. The simplest (and one of the most popular) approach is to use non-overlapping populations with a constant size at each generation [Gold89]. That means that the whole population of a current generation is replaced (moving to the next generation) by an even number of individuals yielded by the genetic operations applied to the existing ones. The cycle of evolution is repeated until a termination criterion is reached. This criterion may be defined in terms of

- a predefined desirable value of fitness to appear OR
- the number of evolutions cycles (generations) to be processed OR
- the degree of variation of individuals between different generations, or even in the same population.

There is no point in continuing the GA process when all the individuals of the current population appear to be, more or less identical. We refer to this situation as convergence to a particular point of the search domain.

In the following sections we present the main genetic operations that are applied in each generation. These are the following:

- Selection
- Crossover (recombination)
- Mutation

These three operators are the basic elements of the simple GAs which apart from yielding good results for many practical problems [Gold89], have been the source of inspiration for more complex GA strategies.

D.3.1 Selection Mechanism
Selection is the process which determines which individuals of the current population will participate in the formation of the new population for the next generation during the following crossover (recombination) cycle, and more specifically, how many times each of those will be selected to mate with another to give offspring. In other words, selection is used for reproduction of more copies of high-fitness individuals rather than low-fitness ones. Although the selection procedure has significant influence on driving the search towards promising regions of search space, the diversity of population must be maintained in order to avoid premature convergence and to reach the globally optimum solution.

In [Bri81] some further modifications are considered: deterministic sampling, remainder stochastic sampling without replacement, stochastic tournament, and remainder stochastic sampling with replacement. The study confirms the superiority of some of these modifications over simple selection. In particular, the remainder stochastic sampling with replacement method, which allocates samples according to the integer part of the expected value of occurrences of each chromosome in a new population and where the chromosomes compete according to the fractional parts for the remaining places in the population, was the most successful one and adopted by many researchers as standard. In [Bak87] a comprehensive theoretical study of these modifications using some well-defined measures is provided, and also a new improved version called stochastic universal sampling is presented. This method uses a single wheel spin. This wheel, which is constructed in the standard way, is spun with a number of equally spaced markers equal to the population size as opposed to a single one. Other methods to sample a population are based on introducing artificial weights: chromosomes are selected proportionally to their
rank rather than actual evaluation values (see e.g., [Bak85], [Whi89]). These methods are based on a belief that the common cause of rapid (premature) convergence is the presence of super individuals, which are much better than the average fitness of the population. Such super individuals have a large number of offspring and (due to the constant size of the population) prevent other individuals from contributing any offspring in next generations. In a few generations a super individual can eliminate desirable chromosomal material and cause a rapid convergence to (possibly local) optimum. Another selection method, tournament selection [Gold91], combines the idea of ranking in very interesting and efficient way. This method (in a single iteration) selects some number k of individuals and selects the best one from this set of k elements to move into the next generation. This process is repeated pop-size number of times. It is clear that large values of k increase selective pressure of this procedure; typical value accepted by many applications is k = 2 (so-called tournament size). Here it is also possible to add a flavour of simulated annealing by considering Boltzmann selection. In [Hof91] selection procedures are divided into dynamic and static methods - a static selection requires that selection probabilities remain constant between generations (for example, ranking selection), whereas a dynamic selection does not have such a requirement (e.g., proportional selection). Additionally, some selection procedures are pure in the sense that parents are allowed to reproduce in one generation only (i.e., the life time of each individual is limited to one generation only regardless of its fitness). Some selections are on-the-fly, meaning that an offspring replaces its parents immediately. Some selections are elitist in the sense that some (or all) of the parents are allowed to undergo selection with their offspring, similar to the elitist model of [DeJ75].

D.3.1.1 Basic Selection Schemes

In GAs two selection procedures, proportional selection and ranking-based selection [Whi89] are usually used. Proportional selection is also known as "roulette wheel" selection, where fitness values of individuals represent the widths of slots on the roulette wheel. After a random spin of the wheel, individuals in slots with larger widths representing high-fitness values, will have a higher chance to be selected. The rank-based reproduction procedure, aims to limit the range of trials allocated to a single individual, so that no one individual generates too many offspring, thus preventing premature convergence to locally optimal solutions. Having decoded the chromosome into the decision variable domain, it is possible to assess the performance, or fitness, of an individual. This is done through an objective function that characterizes an individual’s performance in the problem domain. In the natural world, this would be an individual’s ability to survive in its present environment. Thus, the objective function establishes the basis for selection. Each individual is assigned a fitness value derived from its raw performance measure, given by the objective function. The fitness value for each individual is a non-negative value and serves as a metric (figure of merit) of the quality of its performance (the better it performs, the greater its fitness value). The selection mechanism selects individuals one by one from the current population (one can be selected more than once), and each time, the probability that a particular individual is selected is proportional to its fitness value. There are several schemes for the implementation of the selection operator [Zal97][Mic92].

Rarely can the objective function itself give a non-negative figure of merit value, for each individual. GA can in fact be viewed as a maximization (optimisation) process, where the ultimate target is the maximization of the fitness value of individuals. However, usually cost minimization problems appear (i.e. we view as cost, a property of the system in question, to be minimized: e.g. for a telecommunication network to design, this could be cost or average packet delay.) In such a case the objective function ($0_x = \text{cost}$) can be mapped to fitness form ($f(x)$) using a simple transformation of the form:

$$f(x) = C_{\text{max}} - 0_x \quad \text{when } 0_x < C_{\text{max}}$$

$$0 \quad \text{otherwise}$$

where $C_{\text{max}}$ can be set either as a great positive value or (more appropriate choice) as the maximum $0_x$ value in the current population [Gol89]. Also, there are maximization problems where the property in question (viewed as profit), when expressed in terms of the objective function, does not yield necessary positive values. In such a case, the objective function ($0_x = \text{profit}$) can be mapped to fitness form ($f(x)$) using a simple transformation of the form:

$$f(x) = C_{\text{min}} + 0_x \quad \text{when } 0_x + C_{\text{min}} > 0$$

$$0 \quad \text{otherwise}$$

209
where \( C_{\text{min}} \) can be set either as an input coefficient or (more appropriate choice) as the absolute value of the minimum \( \Theta \), value in the current population.

### D.3.2 Crossover (Recombination)

Crossover (recombination) is the operation that distinguishes GAs from algorithms such as dynamic programming. This is a fundamental GA operation that allows the exchange and combination of information encoded in different individuals. Crossover enhances the ability of GA to converge to an optimum solution since the combined effect of genetic material of fit strings, is highly likely to result in formation of individuals with higher fitness values. In the crossover cycle, two existing individuals (parents) selected by the selection operator are paired in a stochastic way (e.g. completely at random), and exchange genetic information, yielding in this way, two new individuals (offspring). The simplest recombination operator is that of one-point crossover [Gold89]. A crossover point is randomly set and the genetic information about this point is exchanged between the two parent individuals yielding two offspring individuals.

![Figure D.5 - Example of muti (2)-point crossover](image1)

Figure D.5 - Example of muti (2)-point crossover

![Figure D.6 - bit-mutation](image2)

Figure D.6 - bit-mutation

This crossover operation is not necessarily performed on all strings in the population. Instead, it is applied with a probability \( P_{\text{cross}} \), when the pairs are chosen for breeding. Typical values of \( P_{\text{cross}} \) are between 0.6 and 0.9 (the latter is the most frequently used) [DeJ90][Gre86]. This operation does not change the value of the bits. Although one-point crossover was inspired by biological processes, it has one major drawback in that certain combinations of schemata can not be combined in some situations [Mic94]. A multi-point crossover on the other hand, can overcome this problem and improve the performance of generating offspring. The process is depicted in Figure D.5. Another approach is uniform crossover where offspring are generated from the parents based on a randomly generated crossover mask. Arguments over the best crossover technique still continue. In [DeJ75] it is concluded that a two-point crossover seems to be an optimal number for multiple crossovers, however this has since been contradicted by as two-point crossover can perform poorly in situations where the population has largely converged because of reduced crossover productivity. This low-crossover productivity problem, can be solved by the proposal of reduce-surrogate crossover in [Boo87]. Since uniform crossover exchanges bits rather than segments, it can combine features regardless of their relative locations. This ability may outweigh the disadvantage of destroying building blocks and make uniform crossover a superior operator for some problems [Sys89].

The preference of which crossover techniques to use is arguable [Gold89]. When too many crossover points are used for each mating couple, the GA search becomes a random search. The meaningful short low order schemata (building blocks) are not that likely, as in the one-point crossover, to survive and evolve among the population across the generations. Therefore the application of particular crossover technique remains problem oriented and there is no unified view on this front.

### D.3.3 Mutation

This is the next step in the reproduction mechanism, after selection and crossover. It is usually implemented in parallel to the other two operations regardless of whether crossover occurs or not. Mutation has a rather secondary role in the search process, and aims at introduction of new individual configurations. Although selection and crossover manage to search effectively and sometimes recombine constructively useful information stored in individuals, some useful genetic material may still be lost. Mutation plays the role of restoring this information [Mic96][Gold89]. The bit-mutation operation (Figure D.6), is applied to each offspring individually after the crossover exercise. Mutation causes the individual genetic representation to be changed with a small probability \( P_{\text{mutation}} \). In the binary string
representation, mutation will cause a random bit to change its state (0→1 or 1→0). As in crossover, the frequency of mutation is controlled by \( P_{\text{mutation}} \), which is a percentage of the total number of bits in the string. Thus, a child string is reproduced from a single parent string. It helps GA avoid premature convergence, and forces the algorithm to search new areas of the search space. Typical values for \( P_{\text{mutation}} \) are below 0.01 [DeJ90][Gre86]. More specifically for applications with large population size (100 or more) a \( P_{\text{mutation}} = 0.001 \) or less value is proposed [Gold89]. As the mutation probability increases the random nature of the GA is intensified, and eventually it becomes a random walk.

D.3.4 Inversion/Reordering
As stated previously, the ordering of genes in a chromosome is critical. The purpose of reordering is to find gene orders which have better evolutionary potential. Inversion operates on one individual at a time. In [Gold89] for example, inversion technique is proposed, where two points are randomly selected from an individual and the part of the string between those two points is reversed.

D.3.5 Reinsertion
Following the generation of offspring, several strategies can be proposed to replace the old generation. In the case of generation replacement, the chromosomes in the current population are completely replaced by the offspring [Ger86]. Thus the population of size \( N \) will generate an equal number of offspring in this strategy. Since the best chromosome of the population may fail to reproduce in the next generation, this is usually combined with an elitist strategy, such that one or more of the best chromosomes can be copied into the next generation. The elitist strategy may speed up domination of population by a super individual, but seems to improve the performance.

D.4 Improvements over the Basic Technique
D.4.1 Probability Rates Setting
The choice of an optimal probability rate for crossover and mutation operations is an important consideration from viewpoints of both analytical and empirical investigations. The increase of crossover probability causes the recombination of building blocks to rise and at the same time, it increases the disruption of good chromosomes. On the other hand should the mutation probability increase, it would transform the genetic search into a random search, but would help re-introduce lost genetic material. In [Dav85] Davis suggests a linear variation in crossover and mutation probabilities with a decreasing crossover rate during the run while the mutation rate increases. In [Sys91] a fixed schedule is imposed for both, and in [Boo87] a dynamically variable crossover rate is proposed which is dependent on the spread of fitness. References [Dav89][Dav91] introduce modifications to the operator probabilities according to the success of generating good offspring. Despite all the various proposals in the literature, the recommendations in [Pat94][DeJ90] and [Gre86] are still the yardstick to apply.

D.4.2 Fitness Evaluation Schemes and scaling mechanisms
The objective function of a problem provides the means for evaluating the status of each individual. It takes the chromosome as input and produces a number or list of numbers (objective value, generally in least square form) as a measure of the chromosomes performance. However, its range of values varies from problem to problem. To maintain uniformity over various problem domains, objective value is re-scaled to a fitness value [Gold89][Mic94]. Except for very simple problems (where actually we don’t need a GA) simply obtaining a non-negative fitness value for each individual, prior to the selection cycle, is not enough [Gold89].

Thus apart from mechanism for forming a new population from the old one, it seems that some additional measures might be helpful in fighting problems related to the characteristic of the function being optimized. To combat these problems, there are three basic directions. One of them borrows the simulated annealing technique of varying the system’s entropy, (see e.g., [Sir87], where the authors control the rate of population convergence by thermodynamic operators, which use a global temperature parameter). Another direction is based on allocation of reproductive trials according to rank rather than actual evaluation values (as discussed in the previous section), since ranking automatically introduces a uniform scaling across the population. The last direction concentrates on trying to fix the function itself by introducing a scaling mechanism. Following [Gold89, pp. 122-124] scaling mechanisms can be divided into three categories:
1. **Linear Scaling.** Let's denote by $f_i$ the raw fitness of the population individual $i$, and by $f'_i$ its scaled fitness value. In this method the actual chromosomes' fitness is scaled as

$$f'_i = a \cdot f_i + b. \quad (D.1)$$

This mechanism, though quite powerful, can introduce negative evaluation values that must be dealt with. In addition, the parameters $a$, $b$ are normally fixed for the population life and are not problem dependent. The parameters $a$, $b$ are normally selected so that the average fitness is mapped onto itself and the best fitness is increased by a desired multiple of the average fitness, i.e. the coefficients $a$ and $b$, are chosen such that:

- the average scaled fitness, $f'_\text{avg}$, is equal to the average raw fitness, $f_\text{avg}$, to insure that each average population member contributes one expected offspring to the next generation.
- the scaled maximum fitness is $f'_\text{max} = C_{\text{mult}} \cdot f_\text{avg}$ where $C_{\text{mult}}$ is the number of expected copies desired for the best population member (a typical value is $C_{\text{mult}} = 2$).

The above two targets imply the following equations to be in effect, respectively (figure D.7):

$$f'_\text{avg} = a \cdot f_\text{avg} + b \quad (D.2)$$
$$C_{\text{mult}} \cdot f'_\text{avg} = a \cdot f'_\text{max} + b \quad (D.3)$$

To satisfy the above equations, $a$ and $b$ are set as:

$$a = f_\text{avg} \cdot (C_{\text{mult}} - 1) / (f'_\text{max} - f_\text{avg}) \quad (D.4)$$
$$b = f_\text{avg} \cdot (1 - a) \quad (D.5)$$

However, the above settings might sometime cause trouble. If in a generation a few lethals (bad strings) are found to have raw fitness far below the population average and maximum raw fitness values (figure D.8), the above linear scaling will yield negative scaled fitness values for them, which is unacceptable. In such a case, the above $a$ and $b$ values are not appropriate and we choose them such the average scaled fitness is $f'_\text{avg} = a \cdot f'_\text{avg} + b$ (again) while the minimum scaled fitness is set $f'_\text{min} = 0$. These two targets imply the following equations to be in effect, respectively (figure D.6):

$$f'_\text{avg} = a \cdot f'_\text{avg} + b \quad (D.6)$$
$$0 = a \cdot f'_\text{min} + b \quad (D.7)$$

To satisfy the above equations, $a$ and $b$ are set, this time, as:

$$a = f'_\text{avg} / (f'_\text{avg} - f'_\text{min}) \quad (D.8)$$
$$b = f'_\text{avg} \cdot (1 - a) \quad (D.9)$$

To sum up, when the above linear scaling scheme is employed in a GA application, the following steps have to be followed in each generation:

![Figure D.7 - Linear scaling under normal conditions](image1)

![Figure D.8 - Lethals exist, normal linear scaling assigns negative fitness values to them](image2)
Step 1: Compute the average raw fitness value ($f_{avg}$). Identify the maximum, and minimum raw fitness values ($f_{max}$, $f_{min}$).

Step 2: Compute $a$ and $b$ scaling coefficients according to eq.D.4 and eq.D.5, respectively.

Step 3: If for the above settings, the minimum scaled fitness is found negative ($a \times f_{min} + b < 0$), re-compute $a$ and $b$ scaling coefficients, according to eq.D.8 and eq.D.9, respectively.

Step 4: Proceed to the selection cycle.

![Figure D.9 - Alternative scaling for the case where lethals exist](image)

Sigma truncation. This method was designed as an improvement of linear scaling both to deal with negative evaluation values and to incorporate problem dependent information into the mapping itself. Here the new fitness is calculated according to:

$$f_i' = f_i + (f_{ave} - c \cdot \delta)$$

where $c$ is chosen as a small integer (usually a number from the range 1 and 5), $f_{ave}$ is the mean of the objective values and $\delta$ is the population's standard deviation; possible negative evaluations $f'$ are set to zero. Sigma truncation can be deployed prior to a linear scaling scheme to reduce the probability of having to scale according to equations eq.D.8, eq.D.9, which degrades the performance of the linear scaling scheme.

Power Law Scaling. In this method the initial fitness is taken to some specific $k^{th}$ power of the raw objective value, $f_i$:

$$f_i' = f_i^k$$

with some $k$ close to one. The parameter $k$ scales the function $f_i$; and is in general, problem dependent or even varying during the run [Gil85].

D.4.3 Enhanced Selection Schemes

To generate good offspring, efficient parent selection mechanisms are necessary. Basically there are three measures of the performance of selection algorithms: bias, spread and efficiency [Bak87]. Bias defines the absolute difference between actual and expected selection probabilities of individuals. Spread is the range in the possible number of trials that an individual may achieve. Efficiency is related to the overall time complexity of the algorithms. As mentioned previously, there are several selection schemes in the literature, including roulette wheel, tournament selection [Gold89a] and rank-based selection [Bak85][Whi89]. The roulette wheel although the most popular, has the following drawbacks:

- Roulette wheel selection tends to give zero bias but potentially inclines to spread unlimitedly.
- The fittest individual may be lost during selection, due to its stochastic nature.
- Fit individuals may be copied several times, and a fit individual may quickly dominate the population at an early stage, especially for small population sizes.
- The selection operation alone explores no new points in the search space i.e. it can not create new schemata.
Thus new and more effective selection schemes are required. One such scheme used in the hybrid GA of [Pha96] employs a new modified reproduction mechanism. Another is the Stochastic Remainder Sampling Without Replacement scheme [Gold89]. This technique combines minimum spread beyond zero bias, with low computational complexity. Stochastic universal sampling (SUS) of [Bak87] is an other single-phase sampling algorithm with minimum spread, zero bias. There are other methods that can be used such as the ranking scheme [Bak85] which introduces an alternative to proportional fitness assignment. The chromosomes are selected proportionally to their rank rather than actual evaluation values. The ranking scheme has been shown to help in the avoidance of premature convergence and to speed up the search when the population approaches convergence [Whi89].

*Stochastic Remainder Sampling Without Replacement* - In this selection scheme the expected number of trials that an individual (i) of the current population will give (expect(i)) is calculated for each individual:

\[
\text{expect}(i) = \frac{f_i}{f_{avg}}
\]

(D.12)

where \(f_i\) is the fitness value (scaled) of individual \(i\), and \(f_{avg}\) the average fitness throughout the current population. expect(i) is a decimal number. Each string \(i\) is allocated samples according to the integer part of expect(i). The fractional parts of the expected number values are treated as probabilities. One by one, weighted coin tosses (Bernoulli trials) are performed using the fractional parts as success probabilities. For example, a string with an expected number of trials equal to to 1.5 would receive a single trial surely and another (but not more than one) with probability 0.5. This process continues until the population is full.

**D.4.4 Handling Constraints**

Sometimes in an optimisation problem we have to relax one or more constraints beyond optimising a single objective. This situation can be modelled easily in a GA, using a penalty method which transforms a constrained problem into an unconstrained one [Gold89]. Consider for example, the original constrained problem in minimization form:

\[
\text{minimize } g(x) \\
\text{subject to } b_i(x) \leq 0 \text{ for all } i = 1, 2, ..., n, \text{ where } x \text{ is an } m \text{ vector.}
\]

transform this to the unconstrained form i.e.

\[
\text{minimize } g(x) + r \sum_{i=1}^{n} \Phi[b_i(x)]
\]

where \(\Phi\) -penalty function  
\(r\) -penalty coefficient

A number of alternatives exist for the penalty function \(\Phi\). Usually we square the violation of the constraint, \(\Phi[b_i(x)] = b_i^2(x)\), for all violated constraints \(i\). As a practical matter, \(r\) values in generic algorithms are often sized separately for each type of constraint so that moderate violations of the constraints yield a penalty that is some significant percentage of a nominal operating cost. During the last few years several methods have been proposed for handling constraints by genetic algorithms for numerical optimization problems. Most are based on the concept of penalty functions, which penalise infeasible solutions. However, methods differ in many important details, such as how the penalty function is designed and applied to infeasible solutions. For details refer to [Hom94], [Joi94], [Sch93], [Pow83].