Mobility Management in Next Generation All-IP Based Wireless Systems

A Thesis
Presented to
The Academic Faculty

by

Jiang (Linda) Xie

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
April 2004

Mobility Management in Next Generation All-IP Based Wireless Systems

Approved by:

Professor Ian F. Akyildiz,
Committee Chair

Professor Chin-Hui Lee

Professor Chuanyi Ji

Professor Jun Xu

Professor Raghupathy Sivakumar

Date Approved: April 1, 2004

To my parents,

Minggan Xie and Chunze Jiang.

ACKNOWLEDGEMENTS

It is my great pleasure to express my gratitude to all the people who have supported me greatly during the pursuit of my Ph.D. degree. Without their encouragement and help, it would not have been possible for me to complete this dissertation.

First of all, I would like to give my special and sincere thanks to my dissertation advisor, Dr. Ian F. Akyildiz, who has been constant in his valuable guidance and encouragement in my research. Dr. Akyildiz also helped me in any other ways that could lead me to success in my future career. He has been the source of support and inspiration for me to be a professor. I learned from him much more than what is written on this dissertation.

I would like to extend my appreciation to Dr. Chuanyi Ji, Dr. Raghupathy Sivakumar, and Dr. Chin-Hui Lee, who participated in my research proposal and examined the draft of my dissertation. I would also like to thank Dr. Jun Xu for serving on my dissertation defense committee.

I would like to thank my Master's thesis advisor, Dr. Chin-Tau Lea. He guided me into this communications and networking field, motivated my interests in doing research, and taught me how to do research.

I would also like to thank the members in the Broadband and Wireless Networking (BWN) Laboratory at the Georgia Institute of Technology. Their friendship, help, and encouragement deserve my sincere appreciation. I am also grateful to all my friends at Georgia Tech, Hong Kong, and China for their continuous support to me.

I am deeply indebted to my parents, Minggan Xie and Chunze Jiang, and my love, Zhong Chen, whose patience, tolerance, encouragement, and love have been continuously giving me the braveness to face the challenges.

TABLE OF CONTENTS

DEDICATION . . . . . **iii**

ACKNOWLEDGEMENTS . . . . . **iv**

LIST OF TABLES . . . . . **ix**

LIST OF FIGURES . . . . . **x**

SUMMARY . . . . . **xiii**

CHAPTER I INTRODUCTION **1**

1.1 Mobility Management 3

1.1.1 Location Management 3

1.1.2 Handoff Management 5

1.1.3 Mobility Management Based on Different Layers 6

1.2 Research Objectives and Solutions 7

1.2.1 Distributed Dynamic Regional Location Management Scheme for NG Wireless Internet 8

1.2.2 User Independent Paging Scheme for NG Wireless Internet . 9

1.2.3 Paging-Aided Connection Setup for Real-Time Communication in Mobile Internet 11

1.2.4 Location Management in NG Heterogeneous Wireless Overlay Networks 11

1.2.5 Handoff Management in NG Heterogeneous Wireless Overlay Networks 12

1.3 Thesis Outline 13

CHAPTER II DISTRIBUTED DYNAMIC REGIONAL LOCATION MANAGEMENT SCHEME FOR NG WIRELESS INTERNET 14

2.1 Problem and Solution 14

2.2 Distributed And Dynamic Regional Location Management 18

2.2.1 Overview of the Distributed Dynamic Scheme 18

2.2.2 Operations of the Distributed Dynamic Scheme 20

2.2.3 Comparison 21
2.3 Signaling Cost Function 22
2.3.1 Location Update Cost 23
2.3.2 Packet Delivery Cost 28
2.3.3 Total Signaling Cost 30
2.4 Optimal Regional Network Size 30
2.5 Analytical Results 32
2.5.1 Centralized Fixed Scheme vs. Distributed Fixed Scheme . . . 33
2.5.2 Distributed Fixed Scheme vs. Distributed Dynamic Scheme . 35
CHAPTER III USER INDEPENDENT PAGING SCHEME FOR NG WIRELESS INTERNET 44
3.1 Problem and Solution 44
3.2 User Independent Paging Scheme for Mobile IP 48
3.2.1 Overview of the User Independent Paging Scheme 49
3.2.2 Detailed Solution for Obtaining Subnet Mobility Rate 51
3.2.3 Advantages of the User Independent Paging Scheme 55
3.2.4 Tradeoffs of Introducing "Semi-Idle" State 57
3.3 Analytical Model 57
3.3.1 Costs of Obtaining Mobility Rates and Setting Up User Profiles 58
3.3.2 Relationship Between Location Probabilities and Mobility Rates 60
3.3.3 Paging Costs 61
3.4 Performance Evaluation 63
3.4.1 Paging A Single User 64
3.4.2 Paging Multiple Users 71
3.4.3 Confidence of Location Probabilities 73
CHAPTER IV PAGING-AIDED CONNECTION SETUP FOR REAL-TIME COMMUNICATION IN MOBILE INTERNET 77
4.1 Problem and Solution 77
4.2 Background 80

4.2.1 Mobile IP Paging 80
4.2.2 RSVP Signaling 80
4.3 Paging-Aided Connection Setup for Real-time Communication . . . 82
4.3.1 Unicast Case 82
4.3.2 Multicast Case 85
4.4 Performance Evaluation 86
CHAPTER V LOCATION MANAGEMENT IN NG HETEROGENEOUS WIRELESS OVERLAY NETWORKS 89
5.1 Problem and Solution 89
5.2 System Architecture and Problem Formulation 93
5.2.1 System Architecture 93
5.2.2 Problem Formulation 95
5.3 Proposed Location Management Techniques 96
5.3.1 User Preference Call Delivery 97
5.3.2 The Proposed Location Management Schemes 102
5.4 Performance Analysis 108
5.4.1 Total Signaling Cost 109
5.4.2 Numerical Results 111
5.5 Threshold-Based Adjustable Registration 123
CHAPTER VI HANDOFF MANAGEMENT IN NG HETEROGENEOUS WIRELESS OVERLAY NETWORKS 126
6.1 Problem and Solution 126
6.2 System Model 129
6.3 Hybrid Control Scheme for Resource Allocation 130
6.4 Cost Function 132
6.4.1 Cost Function of the Terminal-Based Selection Mechanism . 133
6.4.2 Cost Function of the Network-Based Selection Mechanism . . 134
6.5 Optimization Solution and Adjustment 136
6.5.1 Optimization Solution 136

6.5.2 Adjustment Policy 138
6.6 Numerical Results 141
6.6.1 System and User Parameters 141
6.6.2 Iteration Procedure 142
6.6.3 Hybrid Control Scheme vs. Distributed Scheme 143
CHAPTER VII CONCLUSIONS AND FUTURE RESEARCH WORK 147
7.1 Summary of Research Results 147
7.1.1 Location management in NG wireless Internet 147
7.1.2 Paging in NG wireless Internet 148
7.1.3 Paging-aided connection setup in NG wireless Internet 149
7.1.4 Location management in NG wireless overlay networks 149
7.1.5 Handoff management in NG wireless overlay networks 150
7.2 Future Research Work 151
7.2.1 QoS Issues 151
7.2.2 Location and Handoff Management in Overlay Networks . . . 152
7.2.3 Cross Layer Optimization 152
REFERENCES 154
VITA 162

LIST OF TABLES

Table 1 Performance Analysis Parameters for Location Management in Mobile IP . . . 33

Table 2 Cost Parameters for Evaluating Paging Schemes in Mobile IP . . . 64

Table 3 Performance Analysis of Paging-Aided Connection Setup Scheme . 87

Table 4 Selected Data Sets For c_i . . . 113

Table 5 Selected Data Sets For α_i . . . 117

Table 6 Selected Data Sets For ϕ_i . . . 120

Table 7 Initial User Distribution for the Hybrid Control Resource Allocation Scheme . . . 141

Table 8 System Parameters for the Hybrid Control Resource Allocation Scheme 142

Table 9 Iteration Procedure for Finding the Optimal User Distribution . . . 145

Table 10 Optimal User Distribution for the Hybrid Control Resource Allocation Scheme . . . 146

Table 11 Adjustment Number for Each Network . . . 146

Table 12 New User Distribution after Applying the Hybrid Control Scheme . 146

Table 13 New User Distribution without Central Control . . . 146

LIST OF FIGURES

Figure 1 Next generation heterogeneous wireless overlay networks. 2

Figure 2 Mobility management in NG wireless systems. 4

Figure 3 Research topics on mobility management in NG all-IP based wireless systems. 8

Figure 4 The IETF Mobile IP regional registration. 15

Figure 5 The distributed dynamic Mobile IP regional registration. 19

Figure 6 Protocols of the distributed dynamic scheme for MNs. 21

Figure 7 The distributed fixed Mobile IP regional registration. 22

Figure 8 Process of home location registration. 23

Figure 9 Process of regional location registration. 24

Figure 10 Discrete system mobility model of an MN. 26

Figure 11 Optimal regional network size for centralized and distributed systems. 34

Figure 12 Comparison of total signaling cost for fixed schemes. 35

Figure 13 Comparison of total signaling cost under user-variant residence time. 37

Figure 14 Comparison of total signaling cost under time-variant residence time. 39

Figure 15 Comparison of total signaling cost under time-variant residence time. 40

Figure 16 Comparison of total signaling cost under user-variant packet arrival rate. 41

Figure 17 Comparison of total signaling cost under time-variant packet arrival rate. 42

Figure 18 Comparison of total signaling cost under time-variant packet arrival rate. 43

Figure 19 State transition diagram of Mobile IP paging. 47

Figure 20 User independent paging scheme based on location and mobility rate. 50

Figure 21 State transition diagram of the proposed paging scheme. 52

Figure 22 Flowchart of operations at FAs. 54

Figure 23 The average paging cost for uniform location probability distribution, when paging a single user. 66

Figure 24 The average paging cost for shifted truncated Gaussian location probability distribution, when paging a single user. 67

Figure 25 The average paging cost for variant truncated Gaussian location probability distribution, when paging a single user. 68

Figure 26 The average paging cost for user-variant location probability distribution, when paging a single user. 70

Figure 27 The average paging cost for user-variant location probability distribution, when paging multiple users, $\Omega = 4$. 73

Figure 28 The average paging cost for user-variant location probability distribution, when paging multiple users, $\Omega = 8$. 74

Figure 29 The average paging cost when imperfect location information is used, $\alpha = 80\%$. 75

Figure 30 The average paging cost when imperfect location information is used, $\alpha = 90\%$. 76

Figure 31 Procedure of location registration. 83

Figure 32 Procedure of sending paging request and RSVP PATH messages. . 84

Figure 33 Procedure of sending paging reply and RSVP RESV messages. . . . 85

Figure 34 Flow chart for multicast case. 86

Figure 35 The proposed architecture for mobility management in heterogeneous overlay networks. 94

Figure 36 Call delivery procedure when the registration network is the call delivery network. 98

Figure 37 Call delivery procedure when the registration network is not the call delivery network. 100

Figure 38 Network switching procedure. 103

Figure 39 Procedures of registration network selection for the CHAR scheme. 107

Figure 40 Comparison of the total signaling cost for the LATR scheme and the PPR scheme under different sets of c_i. 114

Figure 41 Comparison of the total signaling cost for the LATR scheme and the CHAR scheme under different sets of c_i. 115

Figure 42 Comparison of the total signaling cost for the CHAR scheme and the PPR scheme under different sets of c_i. 116

Figure 43 Comparison of the total signaling cost for the LATR scheme and the PPR scheme under different sets of α_i. 117

Figure 44 Comparison of the total signaling cost for the LATR scheme and the CHAR scheme under different sets of α_i. 118

Figure 45 Comparison of the total signaling cost for the CHAR scheme and the PPR scheme under different sets of α_i. 119

Figure 46 Comparison of the total signaling cost for the LATR scheme and the PPR scheme under different sets of ϕ_i. 120

Figure 47 Comparison of the total signaling cost for the LATR scheme and the CHAR scheme under different sets of ϕ_i. 121

Figure 48 Comparison of the total signaling cost for the CHAR scheme and the PPR scheme under different sets of ϕ_i. 122

Figure 49 Operation procedures of the THAR scheme. 124

Figure 50 Iterative algorithm for finding the optimal user distribution. 138

Figure 51 Relationship between the total cost and the user distribution. . . . 139

SUMMARY

Next generation wireless systems have an IP-based infrastructure with the support of heterogeneous access technologies. One research challenge for next generation all-IP based wireless systems is to design intelligent mobility management techniques that take advantage of IP-based technologies to achieve global roaming between various access networks. To support global roaming, next generation wireless systems require the integration and interoperation of heterogeneous mobility management techniques. Mobility in a hierarchical structure or multilayered environment should be supported. The objective of this study is to develop new mobility management techniques for global roaming support in next generation all-IP based wireless systems. More specifically, new schemes for location management and paging in Mobile IP for network layer mobility support, and new schemes for location management and handoff management in heterogeneous overlay networks for link layer mobility support are proposed and evaluated. For network layer mobility support, a distributed and dynamic regional location management mechanism for Mobile IP is proposed. Under the proposed scheme, the signaling burden is evenly distributed and the regional network boundary is dynamically adjusted according to the up-to-date mobility and traffic load for each terminal. Next, a user independent paging scheme based on last-known location and mobility rate information for Mobile IP is proposed. The proposed scheme takes the aggregated behavior of all mobile users as the basis for paging. For link layer mobility support, an IP-based system architecture for the integration of heterogeneous mobility management techniques is proposed. Three location management schemes under this IP-based architecture are proposed. All the three schemes support user preference call delivery which is a very important feature of

next generation wireless communications. A threshold-based enhancement method is also proposed to further improve the system performance. Finally, a hybrid resource allocation scheme for handoff management in wireless overlay networks is proposed. Under this scheme, the overall system resources can be optimally allocated when mobile users are covered by multiple overlay networks.

CHAPTER I

INTRODUCTION

Currently, various wireless technologies and networks are existing which capture different needs and requirements of mobile users. For high data rate local area access, wireless LANs (WLANs) [1] are satisfactory solutions. For wide area communications, traditional and next generation (NG) cellular networks may provide voice and data services. For worldwide coverage, satellite networks have been used extensively in military and commercial applications. Since these existing different wireless networks are complementary to each other, their integration will empower mobile users to be "always best connected" [2] by using the best available access network that suits their needs. The integration of different networks generate several research challenges because of the following heterogeneities [3]:

- **Access Technologies:** NG wireless systems will include many heterogeneous networks using different radio technologies. These networks may have overlapping coverage areas and different cell sizes ranging from a few square meters to hundreds of square kilometers, as shown in Figure 1.

- **Network Architectures and Protocols:** NG wireless systems will have different network architectures and protocols for transport, routing, mobility management, etc.

- **Service Demands:** Mobile users demand different services ranging from low data rate non-real-time applications to high speed real-time multimedia applications, offered by various access networks.

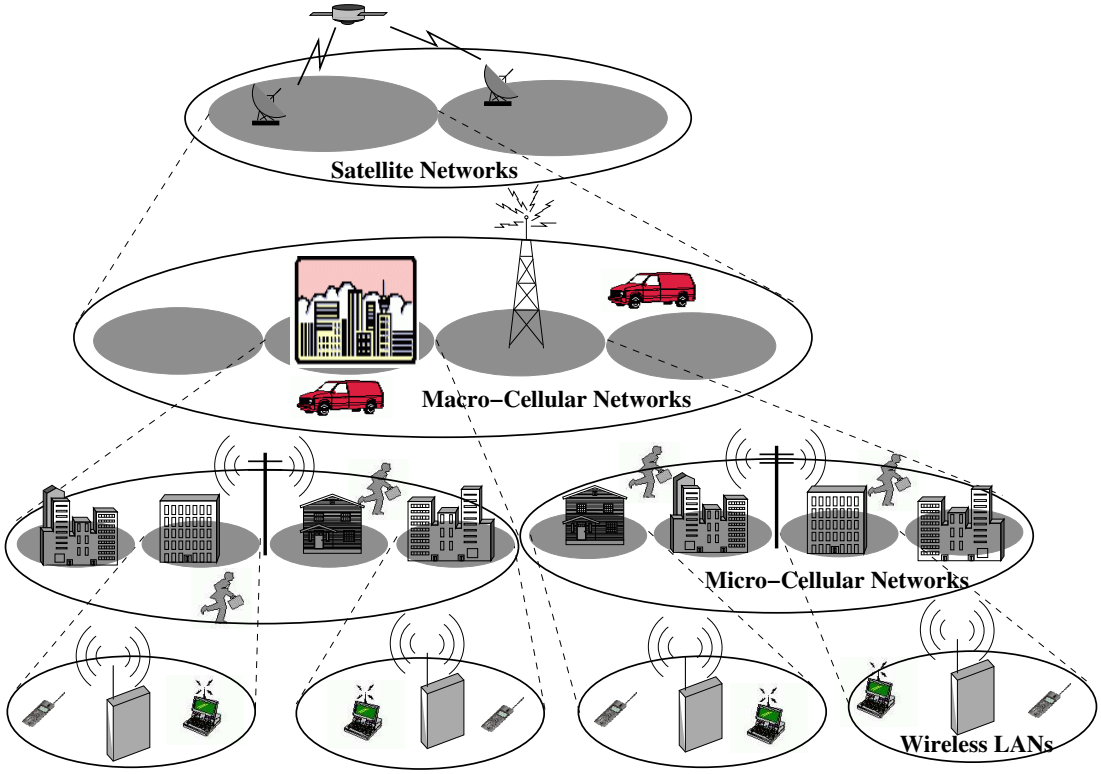

Figure 1: Next generation heterogeneous wireless overlay networks.

The above intrinsic technology heterogeneities ask for a common infrastructure to interconnect multiple access technologies. IP (Internet Protocol) is recognized to become the core part of the NG integrated wireless systems to support ubiquitous communications [4]. For inter-operation of different communication protocols, an adaptive protocol suite is required that will adapt itself to the characteristics of the underlying network, and provide optimal performance across a variety of wireless network environments. Furthermore, adaptive terminals in conjunction with "smart" base stations will support multiple air interfaces and will allow users to seamlessly switch between different access technologies.

One important component of the adaptive protocol suite is the interoperation of mobility management schemes. In this paper, we address the design of intelligent mobility management techniques that take advantage of IP-based technologies to achieve global roaming between heterogeneous networks to satisfy the service and

connection requirements of mobile users [4] [5]. To make this roaming seamless, the integration and interoperation of heterogeneous mobility management techniques [6] with efficient support for both inter-domain and intra-domain mobility management is required. The existing mobility management techniques try to reduce the delay associated with intra-domain mobility management [7] [8] [9]. However, these solutions have high signaling load and handoff delay for inter-domain mobility management, which should be reduced for seamless mobility between different domains. Therefore, we advocate there is a need for new mobility management architecture for heterogeneous environment to reduce both intra-domain and inter-domain signaling load and handoff delay.

1.1 Mobility Management

Mobility management contains two components: location management and handoff management [6]. In the NG wireless systems, there are two types of roaming for mobile terminals (MTs): *intra-system* (intra-domain) roaming and *inter-system* (inter-domain) roaming. Intra-system roaming refers to MTs that move between different cells of the same system. Intra-system mobility management techniques are based on similar network interfaces and protocols. Inter-system roaming refers to MTs that move between different backbones, protocols, technologies, or service providers. Based on intra-system or inter-system roaming, the corresponding location management and handoff management can be further classified into intra- and inter- system location management and handoff management, as shown in Figure 2.

1.1.1 Location Management

Location management enables the system to track the locations of MTs between consecutive communications. It includes two major tasks. The first is the *location registration* or *location update*, where the MT periodically informs the system to update relevant location databases with its up-to-date location information. The

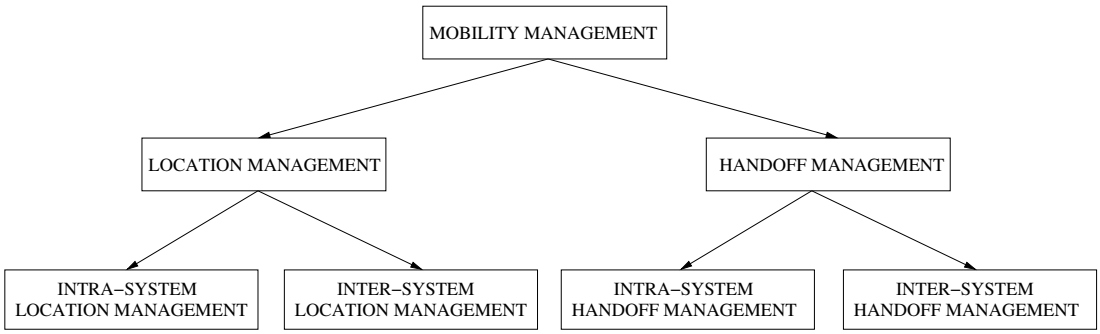

Figure 2: Mobility management in NG wireless systems.

second is the *call delivery*, where the system determines the current location of the MT based on the information available at the system databases when a communication for the MT is initiated. Two major steps are involved in call delivery: determining the serving database of the called MT, and locating the visiting cell/subnet of the called MT. The latter one is also called *paging* procedure where polling messages are sent to all cells/subnets within the residing registration area (RA) of the called MT. Location management is critical to the mobility support and QoS provisioning. In heterogeneous wireless systems, roaming across systems imposes a significant increase in signaling traffic load and call delivery delay [10]. For inter-system roaming, location management techniques have the following challenges:

- Reduction of signaling overheads and latency of service delivery;
- QoS guarantees in different systems;
- If the service areas of heterogeneous wireless networks are fully overlapped, additional issues need to be addressed are:
  - Through which networks an MT should perform location registrations;
  - In which networks and how the up-to-date user location information should be stored;
  - How would the exact location of an MT be determined within a specific time constraint.

1.1.2 Handoff Management

Handoff management is the process by which an MT keeps its connection active when it moves from one access point to another one. The entire handoff process can be sub-divided into four phases: *Initiation, Preparation, Start, and Completion* [5]. In the *Initiation* phase, either the MT or the network identify the need for handoff. Once initiated, both the network and the MT are responsible for the *Preparation* phase, where the new resources are allocated and the rerouting of the ongoing communications is performed. In case of inter-system handoff, the format transformation is also carried during this phase. Format transformation configures the MT with the protocol stack of the new system. Finally, the network decides when to begin the handoff and executes the *Start* phase, which is followed by the *Completion* phase.

Handoff process can be intra-system or inter-system. Intra-system handoff is the handoff in homogeneous networks. The need for intra-system handoff (or horizontal handoff) arises when the signal strength of the serving base station (BS) detoriates below a certain threshold value and there exists a neighboring BS whose signal strength at the MT is above the threshold level. The need for inter-system handoff (or vertical handoff) between heterogeneous networks may arise in the following scenarios [5]:

- When a user is going to move out of the serving network and will enter another overlaying network shortly.
- When a user is connected to a particular network, but chooses to be handed off to the underlying or overlaying network for its future service needs.
- To evenly distribute the overall network load between different systems. This will optimize the capacity and performance of each individual networks.

Handoff management in NG all-IP based wireless systems has the following challenges:

- Reduction of both signaling and power overheads.
- QoS guarantees during the handoff process:
  - Extreme low intra-system and inter-system handoff latency which includes signaling message processing time, resources and routes set up delay, format transformation time, etc.
  - Limited disruption to user traffic.
  - Near-zero handoff failure and packet loss rates.
- Efficient use of network resources.
- Enhanced scalability, reliability, and robustness.

1.1.3 Mobility Management Based on Different Layers

Mobility management techniques in homogeneous networks have been comprehensively surveyed in [6]. In this study, we focus on mobility management techniques in heterogeneous wireless networks. Several protocols are proposed for NG all-IP based wireless systems. These protocols try to provide mobility management from different layers of TCP/IP protocol stack reference model. We classify these mobility management solutions into the following categories [11]:

- Network layer solutions (layer 3 solutions)
- Link layer solutions (layer 2 solutions)
- Cross layer solutions (layer 3 + layer 2 solutions)

Network layer solutions provide mobility-related features at the IP layer. They do not rely on or make any assumption on the underlying wireless access technologies [8]. IP layer (layer 3) location area is defined as a set of IP network attachment points identified by one or more IP addresses [12]. Signaling messages for mobility purpose

are carried by IP traffic. Link layer solutions provide mobility-related features in the underlying radio systems. They ensure uninterrupted communications when MTs change position within the scope of an access router. Additional gateways are usually proposed to handle the interworking and interoperating issues when roaming between heterogeneous access networks. Signaling messages are transmitted through wireless links. Link layer solutions are tightly coupled with the specific wireless technologies. Mobility supported by link layer is also called *access mobility* or *link-layer mobility* [4]. Cross layer solutions are mainly proposed for handoff management techniques. They aim to achieve layer 3 handoff with the help from layer 2. By obtaining signal strength reports and movement detection information from link layer in advance, the system can make better preparation for network layer handoff so that the packet loss is eliminated and the handoff latency is reduced.

1.2 Research Objectives and Solutions

Mobility management is critical to the global roaming support and QoS provisioning guarantee in integrated all-IP based wireless systems. Mobility support can be provided from different layers of TCP/IP protocol stack reference model. The objective of this research is to develop new mobility management techniques for global roaming support in next generation all-IP based wireless systems. More specifically, new schemes for location management and paging in Mobile IP for network layer mobility support, and new schemes for location management and handoff management in heterogeneous overlay networks for link layer mobility support are proposed and evaluated.

In this proposal, four research topics are investigated. Since Mobile IP is a network-layer mobility solution for the global Internet, two topics are investigated in the Mobile IP environment: location management and paging in Mobile IP supporting network layer mobility. Link-layer mobility solutions usually handle interworking

and interoperating issues when mobile terminals have inter-system roaming between heterogeneous access networks. The two topics are investigated in the wireless overlay network environment: location management and handoff management in wireless overlay networks supporting link layer mobility. All these solutions compose of a set of mobility management solutions for NG all-IP based wireless systems and they provide the global roaming feature to mobile terminals in the integrated system. The research work of this thesis is summarized in Figure 3.

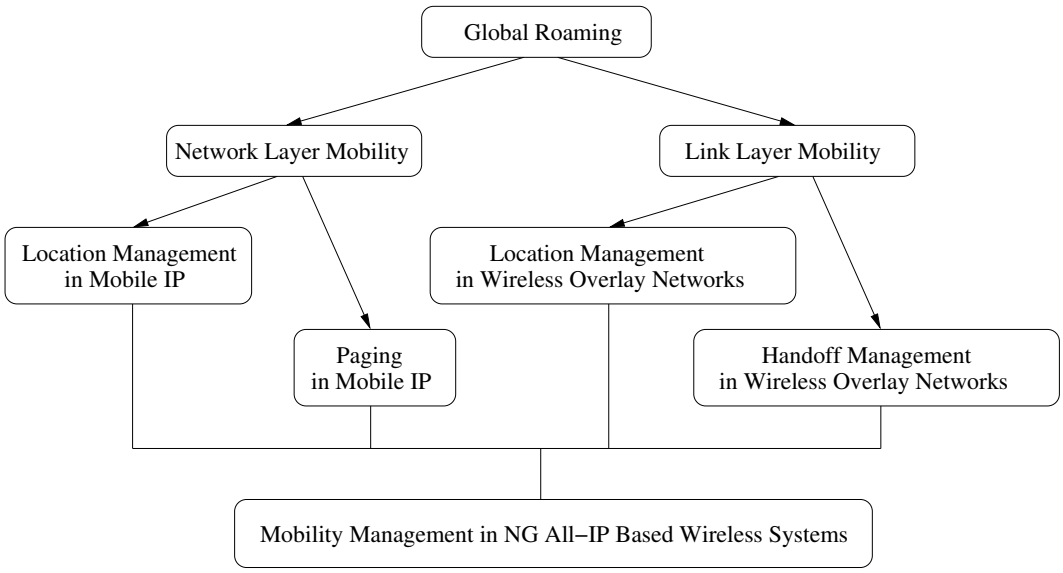

Figure 3: Research topics on mobility management in NG all-IP based wireless systems.

1.2.1 Distributed Dynamic Regional Location Management Scheme for NG Wireless Internet

Several network layer mobility management solutions for NG all-IP based wireless systems are proposed. These solutions can be broadly classified into two categories: *macro mobility* management solutions and *micro mobility* management solutions. Mobile IP is a mobility-enabling protocol for the global Internet. Standards for Mobile IP have been developed by the Internet Engineering Task Force (IETF) and outlined

in Request for Comments (RFC) 3220 [13]. However, Mobile IP leads to high signaling overheads and long signaling delay to the home network. To reduce the signaling load and delay during the movement of users between subnets of a particular domain, many micro-mobility solutions have been proposed. These can be broadly classified into two groups: *tunnel-based* and *routing-based* micro-mobility schemes [14]: Mobile IP regional registration (MIP-RR) [15], hierarchical Mobile IP (HMIP) [16], and intra-domain mobility management protocol (IDMP) [8] are tunnel-based micro-mobility protocols, while Cellular IP (CIP) [7] and handoff-aware wireless access Internet infrastructure (HAWAII) [9] are routing-based micro-mobility protocols.

In this thesis, a novel distributed and dynamic regional location management for Mobile IP is introduced where the signaling burden is evenly distributed and the regional network boundary is dynamically adjusted according to the up-to-date mobility and traffic load for each terminal. In our distributed system, each user has its own optimized system configuration which results in the minimal signaling traffic. In order to determine the signaling cost function, a new discrete analytical model is developed which captures the mobility and packet arrival pattern of a mobile terminal. This model does not impose any restrictions on the shape and the geographic location of subnets in the Internet. Given the average total location update and packet delivery cost, an iterative algorithm is then used to determine the optimal regional network size. Analytical results show that our distributed dynamic scheme outperforms the IETF Mobile IP regional registration scheme for various scenarios in terms of reducing the overall signaling cost.

1.2.2 User Independent Paging Scheme for NG Wireless Internet

A major problem of MNs is their limited battery capacity. In order to save the battery power consumption at MNs, IP paging is proposed as an extension for Mobile IP [17] [18] [19] [20] [21]. Under Mobile IP paging, an MN is allowed to enter a power saving

idle mode when it is inactive for a period of time. During the idle mode, the system knows the location of the MN with coarse accuracy defined by a paging area which is composed of several subnets [17]. Currently, there are three major paging protocols proposed for Mobile IP: *home agent paging* [18], *foreign agent paging* [19] [20], and *Domain paging* [18]. The differences among these paging schemes are which node initiates paging and how the messages exchange between nodes.

Multi-step paging has been widely proposed in personal communications services (PCS) system to reduce the signaling overheads. Similar ideas can be applied to Mobile IP to provide IP paging services. However, current proposed multi-step paging schemes are user dependent under which the partition of paging areas and the selection of paging sequence are different for each user. The performance of a user dependent paging scheme for individual users may be affected by many factors. It is often difficult to achieve perfect performance for each user. In addition, when multiple users are paged at the same time, user dependent paging schemes may consume significant system resources.

In this thesis, a user independent paging scheme is introduced where the paging criterion is not based on individual user information. The goal of user independent paging is to provide satisfactory overall performance of the whole system, when personalized optimal performance for each user is hard to obtain. The user independent paging scheme is proposed for IP mobility for its easy implementation and convenient combination with paging request aggregation. The paging criterion adopted is the mobility rate of each subnet determined by the aggregated movements of all mobile users. In order to implement the proposed scheme, a concept of "semi-idle state" is introduced and the detailed solution for obtaining mobility rate is presented. Analytical results show that when paging one user at a time, the performance of the proposed user independent paging scheme is comparable to that of the paging schemes based on perfect knowledge of user movement statistics. When paging multiple users

simultaneously and when the knowledge on individual user behavior is not perfectly accurate, the proposed scheme has remarkable advantages in terms of reducing the overall paging cost.

1.2.3 Paging-Aided Connection Setup for Real-Time Communication in Mobile Internet

Mobile IP is a solution for mobility on the global Internet. However, the basic Mobile IP does not support paging. The main benefit of providing paging services is to save the battery power consumption at mobile terminals. Next generation Internet is expected to support multimedia communications. For real-time data traffic, Quality of Service (QoS) provision must be guaranteed. The Resource Reservation Protocol (RSVP) was proposed to support the signaling of end-to-end IP QoS. When both IP paging and RSVP are supported in the network, the signaling delay for connection setup is the sum of the paging delay and the time for RSVP path setup.

In this thesis, a new scheme for fast connection setup of real-time communication in Mobile Internet is introduced. The connection is set up with the help of Mobile IP location registration and paging. Performance analysis shows that the proposed scheme reduces the overall signaling delay and the total number of signaling messages.

1.2.4 Location Management in NG Heterogeneous Wireless Overlay Networks

NG wireless system calls for the integration and interoperation of heterogeneous mobility management techniques. When the service areas of heterogeneous networks overlap, new challenges for intelligent location management techniques arise. Multi-tier wireless systems are recognized as an efficient way to improve the capacity and quality of mobile services. The objective is to integrate the higher tier and lower tier systems into a single system to provide the advantages of all tiers in an integrated manner.

In this thesis, a new mobility management architecture for heterogeneous overlay

networks is introduced where various signaling control entities are connected to each other through Internet. Three location management techniques are proposed under this architecture. All the three schemes support user preference call delivery which is an important feature for NG multimedia communications. Lowest Available Tier Registration (LATR) scheme adopts the lowest tier network for location registration. When calls are delivered from variety of networks with balanced amount, the signaling cost is low under this scheme because of the low access cost of the lowest tier network. Under user preference scenario, majority of calls can be expected to come from a specific network. A-Posteriori Probability-based Registration (PPR) scheme can reduce the signaling cost significantly as it chooses the network with most call arrivals for location registration. However, the PPR scheme is not practical in reality because it requires a-posteriori knowledge. Call History-based Adaptive Registration (CHAR) scheme is a feasible solution for the PPR scheme. Under the CHAR scheme, mobile terminals dynamically changes the registration network according to communication histories. Numerical results show that the CHAR scheme maintains the main features of the PPR scheme and may lead to even better performance. Finally, a threshold-based enhancement is proposed for the system to dynamically switch between the LATR and the CHAR schemes to further improve the system performance.

1.2.5 Handoff Management in NG Heterogeneous Wireless Overlay Networks

Handoff management in wireless overlay networks is addressed in [22] [23] [24]. Vertical handoffs in wireless overlay networks is designed in [22] where heterogeneous networks in a hierarchical structure has fully overlapping service areas. Vertical handoff is defined as handoff between BSs that are using different wireless network technologies. Rather than depending on network-specific channel measurements to predict disconnections, the proposed scheme depends on higher-order information such as the presence or absence of beacon and data packets. A policy-enabled handoff system

in wireless overlay networks is later proposed in [23]. It allows users to issue policies and have their mobile devices connected to the most desirable network to them. A performance reporting scheme is designed for the policy-enabled handoff system to estimate current network conditions which serves as input to the policy specification. The goal of the proposed scheme is to make it possible to balance the bandwidth load across networks with comparable performance.

In this thesis, a novel resource management scheme for vertical handoff in wireless overlay networks is introduced where the overall system resources can be allocated in an economical way. This scheme includes a set of access selection criteria and mechanisms that allow mobile terminals to connect to various services through multiple access networks optimally. It is a hybrid control scheme that combines terminal-based and network-based selection mechanisms. An analytical model is also developed to solve the optimal resource allocation problem in the vertical roaming scenario.

1.3 Thesis Outline

This thesis is organized as follows. In Chapter 2, a distributed dynamic regional location management scheme for NG wireless Internet is proposed. In Chapter 3, a user independent paging scheme for NG wireless Internet is presented, followed in Chapter 4, a paging-aided connection setup scheme for real-time communications in mobile Internet is described. In Chapter 5, the mobility management architecture for the integration of heterogeneous wireless networks and three location management schemes under this architecture are presented. Another enhancement method based on the performance of the three proposed location management schemes is also given. In Chapter 6, a new resource allocation scheme for vertical handoff in NG heterogeneous wireless overlay networks is proposed. Finally, the research work of this thesis is summarized in Chapter 7 and future work is also pointed out.

CHAPTER II

DISTRIBUTED DYNAMIC REGIONAL LOCATION MANAGEMENT SCHEME FOR NG WIRELESS INTERNET

2.1 Problem and Solution

The growth of the Internet and the success of mobile wireless networks lead to an increasing demand for mobile wireless access to Internet applications. Mobile IP is a mobility-enabling protocol for the global Internet. Standards for Mobile IP have been developed by the Internet Engineering Task Force (IETF) and outlined in Request for Comments (RFC) 3220 [13] [25].

Mobile IP enables terminals to maintain all ongoing communications while moving from one subnet to another. It is a simple and scalable global mobility solution. However, it is not a satisfactory solution for highly mobile users [26]. When a mobile node (MN) moves among subnets, its location and routes must be updated. Mobile IP requires that an MN sends a location update to its home agent (HA) whenever it moves from one subnet to another one. This location registration is required even though the MN does not communicate with others while moving. The signaling cost associated with location updates may become very significant as the number of MNs increases [27]. Moreover, if the distance between the visited network and the home network of the MN is large, the signaling delay for the location registration is long.

Mobile IP regional registration aims to reduce the number of signaling messages to the home network, and also to reduce the signaling delay when an MN moves from one subnet to another. The detailed protocol specification can be found in [15] and

the general model of operation is illustrated in Figure 4. Regional registration is a solution for performing registrations locally in a regional network. When an MN first arrives at a regional network, it performs a home registration with its HA. During the home registration, the HA registers the care-of address of the MN, which is actually a publicly routable address of another mobility agent called gateway foreign agent (GFA). When an MN changes foreign agent (FA) within the same regional network, it performs a regional registration to the GFA to update its FA care-of address. When it moves from one regional network to another one, it performs a home registration with its HA. During the communication, when packets are sent to the MN by a correspondent node (CN), they are addressed to the HA of the MN first. The HA intercepts these packets and encapsulates them inside packets that are addressed to the care-of address of the MN. These packets are tunneled through the network until they reach the registered GFA of the MN. The GFA checks its visitor list and forwards the packets to the corresponding FA in the visiting subnet of the MN. The FA further relays the packets to the MN.

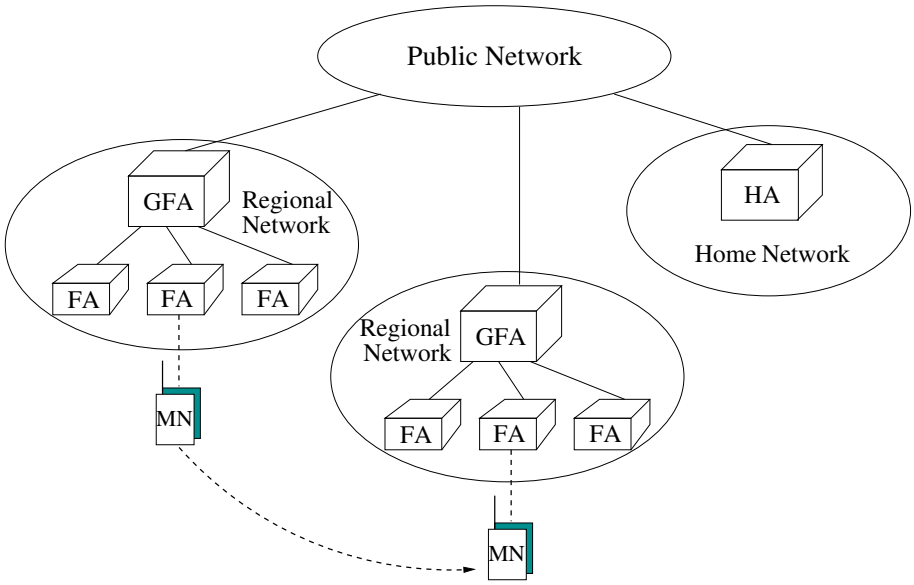

Figure 4: The IETF Mobile IP regional registration.

However, because of the centralized system architecture, i.e., a centralized GFA

manages all the traffic within a regional network, Mobile IP regional registration is more sensitive to the failure of GFAs. The failure of a GFA will prevent packets routed to all the users in the regional network [28]. Another issue that draws our attention is how many FAs should be beneath a GFA within a regional network. The number of FAs under a GFA is very critical for the system performance. A small number of FAs will lead to excessive location updates to the home network and consequently cannot provide the full benefit of regional registration. A large number of FAs will also degrade the overall performance since it will generate a high traffic load on GFAs, which results in a high cost of packet delivery [27].

To improve the system performance, we propose a distributed GFA management scheme where each FA can function either as an FA or a GFA. Whether an agent should act as an FA or a GFA depends on the user mobility. Thus, the traffic load in a regional network is evenly distributed to each FA. Through this approach, the system robustness is enhanced. We also propose a dynamic scheme which is able to adjust the number of FAs under a GFA for each MN according to the user-variant and time-variant user parameters. In this dynamic system, there is no fixed regional network boundary for each MN. An MN decides when to perform a home location update according to its changing mobility and packet arrival pattern.

In order to minimize the signaling traffic, it is desirable to find the optimal number of FAs beneath a GFA in a regional network. This optimal number is user-variant and time-variant. A method for calculating the optimal location area (LA) size in personal communication service (PCS) systems to reach the minimal costs for location update and terminal paging is introduced in [29]. However, there are some differences between the analysis of location management schemes for Mobile IP and those in PCS. First, the cellular network is geographic-oriented. Most researchers adopted structured cell configurations for evaluations [30]. For example, mesh or hexagonal cell configurations are often used in two-dimensional models [31] [32]. But Internet

is more spatial-oriented. We cannot use any geometric shape to accurately abstract a subnet, which increases the difficulty for analysis. Second, in PCS, the geographic distance between two cells is used for analysis [33]. However, the distance between two end points in Internet has nothing to do with the geographic location of these two points. Their distance is usually counted by the number of hops packets travel. This type of distance is called "*virtual*" distance. Third, when an incoming call arrives, the cellular network locates the terminal by simultaneously paging all cells within an LA. Whereas in Mobile IP, HAs or GFAs know the corresponding FA of each MN. But because of the triangular routing, packet delivery introduces extra processing and transmission costs. So there is packet delivery cost instead of paging cost for Mobile IP.

In this chapter, we also introduce a new mathematical model to calculate the optimal number of FAs under a GFA such that the total signaling traffic for location update and packet delivery consumes the minimal network resource. This model does not impose any restrictions on the shape and the geography of system topology. It is a general model which is applicable for all types of subnets. The distance unit in our model is the number of hops packets travel. Based on this model, we obtain the average location update and packet delivery costs. We use an iterative method to determine the optimal number of FAs under a GFA that will result in the minimal average signaling cost. We then incorporate this optimal value to our distributed and dynamic scheme to further enhance the system performance.

The proposed mathematical model was first introduced in [34]. The distributed dynamic regional location management scheme was proposed in [35], and later revised in [36]. This chapter is organized as follows. In Section 2.2, the distributed dynamic regional location management scheme is explained and the protocol for operating the scheme is given. Then, in Section 2.3, the mobility model is described and a method for deriving the total location update and packet delivery cost is introduced. After

that, in Section 2.4, an algorithm for obtaining the optimal number of FAs beneath a GFA is provided. In Section 2.5, analytical results are presented.

2.2 Distributed And Dynamic Regional Location Management

In this section, we introduce our distributed dynamic regional location management scheme. We also present the operational protocols of our distributed dynamic scheme. In the following discussion, we assume that the regional registration protocol supports one level of foreign agent hierarchy beneath the GFA.

2.2.1 Overview of the Distributed Dynamic Scheme

We propose a new distributed system architecture where each FA can function either as an FA or a GFA. Whether an agent should act as an FA or a GFA depends on the user mobility. When an MN enters a regional network, the first FA of the subnet the MN visits will function as the GFA of this regional network. If an agent acts as a GFA, it needs to maintain a visitor list and keeps entries in the list updated according to the regional registration requests sent from other FAs within the regional network. The GFA also relays all the home registration requests to the HA. Other agents in the regional network act as the general foreign agents for the MN. Of course, there should be some authentication setup between mobility agents to guarantee the security of message delivery.

We also propose a dynamic location management mechanism. In this scheme, the number of FAs under a GFA is not fixed but optimized for each MN to minimize the total signaling traffic. The optimal number is obtained based on the incoming packet arrival rate and mobility characteristics of each user. Since the mobility and the packet arrival rate of each user are different and they may also not be constant from time to time, the optimal number of FAs is different for each user and it is adjustable from time to time. Thus, the dynamic system is able to perform optimally

for all users.

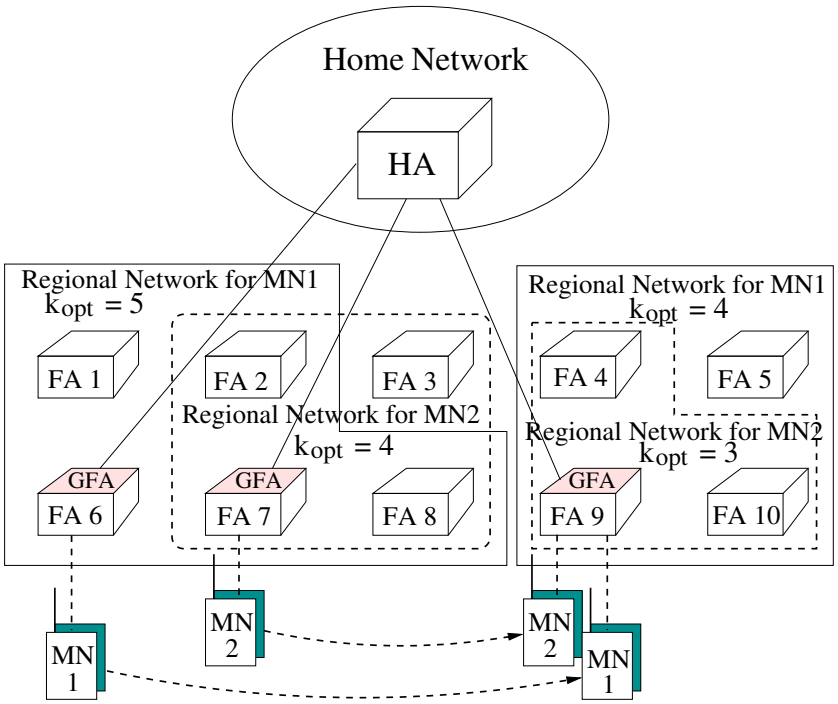

Figure 5: The distributed dynamic Mobile IP regional registration.

The system architecture of our new scheme is shown in Figure 5 where FA6 functions as the GFA for MN1 at first. The optimal regional network size is equal to 5. After visiting 5 *different* subnets, i.e., subnets served by FA1, FA2, FA6, FA7, and FA8, MN1 moves to FA9 and FA9 becomes the GFA in the new regional network for MN1. Then, MN1 updates the new optimal regional network size based on its up-to-date mobility and traffic load values. Similar for MN2, FA7 functions as the GFA in the regional network at first. The optimal regional network size for MN2 is 4. After visiting subnets served by FA2, FA3, FA7, and FA8, MN2 moves FA9 and FA9 becomes the GFA in the new regional network for MN2 also. MN2 adjusts its regional network size and this optimal size will be dynamically changed each time MN2 moves into a new regional network.

Therefore, in our distributed and dynamic system, each user has different network configuration with others: different mobility agents act as the GFA for each user and

different size of a regional network in terms of the number of FAs. The advantages of this distributed dynamic system are: the traffic load for all the users in a regional network is distributed to each mobility agent; the system robustness is enhanced since the failure of a GFA will only effect the packets routing to MNs managed by the failing GFA; and each MN has its own optimized system configuration from time to time.

2.2.2 Operations of the Distributed Dynamic Scheme

Now, we describe how MNs operate in real implementations. In particular, we explain how MNs determine the dynamically adjusted boundaries of regional networks.

Each MN keeps a buffer for storing IP addresses of mobility agents. An MN records the address of the GFA into its buffer when it enters a new regional network and then performs a home registration through the new GFA. After the home registration, the optimal number of FAs for a regional network is computed based on the up-to-date parameters of the MN. The algorithm for deriving the optimal value k_{opt} will be described in the next section. This optimal value k_{opt} is set for the buffer length threshold of the MN. If the MN detects that it enters a new subnet, it does a regional registration by sending a regional registration request to the recorded IP address of the GFA, i.e., the first FA it met in the regional network. The MN then compares the IP address of the FA in the new subnet with the addresses recorded in its buffer. If the address of the current FA has not been recorded in the buffer, then the MN records it. Otherwise, ignores it. If the total number of addresses in the buffer as well as the address of the current FA exceeds the threshold, it means the MN is in a new regional network. The MN deletes all the addresses in its buffer, saves the new one, and requests a home registration. Thus, there is no strict regional network boundary for each MN. An MN may move back and forth between two subnets and it may also visit a subnet more than once. The zigzag effect will not lead to excessive home

location registrations since the MN will know that it has moved out of a regional network only after it has visited k_{opt} *different* subnets.

The protocol descriptions of the distributed dynamic regional location registration for MNs are shown in Figure 6.

```
if (MN enters a new subnet)
   compare the address of the new FA to the addresses in buffer;
   if (the new address ≠ any address in buffer)
      if (# of addresses in buffer + the new address > kopt)
         delete all the addresses in buffer;
         record the new FA address in buffer;
         mark the new FA address as the new GFA address;
         perform a regional registration to the new GFA;
         perform a home registration through the new GFA;
         compute the new kopt;
      else
         record the new FA address in buffer;
         perform a regional registration to the GFA;
      end
   else
      perform a regional registration to the GFA;
   end
end
```

Figure 6: Protocols of the distributed dynamic scheme for MNs.

2.2.3 Comparison

Note that "*distributed* system architecture" and "*dynamic* regional network" are independent. "*distributed*" means that GFAs of different users are distributed among FAs, and "*dynamic*" means changing regional network size k_{opt} from time to time. Consequently, there are four possible combinations as follows:

- *Centralized* system architecture and *fixed* regional network
- *Centralized* system architecture and *dynamic* regional network
- *Distributed* system architecture and *fixed* regional network

- *Distributed* system architecture and *dynamic* regional network

Centralized fixed scheme is the IETF Mobile IP regional registration, which is shown in Figure 4; centralized dynamic scheme is difficult for implementation, since each FA is required to know the entire network configuration in order to be aware of when to send registration requests to which GFA; distributed fixed scheme is shown in Figure 7; and distributed dynamic scheme is our proposed scheme, which is shown in Figure 5. Note that for distributed fixed scheme, the regional network size k_{opt} may be either the same for all users or user-variant. Figure 7 presents the user-variant fixed regional network size for MN1 and MN2. We will compare our distributed dynamic scheme to the centralized fixed scheme, i.e., the IETF Mobile IP regional registration, and the distributed fixed scheme in the following sections.

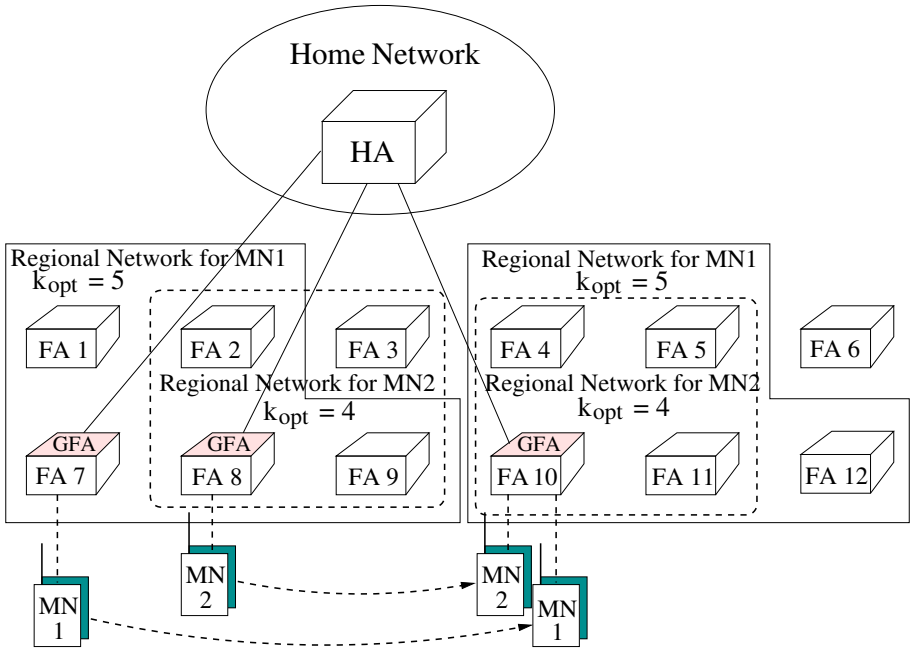

Figure 7: The distributed fixed Mobile IP regional registration.

2.3 Signaling Cost Function

In this section, we derive the cost function of location update and packet delivery to find out the optimal size of a regional network. The total signaling cost in location

update and packet delivery is considered as the performance metric. We do not take the periodic binding updates that an MN sends to mobility agents to refresh their cache into account.

2.3.1 Location Update Cost

Similar to [37], we define the following parameters for location update in the rest of this chapter:

C_{hg} The transmission cost of location update between the HA and the GFA.

C_{gf} The transmission cost of location update between the GFA and the FA.

C_{fm} The transmission cost over the wireless link between the FA and the MN.

a_h The processing cost of location update at the HA.

a_g The processing cost of location update at the GFA.

a_f The processing cost of location update at the FA.

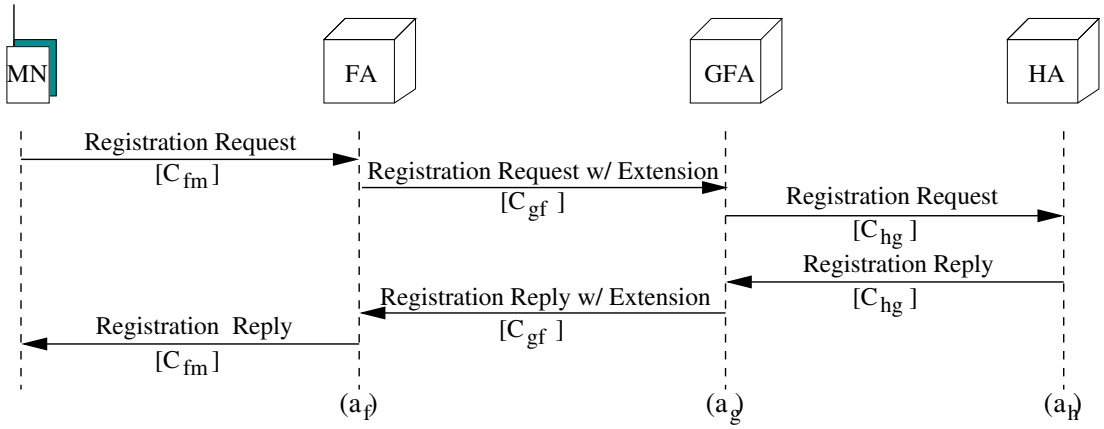

Figure 8: Process of home location registration.

Figure 8 and Figure 9 illustrate the signaling message flows for location registration with the home network and regional registration with the GFA, respectively.

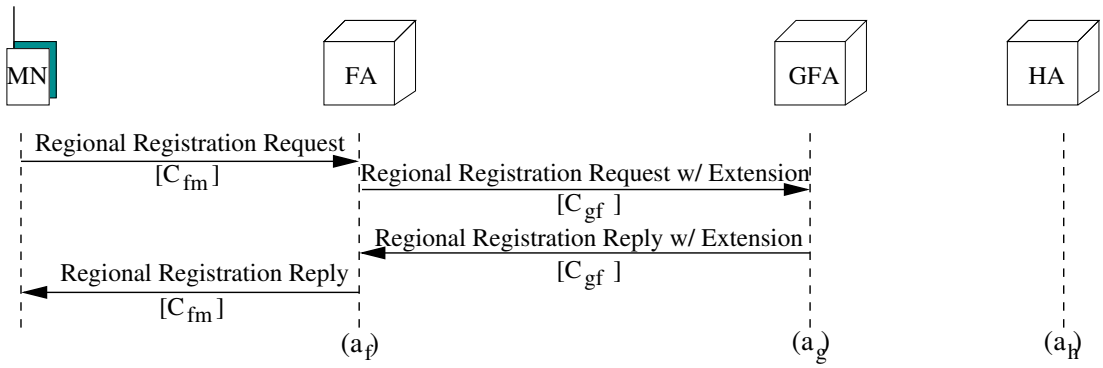

Figure 9: Process of regional location registration.

According to these message flows, the home registration cost and the regional registration cost for each location update can be calculated as follows [34]:

$$C_{Uh} = 2a_f + 2a_g + a_h + 2C_{hg} + 2C_{gf} + 2C_{fm} \quad (1)$$

$$C_{Ur} = 2a_f + a_g + 2C_{gf} + 2C_{fm} \quad (2)$$

Let l_{hg} be the average distance between the HA and the GFA in terms of the number of hops packets travel, and l_{gf} be the average distance between the GFA and the FA. We assume the transmission cost is proportional to the distance between the source and the destination mobility agents and the proportionality constant is δ_U. Thus C_{hg} and C_{gf} can be expressed as $C_{hg} = l_{hg}\delta_U$ and $C_{gf} = l_{gf}\delta_U$. Since usually the transmission cost of the wireless link is generally higher than that of the wired link, we assume that the transmission cost over the wireless link is ρ times higher than the unit distance wireline transmission cost. The transmission cost between the FA and the MN can be written as $C_{fm} = \rho\delta_U$. Then the home registration and regional registration costs can be expressed as:

$$C_{Uh} = 2a_f + 2a_g + a_h + 2(l_{hg} + l_{gf} + \rho)\delta_U \quad (3)$$

$$C_{Ur} = 2a_f + a_g + 2(l_{gf} + \rho)\delta_U \quad (4)$$

Note that for distributed GFA architecture, the first FA of the subnet the MN visits acts as a GFA. When the MN resides in the subnet of the GFA, the regional registration cost is different from the one when the MN is in the subnet not serviced by the GFA. Define this special regional registration as $\widetilde{C}_{Ur}$. Then,

$$\widetilde{C}_{Ur} = a_g + 2C_{fm} = a_g + 2\rho\delta_U \tag{5}$$

Assume an MN may move randomly between N subnets and there are k subnets within a regional network. The MN may visit a subnet more than once and it may also move back and forth between two subnets. We first consider the location update for centralized fixed scheme.

We call the action an MN moving out of a subnet "*a movement*". Define a random variable M so that an MN moves out of a regional network at movement M. We model the movements of an MN as a discrete system. At movement 1, the MN may reside in either subnet 1, 2, $\cdots$ or N. At movement 2, the MN may move to any of the other $N-1$ subnets. We assume the MN will move out to the other $N-1$ subnets with equal probability $\frac{1}{N-1}$.

For centralized fixed scheme, the probability of moving out of a regional network, i.e., the probability of performing a home registration at movement m is:

$$P^m_{h\_cf} = \frac{N-k}{N-1} \cdot \left(\frac{k-1}{N-1}\right)^{m-2}, \quad \text{where } 2 \leq m < \infty \tag{6}$$

where m is an arbitrary integer larger than 1. It can be shown that the expectation of M is:

$$E[M]_{cf} = \sum_{m=2}^{\infty} m P^m_{h\_cf} = 1 + \frac{N-1}{N-k} \tag{7}$$

Assume within a regional network, the average time an MN stays in each subnet before making a movement is T_f. Therefore, the average location update cost for centralized fixed scheme is:

$$C_{LU\_cf} = \frac{E[M]_{cf} C_{Ur} + C_{Uh}}{E[M]_{cf} T_f} \tag{8}$$

For distributed GFA system architecture, the MN will move out of a regional network only after it has visited k *different* subnets. Previous researchers used either Markovian model [38] or random walk model [32] [39] for performance analysis. However, the movement of MNs for distributed scheme is not a Markov process because the decision of whether an MN can move out of a regional network depends on its mobility history, i.e., whether an MN is in another regional network depends on whether it has visited *different* k subnets. This increases the difficulty of analysis.

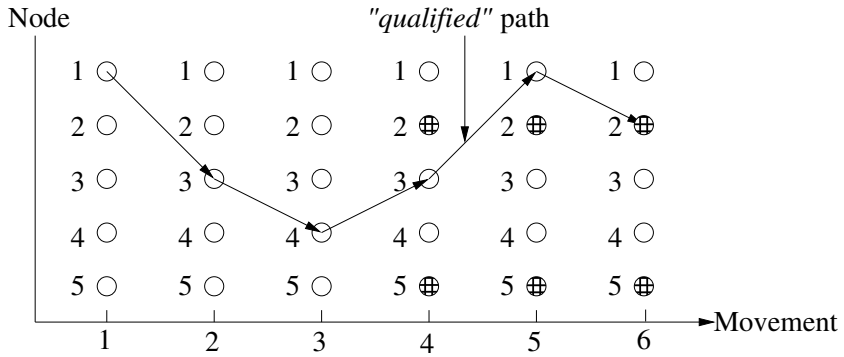

Figure 10: Discrete system mobility model of an MN.

We define the paths by which the MN has visited different k subnets "*qualified*" paths. If an MN moves out of a regional network at movement m, where m is an arbitrary integer larger than k, the path by which the MN has gone through from movement 1 to movement $m-1$ must consist k *and only* k different subnets. Figure 10 shows an example of our discrete system in which $N = 5$ and $k = 3$. In the figure, each node represents a subnet. As shown in the figure, at movement 3, the MN has visited subnet 1, 3, and 4. Therefore, subnet 2 and 5 belong to another regional network for this MN after this moment. If the MN moves out of its regional network to subnet 2 at movement 6, the subnets it visited at movement 4 and 5 are among subnet 1, 3, and 4.

Therefore, the expectation of the moment at which an MN moves out of a regional network for distributed scheme is equal to the expectation of the moment at which an MN has visited *different* k subnets plus the expectation of the time period that an MN

moves within specific k subnets. The latter one is exactly the $E[M]_{cf}$ for centralized fixed scheme. Define $E[M]_{1\to 2}$ the expectation of the number of movements it takes an MN moving from its first subnet to its second *new* subnet, i.e., an MN has visited 2 different subnets. Then

$$E[M]_{1\to 2} = 1 \tag{9}$$

Similarly, when an MN has visited two different subnets, define $E[M]_{2\to 3}$ the expectation of the number of movements it takes an MN moving to its third *new* subnet. Then

$$E[M]_{2\to 3} = \sum_{n=1}^{\infty} n \cdot \left(\frac{1}{N-1}\right)^{n-1} \frac{N-2}{N-1} = \frac{N-1}{N-2} \tag{10}$$

and the expectation of the number of movements it takes an MN moving from its $(k-1)$th subnet to its kth subnet is:

$$E[M]_{k-1\to k} = \sum_{n=1}^{\infty} n \cdot \left(\frac{k-2}{N-1}\right)^{n-1} \frac{N-k+1}{N-1} = \frac{N-1}{N-k+1} \tag{11}$$

Then the expectation of the moment at which an MN moves out of a regional network for distributed fixed scheme and distributed dynamic scheme are:

$$\begin{aligned} E[M]_{df} = E[M]_{dd} &= E[M]_{1\to 2} + E[M]_{2\to 3} + \cdots + E[M]_{k-1\to k} + E[M]_{cf} \quad (12) \\ &= 1 + \frac{N-1}{N-2} + \cdots + \frac{N-1}{N-k+1} + \frac{N-1}{N-k} + 1 \\ &= 1 + (N-1)\sum_{i=1}^{k} \frac{1}{N-i} \end{aligned}$$

Note that the expectation of the moment at which an MN moves out of a regional network for distributed system is always larger than that for centralized system. As a result, the number of home registrations per unit time is reduced. The upper bound of the total location update costs per unit time for distributed fixed scheme and distributed dynamic scheme are:

$$C_{LU\_df} \le \frac{\tilde{C}_{Ur} + (E[M]_{df} - 1)C_{Ur} + C_{Uh}}{E[M]_{df} T_f} \tag{13}$$

$$C_{LU\_dd} \le \frac{\tilde{C}_{Ur} + (E[M]_{dd} - 1)C_{Ur} + C_{Uh}}{E[M]_{dd} T_f} \tag{14}$$

Based on (4)-(14), we may get the average location update cost. Note that our method does not impose any restrictions on the shape and the geographic location of subnets. It is a general model which is applicable to arbitrary subnets.

2.3.2 Packet Delivery Cost

Under Mobile IP regional registration, every IP packet destined for an MN is first intercepted by the HA and is then tunneled to the registered GFA and further forwarded to the current serving FA of the MN. Because of this triangular routing, there are extra costs for packet delivery. The packet delivery cost includes the transmission and processing cost to route a tunneled packet from the HA to the serving FA of an MN. Assume

T_{hg} The transmission cost of packet delivery between the HA and the GFA.

T_{gf} The transmission cost of packet delivery between the GFA and the FA.

v_h The processing cost of packet delivery at the HA.

v_g The processing cost of packet delivery at the GFA.

The cost for packet delivery procedure can be expressed as:

$$C_{PD} = v_h + v_g + T_{hg} + T_{gf} \tag{15}$$

Similar to the assumption for location update case, we assume the transmission cost of delivering data packets is proportional to the distance between the sending and the receiving mobility agents with the proportionality constant δ_D. Then $T_{hg} = l_{hg}\delta_D$ and $T_{gf} = l_{gf}\delta_D$.

The processing cost at GFAs includes decapsulation of the tunneled IP packets from the HA, checking its visitor list to see whether it has an entry for the destination MN, re-encapsulation of the IP packets, and management of routing packets to the FAs. The load on a GFA for processing and routing packets to each FA depends on k,

the number of FAs under a GFA. If k is large, the complexity of the visitor list lookup and IP routing lookup in the GFA is high, and the system performance is degraded. In addition, since the total bandwidth of the network is limited, if the traffic to a GFA is heavy, the transmission delay and the number of retransmissions cannot be bounded. These factors will result in a high processing cost at the GFAs. Assume on average there are ω MNs in a subnet. For centralized system architecture, a GFA serves for all the MNs moving within a regional network, and the total number of MNs in a regional network is ωk on average. Therefore, the complexity of the GFA visitor list lookup is proportional to ωk. On the other hand, for distributed system architecture, different MNs choose different FAs as their GFAs. A GFA only serves the MNs which first enter the subnet managed by this GFA in a regional network. The packet processing load of a GFA in the distributed system is much lower than that in the centralized system because the traffic is allocated evenly among all the FAs in a regional network. Therefore, the complexity of the GFA visitor list lookup for distributed system is proportional only to ω. Since IP routing table lookup is based on the *longest prefix matching* and most implementations use the traditional *Patricia trie* [40], the complexity of IP address lookup is proportional to the logarithm of the length of the routing table k [41]. We define the packet processing cost functions at the GFA for centralized system and distributed system as:

$$v_{g\_cf} = \zeta k \cdot \lambda_a \left(\alpha \omega k + \beta log(k)\right) \tag{16}$$

$$v_{g\_df} = v_{g\_dd} = \zeta k \cdot \lambda_a \left(\alpha \omega + \beta log(k)\right) \tag{17}$$

where λ_a is the packet arrival rate for each MN, α and β are weighting factors of visitor list and routing table lookups, and ζ is a constant which captures the bandwidth allocation cost at the GFA. The larger the ζ is, the more negative effects an MN experiences from not enough network bandwidth available.

The processing cost function at the HA can be defined as: $v_h = \eta\lambda_a$, where η is a packet delivery processing cost constant at the HA. Then the total packet delivery costs per unit time for the three schemes are:

$$C_{PD\_cf} = \eta\lambda_a + \zeta k \cdot \lambda_a \left(\alpha\omega k + \beta log(k)\right) + (l_{hg} + l_{gf})\delta_D \tag{18}$$

$$C_{PD\_df} = C_{PD\_dd} = \eta\lambda_a + \zeta k \cdot \lambda_a \left(\alpha\omega + \beta log(k)\right) + (l_{hg} + l_{gf})\delta_D \tag{19}$$

2.3.3 Total Signaling Cost

Based on the above analysis, we get the overall signaling cost function as:

$$C_{TOT\_(\cdot)}(k, \lambda_a, T_f) = C_{LU\_(\cdot)} + C_{PD\_(\cdot)} \tag{20}$$

where $C_{TOT\_(\cdot)}$, $C_{LU\_(\cdot)}$, and $C_{PD\_(\cdot)}$ represent the total signaling cost, location update cost, and packet delivery cost for the three different schemes, i.e., centralized fixed scheme, distributed fixed scheme, and the proposed distributed dynamic scheme.

2.4 Optimal Regional Network Size

The optimal number of FAs beneath a GFA, k_{opt}, is defined as the value of k that minimizes the cost function derived in Section 2.3. Because k can only be an integer, the cost function is not a continuous function of k. Therefore, it is not appropriate to take derivative with respect to k of the cost function to get the minimum. We use an iterative algorithm. Note that iterative algorithm may result in a local minimum. Solutions to solving the local minimum problem were discussed in [33]. Similar to the algorithm proposed in [29], we define the cost difference function between the system with number k and the system with number $k-1$ ($k \geq 2$), i.e.,

$$\Delta_{cf}(k, \tilde{\lambda}_a, \tilde{T}_f) = C_{TOT\_cf}(k, \tilde{\lambda}_a, \tilde{T}_f) - C_{TOT\_cf}(k-1, \tilde{\lambda}_a, \tilde{T}_f) \tag{21}$$

$$\Delta_{df}(k, \bar{\lambda}_a, \bar{T}_f) = C_{TOT\_df}(k, \bar{\lambda}_a, \bar{T}_f) - C_{TOT\_df}(k-1, \bar{\lambda}_a, \bar{T}_f) \tag{22}$$

$$\Delta_{dd}(k, \lambda_a, T_f) = C_{TOT\_dd}(k, \lambda_a, T_f) - C_{TOT\_dd}(k-1, \lambda_a, T_f) \tag{23}$$

where $\widetilde{\lambda}_a$ and $\widetilde{T}_f$ are the average packet arrival rate and average subnet residence time for all MNs; $\bar{\lambda}_a$ and $\bar{T}_f$ are the average packet arrival rate and average subnet residence time for each MN. Given $\Delta_{(\cdot)}$, the algorithm to find the optimal value of k is defined as follows:

$$k_{opt\_cf}(\widetilde{\lambda}_a, \widetilde{T}_f) = \begin{cases} 1, \text{ if } \Delta_{cf}(2, \widetilde{\lambda}_a, \widetilde{T}_f) > 0 \\ max\{k : \Delta_{cf}(k, \widetilde{\lambda}_a, \widetilde{T}_f) \leq 0\}, \text{ otherwise.} \end{cases} \tag{24}$$

$$k_{opt\_df}(\bar{\lambda}_a, \bar{T}_f) = \begin{cases} 1, \text{ if } \Delta_{df}(2, \bar{\lambda}_a, \bar{T}_f) > 0 \\ max\{k : \Delta_{df}(k, \bar{\lambda}_a, \bar{T}_f) \leq 0\}, \text{ otherwise.} \end{cases} \tag{25}$$

$$k_{opt\_dd}(\lambda_a, T_f) = \begin{cases} 1, \text{ if } \Delta_{dd}(2, \lambda_a, T_f) > 0 \\ max\{k : \Delta_{dd}(k, \lambda_a, T_f) \leq 0\}, \text{ otherwise.} \end{cases} \tag{26}$$

Note that the optimal value of the centralized fixed scheme is the same for all the MNs and is fixed all the time; the optimal value of the distributed fixed scheme is fixed all the time, but each user may have different optimal value; and the optimal value of the proposed distributed dynamic scheme is adapted to each MN and it depends on the up-to-date packet arrival rate and user mobility.

The algorithm for estimating packet arrival rate can be found in [29]. Each MN may use a timer to count the time it spent in each subnet and the average value within a regional network, T_f, is calculated before computing the k_{opt}. T_f can also be estimated if the probability density function (pdf) of the MN residence time in each subnet within a regional network is known. For example, if the pdf of the MN residence time $f_r(t)$ is of Gamma distribution which has Laplace transform $F_r(s) = \left(\frac{\mu\gamma}{s+\mu\gamma}\right)^\gamma$ with mean value $\frac{1}{\mu}$, variance V, and $\gamma = \frac{1}{V\mu^2}$. Then $T_f = \frac{1}{\mu}$. Our algorithm also needs to know the number of hops between the HA and the GFA, l_{hg}, and the number of hops between the GFA and the FA, l_{gf}. If each MN has dedicated paths for transmitting signaling messages from FAs to GFAs and HAs, the number of hops between mobility agents (HA, GFA and FA), l_{hg} and l_{gf}, are fixed numbers. If not,

signaling packets may take different paths each time according to the traffic load and routing algorithms at each mobility agent. Thus, l_{hg} and l_{gf} vary within a certain range. An MN may use the *time-to-live* (TTL) field in IP packet headers to get the number of hops packets travel [42]. Then the average value may be used for optimal number computation.

2.5 Analytical Results

In this section, we demonstrate the performance improvement of the distributed dynamic scheme to the centralized fixed scheme, i.e., the IETF Mobile IP regional registration [15]. Since the distributed dynamic scheme and the centralized fixed scheme are not comparable, first we show the cost saving of the distributed fixed scheme to the centralized fixed scheme. Next we demonstrate the advantages of the proposed distributed dynamic scheme over the distributed fixed scheme.

For the analysis in this chapter, we assume the cost for transmitting signaling messages and the cost for packet processing at mobility agents are available. As discussed in [43], the cost parameters can be expressed in terms of the delay required to process the signaling messages. For example, a_h, a_g, and a_f may represent the delay required by the HA, GFA, and FA to process a location update requested by the signaling message, respectively; δ_U and δ_D may represent the delay for sending the signaling message through the particular path. Other measurements for the cost parameters are possible. For example, the network administration can assign relative costs to the mobility agents based on the current available bandwidth, computation resources in the system, and the expenses required to operate the particular mobility agent. In real implementations, the parameters in our model are designed values. They can be determined based on empirical measurements or some heuristic strategy. For different system architectures, the parameters are different. A table lookup process can be adopted in a particular network implementation, as mentioned in [44]. Given a

particular time of a day, the table located at each FA provides a set of parameters for MNs to determine the optimal regional network size. The parameter table should be updated periodically to reflect the status of the network.

Table 1: Performance Analysis Parameters for Location Management in Mobile IP

Pkt Process Cost			Distance Cost Unit		Wireless Multiple	# of MNs/subnet	Weight		Packet Process Const.	
a_h	a_g	a_f	δ_U	δ_D	ρ	ω	α	β	ζ	η
25.0	15.0	10.0	0.1	0.05	10	15	0.3	0.7	0.01	10.0

Table 1 lists some of the parameters used in our performance analysis. Since the total number of subnets that MNs may access through wireless channels is limited, we assume $N = 30$. For our numerical evaluation, we assume l_{hg} and l_{gf} are fixed numbers. Since the TTL field in IP header is usually initialized by the sender to 32 or 64 [42], i.e., the upper limit on the number of hops through which a packet can pass is 32 or 64, we assume $l_{hg} = 25$ and $l_{gf} = 10$.

2.5.1 Centralized Fixed Scheme vs. Distributed Fixed Scheme

First, we compare the performance of the centralized fixed scheme and the distributed fixed scheme. Similar to the analysis in PCS, we define the call-to-mobility ratio (CMR) as the ratio of the packet arrival rate to the mobility rate, i.e., $CMR = \lambda_a T_f$. Since the cost functions of the two schemes derived in Section 2.3 are different, we focus on compare the total signaling cost of the centralized fixed scheme $C_{TOT\_cf}(k_{opt\_cf}(\widetilde{\lambda}_a, \widetilde{T}_f), \lambda_a, T_f)$ with that of the distributed fixed scheme $C_{TOT\_df}($ $k_{opt\_df}(\bar{\lambda}_a, \bar{T}_f), \lambda_a, T_f)$ when the average values of residence time in each subnet and packet arrival rate of all the MNs are the same, i.e., $\widetilde{T}_f = \bar{T}_f$ and $\widetilde{\lambda}_a = \bar{\lambda}_a$.

Figure 11 plots the optimal k as a function of CMR for the centralized fixed scheme and the distributed fixed scheme. Note that for the two systems, the optimal regional network size k_{opt} is a designed value. It is computed before the communications

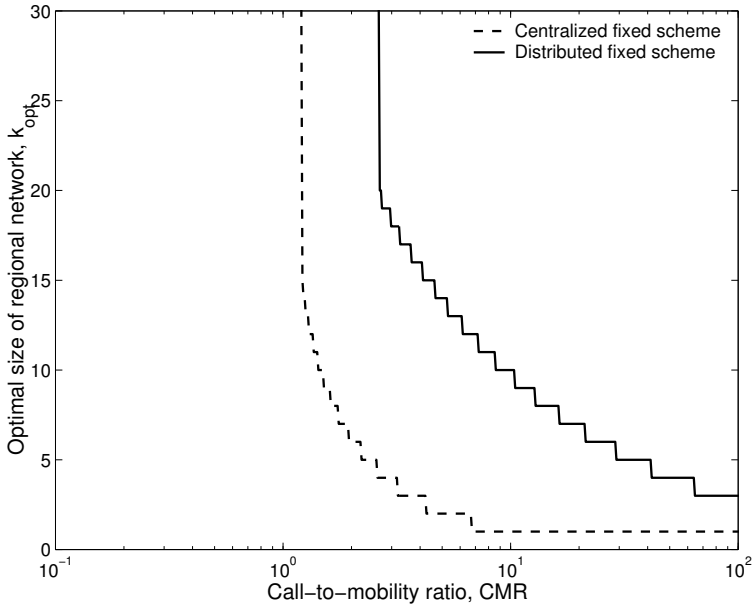

Figure 11: Optimal regional network size for centralized and distributed systems.

based on the average values of user parameters. As shown in the figure, the optimal regional network size decreases as CMR increases for both centralized and distributed systems. When the CMR is low, the mobility rate is high compared to the packet arrival rate and the cost for location update dominates. Systems with larger regional networks may reduce the number of home registrations and provide the benefit of regional registration. When the CMR is high, the packet delivery cost dominates and the saving in packet delivery becomes significant. The saving can be attributed to the smaller regional network size. Note that the optimal regional network size of the distributed system is always larger than or equal to that of the centralized system. This means for the same CMR, the distributed system has larger regional network size and consequently performs less home registrations compared with the centralized system.

Figure 12 shows the total signaling cost as a function of CMR for the two schemes. The dashed line in the figure is the signaling cost of centralized fixed scheme when regional network size is k_{opt_cf}. The dotted line is the signaling cost of distributed fixed scheme with k_{opt_cf} as the regional network size. Note that k_{opt_cf} is the optimal value

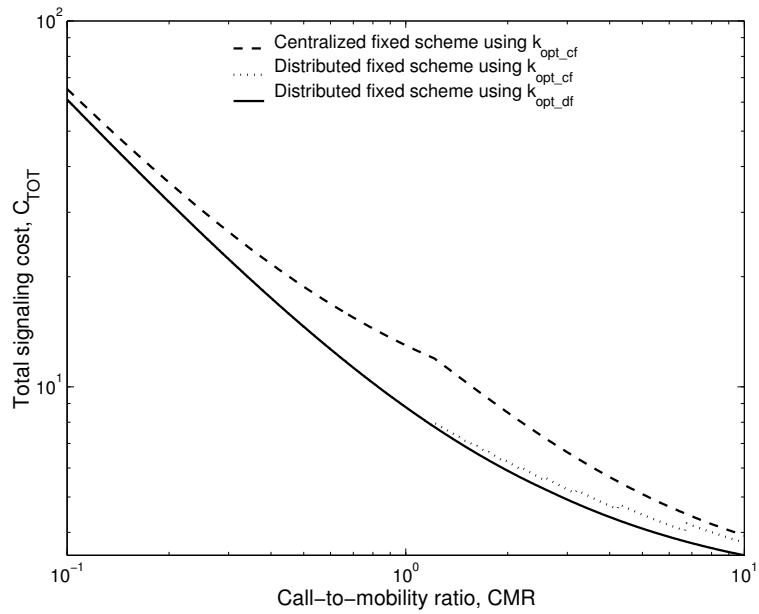

Figure 12: Comparison of total signaling cost for fixed schemes.

for the centralized fixed scheme, in the sense that the minimal cost can be reached. But k_{opt_cf} is not the optimal value for the distributed scheme. The solid line in the figure is the signaling cost of the distributed fixed scheme under k_{opt_df}. This line represents the minimal cost of the distributed fixed scheme. Figure 12 indicates that even under non-optimal regional network size, the distributed scheme always performs better than the centralized IETF Mobile IP regional registration scheme. And the distributed scheme with optimal regional network size can further improve the performance. Up to 36% signaling cost can be saved when using distributed system architecture.

2.5.2 Distributed Fixed Scheme vs. Distributed Dynamic Scheme

Next, we compare the total signaling cost of the distributed fixed scheme C_{TOT_df} $(k_{opt\_df}(\bar{\lambda}_a, \bar{T}_f), \lambda_a, T_f)$ with that of the proposed distributed dynamic scheme $C_{TOT\_dd}($ $k_{opt\_dd}(\lambda_a, T_f), \lambda_a, T_f)$ under various scenarios. Note that $k_{opt\_df}(\bar{\lambda}_a, \bar{T}_f)$ is pre-computed before communications. Once it is set, it will not change. But $k_{opt\_dd}(\lambda_a, T_f)$ is dynamically adapted to the user parameters during the communications. Since the cost

functions of the two schemes are the same, the advantages of the dynamic scheme over the fixed scheme are reflected when the user parameters are different and changing from time to time. Therefore, we investigate the impacts of user-variant and time-variant user parameters.

2.5.2.1 The Impact of User-Variant Residence Time

We first investigate the impact of user-variant mobility. Let packet arrival rate λ_a be a fixed number, i.e., $\lambda_a = \bar{\lambda}_a =$ constant. Similar to [29], we assume there are two groups of MNs. One group represents "*active*" users with average residence time in each subnet $\bar{T}_{f_1} = 1.0$. The other group is for "*passive*" users with average residence time in each subnet $\bar{T}_{f_2} = 100$. The residence time of group 1 users follows an exponential distribution, i.e.,

$$f_1(T_f) = \frac{1}{\bar{T}_{f_1}} e^{-T_f/\bar{T}_{f_1}}, \quad T_f \geq 0 \tag{27}$$

and the residence time of group 2 users follows a Gaussian distribution:

$$f_2(T_f) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(T_f - \bar{T}_{f_2}\right)^2/2\sigma^2}, \quad T_f \geq 0 \tag{28}$$

where $\sigma = 10$. Assume that each group has 50% of total users. The residence time T_f of a randomly selected user has pdf as:

$$f(T_f) = 0.5\left(f_1(T_f) + f_2(T_f)\right) \tag{29}$$

and the overall average residence time is:

$$\bar{T}_f = 0.5\bar{T}_{f_1} + 0.5\bar{T}_{f_2} \tag{30}$$

Therefore, the total signaling cost of the distributed fixed scheme is:

$$\begin{aligned} C_{df} &= 0.5\int_0^\infty f_1(T_f) C_{TOT\_df}(k_{opt\_df}(\bar{\lambda}_a, \bar{T}_{f_1}), \lambda_a, T_f) dT_f \\ &\quad + 0.5\int_0^\infty f_2(T_f) C_{TOT\_df}(k_{opt\_df}(\bar{\lambda}_a, \bar{T}_{f_2}), \lambda_a, T_f) dT_f \end{aligned} \tag{31}$$

where k_{opt} of group 1 users is computed based on their average residence time $\bar{T}_{f_1}$ and k_{opt} of group 2 users is computed based on $\bar{T}_{f_2}$. Note that for distributed fixed scheme, the optimal regional network size may be user-variant or the same for all the users. Figure 7 gives an example of user-variant k_{opt} and (32) indicates that group 1 and group 2 users adopt different fixed optimal regional network size. The total signaling cost of the distributed fixed scheme using fixed k_{opt} for all the users is:

$$\tilde{C}_{df} = \int_0^\infty f(T_f) C_{TOT\_df}(k_{opt\_df}(\bar{\lambda}_a, \bar{T}_f), \lambda_a, T_f) dT_f \tag{32}$$

and the total signaling cost of the distributed dynamic scheme is:

$$C_{dd} = \int_0^\infty f(T_f) C_{TOT\_dd}(k_{opt\_dd}(\lambda_a, T_f), \lambda_a, T_f) dT_f \tag{33}$$

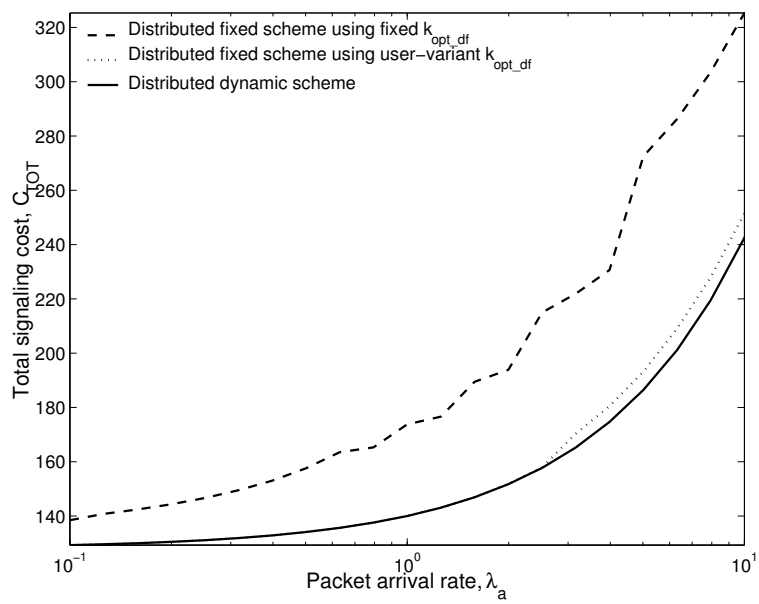

Figure 13: Comparison of total signaling cost under user-variant residence time.

Figure 13 shows the total signaling cost of the distributed dynamic scheme and the distributed fixed scheme under user-variant residence time T_f. The dashed line in the figure is the signaling cost of the distributed fixed scheme using fixed k_{opt_df}, which is actually the case shown in Figure 12 with solid line. It is observed in Figure 13 that the signaling cost of distributed dynamic scheme is less than that of both

the distributed fixed scheme using fixed optimal regional network size and using user-variant optimal size. Our results demonstrate that C_{TOT} is reduced by up to 33% using the dynamic scheme instead of the fixed scheme with fixed k_{opt}. Although the performance improvement of the distributed dynamic scheme is not large compared to the distributed fixed scheme under user-variant k_{opt}, in the following time-variant residence time situation, the dynamic scheme will demonstrate its advantage.

2.5.2.2 The Impact of Time-Variant Residence time

Packet arrival rate λ_a is still a constant. The residence time of all MNs, T_f, is of exponential distribution:

$$f(T_f) = \frac{1}{\bar{T}_f} e^{-T_f/\bar{T}_f} \tag{34}$$

where $\bar{T}_f$ is the mean residence time and $\bar{T}_f$ is time-variant. The overall signaling cost of distributed fixed scheme is:

$$C_{df}(\bar{T}_f) = \int_0^\infty f(T_f) C_{TOT\_df}(k_{opt\_df}, \lambda_a, T_f) dT_f \tag{35}$$

Note that although $\bar{T}_f$ is varying during the communications, the optimal value for the fixed scheme k_{opt_df} is pre-computed as a designed value and is fixed all the time during the communications. The signaling cost of the distributed dynamic scheme is given by (33) using the new pdf function $f(T_f)$ in (34).

Figure 14 and Figure 15 show the total signaling cost as a function of the average residence time $\bar{T}_f$, when $\bar{\lambda}_a = 3.0$. Two cases of the distributed fixed scheme are shown. One is with the optimal regional network size k_{opt_df} pre-computed using $\bar{T}_f = 0.1$ as the average residence time over all users. The other is with the optimal size k_{opt_df} pre-computed using $\bar{T}_f = 100$. Note that the distributed fixed system always pays higher cost than the distributed dynamic system. Our results show that up to 15% cost can be saved by the distributed dynamic scheme compared to the distributed fixed scheme using $\bar{T}_f = 0.1$ for the optimal regional network size computation, and up to 44% cost can be saved compared to the distributed fixed

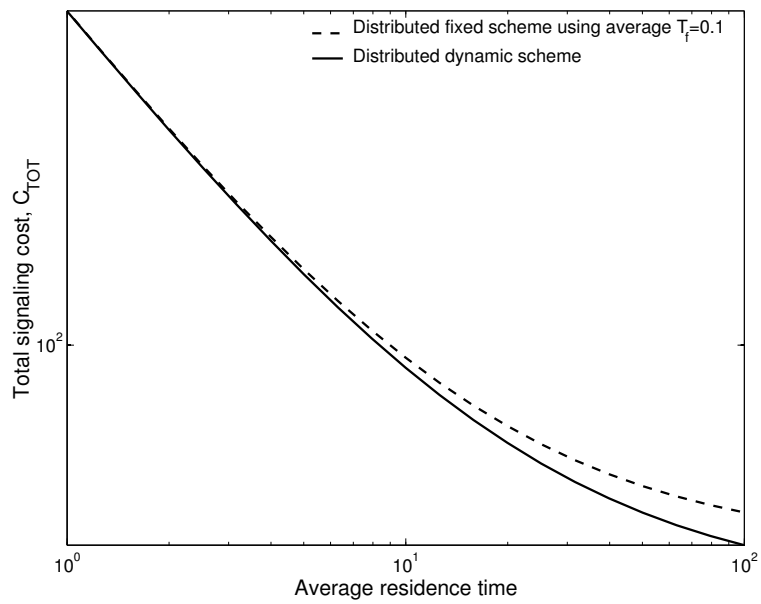

Figure 14: Comparison of total signaling cost under time-variant residence time.

scheme using $\bar{T}_f = 100$ for the computation. We can see from the figures that the distributed fixed system using $\bar{T}_f = 0.1$ for optimal size computation may perform well when the user residence time is small, but when the residence time is large, the fixed scheme consumes more network resource. Similarly, the cost gap of the dynamic system and the fixed system using $\bar{T}_f = 100$ for computation is smaller when $\bar{T}_f$ is large, but the fixed system pays much more extra bandwidth when $\bar{T}_f$ is small. Therefore, it is a difficult task to design an optimal regional network size beforehand for the distributed fixed scheme. If the user mobility has some unusual big changes to its normal average value, the system with a pre-designed fixed regional network size will consume much more bandwidth and the network may be congested.

2.5.2.3 The Impact of User-Variant Packet Arrival Rate

Now we investigate the impact of user-variant packet arrival rate. Let user residence time T_f be a constant, i.e., $T_f = \bar{T}_f =$ constant. Similar to the discussion in Section 2.5.2.1, we assume there are two groups of MNs. One represents normal users with average packet arrival rate $\bar{\lambda}_{a_1} = 0.1$. The other group is for special users with average

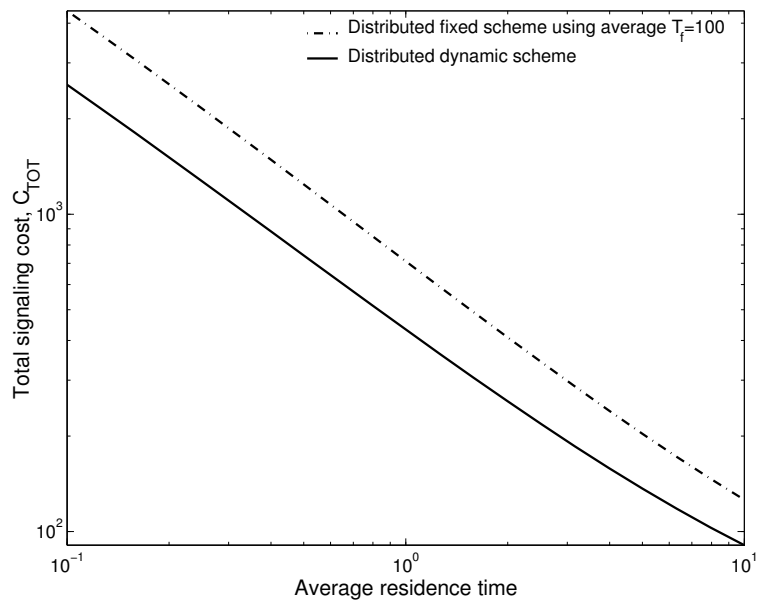

Figure 15: Comparison of total signaling cost under time-variant residence time.

packet arrival rate $\bar{\lambda}_{a_2} = 10.0$. The packet arrival rates of group 1 normal users follow an exponential distribution, i.e.,

$$f_1(\lambda_a) = \frac{1}{\bar{\lambda}_{a_1}} e^{-\lambda_a/\bar{\lambda}_{a_1}}, \quad \lambda_a \geq 0 \tag{36}$$

and the packet arrival rates of group 2 special users follow a Gaussian distribution:

$$f_2(\lambda_a) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\lambda_a - \bar{\lambda}_{a_2}\right)^2/2\sigma^2}, \quad \lambda_a \geq 0 \tag{37}$$

where $\sigma = 4.0$. Assume that each group contributes 50% of total users. For an arbitrary MN, the packet arrival rate has pdf as:

$$f(\lambda_a) = 0.5\left(f_1(\lambda_a) + f_2(\lambda_a)\right) \tag{38}$$

and the overall average packet arrival rate is:

$$\bar{\lambda}_a = 0.5\bar{\lambda}_{a_1} + 0.5\bar{\lambda}_{a_2} \tag{39}$$

Therefore, the total signaling costs of the distributed fixed scheme using fixed k_{opt} for all the MNs and using different k_{opt} for group 1 and group 2 users are:

$$\tilde{C}_{df} = \int_0^\infty f(\lambda_a) C_{TOT\_df}(k_{opt\_df}(\bar{\lambda}_a, \bar{T}_f), \lambda_a, T_f) d\lambda_a \tag{40}$$

$$C_{df} = 0.5\int_0^\infty f_1(\lambda_a)C_{TOT\_df}(k_{opt\_df}(\bar{\lambda}_{a_1},\bar{T}_f),\lambda_a,T_f)d\lambda_a + 0.5\int_0^\infty f_2(\lambda_a)C_{TOT\_df}(k_{opt\_df}(\bar{\lambda}_{a_2},\bar{T}_f),\lambda_a,T_f)d\lambda_a \quad (41)$$

and the total signaling cost of the distributed dynamic scheme is:

$$C_{dd} = \int_0^\infty f(\lambda_a)C_{TOT\_dd}(k_{opt\_dd}(\lambda_a,T_f),\lambda_a,T_f)d\lambda_a \quad (42)$$

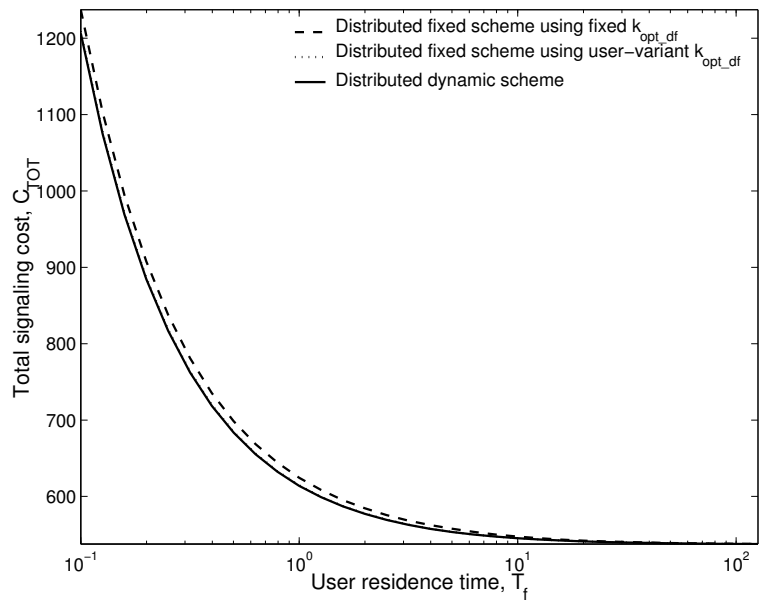

Figure 16: Comparison of total signaling cost under user-variant packet arrival rate.

Figure 16 shows the total signaling cost of the distributed dynamic scheme and the distributed fixed scheme under user-variant packet arrival rate λ_a. The signaling cost of the distributed dynamic scheme is almost the same as that of both the distributed fixed scheme using fixed optimal regional network size and using user-variant optimal size. Only 3% cost can be reduced using the distributed dynamic scheme. It indicates that the optimal regional network size is relatively insensitive to the packet arrival rate. Although different users have widely ranged traffic load, their optimized regional network sizes do not vary much.

2.5.2.4 The Impact of Time-Variant Packet Arrival Rate

Finally, we study the impact of time-variant packet arrival rate. The user residence time T_f is still fixed. The packet arrival rates of all MNs are exponentially distributed:

$$f(\lambda_a) = \frac{1}{\bar{\lambda}_a} e^{-\lambda_a/\bar{\lambda}_a} \tag{43}$$

where $\bar{\lambda}_a$ is the mean arrival rate and $\bar{\lambda}_a$ is time-variant. The overall signaling cost of the distributed fixed scheme is given by:

$$C_{df}(\bar{\lambda}_a) = \int_0^\infty f(\lambda_a) C_{TOT\_df}(k_{opt\_df}, \lambda_a, T_f) d\lambda_a \tag{44}$$

where k_{opt_df} is pre-computed and is fixed all the time. The signaling cost of the distributed dynamic scheme is given by (42) using $f(\lambda_a)$ in (43).

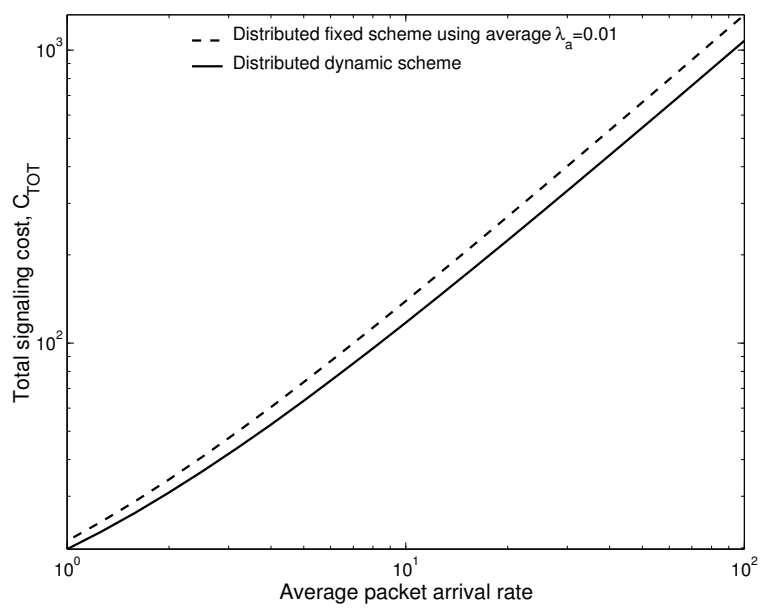

Figure 17: Comparison of total signaling cost under time-variant packet arrival rate.

Figure 17 and Figure 18 plot the total signaling cost as a function of time-variant average packet arrival rate $\bar{\lambda}_a$, when $\bar{T}_f = 10$. The dashed line in Figure 17 is based on k_{opt} calculated using $\bar{\lambda}_a = 0.01$. The dash-dot line in Figure 18 is based on k_{opt} calculated using $\bar{\lambda}_a = 100$. The solid line in both figures is for the proposed distributed dynamic scheme where k_{opt} varies according to the up-to-date parameters. The figures

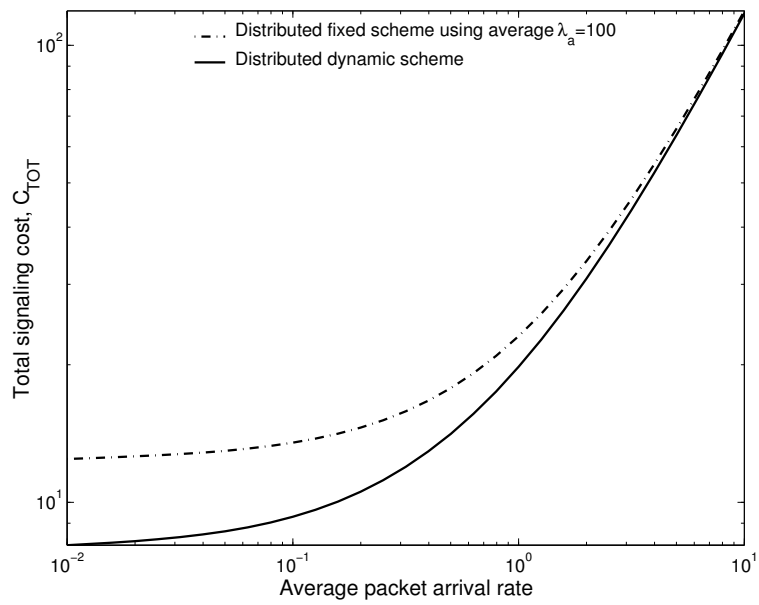

Figure 18: Comparison of total signaling cost under time-variant packet arrival rate.

show that the fixed system always pays higher cost than the dynamic system. The cost gap is larger when $\bar{\lambda}_a < 0.1$ in Figure 18 and when $\bar{\lambda}_a > 10$ in Figure 17. The dynamic system saves up to 19% and 36% cost compared to the fixed system using $\bar{\lambda}_a = 0.01$ and $\bar{\lambda}_a = 100$ for optimal value computation, respectively. This result is similar to that in Section 2.5.2.2. It indicates that the distributed dynamic scheme is more cost-efficient when the user parameters are time-variant.

CHAPTER III

USER INDEPENDENT PAGING SCHEME FOR NG WIRELESS INTERNET

3.1 Problem and Solution

Location tracking of mobile terminals (MTs) in current wireless personal communications services (PCS) system includes two fundamental operations: location update and paging [6]. Location update is the reporting process of the current location area (LA) of an MT. Paging is used by the system to alert an MT of an incoming call by sending poll messages to the cells within the last reported LA. In the current paging strategy of GSM and IS-41 protocols, paging messages are broadcasted to each cell in the registered LA of an MT. The signaling cost of broadcast procedure is maximum, especially when the number of MTs increases. In order to improve the bandwidth utilization, multi-step paging or sequential paging schemes are proposed.

There is a tradeoff between the paging cost and the delay associated with locating an MT using multi-step paging schemes. It is stated in [45] that blocking users from the system due to bandwidth unavailability is much more undesirable from the user's and the operator's viewpoint than the delay of incoming data reaching the user. Therefore, when the system need not find MTs immediately, multi-step paging schemes are preferable. There are numerous link layer multi-step paging schemes proposed to reduce the paging cost [46] [47] [48]. A *shortest-distance-first* (SDF) paging scheme is proposed in [46] where cells closing to the last registered location are paged first. This scheme is associated with a distance-based location update scheme. Under the *highest-probability-first* (HPF) scheme introduced in [47], an LA is divided

into several partitions and each partition consists of a cluster of cells. The sequential paging is performed in decreasing order of cell location probabilities to minimize the mean number of cells being searched. In [49], three methods for dividing an LA are proposed, namely *reverse*, *semi-reverse*, and *uniform* paging. Given the location probabilities, cells are grouped in different ways to reduce the paging cost under delay bounds.

All the above mentioned paging schemes are user-dependent, i.e., the distance to the last registered location and location probabilities are user-variant. Therefore, the partition of paging areas and the selection of paging sequence are different for each user. If the system performs paging optimally for every MT, i.e., the average paging cost for each MT is minimum while satisfying its paging delay requirement, the overall performance of the entire system is also optimal. However, many factors may affect the performance of a paging scheme for individual users. For example, for paging schemes based on cell location probabilities, these schemes assume perfect knowledge on the user mobility statistics, which may not be readily available in practice. Moreover, the cell location probabilities are predicted and estimated values based on user movement history. They cannot reflect the up-to-date user mobility. If the user mobility has some unusual big changes to its normal average value, paging in the decreasing order of statistically average values of location probabilities cannot generate optimal performance for each user. Therefore, to reduce the dependency of the paging scheme on individual user information, but to provide satisfactory overall performance of the whole system is the basic consideration of this research work. Another disadvantage of user-dependent paging schemes is that the consumption of network resource and the signaling overhead may become very significant when a number of users are paged at the same moment, since every paging request is processed separately. This case may happen more often in Mobile IP.

Mobile IP [13] [25] is being developed by Internet Engineering Task Force (IETF).

It introduces three new functional entities: home agent (HA), foreign agent (FA), and mobile node (MN). When an MN moves out of its home network, it obtains a temporary address: care-of address (CoA). This address is used to identify the MN in the local network. When the MN moves from one foreign network to another, it registers its new location, i.e., its new CoA, to its HA. Packets for an MN are sent to its permanent address, i.e., its home address first. The HA intercepts all the IP packets destined to the MN and tunnels them to the serving FA of the MN. The FA decapsulates and forwards these packets to the MN.

A major problem of MNs is their limited battery capacity. In order to save the battery power consumption at MNs, IP paging is proposed as an extension for Mobile IP [17] [18] [19] [20] [21]. Under Mobile IP paging, an MN is allowed to enter a power saving idle mode when it is inactive for a period of time. During the idle mode, the system knows the location of the MN with coarse accuracy defined by a paging area which is composed of several subnets [17]. The MN may also deactivate some of its components for energy-saving purpose. An MN in idle mode does not need to register its location when moving within a paging area. It performs location update only when it changes paging areas. When packets are destined to an MN in idle mode, they are terminated at a paging initiator. The paging initiator buffers the packets and locates the MN by sending IP paging messages to all the subnets within the paging area. After knowing the subnet where the MN is residing, the paging initiator forwards the data packets to the serving FA of the subnet and further to the MN. Since the system and MNs synchronize on the time slots for paging, there are possibilities that multiple paging requests are needed to be sent out in one time slot. In this case, *paging request aggregation* [20] can be adopted to reduce the paging overheads, if the paging criterion is not user dependent. When an MN is in active transmission mode, it operates in the same manner as in Mobile IP and the system keeps the exact updated location information of the MN. The state transition diagram of MNs with paging support is

shown in Figure 19.

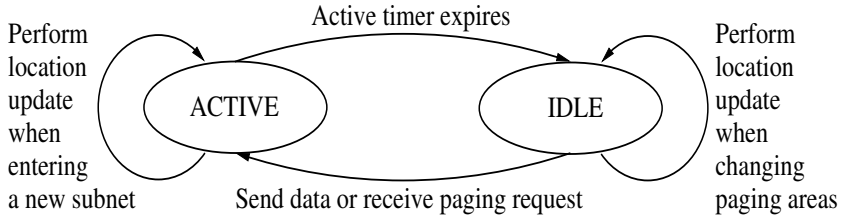

Figure 19: State transition diagram of Mobile IP paging.

Current research activities on Mobile IP paging focus on the *paging architecture* design, i.e., which node initiates paging and how the messages exchange between nodes. They seldom consider the *paging algorithm* design, i.e., how an MN is searched or how the paging requests are sent by the paging initiator [18]. Broadcast procedure is assumed to be used in almost all the proposed paging architectures [17] [18] [19] [20] [21]. The ideas of multi-step link layer paging schemes can be applied to Mobile IP. However, link layer (layer 2) paging and IP layer (layer 3) paging are different. IP layer paging refers to locating the current IP attachment point of an MN within a layer 3 location area. A layer 3 location area is a set of IP subnets identified by IP addresses [12]. Link layer paging is the paging capability of an underlying radio system. It refers to sending poll messages through wireless links to the cells within the last reported link layer location area. Link layer paging is tightly coupled with the specific wireless technology [4]. Note that an IP paging scheme does not make any assumption on how the underlying link layer paging is implemented. When both IP layer paging and link layer paging are supported, a layer 3 location area should be mapped to layer 2 location areas. Moving out of a layer 2 location area does not necessarily imply moving out of a layer 3 location area, and vice versa. Thus, there is possibility that the corresponding subnets of two geographically closed cells are not close together in Internet. Therefore, when applying the ideas of link layer multi-step paging schemes to Mobile IP, differences between layer 2 and layer 3 paging should be paid attention and modifications are needed.

In this chapter, we introduce the concept of user independent paging where the paging criterion is not based on individual user information. The goal of user independent paging is to provide satisfactory overall performance of the whole system, although the paging performance for each user may not be optimal. User independent paging can be applied to both link layer paging and IP layer paging, as long as the selected paging criterion is user independent. In this chapter, we choose the mobility rate of each subnet as the paging criterion for its easy implementation and propose a user independent paging scheme for IP mobility. We focus on how the IP paging messages are sent by the paging initiator to the FAs within a paging area over the wired links. We do not change the link layer paging support by which each FA sends out poll messages through wireless links to locate an MN. The proposed scheme can be employed by the current proposed IP paging architectures: home agent paging [18], foreign agent paging [19] [20], domain paging [18], IDMP-based paging [50], hierarchical paging [17] [21], etc. The proposed scheme is user independent in the sense that the partition of paging areas is determined by the aggregated movements of all mobile users, without the knowledge on the behavior of individual ones. Moreover, we also explain how to obtain the mobility rate in the proposed scheme.

The proposed user independent paging scheme was introduced in [51]. This chapter is organized as follows. In Section 3.2, the proposed paging algorithm is explained. In Section 3.3, the analysis of the paging cost for the proposed scheme and another scheme are given. In Section 3.4, analytical results under various scenarios are presented.

3.2 User Independent Paging Scheme for Mobile IP

In this section, we introduce the new user independent IP paging scheme. We also present a solution for implementing the proposed scheme in real systems.

3.2.1 Overview of the User Independent Paging Scheme

The proposed user independent paging scheme is a multi-step paging scheme which limits the number of paging steps below a pre-designed value while reducing the average paging cost. Instead of broadcasting IP paging messages, the paging initiator pages an idle MN in smaller areas sequentially. Specifically, the registered FA first locates the MN in its last registered subnet. If the MN is in the last registered subnet, the paging procedure is terminated. The MN replies to the paging initiator so that the data packets can be forwarded. If the MN is not in the last registered location, it means the MN has moved since last location update. Then the paging initiator divides the remaining subnets into several partitions based on the up-to-date subnet mobility rates and sends out IP paging messages in the decreasing order of mobility rates in the following steps. In other words, the paging initiator first sends paging messages to the FAs of the subnets with the highest user mobility rates, i.e., the areas with more *new idle* MNs moved in within the latest time period. If there is no response within a timeout interval, the paging initiator sends out paging messages to the subnets with lower mobility rates. We will describe how to obtain the up-to-date user mobility rate of each subnet in the next section. The total number of partitions is determined by the maximum paging delay the user may tolerate. Therefore, the proposed paging scheme is a combination of *last-location-first* paging and *highest-mobility-first* paging. Note that the paging procedure is neither based on geographic distance information which is not suitable for Internet environment, nor based on the user-variant predicted location probabilities which assumes perfect knowledge on the user movement statistics. The subnet partitioning is not dependent on the *individual* behavior of each user. Instead, it considers the *overall* aggregated user mobility as the basis. Therefore, the proposed paging scheme is user independent.

The partitioning method of subnets other than the last registered subnet is flexible. Here, we employ similar idea to that of the uniform paging proposed in [49], i.e., the

number of subnets in each partition is of approximately the same. For example, assume there are totally N subnets in a paging area. If the maximum number of paging steps is $\mathcal{L}$, then after first paging the last registered location, the remaining $N-1$ subnets are evenly divided into $\mathcal{L}-1$ groups based on user mobility rates. Similar to the calculation in [49], assume

$$n = \left\lfloor \frac{N-1}{\mathcal{L}-1} \right\rfloor \tag{45}$$

where $N-1 = n(\mathcal{L}-1)+k$ and k is an integer less than $\mathcal{L}-1$. Then, from the second to the $\mathcal{L}-1-k$ paging steps, n subnets are paged each time and $n+1$ subnets are paged during each of the following k paging steps. The n subnets with the n highest user mobility rates got the paging messages simultaneously.

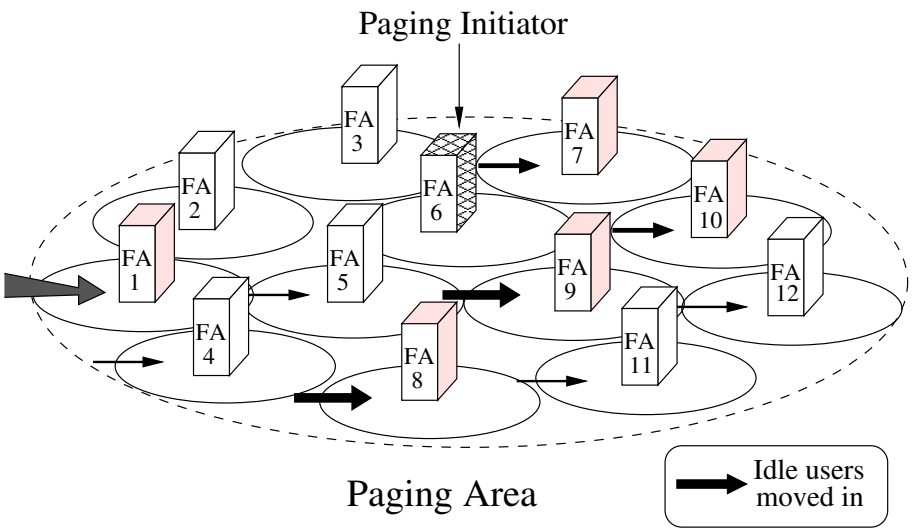

Figure 20: User independent paging scheme based on location and mobility rate.

The proposed paging scheme is illustrated in Figure 20. In the figure, there are totally 12 FAs within a paging area, i.e., $N = 12$. Assume foreign agent paging architecture [19] [20] is used in this case. FA6 is the registered FA and is responsible to initiate paging requests. The width of the solid arrows in the figure is proportional to the number of new idle users moved into each subnet within the latest time period. Assume the total number of paging steps is set to be 3, i.e., $\mathcal{L} = 3$. During the paging

procedure, FA6 first checks whether the paged MN is in its subnet. If not, since $\left\lfloor \frac{N-1}{\mathcal{L}-1} \right\rfloor = 5$ and the mobility rates of subnets 1, 7, 8, 9, and 10 are the five highest ones, FA6 sends paging messages to FA1, FA7, FA8, FA9, and FA10 in the second step. Each of the above FA checks whether the MN is in its subnet. If the MN is found, then the paging procedure is terminated. If not, FA6 continues to send paging messages to the remaining FAs in the paging area.

Note that unlike in PCS system, the number of paging steps is not necessarily proportional to the paging delay in Mobile IP. The transmission of paging messages in IP core network is a multi-hop transmission where queuing delays may influence the transmission time of each paging message, depending on the traffic load of the network. Thus, the paging delay of each paging step varies within a certain range. However, the air interface in the wireless network is treated as a single hop. Therefore, the paging delay of each polling cycle in PCS system is generally considered as a constant. A comprehensive analysis on the calculation and estimation of end-to-end delay bound of Internet services is provided in [52]. Based on the paging delay requirement of each application and the general end-to-end delay bound for IP packet transmission, the system may set the appropriate number of paging steps for each user. In this chapter, we use the term "the maximum number of paging steps" for $\mathcal{L}$, instead of the conventional term "delay bound" as used in PCS system.

3.2.2 Detailed Solution for Obtaining Subnet Mobility Rate

Since MNs in idle mode do not perform location registrations while roaming within a paging area, FAs have no knowledge on how many idle MNs are visiting their subnets. Each FA keeps an updated visitor list of all the active MNs in its subnet as well as idle MNs who have performed idle mode location registrations through the FA when changing paging areas.

In order to implement the new paging scheme, we propose a solution for each FA to

obtain the up-to-date user mobility rate of its subnet. We introduce a new operation mode for MNs named "*semi-idle*" mode. When an MN is in semi-idle mode, the system still does not know the accurate location of the MN. But unlike in idle mode, the MN in semi-idle mode provides minimum user information to the corresponding FA of the subnet it is visiting. Thus, based on this minimum user information, each FA has some knowledge on how many idle MNs are in its subnet. More importantly, MNs in semi-idle mode still save battery consumption compared to working in active mode. The state transition diagram of MNs for the proposed paging scheme is shown in Figure 21 and the detailed procedure is explained below.

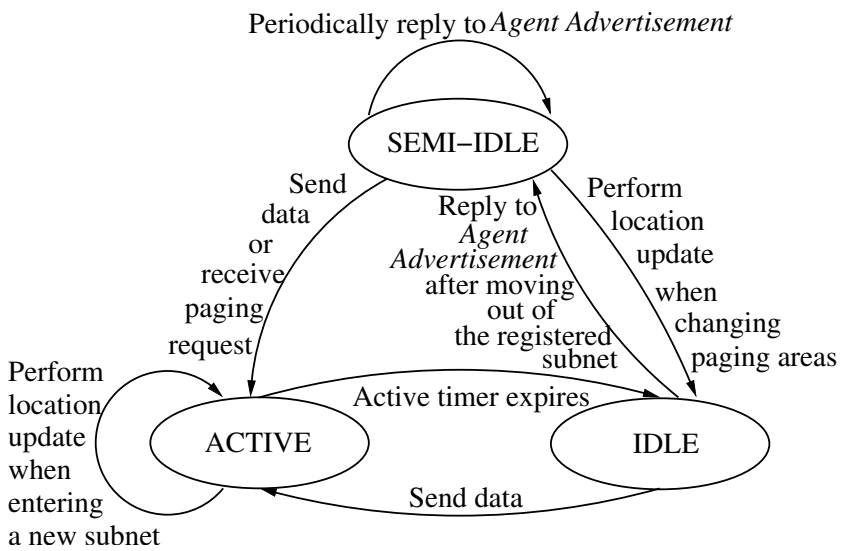

Figure 21: State transition diagram of the proposed paging scheme.

3.2.2.1 Operations at MNs

An MN is able to detect whether it has moved into a new subnet by periodically receiving unsolicited *Agent Advertisement* messages broadcasted from each FA [13]. If paging is supported, the MN and the visited subnet agree on communication time slots used for Agent Advertisement and paging, to restrict link interface power-on time in the MN [17]. Under this mechanism, an MN in idle mode powers on its receiver when an unsolicited Agent Advertisement or a paging request is expected,

and keeps its receiver powered off at other time slots. If an idle MN finds that it is in a subnet other than its registered subnet, the MN extends its power-on time slots a little bit and replies to the Agent Advertisement message by sending its home address and the registered CoA to the corresponding FA of its current subnet. The registered CoA is obtained from the registered FA when the MN changes paging areas and performs idle mode location registration. The content of this reply is the minimum basic information of the MN. Note that replying to the Agent Advertisement message is different from performing a location registration: the idle MN does not get a new CoA from the current FA and it does not send a location update message to the HA. Therefore, the signaling delay and the power consumption of replying to Agent Advertisement message are much less than those of performing a location registration. An MN in semi-idle mode changes to idle mode when it enters a new paging area and performs an idle location update to its HA. The corresponding FA of the registered subnet adds the MN to the visitor list and marks its mode as idle. The MN stays in the idle mode as long as it does not move out of the registered subnet. An MN in idle mode does not need to reply to Agent Advertisement message.

3.2.2.2 Operations at FAs

After receiving the reply message from an MN, the FA compares the CoA of the idle MN with its network prefix. If they are the same, the FA refreshes the idle state of the MN on its visitor list. If they are different, it implies that the MN is in idle mode but registered with another FA. The current FA adds the MN to the visitor list and marks its mode as semi-idle. The MN does not have to reply to every Agent Advertisement message. It may reply to the Agent Advertisement message periodically for every $\mathcal{M}$ advertisement slots to refresh its semi-idle state in the visitor list of the current FA. Same as active and idle states, the semi-idle state is also a soft state. If there is no further refreshment message, the state is expired. Therefore, when the MN leaves the

current subnet, it does not need to send a cancellation message actively to remove its record on the visitor list. As a result, the total number of actual users in each subnet is the number of users in active mode as well as those in idle and semi-idle modes. Each FA keeps a counter. If within a time period $\mathcal{T}$, there is a *new* MN in semi-idle mode added to the visitor list, the counter is incremented by one. At the end of the time period, the counter is reset. So the counter value indicates the mobility rate of new idle MNs within the latest time period $\mathcal{T}$. The operations at FAs are illustrated in Figure 22.

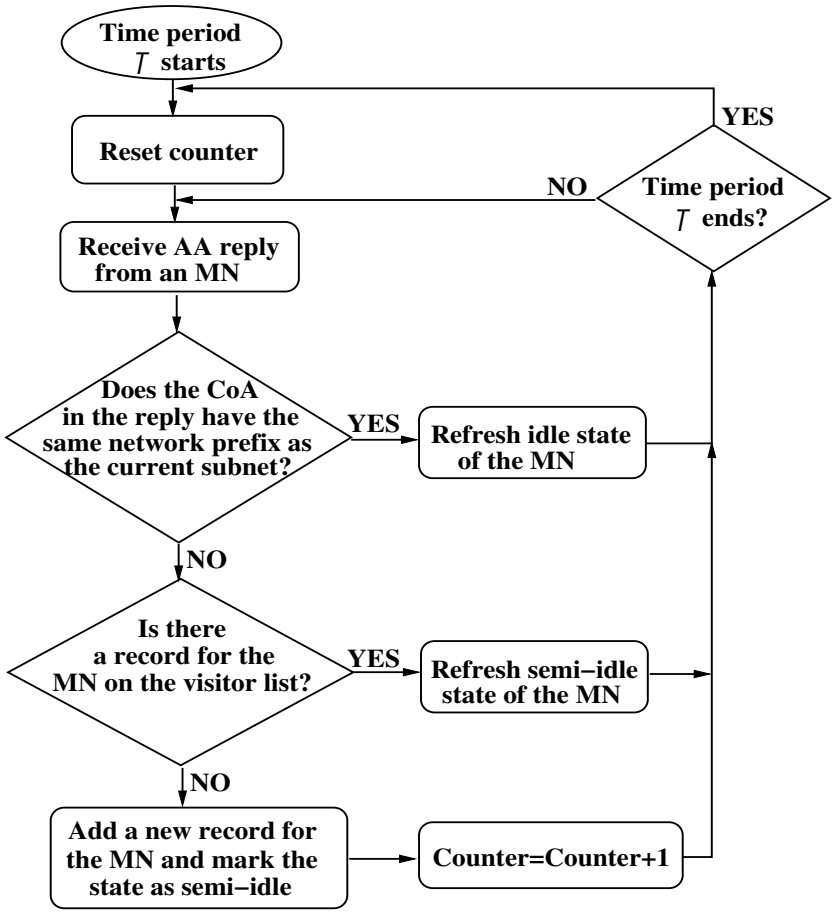

Figure 22: Flowchart of operations at FAs.

At the end of each time period $\mathcal{T}$, all FAs within a paging area exchange the user mobility information. They agree on the time slots of this exchange. When there is a paging request for an idle MN, the registered FA first checks its visitor list. If the

paged MN has a record of idle state on the list, it means the paged MN is in the subnet of the registered FA. The paging procedure is terminated. If the paged MN is not on the list, the paging initiator sends out IP paging messages to subnets in decreasing order of user mobility rates. When an FA receives an IP paging message, it checks the visitor list. If the MN is on the list with semi-idle state, the FA pages the MN over the air so that the MN may reply to the paging initiator and register its current location to the HA.

3.2.3 Advantages of the User Independent Paging Scheme

The most significant feature of the proposed paging scheme is its user independent nature. This feature is reflected in the partition of subnets other than the last registered subnet of an MN. In contrast to the user dependent paging schemes that perform partitioning differently for each user according to the user-variant parameters, the proposed scheme uses one mutual criterion which is related to the aggregated behavior of all users. Although the aggregated mobility rate of each subnet cannot represent the behavior of an individual user precisely sometimes, we will demonstrate in Section 3.4 that the performance of the proposed scheme is comparable to that of user dependent paging schemes based on the assumption that perfect knowledge on user location probabilities are known. After that, we will show that the proposed scheme has remarkable advantages when paging multiple users simultaneously and when the assumption of perfect knowledge on user location probabilities is loose.

Moreover, the proposed paging scheme combines the advantages of last-location-first paging, highest-mobility-first paging, and uniform paging. In the case of networks with low mobility users, searching the last known location of the paged MN can save the paging cost significantly. In the case of networks with high mobility users, uniformly grouping the subnets based on the overall mobility rate of all users and paging each group in the decreasing order of mobility rates can also reduce the paging

load and paging delay.

The advantages of the proposed user independent paging scheme are summarized as follows.

- The proposed scheme has the common advantage of multi-step paging schemes, that is, compared with broadcast paging procedure, the average number of paging messages sent out is reduced under the new scheme, since the system will find the user in the last registered location and high mobility rate areas with high probability.

- The location probability is determined by many factors, such as user movement model and calling pattern [53]. The accuracy of user movement model and the efficiency of the location probability prediction algorithm directly determine the capability of paging schemes. The mobility rate in the proposed scheme is the up-to-date value for each subnet. It represents the exact amount of movements within the latest time period. Therefore, compared with the paging schemes choosing location probabilities as paging criterion [47] [48], the proposed paging scheme results in better system performance, especially when there is some unusual changes to the normal value of user movements.

- Unlike the schemes based on user-variant parameters, the proposed scheme is based on the mobility rate of each subnet for all the users. So paging request aggregation [20] can be adopted and the overall system performance can be further improved when paging multiple users at one moment.

- The proposed scheme is scalable. We will show in Section 3.4 that as the number of MNs in a paging area increases, the average paging cost caused by the UIP scheme does not change much.

3.2.4 Tradeoffs of Introducing "Semi-Idle" State

The introduction of the new operation mode, semi-idle mode, will cause extra signaling overheads due to the periodically replying to Agent Advertisement messages from MNs and message exchanges between FAs. These extra overheads will consume additional bandwidth and battery resources. The impact of the extra signaling overheads can be reduced by setting a relatively large value for the period of message exchanges. There is a tradeoff between the performance of the proposed scheme and the additional resource consumption. The more often the MNs and FAs exchange mobility information, the more accurate the mobility rate values are, and the better performance the proposed scheme may achieve, but the more additional overheads. On the other hand, a small error of the mobility rate values does not necessarily affect the performance of the proposed scheme, since it is the ranking of the mobility rate of each subnet among all the mobility rates that determines the partition of a paging area, instead of the absolute values.

We will show in the following section that compared with a user dependent paging scheme where the location probabilities are given in user profiles, the extra cost of introducing semi-idle mode is comparable to the extra cost of setting up user profiles. Therefore, the overheads caused by the introduction of semi-idle mode can be treated as the maintenance cost for the system to obtain accurate mobility rate information in order to employ the proposed user independent paging scheme.

3.3 Analytical Model

In this section, we derive the expected paging cost of the proposed user independent paging (UIP) scheme. We choose a user dependent paging scheme which is purely based on the location probabilities of each MN for our performance comparison. Here, for IP paging schemes, location probabilities are the probabilities that an MN will visit each subnet. They are different from the cell location probabilities in link layer

paging context. We assume the system has a user profile for each MN. In the user profile, the subnet location probabilities of an MN at different time of a day are provided which can be obtained either through empirical measurements or analysis of user movement models [54]. We call this scheme user profile paging (UPP) scheme.

3.3.1 Costs of Obtaining Mobility Rates and Setting Up User Profiles

In order to implement the proposed UIP scheme, the system provides extra resources to obtain mobility rates of each subnet. Similarly, in order to implement the UPP scheme, the system pays extra costs to set up a user profile for each mobile user and obtain location probability distributions. For both schemes, these extra costs can be summarized as: sampling and data transmission cost; data storage cost; and computation cost. More specifically, the costs of obtaining mobility rates of each subnet in UIP scheme include:

- Radio resource consumption of periodically replying to Agent Advertisement messages from each MN;
- Processing load on each FA of updating the visitor list after receiving the replies to Agent Advertisement messages;
- Wireline bandwidth consumption of mobility rate information exchange between FAs.

The costs of obtaining location probabilities in UPP scheme depend on how the user profile is defined and set up. User profiles can be provided and updated manually by mobile users or determined automatically by monitoring the movement history over a period of time [55]. Methods of obtaining cell location probability distributions in PCS systems are discussed in [49]. Similar methods can be applied to Mobile IP to set up subnet location probabilities. The costs of obtaining subnet location probabilities are:

- Radio resource consumption of periodically sending the location information to the system database from each MN, if the user profile is updated manually by the mobile user;

- Processing load of updating the system database. If the user profile is updated manually by the mobile user [56], after receiving the location information sent from each MN, the system updates the database; If the user profile is updated by the system, the system updates the database whenever the MN initiates communications or the MN receives packets from others;

- Processing load on the system database for MN velocity estimation, movement estimation, subnet residence time estimation, traffic condition estimation, etc., depending on how the user profile is defined. Each of the above computations may consume extra wireline and wireless bandwidth;

- Processing load of estimating and predicting location probabilities using mathematical models. An algorithm for location probability estimation in PCS systems is proposed in [57], where cell location probabilities depend on historic records, current position, velocity, and moving directions of mobile users.

From the above analysis, we may see that to predict location probabilities and to set up mobility profiles for each user require intensive computation. The extra costs of obtaining mobility rates for UIP scheme is comparable or less than the extra costs of obtaining location probabilities for UPP scheme, depending on how the user profile is defined. Hence, in this chapter, we consider the costs of obtaining mobility rates and setting up user profiles as the maintenance cost for the system to employ UIP and UPP schemes, and these costs are not counted in the paging cost comparison in the following.

3.3.2 Relationship Between Location Probabilities and Mobility Rates

In order to compare the expected paging costs of the proposed UIP scheme and UPP scheme, we first derive the relationship between user location probabilities and mobility rates. Let the total number of users in a paging area be M and the total number of subnets in a paging area be N. We assume the movements of a mobile user are independent with those of other users. Assume during the next time period $\mathcal{T}$, the probability that user x will move out of its current subnet to subnet y is $q_{x\to y}$, where $x = 1, 2, \ldots, M$ and $y = 1, 2, \ldots, N$. Now, we find the probability distribution that there are k users moved into subnet y during the next time period.

The probability that there is no user moved to subnet y is equal to the probability that all the users moved to other subnets except subnet y, i.e.,

$$P_y(K=0) = \prod_{x=1}^{M}(1 - q_{x\to y}) \tag{46}$$

where $\sum_{\substack{y=1\\ y\neq v}}^{N} q_{x\to y} = 1$, and v represents the subnet user x is currently visiting. The probability that there is one user moved to subnet y during the next time period is:

$$\begin{aligned} & P_y(K=1) \\ = \; & q_{1\to y}\textstyle\prod_{\substack{x=1\\ x\neq 1}}^{M}(1 - q_{x\to y}) + q_{2\to y}\prod_{\substack{x=1\\ x\neq 2}}^{M}(1 - q_{x\to y}) \\ & + \cdots + q_{M\to y}\textstyle\prod_{\substack{x=1\\ x\neq M}}^{M}(1 - q_{x\to y}) \\ = \; & \textstyle\sum_{l=1}^{M} q_{l\to y} \cdot \prod_{\substack{x=1\\ x\neq l}}^{M}(1 - q_{x\to y}) \end{aligned} \tag{47}$$

Similarly, the probabilities that there are two users and k users moved to subnet y during the next time period are shown in the following, respectively:

$$P_y(K=2) = \sum_{\substack{l_1,l_2=1\\ l_1\neq l_2}}^{M} q_{l_1\to y}q_{l_2\to y} \cdot \prod_{\substack{x=1\\ x\neq l_1,l_2}}^{M}(1 - q_{x\to y}) \tag{48}$$

$$P_y(K=k) = \sum_{\substack{l_1,\ldots,l_k=1\\ l_1\neq\cdots\neq l_k}}^{M} q_{l_1\to y}\cdots q_{l_k\to y} \prod_{\substack{x=1\\ x\neq l_1\neq\cdots\neq l_k}}^{M}(1 - q_{x\to y}) \tag{49}$$

where $k = 1, \cdots, M$.

Given the above probability distribution, we may get the expected number of users moved to subnet y during the next time period as:

$$E\left[P_y(K)\right] = \sum_{k=0}^{M} k \cdot P_y(K = k), \quad \text{where} \;\; y = 1, \cdots, N \tag{50}$$

$E\left[P_y(K)\right]$ can be treated as the mobility rate of subnet y in the proposed paging scheme, i.e., the number of new idle users moved into subnet y in a certain time period.

Note that we do not impose geographic proximity of subnets as the constraint on our mobility model as used in the analysis of PCS paging schemes [46] [48] [33] and other Mobile IP paging schemes [20].

3.3.3 Paging Costs

Now, we derive the paging cost functions for the proposed UIP scheme and the UPP scheme. We define the following parameters for our analysis:

V_{air} The wireless paging cost including broadcasting the polling messages over the air.

V_{wire} The wireline paging cost including transmission and processing of IP paging messages.

U_t The transmission cost of IP paging messages between the paging initiator and any other FA.

U_p The processing cost of IP paging messages at each FA.

$\mathcal{L}$ The maximum number of paging steps.

n_i The number of paged subnets in the ith paging step.

p_{out} The probability that the paged MN is not in the last registered subnet.

p_i The probability that the paged MN is residing in the ith paging group.

ω_i The paging cost spent when the paged MN is successfully located, given that the MN is residing in the ith paging group.

V_{wire} and V_{air} account for the wireline and wireless costs for bandwidth utilization and the computational requirements in order to process the paging messages. U_t may represent the delay cost for sending the paging message through a particular path. U_p may represent the computational cost for an FA to check its visitor list to find whether there is a record for an MN [43] [44]. V_{wire} can be expressed as:

$$V_{wire} = U_t + U_p \tag{51}$$

3.3.3.1 UIP Paging Scheme

The average paging cost of the proposed UIP scheme between two location tracking requests for each MN is:

$$E[C(\mathcal{L})]_{(UIP)} = \begin{cases} NV_{wire} + V_{air} & \mathcal{L} = 1 \\ (1 - p_{out})V_{wire} & \\ +p_{out} \cdot \sum_{i=2}^{\mathcal{L}} p_{i_{(UIP)}} \omega_{i_{(UIP)}} & 2 \leq \mathcal{L} \leq N. \end{cases} \tag{52}$$

$\omega_{i_{(UIP)}}$ can be calculated as:

$$\omega_{i_{(UIP)}} = \sum_{j=2}^{i} n_j V_{wire} + V_{air} \tag{53}$$

where n_j can be obtained from (45).

(49) and (50) give the relationship between location probabilities and mobility rates. Since we use mobility rate of all users in a subnet as the paging criterion in the UIP scheme, the probability that the paged MN is residing in the ith paging group, i.e., $p_{i_{(UIP)}}$ in (52), can be approximated as:

$$p_{i_{(UIP)}} = \frac{\sum_{y \in A_i} E\left[P_y(K)\right]}{\sum_{y=1}^{N} E\left[P_y(K)\right]} \tag{54}$$

where A_i is the set of subnets in the ith paging group. Note that $p_{i_{(UIP)}}$ is the same for all the users.

3.3.3.2 UPP Paging Scheme

The average paging cost of the UPP scheme between two location tracking requests for each MN is:

$$E[C(\mathcal{L})]_{(UPP)} = \sum_{i=1}^{\mathcal{L}} p_{i_{(UPP)}} \omega_{i_{(UPP)}} \tag{55}$$

where $\omega_{i_{(UPP)}}$ is:

$$\omega_{i_{(UPP)}} = \sum_{j=1}^{i} n_j (U_t + V_{air}) \tag{56}$$

Note that $p_{i_{(UPP)}}$ is user-variant. For different MNs, $p_{i_{(UPP)}}$ is different. Assume for user x, the probability that user x is in the ith paging group at the paging moment can be calculated as:

$$p^x_{i_{(UPP)}} = \sum_{y \in A_i} \pi^x_y \tag{57}$$

π^x_y is different from $q_{x \to y}$. π^x_y is the location probability that user x is in subnet y at the paging moment, while $q_{x \to y}$ is the transition probability that user x moves out of its current subnet to subnet y. They can be related as:

$$\pi^x_y = p_{out} \cdot q_{x \to y} \tag{58}$$

Note that for (54), when there is only one user inside the paging area, the value of $E[P_y(K)]$ is between [0, 1], and $\sum_{y=1}^{N} E[P_y(K)] = 1$. Therefore in this case, $p_{i_{(UIP)}}$ calculated based on mobility rate is exactly the same as $p_{i_{(UPP)}}$ in (57) for a specific user.

3.4 Performance Evaluation

In this section, we demonstrate the performance comparison between the UPP scheme and the proposed UIP scheme. Based on (52) and (55), we compare the paging costs of these two schemes for various scenarios. We will first investigate the paging cost based on the assumption that accurate location probability information is provided in each user profile. Then we will consider the case when there is discrepancy between the actual location probabilities and the ones provided in user profiles.

For the analysis in this chapter, we assume the costs for transmitting and processing paging messages are available. These costs account for the wireless and wireline bandwidth utilization and the computational requirements in order to process the paging messages [46]. The methods for determining the cost parameters are discussed in [43] [44]. Table 2 lists the cost parameters used in our performance analysis. For performance comparison purpose, all the cost parameters are normalized to U_t such that $U_t = 1$. We consider two sets of cost parameters. The selected data sets allow us to study the effect of varying the ratio of V_{air} to U_p on the performance of the proposed paging scheme. Since generally the wireless resource is more scarce compared with wireline bandwidth, the transmission cost over the wireless link is several times higher than the wireline transmission cost. In our analysis, we set the paging cost of broadcasting the polling messages over the air, V_{air}, the same and twice higher than the processing cost at each FA in the two sets, respectively. We also assume there are totally 20 subnets in a paging area, i.e., $N = 20$, and the total number of users in a paging area is 100, i.e., $M = 100$. Note that experiments for different values of M are conducted, and for most of the results we show below, the average paging cost of the UIP scheme does not change much when $M \geq 20$.

Table 2: Cost Parameters for Evaluating Paging Schemes in Mobile IP

Cost Parameters	Set 1	Set 2
V_{wire}	3	3
U_t	1	1
U_p	2	2
V_{air}	2	4

3.4.1 Paging A Single User

First, we compare the paging costs when the system pages one user at a time.

In (52), the paging cost is dependent on the user mobility parameter p_{out}. In order to make (52) and (55) comparable, we introduce a "*virtual subnet*" concept.

We assume at the end of the last time period $\mathcal{T}$, all the users are in a virtual subnet which is located outside the whole paging area. During the current time period, all the users move out of the virtual subnet into subnets inside the paging area under consideration. Thus, $p_{out} = 1$ and $\pi_y^x = q_{x \to y}$ in this case. The advantage of the introduction of virtual subnet is that the subnet dependent parameter p_{out} is removed from our analysis and the performance comparison is simply between paging based on user mobility rate and paging based on subnet location probability. However, under this assumption, we lose the chance that each MN stays in its last registered subnet. Therefore, the paging cost calculated under this virtual subnet assumption is the upper bound of the paging cost for UIP scheme. (52) is then changed to:

$$E[C(\mathcal{L})]_{(UIP)} \leq \begin{cases} NV_{wire} + V_{air} & \mathcal{L} = 1 \\ \sum_{i=2}^{\mathcal{L}} p_{i_{(UIP)}} \omega_{i_{(UIP)}} & 2 \leq \mathcal{L} \leq N. \end{cases} \tag{59}$$

$\omega_{i_{(UIP)}}$ and $p_{i_{(UIP)}}$ are obtained from (53) and (54), respectively.

3.4.1.1 Uniform Location Probability Distribution

We first study the relationship between the upper bound of the paging cost of the proposed UIP scheme and the paging cost of UPP scheme, under uniform location probability distribution.

Figure 23 plots the average paging cost as a function of the maximum number of paging steps, $\mathcal{L}$, when the subnet location probabilities of all the users are uniformly distributed. Since all the users are equally likely to be anywhere in the paging area, the total number of users in a paging area, M, does not influence the result.

In Figure 23, the solid lines are for parameter set 1 when $V_{air} = U_p$, while the dashed lines are for parameter set 2 when $V_{air} = 2U_p$. It can be seen from the figure that the average paging costs decreases as the maximum number of paging steps increases. This result is consistent with the conclusions given in [49]. When all the users are uniformly located in each subnet, the probability of UIP scheme calculated

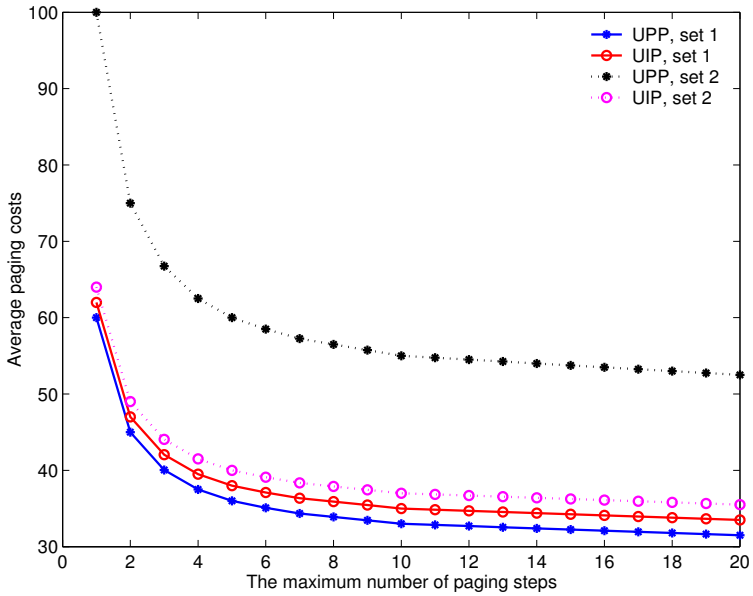

Figure 23: The average paging cost for uniform location probability distribution, when paging a single user.

based on user mobility rate, $p_{i_{(UIP)}}$, is exactly the same as the probability of UPP scheme calculated based on location probability provided in user profiles, $p_{i_{(UPP)}}$. It is observed in Figure 23 that the upper bound of the paging cost of the proposed UIP scheme is slightly higher than that of the UPP scheme, when $V_{air} = U_p$. The maximum difference is only 5%. However, when V_{air} is larger than U_p, UIP scheme may save paging cost significantly. Our results demonstrate that up to 34% cost can be saved by the proposed UIP scheme when $V_{air} = 2U_p$. This is because under UIP scheme, after receiving the paging request from the paging initiator, each FA does not need to send polling messages to all the MNs in its subnet through wireless links. Instead, it consumes processing cost U_p by checking its visitor list of semi-idle users. Therefore, the more precious the wireless resources are, the more cost can be saved by the proposed UIP scheme.

3.4.1.2 Truncated Gaussian Location Probability Distribution

Truncated Gaussian distribution is a typical location probability distribution for systems under isotropic random motion [58]. The discretized version of truncated Gaussian distribution with zero mean mentioned in [47] is expressed as:

$$\pi_y = \frac{1}{\mathrm{erf}\left(\frac{N}{\sigma}\right)} \frac{2}{\sqrt{2\pi\sigma^2}} \int_{y-1}^{y} e^{-\frac{x^2}{2\sigma^2}} dx, \qquad \text{where} \quad y = 1, \cdots, N \tag{60}$$

σ^2 is the variance and $\mathrm{erf}(\cdot)$ is the error function defined as:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{61}$$

Note that the $p_{i_{(UIP)}}$ calculated based on mobility rate is an aggregated result of the motions of all MNs. Theoretically, the distribution of $p_{i_{(UIP)}}$ can be of any type depending on the mobility pattern of each user. The more chaotic and irregular the mobility patterns of all MNs are, the closer the aggregated $p_{i_{(UIP)}}$ approaches to uniform distribution.

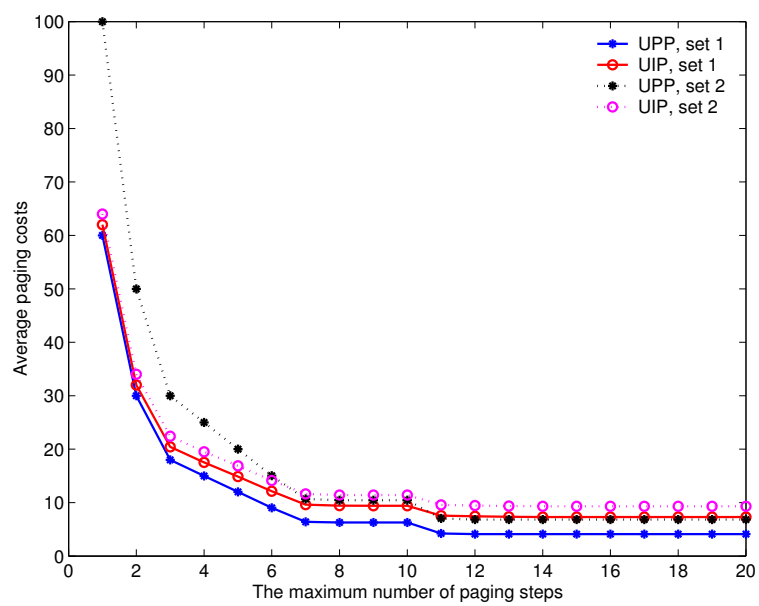

Figure 24: The average paging cost for shifted truncated Gaussian location probability distribution, when paging a single user.

Figure 24 shows the upper bound of the average paging cost of the UIP scheme and the average paging cost of the UPP scheme, when the user location probabilities are

of truncated Gaussian distribution with mean zero and variance one. Since MNs are located in different subnets, we circularly shift the values in the location probability distribution of each MN by sizes ranging in [0, 20]. Thus, the index of the mean of each shifted distribution indicates the most likely subnets an MN will be.

When user location probability is of truncated Gaussian distribution, the average paging cost drops very quickly when $\mathcal{L}$ varies from 1 to 6. This result is also similar to the conclusion in [49]. It is noticed from Figure 24 that the upper bound of the average paging cost of the UIP scheme is still slightly higher than that of the UPP scheme, when $V_{air} = U_p$. The maximum difference is 9%. When $V_{air} = 2U_p$, the UIP scheme results in lower paging cost when $\mathcal{L} \leq 6$. The cost saving is up to 33%. For $\mathcal{L} > 6$, the UIP scheme pays slightly higher cost compared with UPP scheme. This is due to the discrepancy between the aggregated probability $p_{i(UIP)}$ and the individual probability $p_{i(UPP)}$.

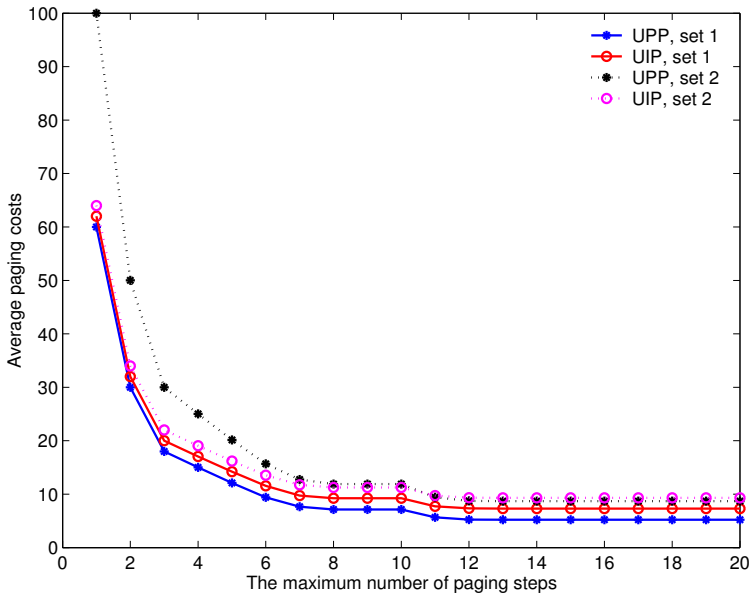

Figure 25: The average paging cost for variant truncated Gaussian location probability distribution, when paging a single user.

Figure 25 gives the upper bound of the average paging cost of the UIP scheme and the average paging cost of the UPP scheme, when the user location probabilities

are of truncated Gaussian distribution with different variances. Standard deviation σ is chosen between [1.0, 2.0] for different MNs. Standard deviation of the subnet probability distribution of the paged MN is assumed to be 1.5.

Figure 25 shows a similar result to that in Figure 24, i.e., when $V_{air} = 2U_p$, the upper bound of the average paging cost of the UIP scheme is reduced by up to 34% when $\mathcal{L} \leq 10$. For $\mathcal{L} > 10$, the upper bound of the paging cost of the UIP scheme is slightly higher than the paging cost of the UPP scheme. We also perform experiments when σ of the subnet probability distribution of the paged MN changes from 1.0 to 2.0. Our results show that as σ of the paged MN increases, more paging cost can be reduced by the proposed UIP scheme compared with UPP scheme. When σ of the paged MN is small, the likely locations of the MN is concentrated on a small portion of subnets with high probabilities. The average paging cost is small in this case. On the other hand, when σ is large, the likely locations of the paged MN will be stretched to more subnets, which leads to a higher average paging cost. The mobility rate employed in the UIP scheme is an aggregated result of mobility behaviors of all MNs. Thus, the performance of the UIP scheme based on the aggregated $p_{i_{(UIP)}}$ should be in between of these two cases.

3.4.1.3 User-Variant Location Probability Distribution

The location probability distribution of each MN may not be the same. Next, we investigate the impact of user-variant location probability distribution. We assume there are two groups of users. The location probability of the first group users follows a truncated Gaussian distribution with mean zero and variance one. The location probability of the other group users follows a uniform distribution. Assume that each group has 50% of the total users. The paged MN is randomly chosen from all the users. Its location probability may follow either a truncated Gaussian distribution or a uniform distribution.

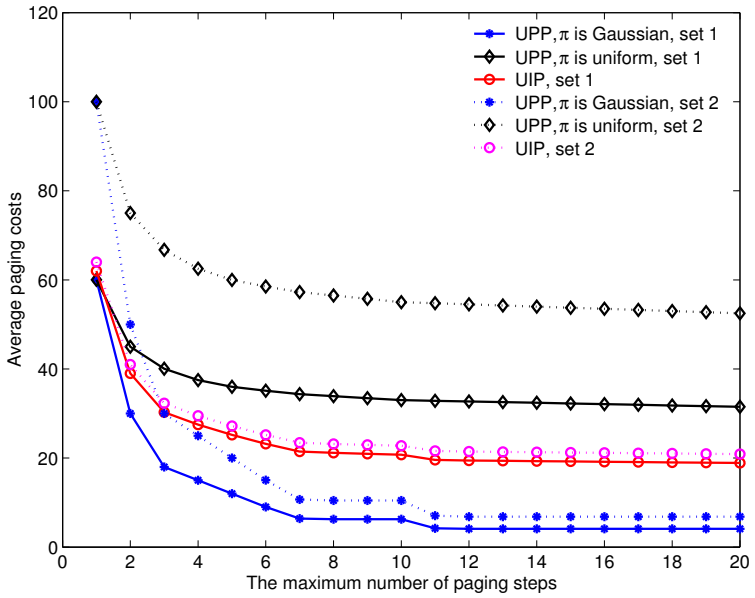

Figure 26: The average paging cost for user-variant location probability distribution, when paging a single user.

Figure 26 plots the average paging costs for user-variant location probability distribution. Two cases of the UPP scheme are considered. One is for the paged MN with a truncated Gaussian distribution. The other is for the paged MN with a uniform distribution. Our results show that the upper bound of the paging cost of the UIP scheme is larger than the paging cost of the UPP scheme when the subnet probability of the paged MN follows a truncated Gaussian distribution, but less than the case when the location probability of the paged MN is uniformly distributed. It is concluded in [47] that the uniform distribution achieves the worst performance of any distribution in the sense that it results in more paging cost. When the location probabilities of a specific user are of uniform distribution, the UPP scheme achieves the worst performance. It is equivalent to the case that the system does not have any information on the future locations of the user at all. This may possibly happen in real systems. However, for UIP scheme, the worst case corresponds to the uniform distribution of mobility rates which is equivalent to the case that all the subnets have the same mobility rates. In other words, within the same time period, all the subnets

have the same number of new idle users moved in. This is unlikely to happen.

3.4.1.4 Summary

Note that the performance results of the proposed UIP scheme are obtained under the worst case: the average paging costs shown are the upper bound, i.e., the information of the last registered location is not considered, and the cost of paging through wireless links is the same or only twice higher than the processing cost at each FA. The actual paging cost of the UIP scheme is lower than that in above results if the last registered location information is incorporated or the wireless resources are more precious.

From the above results, we may conclude that for the case of paging one user at a time, the performance of the proposed UIP scheme is comparable with that of the UPP scheme. When there is not much information on individual user behavior, the proposed UIP scheme performs better than the UPP scheme.

3.4.2 Paging Multiple Users

Since the system and MNs synchronize on the time slots for paging, there are possibilities that multiple paging requests are needed to be sent out in one time slot. When there are multiple users to be paged at one moment, the paging initiator may aggregate all paging requests into a single paging message. This optimization method is described in [20] and the paging message format for paging multiple MNs is also illustrated. Paging request aggregation helps to reduce the paging overhead as the number of paged MNs increases. However, this method is not suitable for UPP scheme since each paged MN follows different location probability distributions. Different partitions of the paging area are used for each user.

We compare the paging costs of UPP and UIP schemes for paging multiple MNs at one moment. We assume the paging initiator and each FA have the ability to aggregate multiple paging requests and send out one single paging message to other FAs and through wireless links, respectively.

Since it is hard to implement paging request aggregation for UPP paging scheme, the system processes each paging request separately when there are more than one users to be paged at a time. The total average paging cost of the UPP paging scheme is:

$$E[C(\mathcal{L})]_{(UPP)} = \sum_{x=1}^{\Omega} \sum_{i=1}^{\mathcal{L}} p^{x}_{i_{(UPP)}} \omega_{i_{(UPP)}} \tag{62}$$

where Ω is the total number of MNs to be paged at one time. $\omega_{i_{(UPP)}}$ can be calculated according to (56). $p^{x}_{i_{(UPP)}}$ is user-variant and it is defined in (57). Here we assume the number of partitions for all the paged MNs are the same, i.e., the paged MNs have the same paging delay requirement.

We still incorporate the "*virtual subnet*" concept and compute the upper bound of the paging cost of UIP scheme. When multiple MNs are found on the visitor lists of FAs during a paging step, those FAs poll the single or multiple users wirelessly in their subnet. So the worst case is that all the n_i FAs in the ith paging group send out a polling message through wireless links. The upper bound of the average paging cost of the UIP scheme is:

$$E[C(\mathcal{L})]_{(UIP)} \leq \begin{cases} N(V_{wire} + V_{air}) & \mathcal{L} = 1 \\ \sum_{i=2}^{\mathcal{L}} p_{i_{(UIP)}} \omega_{i_{(UIP)}} & 2 \leq \mathcal{L} \leq N. \end{cases} \tag{63}$$

$p_{i_{(UIP)}}$ is the same for all the users and it is defined in (54). Here $\omega_{i_{(UIP)}}$ is:

$$\omega_{i_{(UIP)}} = \sum_{j=2}^{i} n_j (V_{wire} + V_{air}) \tag{64}$$

Note that the upper bound of the average paging cost of the UIP scheme is independent of the number of the paged MNs at one time, Ω. Therefore, the UIP scheme is very efficient in saving the system resource when paging multiple MNs, since the upper bound of the average paging cost does not change as the number of paged users grows, if the last location information is ignored.

Figure 27 and Figure 28 show the average paging cost for user-variant location probability distribution, when paging multiple MNs at a time. The upper bound of

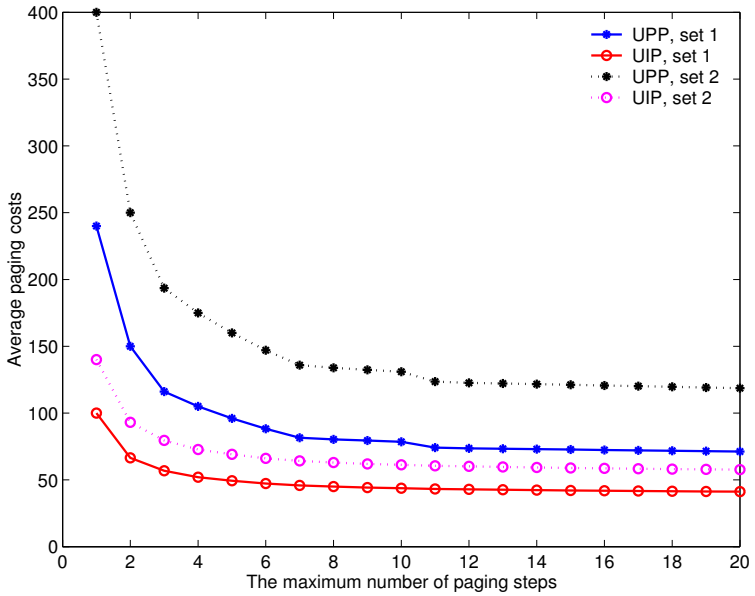

Figure 27: The average paging cost for user-variant location probability distribution, when paging multiple users, $\Omega = 4$.

the paging cost of the UIP scheme does not change when the total number of paged MNs, Ω, changes from 4 to 8 in the two figures. From the figures we may see that the UIP scheme reduces the paging cost drastically as Ω increases. When $\Omega = 4$, the UIP scheme saves up to 50% and 58% cost compared to the UPP scheme for parameter set 1 and set 2, respectively. When $\Omega = 8$, the maximum cost reduction is 75% and 79% for set 1 and set 2, respectively.

The advantage of the UIP scheme is obvious in this case. The user independent nature of the UIP scheme determines that it is very convenient and cost-efficient to page multiple users using UIP scheme and paging request aggregation. Considering the complexity of performing subnet partition and polling paging request individually for each MN when using UPP scheme, the UIP scheme is more preferable.

3.4.3 Confidence of Location Probabilities

For the above results, we assume perfect user location information is provided in the user profiles, i.e., accurate subnet probability distribution as the basis for the UPP scheme. Now, we study the impact of using erroneous location information.

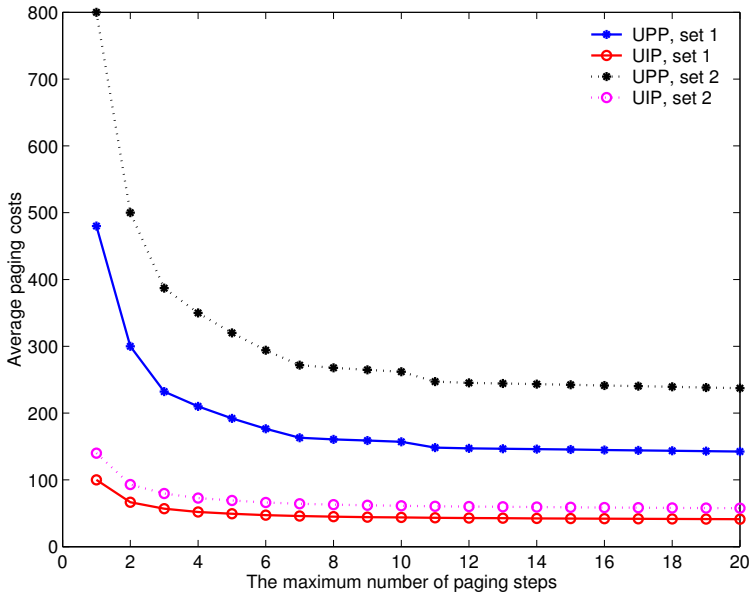

Figure 28: The average paging cost for user-variant location probability distribution, when paging multiple users, $\Omega = 8$.

Assume the confidence level of the prediction of the location probability is α. Confidence level implies the uncertainty of the predicted parameter. It provides a range of plausible values for the unknown parameter. In order to investigate the effect of confidence level on the paging schemes, we consider a simple model where the actual locations of each MN follow the truncated Gaussian distribution with mean zero and variance one. However, since the user profiles give imperfect location information with confidence α, there is $(1-\alpha)$ possibility that the UPP scheme uses erroneous location information for paging. When erroneous location information is used, it is equivalent that the system does not have any information of users at all. Thus, $(1-\alpha)$ percent of the total paging cost comes from the calculation based on uniform location distribution, i.e.,

$$E[C(\mathcal{L})]_{(UPP)} = \alpha E[C(\mathcal{L})]^{g}_{(UPP)} + (1-\alpha)E[C(\mathcal{L})]^{u}_{(UPP)} \tag{65}$$

where $E[C(\mathcal{L})]^{g}_{(UPP)}$ is the paging cost computed based on shifted truncated Gaussian distribution and $E[C(\mathcal{L})]^{u}_{(UPP)}$ is the paging cost computed based on uniform distribution. Since the paging cost of the UIP scheme is calculated based on the actual

user mobility rate, the confidence level does not influence its result and the upper bound of the paging cost is expressed in (52).

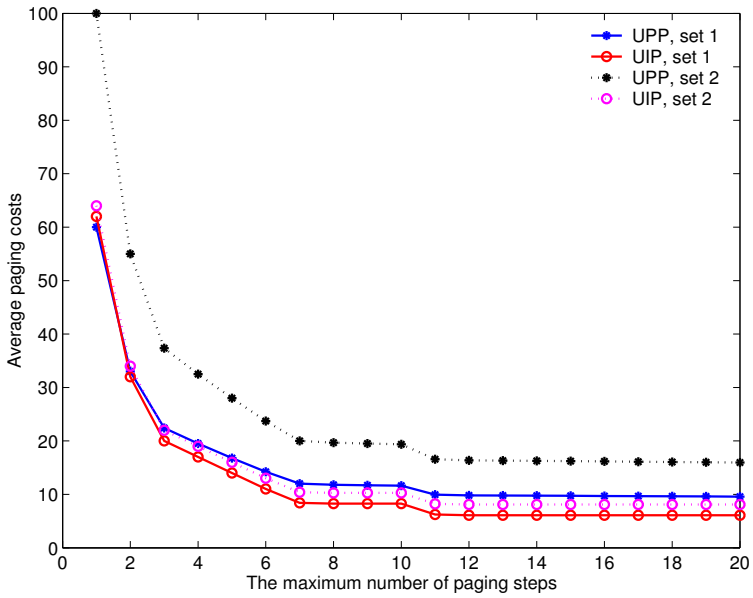

Figure 29: The average paging cost when imperfect location information is used, $\alpha = 80\%$.

Figure 29 and Figure 30 plot the average paging cost of the two paging schemes, when $\alpha = 80\%$ and $\alpha = 90\%$, respectively. Comparing with Figure 24, the results indicate that when imperfect location information is used, the paging cost of the UPP scheme increases. The increase percentage is 16% when $\alpha = 80\%$ and 10% when $\alpha = 90\%$. In both figures, the UIP scheme pays less cost compared with the UPP scheme. Therefore, UIP scheme provides relatively good performance consistently.

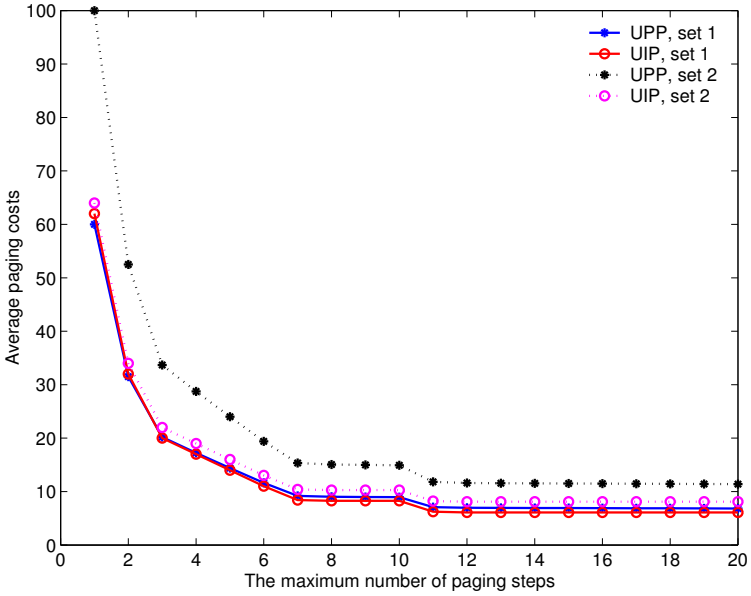

Figure 30: The average paging cost when imperfect location information is used, $\alpha = 90\%$.

CHAPTER IV

PAGING-AIDED CONNECTION SETUP FOR REAL-TIME COMMUNICATION IN MOBILE INTERNET

4.1 Problem and Solution

Mobile IP is a solution for mobility on the global Internet [13] [25]. It has been standardized by Internet Engineering Task Force (IETF) to provide continual Internet connectivity to mobile users. Mobile IP introduces three new functional entities: home agent (HA), foreign agent (FA), and mobile node (MN). When an MN moves out of its home network, it obtains a temporary address: care-of address (CoA). This address is used to identify the MN in the local network. When the MN moves from one foreign network to another, it registers its new location, i.e., its new CoA, to its HA. Packets for an MN are sent to its permanent address, i.e., its home address first. The HA intercepts all the IP packets destined to the MN and tunnels them to the serving FA of the MN. The FA decapsulates and forwards these packets to the MN.

A major problem of mobile terminals is their limited battery capacity. Mobile IP requires that an MN registers its new location to its HA whenever it enters a new subnet. Statistics indicate that the power of actively communicating MNs spent in location updating is an order of magnitude greater than the power spent in standby mode, where MNs perform location updates less frequently [59]. In order to save the battery power consumption at mobile terminals, IP paging is proposed as an extension for Mobile IP [17] [21].

As described in Chapter 3, under Mobile IP paging, an MN is allowed to enter a

power saving idle mode when it is inactive for a period of time. During the idle mode, the system knows the location of the MN with coarse accuracy defined by a paging area which is composed of several subnets [17]. The MN may also deactivate some of its components for energy-saving purpose. An MN in idle mode does not need to register its location when moving within a paging area. It performs location update only when it changes paging areas. When packets are destined to an MN in idle mode, they are terminated at a paging initiator. The paging initiator buffers the packets and locates the MN by sending out paging requests within the paging area. After knowing the exact location of the MN, i.e., the subnet where the MN is residing, the paging initiator forwards the data packets to the serving FA of the subnet and further to the MN. When an MN is in active transmission mode, it operates in the same manner as in Mobile IP and the system keeps the exact updated location information of the MN. The state transition diagram of MNs with paging support is shown in Figure 19.

Next generation Internet is expected to support multimedia communications. For real-time data traffic such as Internet telephony, video conferencing, audio library, and news-on-demand, Quality of Service (QoS) provision must be guaranteed so that the real-time traffic may get predictable service [60]. There has been a lot of research on the provision of QoS guarantees in the environment of wireless and mobile Internet [61] [62] [63] [64]. The Resource Reservation Protocol (RSVP) was developed by IETF to support the signaling of end-to-end IP QoS [65]. It allows a host on behalf of a real-time application to request a given QoS from the network. Mobile RSVP (MRSVP) was later proposed to resolve the impact of mobility on RSVP in mobile computing environments [66]. Under RSVP, signaling messages exchange along the path between the source and the destination to reserve the requested resource for the real-time traffic. After resource reservations are established in each router along the path, the application makes use of these reservations to send real-time traffic.

When there is a real-time communication request from a correspondent node (CN)

to an MN in idle mode, a connection between the CN and the MN along which the requested resources are reserved should be set up for the real-time data traffic. RSVP signaling messages are first sent from the CN to the HA of the MN. When both Mobile IP paging and RSVP are supported in the network, the HA operates in two phases sequentially before the real-time communication. During the first phase, the HA pages the MN to find its exact location. Then, the HA sets up a RSVP path and reserves the requested resources along the path between the HA and the corresponding FA of the MN. Therefore, the signaling delay before the data communication is the sum of the paging delay and the connection setup time of the RSVP path.

In this chapter, we introduce a new scheme for fast connection setup of real-time communication. Under the proposed scheme, the total signaling delay is reduced compared with the traditional scheme. We make the following assumptions in the rest of this chapter:

- The receivers of the real-time traffic are mobile users roaming across the network.
- Both Mobile IP paging and RSVP are supported in the network.

The focus of this chapter is the connection setup phase before the real-time communication. How to maintain the real-time communication during the handoff procedure when an MN moves from one subnet to another is beyond the scope of this chapter.

The proposed scheme was first described in [67]. This chapter is organized as follows. In Section 4.2, the related work on Mobile IP paging and RSVP is reviewed in detail. In Section 4.3, the proposed paging-aided connection setup scheme is presented. The protocols for unicast and multicast communications are described. After that, in Section 4.4, the performance of the proposed scheme is evaluated.

4.2 Background

In this section, we introduce the related work on Mobile IP paging. We also explain the details of RSVP.

4.2.1 Mobile IP Paging

Currently, there are three major paging protocols proposed for Mobile IP. In *home agent paging* [18], the HA acts as the paging initiator and buffers the data packets to MNs before paging. When an MN registers with its HA, it also sends a multicast address of all the FAs in its current paging area to the HA. This multicast address is used for HA to send paging requests. After receiving paging requests, all the FAs in a paging area broadcast paging messages to MNs in their subnets through wireless links. The paged MN sends a paging reply to the paging initiator through its serving FA. The HA updates the current location of the MN and forwards all the buffered packets to the MN. In *foreign agent paging* [19] [20], the paging initiator is the registered FA, which is the FA that an MN registers with when entering a new paging area. Note that the registered FA of an MN is not necessarily to be the current serving FA of the subnet the MN is residing. The registered FA buffers data packets destined to an MN in idle mode and sends paging requests to all other FAs in the paging area. *Domain paging* is a distributed paging architecture, where the paging initiator is dynamically selected from the routers along the path from the domain root router to the last serving FA of the MN [18]. The decision of the paging initiator depends on the paging load of each router.

4.2.2 RSVP Signaling

RSVP signaling messages are carried directly within IP packets following the same paths between the source and the destination as the associated application data packets. The two primary messages are PATH (path establishment) and RESV (reservation). In order to determine and record the path through which the application

data will traverse, the sender transmit periodic PATH messages to the receiver which contains: the IP address of the node sending the message; the QoS being requested; and a *flow* that defines which packets are to receive the specified QoS. The flow is specified as a set of protocol header fields that can be used by a node to distinguish the application data packets from all others. Routers along the path modify the PATH messages by swapping its own IP address with that in the sender field, and forwards the message to the next hop. In order to actually reserve resources along the path from the sender to the receiver, the receiver responds to PATH messages with RESV messages. Routers along the path correlate RESV messages with previously seen PATH messages, examining the QoS on a hop-by-hop basis to determine whether there are sufficient resources to fulfill the request. If so, a router reserves the necessary resources and sends the RESV message to the node from which it received a PATH message. Once resources have been reserved along the entire path, i.e., the connection between the source and the destination is set up, the sender begins to transmit real-time data packets.

Some additional features of RSVP are:

- An RSVP reservation is unidirectional. Bidirectional real-time flows require two reservations.

- Reservations are initiated by the receiver. This allows RSVP to accommodate multicast groups with large and changing group membership.

- RSVP does not perform its own routing. It uses information provided by underlying routing protocols to determine the paths along which to request the QoS.

- RSVP makes reservations at each router for a limited lifetime. Each data session's PATH and RESV messages must be retransmitted periodically to refresh

the state information held by routers along the path. This *soft-state signaling* ensures that reserved resources along any given path will be automatically released if routes change during the lifetime of a data session.

- RSVP may merge resource reservation requests from different branches of a multicast tree to a single reservation request. The outgoing reservation request must satisfy all the requirements of the incoming requests.

4.3 Paging-Aided Connection Setup for Real-time Communication

In this section, we propose a new scheme for fast connection setup of real-time communication in mobile Internet. Since the signaling messages of location registration and paging are not avoidable for MNs roaming across Internet, it is desirable to utilize these messages to achieve connection setup concurrently. Thus, the total signaling delay before the data communication is reduced. We explain the detailed operations in the following.

4.3.1 Unicast Case

First, we assume the communication is unicast in which one sender and one receiver are involved.

The proposed scheme employs home agent paging architecture [18]. When an MN in idle mode moves from one paging area to another, it performs idle mode location registration to its HA. During the location registration procedure, besides the home address and the newly obtained CoA of the MN, the FA of the subnet where the MN is currently visiting inserts a multicast address of all the FAs within the current paging area into the location registration message. This multicast address is used to identify the paging area. The extended location registration message is sent to the HA of the MN, as shown in Figure 31.

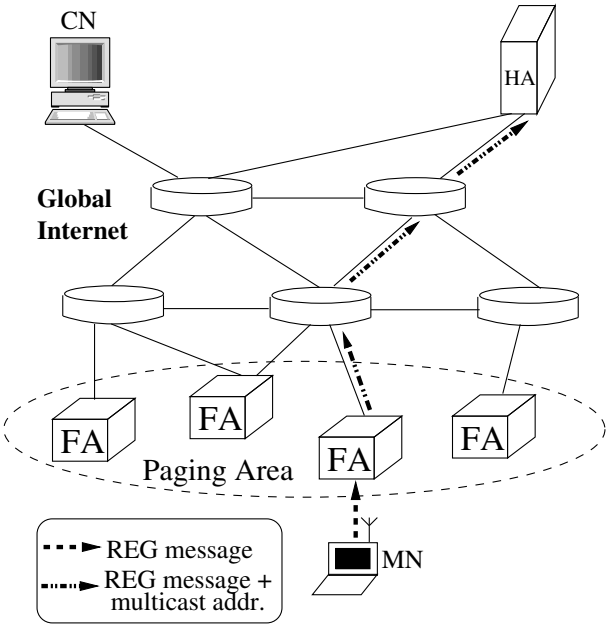

Figure 31: Procedure of location registration.

If a CN wants to have a real-time communication with an MN in idle mode, the CN sends out RSVP PATH messages to set up a connection. These RSVP signaling messages are sent to the HA of the MN first. The HA needs to find the exact location of the MN. It checks its record of the MN and sends out combined messages of paging request and RSVP PATH to the identity of the paging area, i.e., the multicast address. In other words, the combined messages are sent to all the FAs within the paging area, as illustrated in Figure 32. These messages have the information for paging: the multicast address of the paging area and the home address of the paged MN. They also contain the information to support RSVP PATH function. The combined paging and PATH messages follow the *shortest-path* IP route toward the destinations, installing path and QoS states in each router as they go.

After receiving the combined messages, the FAs extract the paging information from the signaling messages and page the MN in their subnets over the air. The paged MN sends paging reply message to its serving FA after receiving the paging request. The corresponding FA of the MN sends the combined message of paging

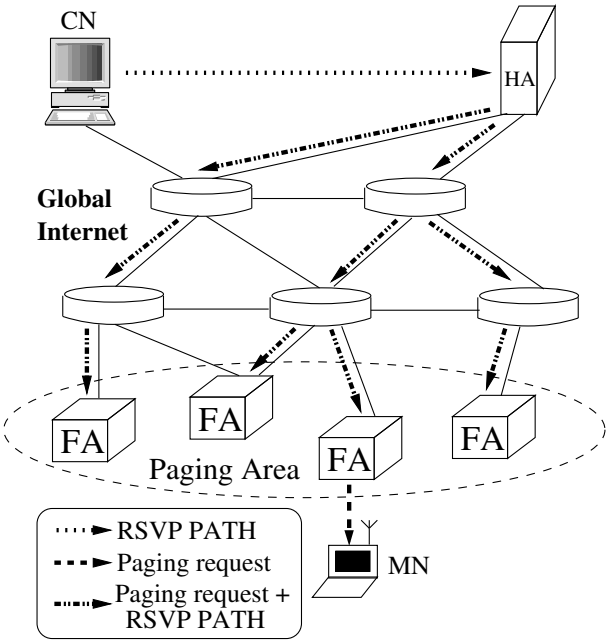

Figure 32: Procedure of sending paging request and RSVP PATH messages.

reply and RSVP RESV to the HA. This message retraces the steps of the matching PATH messages, establishing resource reservation in each router along the path, as shown in Figure 33. Note that other FAs do not reply to the paging requests. The soft states in the routers along the paths between the HA and other FAs will be expired after a certain period. Finally, if the combined message reaches the HA, resources are established along the entire path and the HA obtains the location information of the paged MN at the same time. The HA updates the record of the MN and sends the RSVP RESV message back to the CN to finish the real-time connection setup.

Note that by employing home agent paging architecture, the established connection from the HA to the serving FA of the MN follows the shortest IP routing path. If other paging architectures are employed, the shortest-path property cannot be guaranteed. For example, if foreign agent paging architecture [19] [20] is employed, i.e., the registered FA functions as the paging initiator to send out paging requests. Under the proposed scheme, the path along which the resources are reserved will be from the HA to the registered FA and further to the corresponding FA of the paged MN,

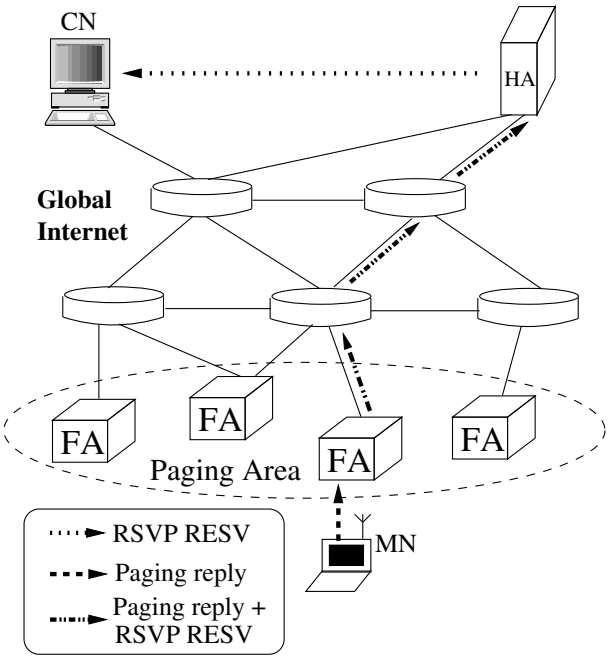

Figure 33: Procedure of sending paging reply and RSVP RESV messages.

which is not necessarily to be the shortest path.

4.3.2 Multicast Case

RSVP may scale to very large multicast groups because it uses receiver-oriented reservation requests that merge as they travel up the multicast tree. The proposed scheme is also applicable to multicast communication by using paging request aggregation [20] and RSVP reservation request aggregation [65]. The flow chart of the operation on multicast communication is shown in Figure 34. In the figure, we assume there are two real-time traffic receivers. If the receivers of the real-time traffic do not belong to the same home network, the operations are similar to the case of the unicast communication. Each connection between the CN and one of the receivers is setup separately. If the receivers belong to the same network but are not located in the same paging area, one aggregated RSVP PATH message is sent from the CN to the HA of all the receivers. The HA operates the same as for unicast communication to each receiver. If the receivers are located in the same paging area, the HA sends an aggregated

message with paging request and RSVP PATH function to the multicast address of the FAs within the paging area. This aggregated message lists the home addresses of all the paged MNs. The message aggregation helps to reduce the signaling overhead as the number of receivers increases.

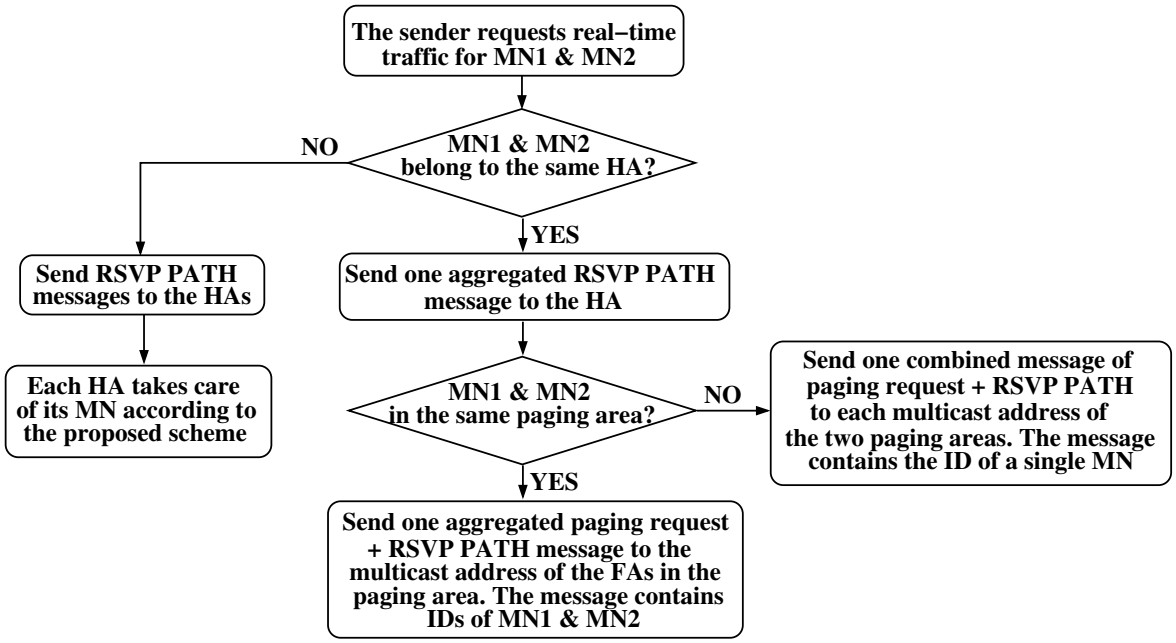

Figure 34: Flow chart for multicast case.

4.4 Performance Evaluation

In this section, we evaluate the proposed scheme in terms of connection setup time, signaling overhead, and processing overhead. We compare the proposed scheme with the traditional scheme: connection setup without the help of location registration and paging. Table 3 lists the performance analysis of the two schemes for unicast communication. We assume there are N FAs within a paging area. When comparing the connection setup time of the two schemes, we compare the time to set up a real-time connection between the HA and the serving FA of the receiver. When comparing the number of signaling messages, we do not take the periodic updates of the PATH and RESV messages to refresh their soft states in each router into account.

Table 3: Performance Analysis of Paging-Aided Connection Setup Scheme

	Connection setup without the help of paging	Paging-aided connection setup
Connection Setup Time (from HA to the serving FA)	Time for sending and processing (paging request + paging reply + PATH message + RESV message)	Time for sending and processing (combined message of paging request & PATH + combined message of paging reply & RESV)
Number of Signaling Messages	N paging requests (flooding) + 1 paging reply (unicast) + 1 PATH message (unicast) + 1 RESV message (unicast)	N combined messages of paging request & PATH (flooding) + 1 combined message of paging reply & RESV (unicast)
Processing Overhead	1 path is setup "softly" + 1 path is setup "hardly"	N paths are setup "softly" + 1 path is setup "hardly"

From the table, we may see that the proposed paging-aided connection setup scheme has advantages over the traditional scheme in terms of fast connection setup and less signaling messages. But it requires more processing resources in routers. Note that sending and processing RSVP messages usually require more time than paging messages, since each router along the communication path needs to modify the PATH messages, setup QoS states, and reserve the requested resources. The time of sending and processing combined paging and RSVP messages is approximately equal to that of sending and processing pure RSVP messages. Therefore, under the proposed scheme, the time for paging can be saved. The total signaling delay for connection setup is approximately equal to the time for RSVP path setup under the traditional scheme. The total number of signaling messages are also reduced under the proposed scheme. But note that there is additional overhead of each combined signaling message due to the increase of packet size. Under the proposed scheme, more path and QoS states are installed in the routers as the RSVP PATH messages travel. But these states are soft states and will be expired later. In addition, no resources are actually reserved if no further RESV messages are sent by the receiver.

These temporarily soft states also do not influence the resource reservation by other data sessions.

CHAPTER V

LOCATION MANAGEMENT IN NG HETEROGENEOUS WIRELESS OVERLAY NETWORKS

5.1 Problem and Solution

NG wireless system is aiming to develop a framework of ubiquitous and integrated networks for mobile users using a wide variety of wireless technologies to access the worldwide information infrastructure [8]. Currently, various wireless technologies and networks have been deployed and cover different needs and requirements [68]. Wireless LANs (WLANs) [1] are good for local area access to high-speed and low mobility data communications. Traditional and NG cellular networks may provide voice and data services for wide areas. Satellite networks have been used extensively in various military and commercial applications for worldwide coverage. NG wireless system is envisioned to be able to satisfy diverse communication requirements simultaneously via a common infrastructure [69].

To support global roaming, NG wireless system requires the integration and interoperation of heterogeneous mobility management techniques [6]. Mobility in a hierarchical structure or multilayered environment should be supported. A basic requirement of the integration of heterogeneous networks is *downward compatibility* [70], which means, mobile users subscribing to multiple networks will receive services from the integrated wireless system; at the same time, the original users will still receive services from their individual networks without being affected by the integration. This will be achieved by integrating the inherited mobility management schemes of

each individual network [71].

One possible mobility management architecture for the integration of heterogeneous networks is to build a global common home location register (HLR) which connects to all visitor location registers (VLRs) in each individual network. This HLR has a global knowledge of the whole system and stores user profiles of all users. However, this architecture has the following problems. First, since individual networks use different signaling formats, authentication procedures, and registration operations, it is difficult to merge heterogeneous HLRs of different networks into a single HLR [70]. Second, business interaction between different wireless service providers is another issue influencing the practicality of this architecture. Each individual network stores user profiles of the subscribed users in their own HLRs. It is not easy to ask heterogeneous networks to share user profiles in a single database. In addition, this architecture changes the inherited mobility management architectures of each network by asking each VLR to communicate with the global HLR, instead of the already existing HLR in each individual network. In other words, this architecture is not downward compatible.

To achieve downward compatibility, the new interworking entities handling inter-system roaming between heterogeneous access networks should not replace existing mobility management architectures in each network, even though some of the functions or signaling in the present networks may be affected [72]. Different interworking units (IWUs) are proposed to facilitate roaming between some specific pairs of practical networks, such as solutions for interworking IS-41 with GSM [73], integrating satellite and terrestrial environment [74], and integrating WLAN and Third-Generation (3G) wireless networks [75]. These IWUs form an additional level of the existing mobility management hierarchy of each network and provide functions such as format transformation, address translation, as well as assistance on signaling message transmission and connection setup. Some recent research efforts attempt to design

general location management mechanisms for the integration and interworking of any heterogeneous networks. A dynamic inter-system location management scheme for any pair of adjacent networks was presented in [31]. The boundary location register is proposed to facilitate roaming between different wireless networks. However, this scheme is designed for heterogeneous networks with partially overlapping coverage at the boundaries. It is not applicable for overlay networks where multiple networks are fully overlapped.

When the service areas of heterogeneous wireless networks are fully overlapped, a mobile terminal (MT) is reachable via multiple access networks to which it subscribes. Under this heterogeneous overlay environment, the problem that in which networks the user location information should be stored becomes critical. One intuitive solution is to let an MT update its location information in all wireless networks it has subscribed when it roams in multiple overlay networks. A Meta-HLR database was proposed for this solution in [76] to maintain the mapping between each MT and the HLR addresses of the networks the user subscribes to. Similar solutions have been proposed in the study of location management in multitier personal communications services (PCS) system. The multitier HLR (MHLR) approach was introduced in [77] [78]. A tier manager is built to be connected with all heterogeneous HLRs to indicate in which networks the location information of MTs is stored. Based on this MHLR approach, two location registration strategies were proposed in [70], namely *single registration* (SR) and *multiple registration* (MR). Under these two protocols, services are always delivered through the lowest available tier network to users. The performance modeling and comparison of these two location registration schemes were presented in [79] [80]. However, it is not clear that where the Meta-HLR or the MHLR should be built. In addition, the above schemes did not implement user preference call delivery, which is very important for NG wireless system supporting multimedia services.

Mobile users may subscribe to multiple networks and they may have their own preferences on which type of service delivered through which specific network. Many other parameters including network conditions, power consumption, and the reliability of the reachable network will also influence the decision on the "best" network to deliver services. It is not practical to always deliver multimedia traffics through the lowest tier networks to all mobile users.

In this chapter, we introduce a new architecture for location management in NG heterogeneous overlay networks. Under the proposed architecture, the location management procedure in each heterogeneous network does not change, i.e., the downward compatibility requirement is satisfied. User profiles are still kept in the individual HLRs and not shared by different networks. Three location management techniques with user preference call delivery implementation are presented under the proposed architecture for the integration of heterogeneous networks. Calls to MTs can be delivered through any network without restrictions.

The proposed mobility management architecture was described in [81]. The proposed location management schemes were introduced in [82]. This chapter is organized as follows. In Section 5.2, the proposed system architecture and problem formulation for location management in heterogeneous overlay networks are described. Then, in Section 5.3, three proposed location management techniques are explained and the details of the signaling protocols are introduced. In Section 5.4, the analysis of the signaling cost for the proposed schemes are presented. Numerical results are also provided in this section to demonstrate the performance of the new protocols. In Section 5.5, an enhancement method is described.

5.2 System Architecture and Problem Formulation

5.2.1 System Architecture

The heterogeneous overlay networks we consider in this chapter include many dissimilar networks using different radio technologies and different network management techniques. These heterogeneous networks have fully overlapping areas of coverage and significantly different cell sizes ranging from a few square meters to hundreds of square kilometers, as shown in Figure 1. For instance, WLANs may adopt IEEE 802.11a and 802.11b standards installed in public indoor locations. Conventional cellular networks using standards GSM, IS-95, IS-54/136 are deployed in different countries. Geostationary Earth Orbit (GEO), Medium Earth Orbit (MEO), and Low Earth Orbit (LEO) satellite systems are supposed to have a worldwide coverage. MTs with multiple physical or software-defined interfaces may seamlessly roam between different access networks. Note that, the service area of a specific network may not be continuous. Networks supported by different network operators may have similar cell size.

We propose a new mobility management architecture for the integration of heterogeneous mobility management techniques. Under the proposed architecture, each network keeps its own location management hierarchy (HLR/VLR) and registration procedure unchanged. User profiles are maintained in the home database of each network. As a result, there is no information sharing. We propose a Network Interworking Agent (NIA) which is connected to heterogeneous home databases in all networks. The heterogeneous home databases could be the home agent (HA) for WLANs, or HLR for cellular networks. An NIA is a database cache to maintain the location information of MTs who subscribe to multiple network services. Thus, the NIA is only responsible for supporting inter-system roaming to users who are able to communicate with multiple networks. For users who only subscribe to one network,

their roaming and location tracking will be taken care of by the location management technique of the subscribed network itself. Therefore, the downward compatibility requirement is satisfied under the proposed architecture. Note that the NIA only stores the necessary information for users who have global roaming requests. It does not maintain user profiles. The proposed system architecture is shown in Figure 35.

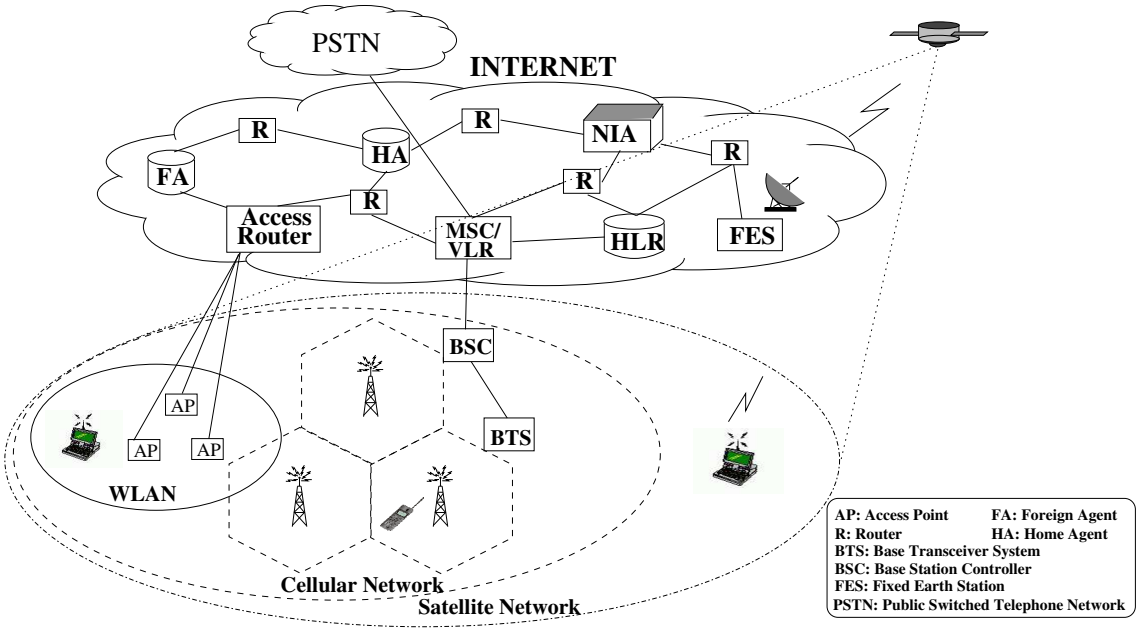

Figure 35: The proposed architecture for mobility management in heterogeneous overlay networks.

For NG wireless networks, the NIA is located in the backbone network — Internet. In other words, various heterogeneous signaling control entities for mobility management, such as HA/FAs for WLANs, HLR/VLRs for cellular networks, and fixed Earth stations (FESs) for satellite networks, are connected to each other through Internet, instead of a direct connection between themselves. The advantages of this architecture are:

- Operators of heterogeneous networks do not need to have direct service level agreements with each other. The operator of the NIA will take care of the issues related to global roaming, such as billing, signaling, authentication, and security. Thus, this approach is independent of individual system operators.

- Installation and operating cost of the NIA can be shared among all the networks. In addition, the NIA can help to enhance the system performance and reliability by implementing high quality security and authentication functions in it.

- This method is scalable and easy to integrate any number of networks belonging to different operators. If the number of heterogeneous networks increases or the number of mobile users with global roaming requests increases, the NIA can be built in a hierarchical structure to make it more scalable. The hierarchical structure may incorporate locality information of user movement pattern. For example, roaming with localized mobility can be taken care of by the NIAs at the lower level in the hierarchy, while roaming from one continent to another continent should involve the NIA at the highest level in the hierarchy.

5.2.2 Problem Formulation

Location management includes two major tasks: location registration and call delivery [39]. Location registration procedures periodically update the relevant location databases with the up-to-date location information of an MT. The call delivery procedures locate the MT based on the information available at system databases when a call for an MT is initiated. Two major steps are involved in call delivery: determining the serving VLR of the called MT, and locating the visiting cell of the called MT. The latter one is called paging. In heterogeneous overlay networks, if a user subscribes to multiple networks, new challenges for intelligent location management techniques include: through which networks an MT should perform location registrations; in which networks the up-to-date user location information should be stored; and after the system decides the "best" network to deliver a call to an MT, how to locate the serving VLR of the MT in the "best" network.

In the following section, we propose three new location management techniques which address the above challenges. We focus on the integration and interoperation

of heterogeneous mobility management schemes. We do not change the inherited mobility management schemes of each individual network, i.e., when an MT updates its location in a specific network, or when a call is delivered through a specific network, the location registration procedure or the call delivery procedure follows the legacy mechanism of the specific network.

5.3 Proposed Location Management Techniques

As described previously, one possible approach for location management is to let an MT update its location in all subscribed networks. However, performing location updates in all subscribed networks will generate significant signaling overheads as the number of MTs increases. It will also set high requirements to the design of physical hardware or software-defined interfaces at MTs. Moreover, the battery power consumption at MTs is high.

In this section, we propose three location management schemes for NG heterogeneous overlay networks, namely *Lowest Available Tier Registration* (LATR), *a-Posteriori Probability-based Registration* (PPR), and *Call History-based Adaptive Registration* (CHAR). Under all three schemes, each MT updates its location only in one network at any time. We call the network that maintains the up-to-date location information of an MT as the *registration network* in this chapter. The NIA has a record for each MT who has subscribed to multiple networks and has global roaming requests. It keeps the updated information of the current HLR that has the latest location information of an MT, i.e., the NIA knows to which network an MT is performing location registrations.

In the remainder of the paper, we assume an MT has the capability to monitor the signals from all subscribed networks in the standby mode so that it has the knowledge on the availability of any network at any time. We assume there is a network entity which can perform the "best" network selection for call delivery for each MT. This

function can be implemented at each terminal or inside the NIA. In [23], a similar policy module was proposed to select the "best" reachable network for handoff. Cost, network conditions, power consumption, connection setup time, and user preferences are considered as the parameters for the policy module. The design of the policy module is beyond the scope of this chapter.

If the policy module is inside the NIA, it requires network availability information to make decisions of the "best" network selection for each MT. Each MT knows the availability of any network it subscribes to at any time by monitoring the signals from all subscribed networks. When the MT moves out or moves into the service area of a new network, it sends the network availability information to the NIA. In this chapter, we assume the policy module is implemented inside the NIA.

Before describing the three proposed schemes, we first introduce the implementations of network availability transmission and user preference call delivery which are supported in all proposed schemes.

5.3.1 User Preference Call Delivery

User preference call delivery is an important feature of NG wireless communications. It indicates that mobile users have their own preferences on which type of service should be delivered through which subscribed network.

The NIA maintains the records on which network stores the up-to-date location information of each MT. During the call delivery procedure, the call is first delivered to the NIA through the calling HLR. The NIA consults with the policy module and obtains the "best" network to deliver the call to the called MT. Depending on which network the registration network of the called MT is, i.e., which network stores the up-to-date location information of the called MT, there are two possible scenarios to deliver the call through the "best" network.

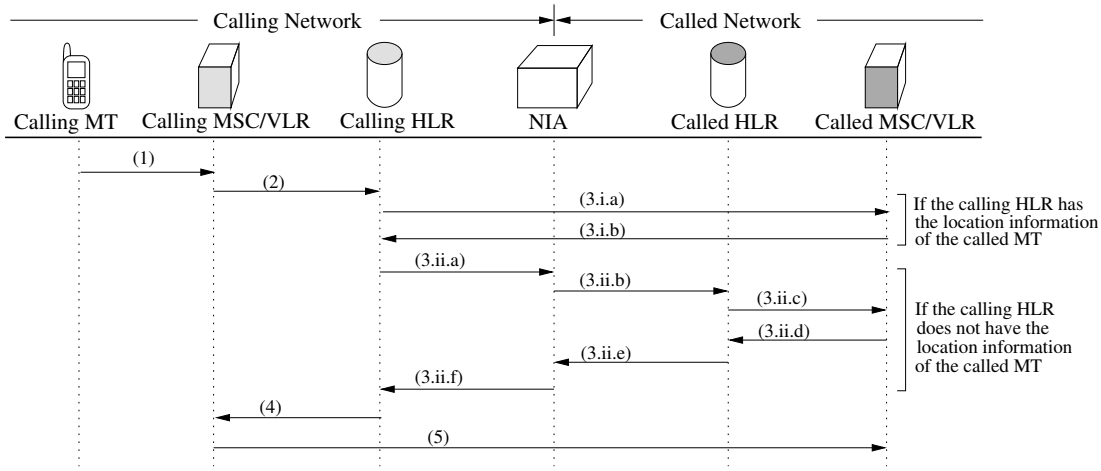

Figure 36: Call delivery procedure when the registration network is the call delivery network.

5.3.1.1 The registration network of the called MT is the "best" network

In this case, the NIA queries the HLR of the network which has the up-to-date location information of the called MT. This network is also the "best" network to deliver the call. The call delivery procedure as shown in Figure 36 is as follows.

(1) A call is initiated by an MT in its communication network and it is forwarded to its serving mobile switching center (MSC)/VLR.

(2) The MSC sends a location request message to the HLR asking for the routing information of the called MT.

(3) There are two possible scenarios:

  (i) If the calling HLR finds the serving MSC/VLR of the called MT on its record, then

    (a) The calling HLR sends a request message of routing information to the serving MSC. The procedure follows the call delivery protocol of the calling network.

    (b) The serving MSC of the called MT responds a routing number to the HLR.

(ii) If the HLR cannot find the location information of the called MT, it means the called MT is performing location registrations with another network.

(a) The calling HLR sends a location request message to the NIA.

(b) The NIA determines the HLR which has the up-to-date location information of the called MT and forwards the location request message to the called HLR.

(c) The called HLR asks the serving MSC of the called MT for the routing information.

(d) The serving MSC of the called MT responds a routing number to the HLR.

(e) The called HLR sends the routing information to the NIA.

(f) The NIA forwards this information to the calling HLR.

(4) The calling HLR sends the routing information to the calling MSC.

(5) The call connection is setup between the two MSCs in two networks.

5.3.1.2 The registration network of the called MT is not the "best" network

In this case, the system does not have any location information of the called MT in the "best" network. The system only knows where the called MT is in its registration network. We propose a *forced registration* operation which is similar to the method mentioned in [83] to let the system obtain the location information of the called MT in the "best" network to deliver the call. The goal of the forced registration is to restore the VLR record in the "best" network before call setup. First, the NIA tells the HLR of the registration network to initiate paging procedure in the registration area (RA) of the called MT. If paging is successful, the MSC controlling the RA sends a forced registration message to ask the called MT to initiate a location update in the "best" network. After the forced registration, the location information is restored in the

"best" network and the call connection can be setup between the calling network and the "best" network. Figure 37 shows the forced registration and call setup procedures for this scenario. The steps are described as follows.

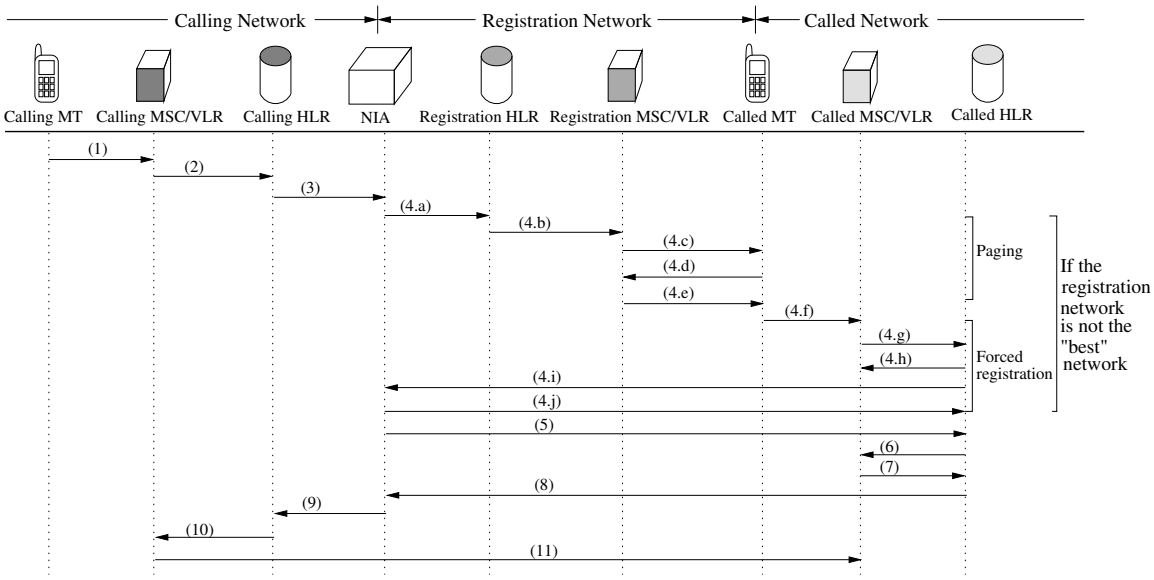

Figure 37: Call delivery procedure when the registration network is not the call delivery network.

(1) The calling MT sends a call initiation signal to the serving MSC in its communication network through a nearby base station.

(2) The MSC sends a location request message to the HLR asking for the routing information of the called MT.

(3) The calling HLR cannot find the location information of the called MT. It sends a location request message to the NIA.

(4) The NIA obtains the "best" network for delivering the call to the called MT. It compares the "best" network with the registration network of the called MT. If the "best" network and the registration network are different, then

   (a) The NIA sends a paging request message and the "best" network information to the HLR of the registration network of the called MT.

(b) The HLR in the registration network sends a paging request message to the MSC serving the called MT.

(c) The MSC in the registration network pages the called MT to determine the cell location.

(d) The called MT replies to the paging message through a nearby base station.

(e) The serving MSC sends a forced registration message to ask the called MT to initiate a location update in the "best" network.

(f) The called MT sends a location update message to the serving MSC in the "best" network through a nearby base station.

(g) The called MSC in the "best" network updates its associated VLR indicating that the MT is residing in its area and sends a location registration message to the called HLR.

(h) The HLR in the "best" network updates its record indicating the current serving MSC of the called MT and sends a registration acknowledgment message to the called MSC.

(i) The called HLR sends a network update request to the NIA.

(j) The NIA updates its record and sends an acknowledgment message to the called HLR.

(5) The NIA sends a location request message to the called HLR asking the routing information of the called MT.

(6) The called HLR asks the serving MSC of the called MT for the routing information.

(7) The called MSC responds a routing number to the called HLR.

(8) The called HLR returns this routing number of the called MT to the NIA.

(9) The NIA forwards the routing information to the calling HLR.

(10) The calling HLR further forwards the routing information to the calling MSC.

(11) The call connection is setup between the two MSCs in two networks.

5.3.2 The Proposed Location Management Schemes

5.3.2.1 Lowest Available Tier Registration (LATR)

Because of the low access cost and high capacity, updating locations in the lowest tier network is cheap and more bandwidth can be assigned for signaling traffic. Therefore, under the LATR scheme, an MT always updates locations in the lowest available tier network to save signaling cost. The lowest tier network has the smallest cell size. When the service area of the lowest tier network is not available, the MT performs *network switching*: switches location registration procedure to the new lowest available tier network, i.e., updates the HLR in the new network with its current location; meanwhile, the record of the MT in the NIA should be updated to indicate the new network which has the up-to-date location information of the MT. Whenever the MT moves out or moves into the lowest available tier network, one network switching is required to update the information in the NIA. Note that the location registration procedure of the LATR protocol is similar to the SR protocol proposed in [70]. But the SR protocol does not implement user preference call delivery. Under the SR protocol, users always receive services from the lowest available tier network, which may not be practical in NG multimedia wireless system. The signaling messages of network switching procedure as shown in Figure 38 are described as follows.

(1) The MT sends a location update message to the MSC controlling the RA where the MT is residing in the new network.

(2) The MSC updates its associated VLR indicating that the MT is residing in its area and sends a location registration message to the HLR of the new network.

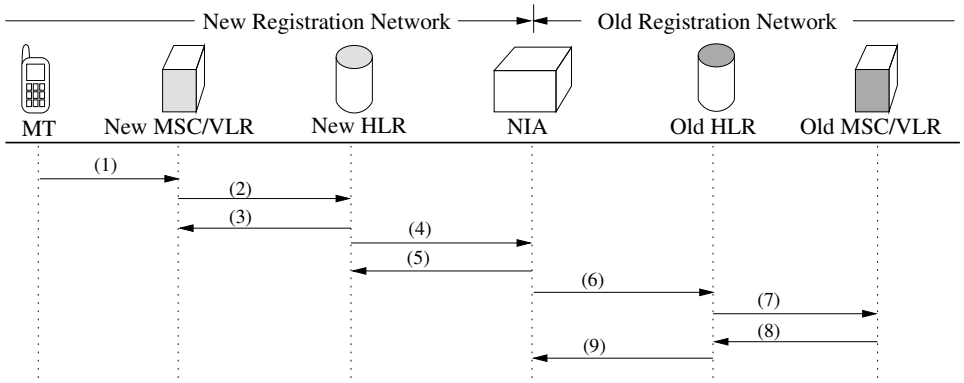

Figure 38: Network switching procedure.

(3) The HLR of the new network updates its record indicating the current serving MSC of the MT and sends a registration acknowledgment message to the MSC.

(4) The HLR of the new network sends a network update request to the NIA.

(5) The NIA updates the record of the MT indicating the new network which has the up-to-date location information of the MT and sends an update acknowledgment message to the new HLR.

(6) The NIA sends a registration cancellation message to the HLR in the old network.

(7) The old HLR deletes the location information of the MT and forwards the registration cancellation message to the MSC controlling the RA where the MT is residing in the old network.

(8) The old MSC deletes the record of the MT in its associated VLR and sends a cancellation acknowledgment message to the old HLR.

(9) The old HLR sends a cancellation acknowledgment message to the NIA.

The call delivery procedure with user preference call delivery implementation of the LATR protocol follows either the procedure described in Section 5.3.1.1 or in Section 5.3.1.2. If the "best" network for call delivery is the current lowest available

tier network, the procedure in Section 5.3.1.1 is followed. No paging in the registration network and forced registration in the "best" network are involved. If the "best" network for call delivery is not the current lowest available tier network, the procedure in Section 5.3.1.2 is followed.

The advantages of the LATR scheme are: since it is usually the cheapest to perform location registrations in the lowest tier network, the LATR scheme causes low registration cost. Besides, when a call is delivered from a network other than the registration network, paging in the lowest tier network also leads to the lowest cost, compared with paging in other available networks.

5.3.2.2 A-Posteriori Probability-based Registration (PPR)

There are several problems with the LATR scheme. First, this scheme classifies the communication networks into a layered structure according to the size of cells. If two heterogeneous networks have similar cell size, which network is chosen for location registration should be pre-defined. Second, since the lowest tier network has the smallest cell size, for high mobility MTs, the cell crossing rate is high, which leads to a high frequency of location registrations. In addition, if the coverage area of the lowest tier network is not continuous, the frequency of performing network switchings is also high. Third, if most calls are not delivered through the lowest tier network, excessive pagings in the lowest tier network will degrade the system performance.

One method to solve the above problems is to let the MT update its locations in the network through which most calls will be delivered in the near future. Thus, less pagings in the registration network are needed. Considering this, we introduce another scheme called a-Posteriori Probability-based Registration (PPR) scheme. This scheme is based on the assumption that the system has the knowledge on which network will deliver most of the calls to an MT in the future. Given this information, under the PPR scheme, an MT performs location registrations in the network through

which most of the future calls will be delivered.

As described previously, each mobile user may have the preference on which call should be delivered through which heterogeneous network. However, this preferred network is not necessarily to be the registration network for the PPR scheme. The policy module decides which network is the "best" network for call delivery. The decision is based not only on the user preference, but also on many other factors, such as network conditions, connection setup time, and power consumption. The probability that future calls for an MT will be delivered from a network is the a-posteriori probability, that is, the probability obtained after future call deliveries are performed. Therefore, to implement the PPR scheme, the NIA should have the a-posteriori knowledge on which network is the "best" network for future call deliveries for each MT. However, in real systems, this knowledge is impossible to obtain in advance.

Under the PPR scheme, except the criterion for choosing the registration network, other operations are similar as under the LATR scheme. When the current registration network is not reachable to an MT, the MT consults the NIA for another "best" network for location registration. Then, the MT performs a network switching and updates the record in the NIA. When a call is delivered from a network different from the current registration network, the system pages the MT in its registration network and the MT updates its location in the call delivery network.

Since user preference is one important factor influencing the decision of call delivery network, most of the future calls will be delivered through the user preferred network with high probability. In this case, the PPR scheme may save great system resources on call delivery. However, the tradeoff is the location registration cost may be high, if the registration network is not the lowest tier network.

5.3.2.3 Call History-based Adaptive Registration (CHAR)

To implement the PPR scheme, the system must know the future call arrival pattern in advance. The PPR scheme is based on the knowledge on future call delivery, which is not available in practice. Next, we propose the third scheme which resolves the practicality problem of the PPR scheme, but still keeps its advantages. We propose the Call History-based Adaptive Registration (CHAR) scheme, which incorporates communication histories and call preferences into the design. Under the CHAR scheme, when an MT roams between multiple networks, it still performs location registrations in one network. However, the registration network is not fixed all the time. It is dynamically changed according to the communication history of the MT.

An MT chooses the network with which it has the latest communication to perform location registrations: either the network which has the latest call delivered to the MT, or the network through which the MT initiates the latest communication. Since user preference call delivery is supported, if an MT performs location registration in the network of the latest call delivery, the probability that the next call will be delivered through the same network is high. Thus, the signaling cost for call delivery can be reduced because the location information is maintained in the network responsible for call delivery and no forced registration is involved. On the other hand, after an MT has communications through a specific network, the network has the updated location information of the MT. Then, the MT continues to perform location registrations in this network without changing to another network to register. Thus, the signaling cost for location registration can be reduced.

The procedures of registration network selection is shown in Figure 39. When an MT is first turned on, it consults with the policy module to get the "best" network for itself and performs location registrations in this "best" network. The HLR of the "best" network maintains the up-to-date location information of the MT. The NIA keeps a record for each MT on the current HLR that stores the updated location

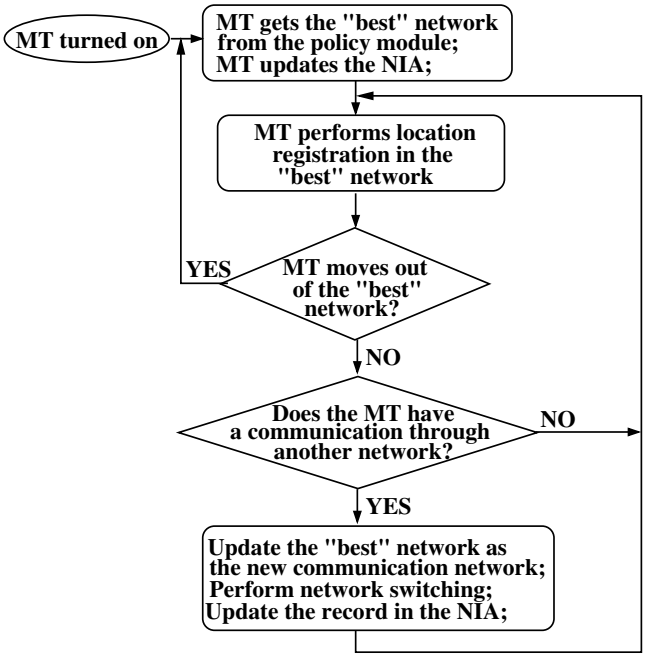

Figure 39: Procedures of registration network selection for the CHAR scheme.

information. As long as the MT roams in the service area of this "best" network, it updates its location in this network until the MT has communications through another network. When the MT moves out of the service area of the current "best" network, it asks the policy module for a new "best" network and updates the record in the NIA. When there is a call delivered to the MT through another network, or when the MT initiates a communication through another network, it means the current "best" network changes to the new communication network. The MT switches the registration network to the new network. Meanwhile, a network switching as shown in Figure 38 is performed to update the record in the NIA. As a result, the MT dynamically changes its registration network based on its communication history.

The call delivery procedure with user preference call delivery implementation of the CHAR protocol also follows the procedures described in Section 5.3.1.1 or Section 5.3.1.2. Note that the probability that the next call delivery network is the same as the current registration network is very high, if user preference call delivery is supported.

The CHAR scheme is a feasible solution in real systems. It does not need to predefine the lowest tier network. It also does not require any a-posteriori knowledge on user mobility and call patterns. When most calls are consecutively delivered from one network, the CHAR scheme is able to keep the advantages of the PPR scheme, while still resolves the problems caused by the LATR scheme.

5.4 Performance Analysis

In this section, we compare the performance of the three proposed location management schemes, i.e., LATR scheme, PPR scheme, and CHAR scheme. The total signaling cost of location registration and call delivery is considered as the performance metric. In order to determine the "best" network for call delivery, each MT transmits the network availability information to the policy module when a network is available or unavailable to it under all three schemes. Therefore, the signaling overhead caused by network availability transmission is not considered in the total signaling cost for performance comparison.

To simplify the analysis, we assume within a certain observation period T, an MT is in the service area of three heterogeneous overlay networks: network 1, network 2, and network 3. Extensions of analysis to more networks can be conducted in a similar way. The cell size of network 1 is the smallest, while network 3 has the largest cell size. These three networks are overlaid to each other.

We consider the MT moves under a certain mobility pattern. The length of the observation period T is chosen that the MT is always inside the lowest available tier network, network 1, during T, and the MT is reachable through three overlay networks. We conduct our analysis specifically for this selected period. Note that when the MT moves out of the current lowest tier network, another observation period starts. The assumptions we made above will not influence the generality of the conclusions we get from the analysis below.

We define the following parameters for our analysis:

c_i *Average* cost of performing location registrations only in network i during the observation period T.

ϕ_i Cost of *each* call delivery through network i without paging and forced registration, i.e., the total cost of steps (1)-(5) in Figure 36.

α_i Cost of *each* paging through registration network i, i.e., the total cost of steps (4.a)-(4.e) in Figure 37.

β_i Cost of *each* forced registration through call delivery network i, i.e., the total cost of steps (4.f)-(4.j) in Figure 37.

where $i = 1, 2, 3$.

Assume within the observation period T, there are totally N calls delivered to an MT. Define n_{ij} as the number of calls delivered through network j to the MT, when the MT is currently registering with network i, where $i, j = 1, 2, 3$. Assume the probability that a call is delivered from a specific network other than the lowest tier network is p. Assume this specific network is the network with the highest a-posteriori probability for call delivery under the PPR scheme. For the analysis in this chapter, we assume this network is network 2. Hence, the probability that a call is delivered from the other two networks is $1 - p$. Assume a call is from the other two networks with equal probability, i.e., a call is from network 1 and network 3 with probability $\frac{1-p}{2}$, respectively.

5.4.1 Total Signaling Cost

We first calculate the total signaling cost of the three proposed schemes.

5.4.1.1 LATR Scheme

For the LATR scheme, the MT always performs location registrations in the lowest tier network, i.e., network 1, during the period T. The total signaling cost of location

registration and call delivery of the LATR scheme is:

$$C_{LATR} = c_1 + \phi_1 n_{11} + (\phi_2 + \alpha_1 + \beta_2) n_{12} + (\phi_3 + \alpha_1 + \beta_3) n_{13} \tag{66}$$

where $n_{11} + n_{12} + n_{13} = N$, $\frac{n_{12}}{N} = p$, and $\frac{n_{11}}{N} = \frac{n_{13}}{N} = \frac{1-p}{2}$. Therefore, the total signaling cost of the LATR scheme during the period T can be expressed as in (67).

$$\begin{aligned} C_{LATR} &= c_1 + \left[\phi_1 \cdot \frac{1-p}{2} + (\phi_2 + \alpha_1 + \beta_2)p + (\phi_3 + \alpha_1 + \beta_3)\frac{1-p}{2}\right] \cdot N \\ &= c_1 + \left(\frac{\phi_1 + \phi_3 + \alpha_1 + \beta_3}{2}\right) N + \left(-\frac{\phi_1}{2} + \phi_2 - \frac{\phi_3}{2} + \frac{\alpha_1}{2} + \beta_2 - \frac{\beta_3}{2}\right) Np \end{aligned} \tag{67}$$

5.4.1.2 PPR Scheme

Similar to the analysis for the LATR scheme, the total signaling cost of location registration and call delivery of the PPR scheme is:

$$C_{PNR} = c_2 + (\phi_1 + \alpha_2 + \beta_1)\tilde{n}_{21} + \phi_2 \tilde{n}_{22} + (\phi_3 + \alpha_2 + \beta_3)\tilde{n}_{23} \tag{68}$$

where $\tilde{n}_{21} + \tilde{n}_{22} + \tilde{n}_{23} = N$, $\frac{\tilde{n}_{22}}{N} = p$, and $\frac{\tilde{n}_{21}}{N} = \frac{\tilde{n}_{23}}{N} = \frac{1-p}{2}$. Therefore, the total signaling cost of the PPR scheme during the period T can be expressed as in (69).

$$\begin{aligned} C_{PNR} &= c_2 + \left[(\phi_1 + \alpha_2 + \beta_1) \cdot \frac{1-p}{2} + \phi_2 p + (\phi_3 + \alpha_2 + \beta_3)\frac{1-p}{2}\right] \cdot N \\ &= c_2 + \left(\frac{\phi_1 + \phi_3 + \beta_1 + \beta_3}{2} + \alpha_2\right) N + \left(-\frac{\phi_1}{2} + \phi_2 - \frac{\phi_3}{2} - \alpha_2 - \frac{\beta_1}{2} - \frac{\beta_3}{2}\right) Np \end{aligned} \tag{69}$$

5.4.1.3 CHAR Scheme

For CHAR scheme, the MT performs location registrations in the network which has the latest communication with the MT. After a call is delivered through a network other than the current registration network, the MT changes its registration network to the new communication network. Therefore, the value of the average location

registration cost during the period T, $\bar{c}$, is between the maximum and the minimum values of c_i. The total signaling cost of location registration and call delivery of the CHAR scheme is:

$$\begin{aligned} C_{CHAR} &= \bar{c} + \phi_1\hat{n}_{11} + (\phi_2 + \alpha_1 + \beta_2)\hat{n}_{12} + (\phi_3 + \alpha_1 + \beta_3)\hat{n}_{13} + (\phi_1 + \alpha_2 + \beta_1)\hat{n}_{21} \\ &\quad + \phi_2\hat{n}_{22} + (\phi_3 + \alpha_2 + \beta_3)\hat{n}_{23} + (\phi_1 + \alpha_3 + \beta_1)\hat{n}_{31} + (\phi_2 + \alpha_3 + \beta_2)\hat{n}_{32} + \phi_3\hat{n}_{33} \\ &= \bar{c} + \sum_{i=1}^{3}\phi_i\hat{n}_{ii} + \sum_{i=1}^{3}\sum_{\substack{j=1 \\ j\neq i}}^{3}(\phi_j + \alpha_i + \beta_j)\hat{n}_{ij} \end{aligned} \tag{70}$$

where $\sum_{i=1}^{3}\sum_{j=1}^{3}\hat{n}_{ij} = N$, and

$$\frac{\hat{n}_{ij}}{N} = \begin{cases} \left(\frac{1-p}{2}\right)^2 & i,j = 1,3 \\ p\left(\frac{1-p}{2}\right) & i = 1,3 \text{ and } j = 2 \\ p\left(\frac{1-p}{2}\right) & i = 2 \text{ and } j = 1,3 \\ p^2 & i = 2 \text{ and } j = 2 \end{cases}$$

Therefore, the total signaling cost of the CHAR scheme during the period T can be expressed as in (71).

$$\begin{aligned} C_{CHAR} &= \bar{c} + \left[\phi_2 p^2 + \left(\sum_{\substack{j=1 \\ j\neq 2}}^{3}(\phi_j + \alpha_2 + \beta_j) + \sum_{\substack{i=1 \\ i\neq 2}}^{3}(\phi_2 + \alpha_i + \beta_2)\right) p\left(\frac{1-p}{2}\right)\right. \\ &\quad \left. + (\phi_1 + (\phi_3 + \alpha_1 + \beta_3) + (\phi_1 + \alpha_3 + \beta_1) + \phi_3)\left(\frac{1-p}{2}\right)^2\right] N \\ &= \bar{c} + \left(-\frac{\alpha_1}{4} - \alpha_2 - \frac{\alpha_3}{4} - \frac{\beta_1}{4} - \beta_2 - \frac{\beta_3}{4}\right)Np^2 + \left(-\frac{\phi_1}{2} + \phi_2 - \frac{\phi_3}{2} + \alpha_2 + \beta_2\right)Np \\ &\quad + \left(\frac{\phi_1}{2} + \frac{\phi_3}{2} + \frac{\alpha_1}{4} + \frac{\alpha_3}{4} + \frac{\beta_1}{4} + \frac{\beta_3}{4}\right)N \end{aligned} \tag{71}$$

5.4.2 Numerical Results

Now, we conduct some quantitative analysis by comparing the total signaling costs of the three proposed schemes based on (67), (69), and (71). We assume that the costs

of each location registration, call delivery, paging, and forced registration are available. These costs account for the wireless and wireline bandwidth utilization and the computational requirements in order to process signaling messages [46]. The methods for determining the cost parameters are discussed in [43] [44]. We set different values for signaling cost parameters to simulate scenarios of different user mobilities and call arrivals. We study the impacts of varying parameters on the performance of the proposed location management schemes.

For our numerical evaluation, we assume the total number of call arrivals during T is 10, i.e., $N = 10$. According to the signaling message flows shown in Figure 36 and Figure 37, the cost of each location registration in a specific network is set to be equal to half of the cost of each call delivery, i.e., $\beta_i = \frac{1}{2}\phi_i$.

5.4.2.1 The Impact of c_i

We first investigate the impact of the average location registration cost c_i in each network during the observation period T. Other parameters are normalized to the values of network 1 such that $\alpha_1 = \phi_1 = 1$. We set $\alpha_i = \phi_i = 1, 2, 3$ for $i = 1, 2, 3$.

The value of the average location registration cost during the period T under the CHAR scheme, $\bar{c}$, is a function of the call arrival pattern. More specifically, the factors which influence the value of $\bar{c}$ are: the average location registration cost, c_i, in each heterogeneous network during T; from which network each call is delivered; the inter-arrival time of each call; and the sequence of call arrivals. Within the period T, $\bar{c}$ is bounded by the minimum and the maximum registration cost value of all heterogeneous networks. The actual value of $\bar{c}$ depends on the call delivery network and the inter-arrival time of each call. If most calls are delivered from the lowest tier network, or the inter-arrival time between a call from the lowest tier network and the next call is long, the MT will update its locations in the lowest tier network for relatively longer time. Hence, $\bar{c}$ will be close to the value of the cost if updating

locations only in the lowest tier network within a specific period. On the other hand, if calls are from all the available networks with equal probability and the inter-arrival time of all calls is the same, $\bar{c}$ will be close to the average value of c_i. In addition, the average location registration cost of each network, c_i, is also a factor influencing the "best" delivery network selection, which in turn changes the probability that a call is delivered from a specific network. Due to the interaction of all the factors involved, it is difficult to give quantitative analysis of $\bar{c}$. To simplify the analysis, we let $\bar{c}$ only depend on the probability that a call is from the preferred delivery network of a user, i.e., $\bar{c} = (c_1 + c_3)\left(\frac{1-p}{2}\right) + c_2 p$.

Table 4: Selected Data Sets For c_i

Data Set	1	2	3
c_1	1	30	40
c_2	2	8	80
c_3	3	10	120

We consider three sets of values for c_i given in Table 4. Data set 1 represents the MT with low mobility. The MT performs few location registrations during the observation period T. Since normally it is the cheapest to access the lowest tier network, c_1 has the lowest value, provided the frequency of location registrations in each network is the same. Data set 2 and 3 represent the MT with high mobilities. However, in the case of data set 2, the roaming range of the MT is small. We may imagine under this scenario, the movements of the MT cause a lot of cell crossings in the lowest tier network. But the movements are always within a cell coverage of network 2 and network 3. Consequently, if the MT performs location registrations in network 1, the registration cost is very high. In the case of data set 3, the movements of the MT cause a lot of cell crossings in network 1, and also in network 2 and network 3. Therefore, the average location registration costs in all networks are very high.

Figure 40 shows the total signaling cost of the LATR scheme and the PPR scheme

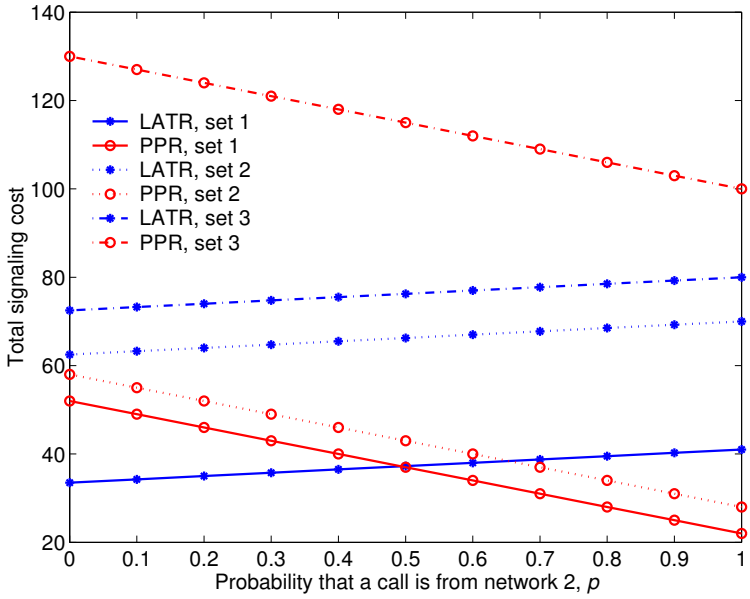

Figure 40: Comparison of the total signaling cost for the LATR scheme and the PPR scheme under different sets of c_i.

under different data sets of c_i. It is observed that under the selected data sets, the total signaling cost of the LATR scheme increases as p increases, while the total signaling cost of the PPR scheme decreases when p increases. When p is large, more calls are delivered from network 2. Under the LATR scheme, the system pages the MT in the lowest tier network frequently to ask the MT to update its location in network 2, which generates high signaling traffic. On the other hand, under the PPR scheme, network 2 always has the up-to-date location information of the MT. Less pagings and forced registrations are performed. From the figure we may notice that the total signaling cost of the LATR scheme can be either higher than or lower than that of the PPR scheme. When the average location registration cost in network 1 during T is very high as represented by data set 2, the registration cost dominates and always registering with the lowest tier network is very expensive. In this case, the PPR scheme outperforms the LATR scheme. When the location registration costs in all networks are very large as shown by data set 3, the PPR scheme does not have advantages over the LATR scheme. On the contrary, when the location registration

costs in all networks are comparable to each other, the two curves of the LATR scheme and the PPR scheme have a crossing point at $p = 0.5$. The larger the p is, the more cost the PPR scheme saves. This result is consistent with our design, since the implementation of the PPR scheme is based on the assumption that the selected registration network is the most likely network for future call delivery.

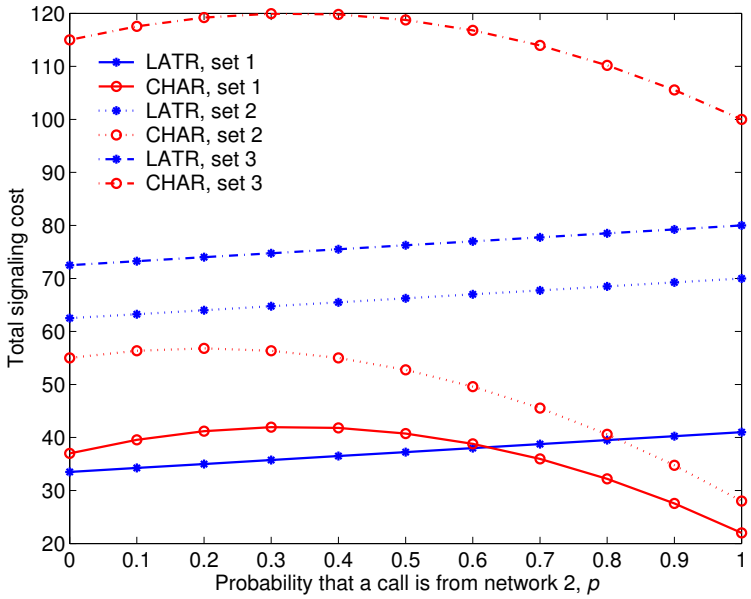

Figure 41: Comparison of the total signaling cost for the LATR scheme and the CHAR scheme under different sets of c_i.

Figure 41 plots the total signaling cost of the LATR scheme and the CHAR scheme. From the figure we see that the relationship between the LATR scheme and the CHAR scheme under the selected data sets is similar to that between the LATR scheme and the PPR scheme. For data set 1, the total signaling costs of the two schemes are equal at $p = 0.62$. For data set 2, the total signaling cost of the CHAR scheme is always lower than that of the LATR scheme, while it is the opposite for data set 3. This means the CHAR scheme keeps the main features of the PPR scheme. Unlike the PPR scheme, the total signaling cost of the CHAR scheme does not linearly decrease when p increases.

Figure 42 gives the relationship of the CHAR scheme and the PPR scheme. For

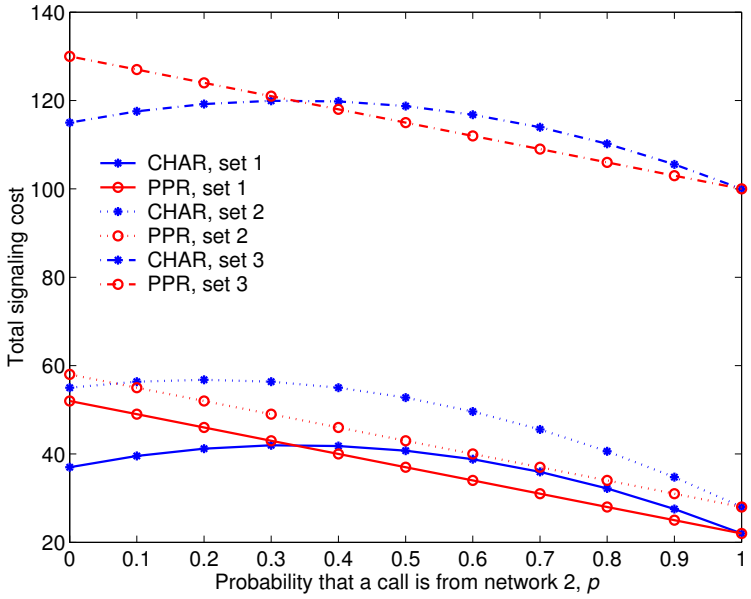

Figure 42: Comparison of the total signaling cost for the CHAR scheme and the PPR scheme under different sets of c_i.

small p, the CHAR scheme causes less signaling cost compared with the PPR scheme, while for large p, the PPR scheme performs slightly better than the CHAR scheme. But the overall cost gap between the two schemes is small. Therefore, the CHAR scheme can be considered as an approximation of the PPR scheme in practical systems. When $p = 1$, both schemes result in the same signaling costs. This observation verifies our analysis. When all calls are delivered from network 2, the MT will always update its locations in network 2 under both the PPR and the CHAR schemes.

5.4.2.2 The Impact of α_i

Now, we study the impact of the paging cost α_i in each network. Parameters ϕ_i and β_i are fixed at $\beta_i = \frac{1}{2}\phi_i = 1, 2, 3$, for $i = 1, 2, 3$. c_i are set to be 10 for $i = 1, 2, 3$. Four sets of values are considered for α_i as shown in Table 5. For data sets 1, 2, and 3, paging costs in the three networks have the same relative ratio, that is, $\alpha_2/\alpha_1 = 2$ and $\alpha_3/\alpha_1 = 3$. The difference between these three sets is their relative values to the registration costs, β_i. The values of data set 1 are much smaller than their corresponding registration costs, while the values of data set 3 are much larger.

Paging costs in data set 4 have a larger relative ratio: $\alpha_2/\alpha_1 = 5$ and $\alpha_3/\alpha_1 = 10$. It implies that paging in the lowest tier network is relatively cheaper.

Table 5: Selected Data Sets For α_i

Data Set	1	2	3	4
α_1	0.1	1	10	2
α_2	0.2	2	20	10
α_3	0.3	3	30	20

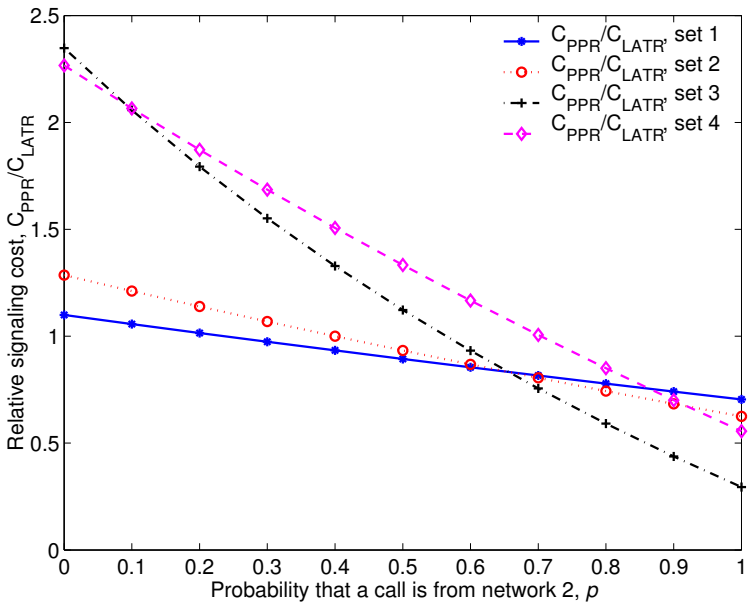

Figure 43: Comparison of the total signaling cost for the LATR scheme and the PPR scheme under different sets of α_i.

Figure 43 plots the relative cost ratio of the total signaling cost for the PPR scheme to that of the LATR scheme, C_{PPR}/C_{LATR}, under different data sets for α_i. A relative cost of 1 means that the costs under both schemes are the same. As shown in the figure, for all four data sets, when the probability that a call is from network 2, p, is small, the LATR scheme outperforms the PPR scheme in terms of reducing the signaling cost. When less calls are delivered from network 2, the probability that the call delivery network is different from the registration network under the PPR scheme is very high. As a result, the system needs to perform more

pagings in network 2 and forced registrations in the delivery network. Since paging in network 1 is the cheapest, for small p, the LATR scheme is more advantageous, especially when the relative paging cost ratio of other networks to the lowest tier network is large, as shown by data set 4. On the other hand, when p is large, the PPR scheme saves signaling cost significantly. Up to 70% cost can be saved by the PPR scheme, compared with the LATR scheme. This is because that less pagings in the registration network are needed when the call delivery network is the same as the registration network with high probability. Note that when the paging costs increase from set 1 to set 3, the crossing points where the two schemes have the same cost shift from $p = 0.25$ to $p = 0.56$ in the figure. It indicates that when the paging costs are large, p should be large for the PPR scheme to perform well. We also notice that the cost gap between the two schemes increases as α_i increases. This implies that when less calls are delivered from network 2, the larger the paging cost in each network is, the better the LATR scheme performs. On the contrary, when p is large, the PPR scheme is more cost-efficient when the paging cost in each network is large.

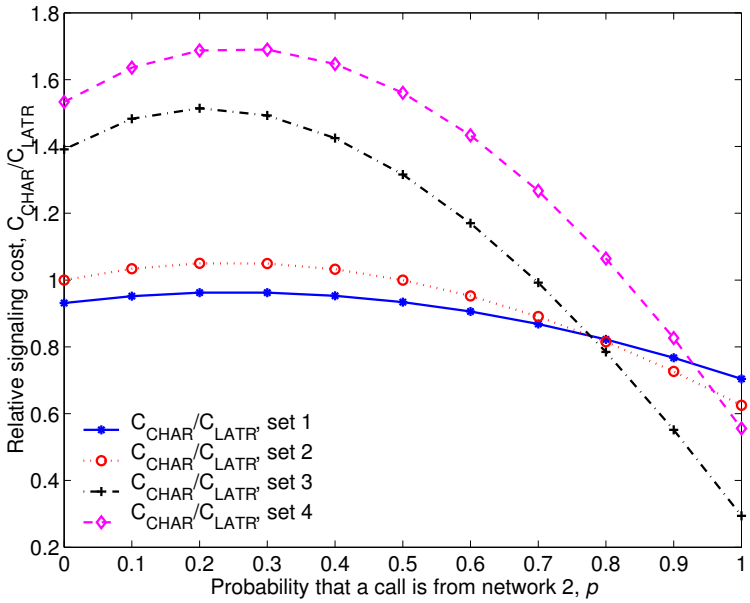

Figure 44: Comparison of the total signaling cost for the LATR scheme and the CHAR scheme under different sets of α_i.

Figure 44 shows the relative cost ratio between the CHAR scheme and the LATR scheme, C_{CHAR}/C_{LATR}. The results shown in the figure are similar to those shown in Figure 43, i.e., for small p, the LATR scheme performs better than the CHAR scheme, especially for data set 4, while for large p, the CHAR scheme is more cost-efficient. Comparing Figure 43 and Figure 44, we may notice that when the paging costs in all networks are very small as represented by data set 1, the CHAR scheme may improve the system performance more, compared with the PPR scheme. In other words, the CHAR scheme always results in lower signaling cost than the LATR scheme for data set 1, as shown in Figure 44.

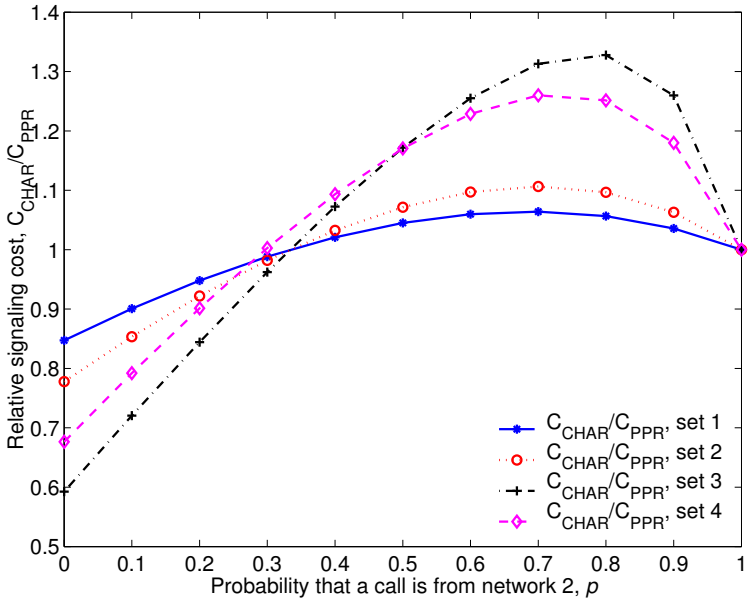

Figure 45: Comparison of the total signaling cost for the CHAR scheme and the PPR scheme under different sets of α_i.

Figure 45 compares the total signaling cost for the CHAR scheme and the PPR scheme. Similar to the results shown in Figure 42, when p is small, the CHAR scheme is more favorable. When p is large, the PPR scheme may save more cost. The two schemes have the same cost value when $p = 1$. Note that when the paging cost in each network increases from data set 1 to data set 3, the cost gap between the two schemes increases. Therefore, the performance of the CHAR scheme approximates that of the

PPR scheme well when the paging costs of all networks are relatively small.

5.4.2.3 The Impact of ϕ_i

Table 6: Selected Data Sets For ϕ_i

Data Set	1	2	3	4
ϕ_1	0.1	1	10	2
ϕ_2	0.2	2	20	20
ϕ_3	0.3	3	30	100

Finally, we study the impact of the call delivery cost in each network, ϕ_i. Paging costs α_i are fixed at $\alpha_i = 1, 2, 3$ for $i = 1, 2, 3$. c_i are still set to be 10 for $i = 1, 2, 3$. Four sets of values are considered for ϕ_i given in Table 6. Note that the cost of each forced registration in a specific network, β_i, is set to be half of each call delivery cost, ϕ_i. The selected data sets are similar to those in Section 5.4.2.2. For data sets 1, 2, and 3, the cost ratios are the same, i.e., $\phi_2/\phi_1 = 2$ and $\phi_3/\phi_1 = 3$. But their relative values to the paging costs are different. For data set 4, the cost ratios are larger: $\phi_2/\phi_1 = 10$ and $\phi_3/\phi_1 = 50$.

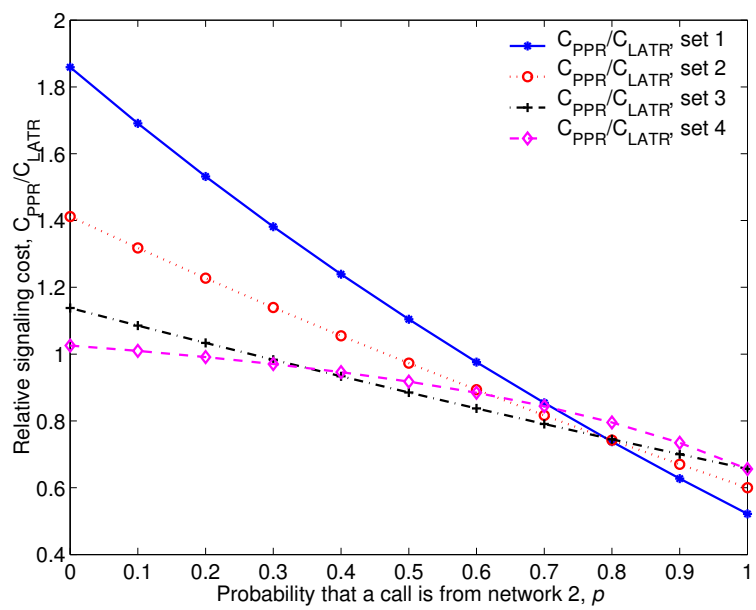

Figure 46: Comparison of the total signaling cost for the LATR scheme and the PPR scheme under different sets of ϕ_i.

Figure 46 shows the relative cost ratio, C_{PPR}/C_{LATR}, under different data sets for ϕ_i. Similar to the conclusions obtained from Figure 43, when p is small, the LATR scheme outperforms the PPR scheme. When p is large, the PPR scheme is more cost-efficient. Up to 50% cost can be reduced by the PPR scheme, compared with the LATR scheme. The points where the relative cost ratio is equal to 1 shift from $p = 0.58$ to $p = 0.25$, when the values of ϕ_i increases from set 1 to set 3. It suggests that a smaller p will lead to a better performance for the PPR scheme, when call delivery cost ϕ_i increases. Figure 46 differs with Figure 43 in that the cost gap between the two schemes decreases as ϕ_i increases. When p is less than the value at the crossing point where the two schemes lead to the same total cost, the smaller the call delivery cost is, the better the LATR scheme performs. When p is larger than the value at the crossing point, the PPR scheme can reduce more cost when ϕ_i is small. Moreover, when the call delivery cost ratios of other networks to network 1 is large, the performance difference between the two schemes are small compared with the case of small cost ratios.

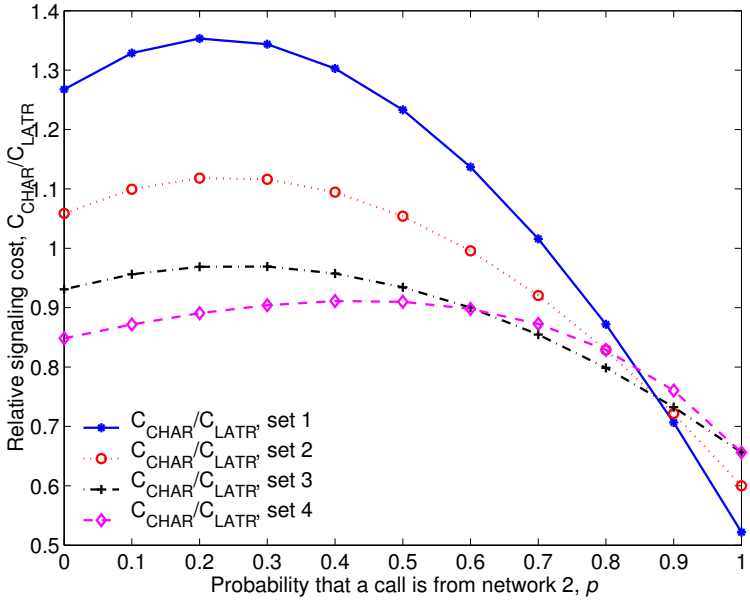

Figure 47: Comparison of the total signaling cost for the LATR scheme and the CHAR scheme under different sets of ϕ_i.

Figure 47 plots the relative cost ratios of the total signaling cost for the CHAR scheme to that of the LATR scheme, C_{CHAR}/C_{LATR}. Similar conclusions as shown in Figure 44 can be obtained. The CHAR scheme maintains the main features of the PPR scheme. Its performance relationship with the LATR scheme is similar to the performance relationship between the PPR scheme and the LATR scheme. For data set 3 and 4, the CHAR scheme performs better than the PPR scheme, since the cost ratios C_{CHAR}/C_{LATR} are always less than 1.

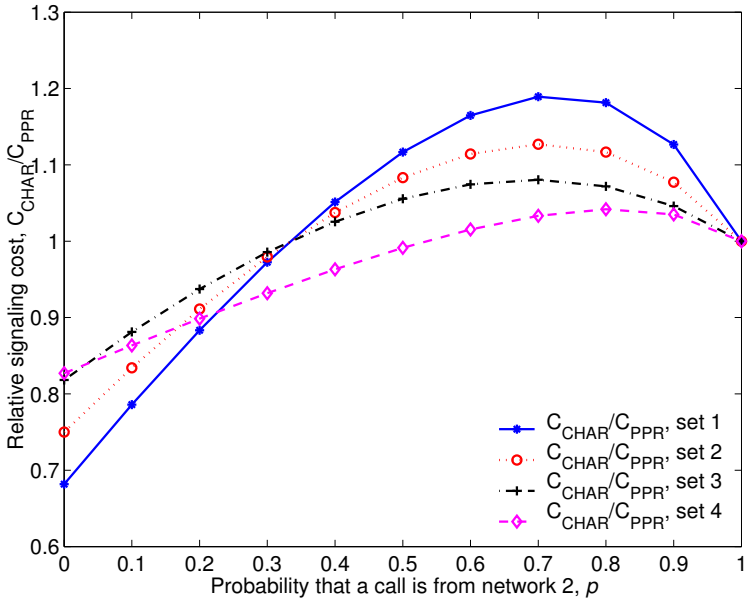

Figure 48: Comparison of the total signaling cost for the CHAR scheme and the PPR scheme under different sets of ϕ_i.

Figure 48 gives the cost comparison between the CHAR scheme and the PPR scheme. As shown in the figure, when ϕ_i increases, the cost difference between the two schemes decreases. What is more, when the relative call delivery cost ratio of other networks to the lowest tier network is large as shown by data set 4, the performance of the CHAR scheme is closer to that of the PPR scheme, compared with the cases of small relative cost ratios.

5.4.2.4 Summary

From the above analysis, we see that both the LATR scheme and the PPR scheme have their advantages under different scenarios. When not many calls are delivered from a specific network, the LATR scheme performs better. When most calls are consecutively from one network, the PPR scheme is more cost-efficient. In addition, when the average location registration cost or the paging cost in the lowest tier network are small, the LATR scheme is more favorable in terms of lower total signaling cost. When the average location registration cost of the lowest tier network is much more than that of the preferred call delivery network, or the paging cost of the lowest tier network is comparable with those of other networks, the PPR scheme may improve the system performance. We also find that the CHAR scheme has similar performance as the PPR scheme. It is a good approximation of PPR scheme in practical systems. When the paging costs in all networks are small, or the call delivery costs are large, the CHAR scheme improves the system performance more than the PPR scheme. The CHAR scheme maintains the main advantages of the PPR scheme, but does not require special a-posteriori knowledge to perform well.

5.5 Threshold-Based Adjustable Registration

It is demonstrated in Section 5.4.2 that the performance of the LATR and the CHAR scheme depends on the probability that a call is delivered from a specific network other than the lowest tier network, p. For small p, the LATR scheme is more favorable, while for large p, the CHAR scheme is more advantageous. Based on this conclusion, we introduce an enhancement method called *THreshold-based Adjustable Registration* (THAR). Under this method, each MT keeps the records on which network delivers each call. The highest frequency of calls from a network other than the lowest tier network is used as the criterion for selecting location management schemes. If the highest frequency is less than a pre-defined value, the MT adopts the LATR

scheme and chooses the lowest available tier network for location registrations. When the frequency is larger than the threshold, the MT turns to the CHAR scheme and dynamically changes the registration network according to its communication history of the MT. Thus, based on the threshold, the MT adaptively chooses which location management scheme to follow.

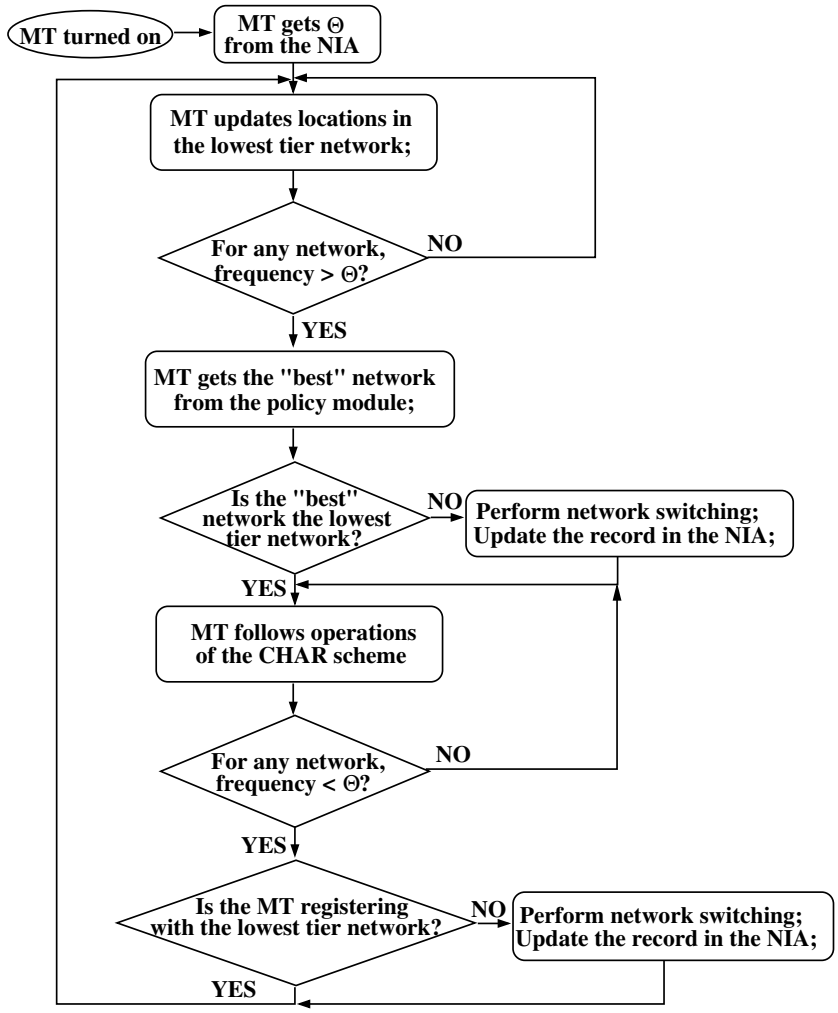

Figure 49: Operation procedures of the THAR scheme.

The operation procedure of the THAR scheme is shown in Figure 49. When an MT is first turned on, it gets the threshold value Θ from the NIA. The NIA has a table for each MT which gives a pre-calculated threshold Θ. Given the cost values of all the available networks, the threshold is obtained when $C_{LATR} = C_{CHAR}$ in (67)

and (71). Note that for different networks, Θ is different. The MT performs location registrations in the lowest tier network at first. It records how many calls are delivered from each network. Whenever a call is delivered, the MT calculates the frequency of calls from each network by dividing the number of calls delivered from each network to the total number of call arrivals. Note that this frequency can be considered as the a-priori probability of call arrivals. If the highest value of the frequency is larger than the threshold Θ of the corresponding network, the MT obtains the "best" network from the policy module as its registration network. It performs a network switching to update the record in the NIA. Then the MT follows the operations of the CHAR scheme. If after some call arrivals, the highest value of the frequency is lower than Θ of the corresponding network, the MT changes to the lowest tier network for location registrations. Thus, based on the threshold, the system dynamically switches between the LATR scheme and the CHAR scheme. Under the THAR method, the system can always have the better overall performance of both schemes.

CHAPTER VI

HANDOFF MANAGEMENT IN NG HETEROGENEOUS WIRELESS OVERLAY NETWORKS

6.1 Problem and Solution

Handoff management is the process by which an MT keeps its connection active with the system when it changes from one access point to another one. Horizontal handoff (intra-system handoff) occurs when an MT is moving out of the coverage area of a cell into the coverage area of another cell. In heterogeneous wireless overlay networks, vertical handoff (inter-system handoff) occurs in two different scenarios. The first one is when an MT is moving out of the current serving network into an overlaying network. This scenario is similar to that in horizontal handoff and we call the corresponding handoff *forced handoff*. The second scenario is when an MT is covered by several overlay networks and it chooses to be handed off from its current serving network to an underlying or overlaying network for better performance. Under this scenario, an MT performs vertical handoff not because of loss of connection in the current serving network. We call this type of handoff *unforced handoff*.

Several proposals for inter-system handoff have been explored. Some techniques [84] [85] [86] addressed handoff between different tiers or different technologies used within an existing architecture, such as the International Mobile Telecommunication System 2000 (IMT-2000) or the Universal Mobile Telecommunication System (UMTS). Other proposals developed new architectures to support inter-system roaming between different networks [87] [22] [23] [24]. Some recent research activities

focus on the inter-system handoff management in the integrated 3G/WLAN environment [75] [88] [89] [90]. Different approaches have been proposed to interconnect these two systems, which can be broadly classified into *tight coupling* (also known as *emulator approach*), *loose coupling* (also known as *Mobile IP approach*), and *no coupling* (also known as *gateway approach*) [91] [92].

Handoff management in heterogeneous wireless overlay networks is addressed in [22] [23] [24]. Vertical handoff in wireless overlay networks is designed in [22] where heterogeneous networks in a hierarchical structure has fully overlapping service areas. Vertical handoff is defined as handoff between BSs that are using different wireless network technologies. Rather than depending on network-specific channel measurements to predict disconnections, the proposed scheme depends on higher-order information such as the presence or absence of beacon and data packets. A policy-enabled handoff system in wireless overlay networks is later proposed in [23]. It allows users to issue policies and have their mobile devices connected to the most desirable network to them. A performance reporting scheme is designed for the policy-enabled handoff system to estimate current network conditions which serves as input to the policy specification. The goal of the proposed scheme is to make it possible to balance the bandwidth load across networks with comparable performance.

For horizontal handoff, the choice of the "best" BS is purely based on the signal strength an MT receives from neighboring BSs. However, in wireless overlay networks, the choice of the "best" network for handoff places a new challenge. It cannot be determined only by channel-specific factors such as signal strength, because different overlay levels may have widely varying characteristics [22]. For both forced and unforced handoff, policies on what the "best" reachable network can be complex to specify. A single hard coded policy is suboptimal [23].

Several factors influence the design of policies on the "best" network for vertical handoff. Cost, network conditions, power consumption, connection setup time, and

user activity history are considered as the parameters for the policy module. In addition, in NG multimedia communications environment, QoS maintenance must be guaranteed after an MT is handed off to a new network. Different traffic types have different bandwidth and delay requirements. Therefore, the required QoS from applications is also an input parameter for the policy design. Moreover, the "best" network selection also influences the distribution of the overall network load. If all the MTs are handed off to one network at the same time, this network is likely to get congested and the resources in other networks are wasted. On the other hand, if the number of MTs communicating in each network is the same, the overall resources may still not be optimally distributed since the network conditions and cost of different networks are different.

In [23], a terminal-based decision making mechanism on the "best" network for handoff was proposed. MTs periodically collect dynamic network conditions and determine the "best" reachable network for handoff. Terminal-based mechanism is a distributed scheme. It is scalable and easy for implementation. The decision making module is located inside each MT and each MT may easily monitor the dynamically changing input parameters. However, under the distributed mechanism, each MT makes the handoff decision without considering the overall performance of the whole system. Several MTs in the same vicinity may discover the same better network and switch to it simultaneously, causing its load to increase dramatically. In addition, due to the lack of a centralized control, the resource of the entire system may not be optimally allocated under the distributed mechanism. On the other hand, under a network-based decision making mechanism, each network may provide network-specific information that may be hard for the MT itself to acquire [2]. Network-based decision making mechanism is a centralized scheme. It may select the "best" network for each MT based on global observation and achieve optimal performance for the whole system. The decision is made in order to optimally utilize the limited resources

of the entire system, to provide satisfactory overall performance, and to keep the low cost to each MT. The drawback of the centralized mechanism is that a central module needs to periodically gather the dynamically changing network conditions in order to make decisions.

In this chapter, we design an efficient mechanism for decision making on what the "best" network is. The objective of this mechanism is to provide a satisfactory overall performance of the whole system. Each decision is made so that the cost for each MT is low, and the most important, the resource of the system is optimally allocated and the load on each network is balanced.

This chapter is organized as follows. In Section 6.2, the system model for the proposed mechanism is described. In Section 6.3, the hybrid control resource allocation scheme for vertical handoff in wireless overlay networks is explained in detail. Then, in Section 6.4, the cost function and the analytical model are derived to solve the optimization problem of resource allocation. After that, in Section 6.5, an algorithm for finding the optimal solution is provided. In Section 6.6, numerical results are presented.

6.2 System Model

The wireless overlay networks we consider in this chapter include various heterogeneous networks using different radio technologies and different network management techniques. These networks have fully overlapping areas of coverage and are organized in a hierarchical structure, as shown in Figure 1. Networks at lower levels in the hierarchy are comprised of high bandwidth wireless cells that cover a relatively small area, while networks at higher levels provide a lower bandwidth per unit area connection over a larger geographic area. MTs with multiple physical or software-defined interfaces may communicate in all the networks they have subscriptions. Our design goal is that each MT selects the "best" reachable network for communications. The

"best" is defined in the sense that each MT pays the minimum cost for communications as well as the overall system pays the minimum cost for supporting all users in the system.

Since handoff is the procedure by which an MT keeps its connection with the system when it chances its access point, we assume all the MTs considered in this chapter are actively communicating with others. Depending on the mobility of MTs, there are three possible scenarios:

1. All the MTs under consideration are staying inside its current communication network without moving to another network, i.e., handoff happened in this scenario is unforced handoff.

2. Part of the MTs under consideration are moving out of the current serving network into an overlaying network, i.e., handoff happened in this scenario can be either forced handoff or unforced handoff.

3. Besides the roaming issue in the second scenario, new users may initiate their communications and some MTs may finish their communications during the time period of consideration, i.e., the total number of MTs that are actively communicating with others may change.

In this chapter, we focus on the first scenario. We assume the total number of MTs under consideration is fixed during the time period of consideration and no roaming is involved. We leave the other two scenarios to the future work.

6.3 Hybrid Control Scheme for Resource Allocation

We propose a hybrid control resource allocation scheme for vertical handoff in heterogeneous wireless overlay networks. This scheme includes a set of access selection criteria and mechanisms that allow MTs to connect to various services through multiple

access networks optimally. It is a hybrid control scheme that combines terminal-based selection and network-based selection mechanisms.

The proposed scheme is a two-level decision making scheme. At the first level, each MT monitors and periodically collects the dynamically changing input parameters for decision making at the terminal side. These parameters are network characteristics which include available bandwidth, network cost, connection setup time, reliability, etc. and power consumption at the terminal. In addition, a user profile should be stored inside the terminal, which contains the user's personal preferences for the choice of access network and user activity history. When booting up, the MT has no connection to the network-based functionality for access selection. Therefore, the MT needs a stored profile, a priority list, or a default setting for choosing access network. This also applies if the terminal loses connectivity over the currently used network. In order to regain connectivity to the application servers, the MT must choose another access without support from the network [2]. The decision making module inside each MT determines the "best" reachable access network based on the input parameters so that each MT pays the minimum total cost. This decision is periodically made depending on the reporting frequency of the input parameters. Note that this decision is a local decision in the sense that the terminal has no idea about the overall resource allocation. At the second level, a central module for the entire system periodically collects the changing information on each network conditions. This information is related to the number of users that are communicating in each network. The central module can be located inside a central controller like the Network Interoperating Agent (NIA) we proposed before [82], which takes care of the interworking issues related to global roaming. The central module finds the optimal user distribution based on a global cost function. This optimal distribution will be used to determine the adjustment for the distributed decision at each terminal.

The proposed scheme is a hybrid control scheme. Each distributed module inside

each MT gives the "best" access network for each MT based on local collected information. Each MT decides which access network it will select so that it will pay the minimum total cost for communications. The central module gives adjustment to the decision made by the distributed module based on global information and a cost function for the entire system. The objective of the decision made by the central module is that the overall system pays the minimum total cost for supporting all the users communicating in the system. The adjustment command given by the central module will reduce the overall system cost and make all MTs gradually move to the optimal distribution status, instead of abruptly forcing a certain amount of MTs to change their access networks. Therefore, the proposed scheme is a hybrid control scheme that combines terminal-based selection and network-based selection mechanisms. Note that the cost function for decision-making at each terminal is different from that at the central module. The input parameters for the cost function at each terminal are local information of the networks that are reachable to the MT, while the input parameters for the cost function at the central module are global information on each network conditions. We will explain the meaning of these parameters in detail in the next section.

6.4 Cost Function

In this section, we design the cost functions of the terminal-based selection mechanism and the network-based selection mechanism, respectively. The cost function of the terminal-based selection mechanism is to help each MT select the "best" access network that results in the minimum total cost. The cost function of the network-based selection mechanism is to help the system to find an optimal user distribution that results in the minimum overall system cost, i.e., to find how many users should be communicating in each network.

6.4.1 Cost Function of the Terminal-Based Selection Mechanism

We adopt the same cost function proposed in [23] for the terminal-based selection mechanism. The cost of using a network i at a certain time for each MT is a function of three parameters: the average bandwidth network i can offer (B_i), the power consumption at each MT of using the network i (P_i), and the price of this network (M_i), i.e., the cost at Level 1 control is:

$$Cost_i^{L1} = f(B_i, P_i, M_i) \tag{72}$$

The bandwidth parameter estimates the current network condition. A performance agent was proposed in [23] that collects the information on current bandwidth usage at BSs and periodically announces this information to its coverage area. Power consumption and price are parameters with fixed budgets, namely, the battery consumption and the amount of money the user will spend for a period of time, respectively.

Normalization of the above cost function is needed to ensure that the sum of the values in different units is meaningful. The normalized cost function at Level 1 is:

$$Cost_i^{L1} = \omega_b \cdot \log \frac{1}{B_i} + \omega_p \cdot \log P_i + \omega_m \cdot \log M_i \tag{73}$$

where ω_b, ω_p, and ω_m are weights of each parameters and $\omega_b + \omega_p + \omega_m = 1$. Users may specify the weights to show their preferences. For those parameters that are not of concern, their weights can be set to 0. For example, if a user wants to be connected to the cheapest network at all time, then $\omega_m = 1$ and other weights are set to 0. Furthermore, weights can also be modified by users or the network at run-time to reflect the changing importance.

Each MT periodically compare the reachable access networks by calculating the cost functions of each network, $Cost_i^{L1}$. It then makes the decision by selecting the network with the lowest value of the cost function.

6.4.2 Cost Function of the Network-Based Selection Mechanism

The cost function at the central module for decision-making is the sum of costs each network needs to pay to support the users communicating in each network, i.e., the cost at Level 2 control is:

$$Cost^{L2}(N_1, N_2, \ldots, N_M) = \sum_{i=1}^{M} C_i(N_i) \tag{74}$$

where C_i is the cost network i needs to pay to support MTs using service in this network and M is the total number of networks in the system. In other words, in order to support all the users in the network, network i will spend a certain amount of resources (available bandwidth, processing capability, computational resource, etc.). These resources paid by network i are quantified as cost C_i. Note that C_i is a function of the total number of users using network i, N_i. The larger the N_i, the more resources the network needs to spend, and the larger the C_i.

The cost network i needs to pay to support its users is a function of two parameters: the total bandwidth offered to all the users ($\mathcal{B}_i$) and the service quality. Both these two parameters are proportional to the total number of users in the network, N_i. If N_i increases, the requested bandwidth from users also increases. Since the total amount of bandwidth a network can offer is limited, when N_i increases, the service quality provided to each user may not be guaranteed. We choose the average number of errors occurred during the communications within the time period under consideration ($E_i(err)$) as the metric of service quality. Note that this $E_i(err)$ is the sum of all kinds of errors caused by the increase of the total number of active users, that is, due to the limited system resources and the increasing number of users, the network get congested, e.g., the transmission error, the long service request delay, and the disconnection of a communication. $E_i(err)$ does not refer to the error caused by the poor channel link or other factors that are not related to the increase of N_i.

Therefore, we design C_i as:

$$C_i(N_i) = c_i(b)\mathcal{B}_i + c_i(e)E_i(err) \tag{75}$$

where $c_i(b)$ and $c_i(e)$ are the cost of offering unit bandwidth and the cost of correcting each error. They are constants and can be chosen to make the sum of the valued in different units meaningful.

According to the performance agent proposed in [23], each BS periodically announces the current bandwidth usage in its coverage area. Hence, $\mathcal{B}_i$ is known to network i. $\mathcal{B}_i$ is proportional to N_i. We may write $\mathcal{B}_i$ as $\mathcal{B}_i = Avg(B_i) \cdot N_i$, where $Avg(B_i)$ is the average bandwidth requested by each user in network i.

$E_i(err)$ is the expectation of the number of errors occurred within a certain time period, i.e., $E_i(err)$ is:

$$E_i(err) = \sum_{n=1}^{\infty} np_i(n) \tag{76}$$

where $p_i(n)$ is the probability of having n errors during the communications in network i. Assume the probability of having one error during the whole communications within the considered time period is q_i. Then, $p_i(n) = q_i^n$. The above equation of $E_i(err)$ changes to:

$$\begin{aligned} E_i(err) &= \sum_{n=1}^{\infty} nq_i^n \\ &= q_i + 2q_i^2 + 3q_i^3 + \cdots + nq_i^n + \cdots \end{aligned} \tag{77}$$

We may calculate $E_i(err)$ as:

$$E_i(err) = \frac{q_i}{(1-q_i)^2} \tag{78}$$

Now, we find the relationship of q_i and N_i. Assume within one unit time, the probability that a user's communications have errors due to the congestion of the network is a_i. Then, the probability that N_i users' communications have errors within one unit time is:

$$q_i = 1 - (1 - a_i)^{N_i} \tag{79}$$

Note that N_i increases, the network gets more congested and the probability of having errors, q_i will also increase.

(75) can be expressed as:

$$\begin{aligned} C_i(N_i) &= \alpha c_i(b) \cdot Avg(B_i)N_i + \beta c_i(e)\frac{q_i}{(1-q_i)^2} \\ &= \alpha c_i(b)N_i + \beta c_i(e)\frac{q_i}{(1-q_i)^2} \end{aligned} \tag{80}$$

Here, we absorb $Avg(B_i)$ into the cost constant $c_i(b)$. α and β are weights of bandwidth and error parameters in the cost function similar to (73) and $\alpha + \beta = 1$. α and β can reflect the importance of each parameter.

Given the cost function at the central module $Cost^{L2}(N_1, N_2, \ldots, N_M)$, we may find the optimal user distribution in each network, $(N_1, N_2, \ldots, N_M)$, that minimized the overall system cost.

6.5 Optimization Solution and Adjustment

In this section, we derive the optimal user distribution based on the cost function in (74). Then, we design the adjustment policy for each MT to adjust the decision made by the distributed module inside the terminal.

6.5.1 Optimization Solution

6.5.1.1 Mathematical Formulation

Assume the total number of MTs in the entire system is fixed within the time period of consideration, i.e., $N_1 + N_2 + \cdots + N_M = N$, where N is a fixed number. Since the cost function in (74) is a non-linear function and N_i must be an integer, the optimization problem is a non-linear integer problem. It is very hard to find the explicit solution of non-linear integer optimization problem. In order to simplify and expedite the online choice of user distribution, we propose an iterative method.

We assume the cost constants $c_i(b)$ and $c_i(e)$ are available at the central module. $c_i(b)$ contains the value for $Avg(B_i)$ which can also be available to the system, since

the total bandwidth usage in each network is known according to [23]. Assume the average error probability per unit time of each network, a_i, is also known. It can be determined based on empirical measurements. The integer problem for the proposed hybrid control scheme is summarized as follows:

$$\begin{array}{ll} \text{GIVEN} & \alpha,\ \beta,\ c_i(b),\ c_i(e),\ a_i \\ \text{FIND} & (N_1, N_2, \ldots, N_M) \quad \text{M integer variables} \\ \text{MINIMIZE} & Cost^{L2}(N_1, N_2, \ldots, N_M) = \sum_{i=1}^{M} C_i(N_i) \\ \text{SUBJECT TO} & N_1 + N_2 + \cdots + N_M = N \end{array} \tag{81}$$

6.5.1.2 Iterative Algorithm

For the system of small size (small M), or for a small number of users (small N), the online use of an optimal tool is a fast and accurate way. However, for a large system and a large number of users, an iterative algorithm that could approximate the optimal result would be preferable.

In order to simplify the online choice of user distribution, we propose a new iterative algorithm to find the optimal user distribution that minimized the overall system cost. Under the proposed algorithm, during each iteration step, only two integer variables of N_i change their values. Since the high cost of a network is caused by the large number of users communicating inside this network, the number of users in the network that has the highest cost should be reduced. In addition, since the total number of users in the whole system is fixed, the number of users in the network that has the lowest cost should be increased. δ of user change is applied to these two networks, where δ is the changing number of users during each step. After each iteration, the costs of these two networks C_i are re-computed and the cost values of all the networks are re-ordered. During the next iteration step, the new networks with the highest cost and the lowest cost change their user numbers. This iteration is continued until the new total cost value $Cost^{L2}$ is larger than the old value before

the iteration step. Figure 50 shows the details of the iterative algorithm.

```
Initialize N_i;
Cost^{L2}_{old} = 0;
Compute C_i(N_i);
Compute Cost^{L2}_{new} = Σ_{i=1}^{M} C_i(N_i);
while Cost^{L2}_{old} − Cost^{L2}_{new} ≥ 0
   Sort C_i(N_i) and find C_max(N_max), C_min(N_min);
   N_max = N_max − δ;
   N_min = N_min + δ;
   Compute C_i(N_i);
   Compute Cost^{L2}_{new} = Σ_{i=1}^{M} C_i(N_i);
end
```

Figure 50: Iterative algorithm for finding the optimal user distribution.

After finishing the algorithm, the optimal user distribution $(N_1, N_2, \ldots, N_M)$ is found that minimizes the overall system cost $Cost^{L2}(N_1, N_2, \ldots, N_M)$. Note that the iterative algorithm may result in a local minimum. As can be seen from the following sections, the goal of the optimal user distribution is to generation an adjustment for each network so that the users using the services from this network may incorporate this adjustment into their decisions of the "best" network selection made by the distributed module. Whether an MT will be handed off to a new network depends on the user preference as well as the global adjustment suggestion from the central module. The objective of the proposed hybrid control scheme is to gradually reduce the overall system cost through the adjustment. Therefore, the accuracy of the optimal solution is not critical to the system performance.

6.5.2 Adjustment Policy

Before describing the designed adjustment policy, let us analyze an example first, which is shown in Figure 51. Assume the total number of networks in the entire system is two. In the figure, X-axis represents the number of users in network 1, while Y-axis represents the number of users in network 2. Z-axis represents the total

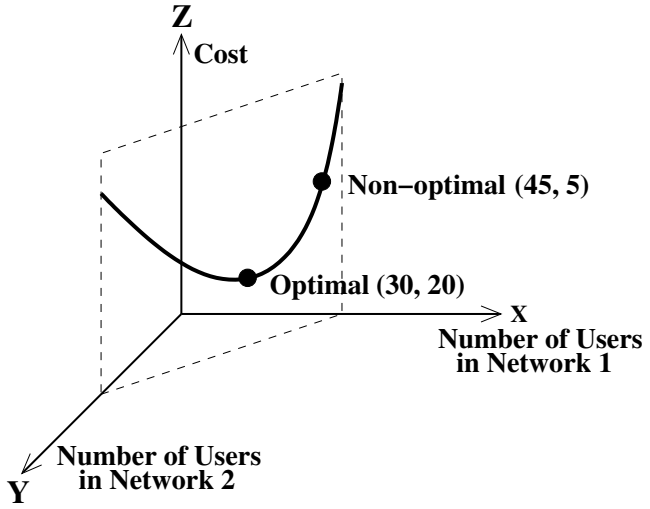

Figure 51: Relationship between the total cost and the user distribution.

system cost, i.e., the sum of the costs of the two networks. Since the total number of users in the system is fixed during the time period of consideration, the cost curve is on a plane which is parallel to the Z-axis. The integer optimization problem at the central module is equivalent to finding the optimal user distribution that results in the minimal cost, i.e., the minimum point on the cost curve in the figure. Assume the current user distribution leads to an non-optimal point on the cost curve, that is, the current system total cost is larger than the minimal cost. The goal of the adjustment policy design is to find an efficient way so that the system status can gradually transfer from the non-optimal point to the optimal point on the cost curve.

Under the proposed hybrid control scheme, each user's own preference on the "best" access network selection is important. Users may specify the weights value in (73) to show their preferences. The global control from the central module is also important for the total system cost control. However, the adjustment to each network should be smooth and gradual, not abrupt. After finding the optimal user distribution, the central module should not tell each network to force several users to change their communication networks.

We propose a new adjustment policy. Based on the optimal user distribution, the

central module finds an adjustment number, Δ, for each network. Δ for network i is computed by the following equation:

$$\Delta_i = \frac{N_{i\_opt} - N_{i\_non-opt}}{\sqrt{(N_{1\_opt} - N_{1\_non-opt})^2 + (N_{2\_opt} - N_{2\_non-opt})^2 + \cdots + (N_{M\_opt} - N_{M\_non-opt})^2}} \tag{82}$$

where N_{i_opt} and $N_{i\_non-opt}$ are the values of N_i in the optimal user distribution and non-optimal user distribution, respectively. The denominator part of the above equation is actually the Euler Distance between the optimal point and the non-optimal point on the cost curve.

The adjustment number Δ is added to the cost calculated by the distributed module inside each terminal, i.e., the new cost function for network i at the terminal side changes to:

$$Cost_i^{L1} = Cost_i^{L1}(B_i, P_i, M_i) + \Delta_i \tag{83}$$

where $Cost_i^{L1}(B_i, P_i, M_i)$ is calculated by (73).

We use the example in Figure 51 to explain the function of the penalty number. In this example, the current user distribution is (45, 5), that means, there are 45 users communicating in network 1 and 5 in network 2. The total number of active users in the system is 50. Assume the central module finds the optimal user distribution should be (30, 20), i.e., 30 users use network 1 and 20 for network 2. Then, the central module calculates an adjustment number for each network based on the optimal user distribution and the current user distribution, which is $\Delta_1 = \frac{30-45}{\sqrt{(30-45)^2+(20-5)^2}} = -0.71$ for network 1 and $\Delta_2 = \frac{20-55}{\sqrt{(30-45)^2+(20-5)^2}} = 0.71$ for network 2. The adjustment value is then added to the distributed cost function. Note that the adjustment value Δ is negative for network 1. It means that network 1 discourage MTs to be handed off to it since it already has too many users. Similarly, the positive value of Δ_2 implies that network 2 encourages users to communicate in it.

After the adjustment, each MT chooses the "best" access network for communication. The "best" network is determined periodically and the proposed hybrid

control scheme is applied each time a decision is made. Note that under the proposed adjustment policy, the central control is reflected by the adjustment number for each network.

6.6 Numerical Results

In this section, we demonstrate the performance improvement of the propose hybrid control resource allocation scheme for vertical handoff. We first present the iteration procedure at the central module and show the optimal user distribution and the minimal cost after the iteration. Next, we compare the performance of the hybrid control scheme and the scheme without the central control proposed in [23].

6.6.1 System and User Parameters

We assume there are totally four heterogeneous access networks in the system, i.e., $M = 4$. These four networks are overlay networks. Assume 1000 MTs are covered by the four networks at the same time and they are communicating with others using these four networks, i.e., $N = 1000$. At the initial phase, each MT randomly picks a network for communication, that is, the probability that an MT chooses a specific network is equal to $\frac{1}{4}$. Table 7 shows the initial user distribution, i.e., $(N_1, N_2, N_3, N_4) = (227, 271, 230, 272)$.

Table 7: Initial User Distribution for the Hybrid Control Resource Allocation Scheme

	Network 1	Network 2	Network 3	Network 4
Number of Users	227	271	230	272

Table 8 lists the values of system parameters: cost constant of bandwidth $c_i(b)$, cost constant of correcting errors $c_i(e)$, and the probability that a user's communications have errors within one unit time a_i. We assume the values of weights of bandwidth and error parameters are equal, i.e., $\alpha = \beta = 10000$. This means that offering bandwidth and correcting communication errors are of equal importance to

each network.

Table 8: System Parameters for the Hybrid Control Resource Allocation Scheme

	$c_i(b)$	$c_i(e)$	a_i
Network 1	1.2×10^{-6}	0.12	2.0×10^{-6}
Network 2	1.23×10^{-6}	0.113	1.0×10^{-7}
Network 3	1.0×10^{-6}	0.12	1.0×10^{-5}
Network 4	1.25×10^{-6}	0.11	8.0×10^{-9}

We randomly generate the values of user parameters for each user. These parameters are: the initial network in which the user is communicating, the bandwidth each network can offer B_i, and the price of each network M_i. We assume the power consumption P_i is not important when each user considers the cost of each network and set $\omega_p = 0$. Then $\omega_m = 1 - \omega_b$. We also randomly generate the value of ω_b for each user, where $0 \leq \omega_b \leq 1$.

6.6.2 Iteration Procedure

First, we demonstrate the procedure of the iterative algorithm at the central module for finding the optimal user distribution. The initial value of $Cost^{L2} = 9.42$. Table 9 shows the iteration procedure of the optimization. In the table, MAX refers to the network with the highest cost and its user number is reduced by one, while MIN refers to the network with the lowest cost and its user number is incremented by one.

The iteration stops after 112 steps when the newly calculated $Cost^{L2}$ is larger than the previous value. As we can see from the table, the value of $Cost^{L2}$ continuously goes down during the iteration. The number of users using network 3 should be reduced since the error probability of network 3, a_3, is large compared with those of other networks, which leads to high network cost C_3. After the iteration, we get the minimal cost value $Cost^{L2}_{min} = 8.26$. The corresponding optimal user distribution is shown in Table 10, i.e., $(N_1, N_2, N_3, N_4) = (240, 319, 121, 320)$.

6.6.3 Hybrid Control Scheme vs. Distributed Scheme

Next, we compare the performance of the proposed hybrid control resource allocation scheme and the scheme without the central control proposed in [23], i.e., the scheme under which each user determines the "best" access network only based on its own preferences and local information on network conditions.

6.6.3.1 Hybrid Control Scheme

For the proposed hybrid control scheme, after finding the optimal user distribution, the central module computes the adjustment number for each network, which is shown in Table 11.

Each MT receives the adjustment numbers from the central module. It adds these numbers to the cost value $Cost_i^{L1}$ and decides which network it will be handed off to or stay as no change. After the adjustment, the new user distribution is $(N_1, N_2, N_3, N_4) = (249, 360, 117, 274)$, which is shown in Table 12. And the new total cost $Cost^{L2} = 14.45$.

Note that the new user distribution is different from the optimal user distribution, but it follows the changing trend, i.e., the number of users using network 3 should be reduced, while the population in other networks should be increased. This indicates that although the central module gives the suggestions of user distribution, each user's own preference still places important role in the decision-making. The new total cost $Cost^{L2}$ is higher than the initial total cost before applying the hybrid control for two reasons. First, the initial "best" network selection of each user is random without considering user preferences. This will rarely happen in the real system. Second, the parameters determining the $Cost^{L1}$ and $Cost^{L2}$ are different. When users value more on the user preference parameters (power consumption and price) which are not reflected in the cost function at the central module $Cost^{L2}$, users' choices of "best" network will deviate the total system cost from the optimal value.

6.6.3.2 Distributed Scheme

The distributed scheme refers to the scheme without applying the central control from the system. Each user determines the "best" network only based on the values of $Cost_i^{L1}$. The new user distribution without applying the central adjustment is $(N_1, N_2, N_3, N_4) = (222, 175, 480, 123)$, which is shown in Table 13. And the new total cost $\widetilde{Cost}^{L2} = 17.88$.

Note that the new user distribution under the distributed scheme does not following the changing trend suggested from the optimal user distribution. The number of users communicating in network 3 is not reduced but increased. This is due to the low price offered by network 3 to most users and many users value the price factor more when selecting the "best" communication network. Moreover, without the central control, under the distributed scheme, the total cost $\widetilde{Cost}^{L2}$ is 24% higher than $Cost^{L2}$ under the hybrid control scheme.

6.6.3.3 Summary

From the numerical results of the two schemes, we may conclude that the user distribution is closer to the optimal user distribution under the hybrid control scheme, compared with the scheme without the central control. In addition, the total system cost can be reduced by applying the hybrid control scheme and the overall system resources can be allocated close to the optimal solution. At the same time, the user preferences are still retained.

Table 9: Iteration Procedure for Finding the Optimal User Distribution

Step	$Cost^{L2}$	MAX	MIN	Step	$Cost^{L2}$	MAX	MIN	Step	$Cost^{L2}$	MAX	MIN
1	9.42	3	2	39	9.01	3	2	77	8.61	3	2
2	9.41	3	4	40	9.00	3	4	78	8.60	3	1
3	9.40	3	2	41	8.99	3	2	79	8.59	3	4
4	9.39	3	4	42	8.98	3	4	80	8.58	3	2
5	9.38	3	2	43	8.97	3	2	81	8.57	3	1
6	9.37	3	4	44	8.96	3	4	82	8.56	3	4
7	9.35	3	2	45	8.95	3	2	83	8.55	3	2
8	9.34	3	4	46	8.93	3	4	84	8.54	3	1
9	9.33	3	2	47	8.92	3	2	85	8.53	3	4
10	9.32	3	4	48	8.91	3	4	86	8.52	3	2
11	9.31	3	2	49	8.90	3	2	87	8.51	3	4
12	9.30	3	4	50	8.89	3	4	88	8.50	3	2
13	9.29	3	2	51	8.88	3	2	89	8.49	3	1
14	9.28	3	4	52	8.87	3	4	90	8.48	3	4
15	9.27	3	2	53	8.86	3	2	91	8.47	3	2
16	9.26	3	4	54	8.85	3	4	92	8.46	3	1
17	9.25	3	2	55	8.84	3	2	93	8.45	3	4
18	9.24	3	4	56	8.83	3	4	94	8.44	3	2
19	9.23	3	2	57	8.82	3	2	95	8.43	3	1
20	9.21	3	4	58	8.81	3	4	96	8.42	3	4
21	9.20	3	2	59	8.79	3	2	97	8.41	3	2
22	9.19	3	4	60	8.78	3	4	98	8.40	3	4
23	9.18	3	2	61	8.77	3	2	99	8.39	3	2
24	9.17	3	4	62	8.76	3	1	100	8.37	3	1
25	9.16	3	2	63	8.75	3	4	101	8.37	3	4
26	9.15	3	4	64	8.74	3	2	102	8.36	3	2
27	9.14	3	2	65	8.73	3	4	103	8.34	3	1
28	9.13	3	4	66	8.72	3	2	104	8.34	3	4
29	9.12	3	2	67	8.71	3	1	105	8.33	3	2
30	9.11	3	4	68	8.70	3	4	106	8.31	3	1
31	9.10	3	2	69	8.69	3	2	107	8.31	3	4
32	9.09	3	4	70	8.68	3	1	108	8.30	3	2
33	9.07	3	2	71	8.67	3	4	109	8.28	3	4
34	9.06	3	4	72	8.66	3	2	110	8.27	3	2
35	9.05	3	2	73	8.65	3	1	111	8.26	2	3
36	9.04	3	4	74	8.64	3	4	112	8.27	3	2
37	9.03	3	2	75	8.63	3	2				
38	9.02	3	4	76	8.62	3	4				

Table 10: Optimal User Distribution for the Hybrid Control Resource Allocation Scheme

	Network 1	Network 2	Network 3	Network 4
Number of Users	240	319	121	320

Table 11: Adjustment Number for Each Network

	Network 1	Network 2	Network 3	Network 4
Adjustment Number	0.100724	0.371903	-0.844530	0.371903

Table 12: New User Distribution after Applying the Hybrid Control Scheme

	Network 1	Network 2	Network 3	Network 4
Number of Users	249	360	117	274

Table 13: New User Distribution without Central Control

	Network 1	Network 2	Network 3	Network 4
Number of Users	222	175	480	123

CHAPTER VII

CONCLUSIONS AND FUTURE RESEARCH WORK

7.1 Summary of Research Results

The research work in this thesis was focused on the development of new mobility management techniques for NG all-IP based wireless systems. Research contributions have been made in the following areas:

1. Location management in NG wireless Internet
2. Paging in NG wireless Internet
3. Paging-aided connection setup in NG wireless Internet
4. Location management in NG wireless overlay networks
5. Handoff management in NG wireless overlay networks

7.1.1 Location management in NG wireless Internet

In Chapter 2, a distributed and dynamic regional location management mechanism for Mobile IP was introduced. We proposed a distributed GFA system architecture where each FA can function either as an FA or a GFA. This distributed system may allocate signaling burden more evenly. A dynamic scheme is adopted by the distributed system to dynamically optimize the regional network size of each MN according to its current traffic load and mobility. We also presented the operation protocols of the distributed dynamic scheme for MNs. The proposed distributed and dynamic scheme is able to perform optimally for all users from time to time and the system robustness

is enhanced. Since the movement of MNs does not follow a Markov process, we introduced a novel discrete analytical model for cost analysis and an iterative algorithm to find out the optimal number of FAs in a regional network which consumes the minimal network resource. Our model does not have constraints on the shape and the geographic location of Internet subnets. Analytical results demonstrated that the signaling bandwidth is significantly reduced through our proposed distributed system architecture compared with the IETF Mobile IP regional registration scheme. It is also demonstrated that our dynamic scheme has great advantages under time-variant user parameters when it is not obvious to pre-determine the optimal regional network size.

The proposed distributed dynamic location management scheme requires that all FAs are capable of functioning as both an FA and a GFA. It increases the requirement of the processing capability on each mobility agent. There is additional processing load on the mobile terminals, such as the estimation of the average packet arrival rate and subnet residence time.

7.1.2 Paging in NG wireless Internet

In Chapter 3, a user independent paging scheme based on last-known location and mobility rate information for Mobile IP was introduced. User independent paging can be applied to both link layer paging and IP layer paging, as long as the selected paging criterion is user independent. In this paper, we proposed an efficient IP paging scheme which can be employed by all the proposed paging architectures. In contrast to the user dependent paging schemes that choose the user-variant parameters as the paging criterion, the proposed scheme takes the aggregated behavior of all mobile users as the basis for paging. It combines the advantages of last-location-first paging, highest-mobility-first paging, and uniform paging. In order to obtain the mobility rate of each subnet, a new operation mode named "semi-idle" mode is introduced. Analytical

results demonstrated that when paging a single user at a time, the performance of the proposed paging scheme is comparable to that of the paging scheme based on perfect knowledge of user movement statistics. However, when paging multiple users at one moment and when the assumption of perfect knowledge is loose, the proposed paging scheme saves signaling bandwidth significantly.

The proposed user independent paging scheme may be considered as an alternative approach. If the system may perform paging optimally for every user, user dependent paging scheme is a good solution to provide personalized service. However, when it is hard to achieve perfect performance for each user, or obtaining user mobility profiles causes great cost, user independent paging scheme may provide satisfactory overall performance for the whole system.

7.1.3 Paging-aided connection setup in NG wireless Internet

In Chapter 4, a new paging-aided connection setup scheme for real-time communication in Mobile Internet was introduced. The proposed scheme employs home agent paging architecture in order to keep the property of shortest IP routing path. Instead of performing paging and RSVP path setup sequentially, we proposed to perform the two procedures concurrently. We explained the operation details for both unicast and multicast communications under the proposed scheme. Performance analysis demonstrated that the new paging-aided connection setup scheme outperforms the traditional scheme in terms of reducing the overall connection setup time and the total number of signaling messages.

7.1.4 Location management in NG wireless overlay networks

In Chapter 5, a new mobility management architecture for NG heterogeneous overlay networks was introduced. Under the proposed architecture, each heterogeneous network keeps its own location management hierarchy and registration procedure unchanged. With the Internet as the common backbone network for signaling message

transmission, this architecture is more cost efficient and can be built in a hierarchical structure to make it more scalable. We proposed three location management schemes under this architecture. Under all the three schemes, each MT updates its location in one network at any time and has user preference call delivery support. Calls can be delivered through any network without restrictions. Numerical results showed that the LATR scheme and the PPR scheme have their advantages under different scenarios. When most calls are delivered from one network, the PPR scheme may reduce signaling cost significantly. However, the PPR scheme is implemented based on the a-posteriori knowledge on future call deliveries, which limits its usage in real systems. The CHAR scheme is a feasible solution in practice and a good approximation of the PPR scheme. It does not require any a-posteriori knowledge of user mobility and call arrival patterns. It keeps the main advantages of the PPR scheme and may further improve the system performance under certain scenarios. We also proposed a threshold-based enhancement method. According to the communication history, the system dynamically switches between the LATR scheme and the CHAR scheme so that the overall performance is always the better one of both schemes.

The future work lies in the further investigation of the THAR scheme. The details of the implementation and the evaluation of the performance improvement of the THAR scheme are worthy of more research.

7.1.5 Handoff management in NG wireless overlay networks

In Chapter 6, a new hybrid control resource allocation scheme for vertical handoff in heterogeneous wireless overlay networks was introduced. Under the proposed scheme, each MT may select the best access network for communications based on its own preferences as well as the adjustment suggested by a central controller. This scheme combines the terminal-based selection and network-based selection mechanisms. It is

a two-level decision-making scheme. The first level decision is made with user preferences taken into account. The second level decision is made in order to optimally allocate the entire system resources so that the overall system cost is minimized. Instead of abruptly forcing users to change their communication networks without considering their own preferences, we proposed a novel adjustment policy. Under this policy, the user preferences are valued and at the same time, the total system cost can be reduced. In order to find the solution of the non-linear integer optimization problem, we also proposed a new iterative algorithm to find the optimal user distribution. Numerical results show that after applying the proposed hybrid control scheme, the overall system cost is reduced and the user distribution moves closer to the optimal one.

7.2 Future Research Work

The future wireless systems will be based on IP infrastructure with packet-switching techniques for multimedia services. However, they will still suffer from the diverse standards that limit the roaming of mobile users between different networks [93]. The NG wireless systems will be able to support global roaming between different access technologies. Mobility management will continuously play an important role to provide seamless services. There are many challenging research issues related mobility management for NG all-IP based wireless systems.

7.2.1 QoS Issues

The NG all-IP wireless systems will have to provide guaranteed QoS to mobile terminals carrying multimedia applications including best effort and real-time traffic. These applications have varying requirements which challenge the best effort service model of the original framework for IP. Bandwidth, throughput, timeliness, reliability, perceived quality, and costs are the foundations of QoS. There have been some proposed QoS architectures for wired networks. However, QoS provisioning in heterogeneous

mobile computing environment introduces new problems to mobility management, such as location management for efficient access and timely service delivery, QoS negotiation during inter-system handoff, etc. In addition, there has been very little work on a suitable QoS model for combined macro and micro mobility [94].

7.2.2 Location and Handoff Management in Overlay Networks

Although mobility management for inter-system roaming of adjacent networks with partially overlapping area will continuously be an important research issue, mobility management in overlay networks which have fully overlapping coverage should be paid more attention. Future wireless system has a hierarchical architecture where different access networks have dramatically different coverage areas. Mobile users will expect to receive personalized end-to-end services no matter where they go and which network is providing services. In this vertical roaming scenario, mobile terminals have a more active role and initiate the specific mobility management mechanisms [71]. Mobility management techniques should allow mobile users to roam among multiple wireless networks in a manner that is completely transparent to applications and disrupts connectivity as little as possible. Moreover, in overlay networks, the choice of the "best" network for location and handoff management places a new challenge because different overlay levels may have widely varying characteristics [22].

7.2.3 Cross Layer Optimization

In the global roaming scenario, mobility can be supported from both the network layer (i.e., IP mobility) and the link layer (i.e., access mobility). Cooperation between the network layer and the link layer can improve the performance of mobility management in IP-based heterogeneous communication environments. Information from the link layer, such as signal strength and velocity of mobile terminals, may help the decision-making of mobility management techniques at the network layer. Signaling cost and delay should also be reduced in both the Internet domain and underlying

radio systems. Therefore, the cross-layer mobility protocol design requires more exploration. Other possible research issues are how the cooperation is implemented, how tight the cooperation is, and how much information is exchanged between the two layers.

REFERENCES

[1] J. Ala-Laurila, J. Mikkonen, and J. Rinnemaa, "Wireless LAN access network architecture for mobile operators," *IEEE Communication Magazine*, pp. 82–89, November 2001.

[2] E. Gustafsson and A. Jonsson, "Always best connected," *IEEE Wireless Communications Magazine*, pp. 49–55, February 2003.

[3] G. Araniti, A. Iera, S. Pulitano, and A. Molinaro, "Managing IP traffic in radio access networks of next-generation mobile systems," *IEEE Wireless Communications*, pp. 36–43, August 2003.

[4] F. M. Chiussi, D. A. Khotimsky, and S. Krishnan, "Mobility management in third-generation all-IP networks," *IEEE Communications Magazine*, pp. 124–135, September 2002.

[5] S. Mohanty and I. F. Akyildiz, "An architecture and associated protocols for inter-system handover between 3G and WLAN," *submitted for publication*, 2004.

[6] I. F. Akyildiz, J. McNair, J. S. M. Ho, H. Uzunalioglu, and W. Wang, "Mobility management for next generation wireless systems," *Proceedings of IEEE*, vol. 87, no. 8, pp. 1347–1384, August 1999.

[7] A. T. Campbell, J. Gomez, S. Kim, A. G. Valko, C.-Y. Wang, and Z. R. Turanyi, "Design, implementation, and evaluation of Cellular IP," *IEEE Personal Communications Magazine*, pp. 42–49, August 2000.

[8] A. Misra, S. Das, A. Dutta, A. McAuley, and S. Das, "IDMP-based fast handoffs and paging in IP-based 4G mobile networks," *IEEE Communication Magazine*, pp. 138–145, March 2002.

[9] R. Ramjee, K. Varadhan, L. Salgarelli, S. R. Thuel, S.-Y. Wang, and T. L. Porta, "HAWAII: A domain-based approach for supporting mobility in wide-area wireless networks," *IEEE/ACM Trans. Networking*, vol. 10, no. 3, pp. 396–410, June 2002.

[10] V. Garg and J. E. Wilkes, "Interworking and interoperability issues for North American PCS," *IEEE Communications Magazine*, vol. 34, no. 3, pp. 94–99, March 1996.

[11] I. F. Akyildiz, J. Xie, and S. Mohanty, "A survey on mobility management in next generation all-IP based wireless systems," *to appear in IEEE Wireless Communications*, 2004.

[12] T. Zhang, S.-W. Li, Y. Ohba, and N. Nakajima, "A flexible and scalable IP paging protocol," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM 2002)*, vol. 1, 2002, pp. 630–635.

[13] C. E. Perkins, "IP mobility support for IPv4," Request for Comments (RFC) 3220, January 2002.

[14] A. T. Campbell and J. Gomez-Castellanos, "IP micromobility protocols," *ACM SIGMOBILE Mobile Computing and Communication Review*, vol. 4, no. 4, pp. 45–54, October 2001.

[15] E. Gustafsson, A. Jonsson, and C. Perkins, "Mobile IPv4 regional registration (work in progress)," Internet Draft, Internet Engineering Task Force, draft-ietf-mobileip-reg-tunnel-07.txt, November 2003.

[16] H. Soliman, C. Castelluccia, K. El-Malki, and L. Bellier, "Hierarchical mobile IPv6 mobility management (HMIPv6) (work in progress)," Internet Draft, Internet Engineering Task Force, draft-ietf-mobileip-hmipv6-07.txt, October 2002.

[17] H. Haverinen and J. Malinen, "Mobile IP regional paging (work in progress)," Internet Draft, Internet Engineering Task Force, draft-haverinen-mobileip-reg-paging-00.txt, June 2000.

[18] R. Ramjee, L. Li, T. L. Porta, and S. Kasera, "IP paging service for mobile hosts," *ACM Wireless Networks (WINET)*, vol. 8, no. 5, pp. 427–441, September 2002.

[19] X. Zhang, J. G. Castellanos, A. T. Campbell, K. Sawada, and M. Barry, "P-MIP minimal paging extensions for Mobile IP (work in progress)," Internet Draft, Internet Engineering Task Force, draft-zhang-pmip-00.txt, July 2000.

[20] X. Zhang, J. G. Castellanos, and A. T. Campbell, "P-MIP: paging extensions for Mobile IP," *ACM Mobile Networks and Applications (MONET)*, vol. 7, no. 2, pp. 127–141, April 2002.

[21] B. Sarikaya and *et al.*, "Mobile IPv6 hierarchical paging (work in progress)," Internet Draft, Internet Engineering Task Force, draft-sarikaya-seamobymipv6hp-00.txt, September 2001.

[22] M. Stemm and R. H. Katz, "Vertical handoffs in wireless overlay networks," *ACM Mobile Networks and Applications (MONET)*, vol. 3, pp. 335–350, 1998.

[23] H. J. Wang, R. H. Katz, and J. Giese, "Policy-enabled handoffs across heterogeneous wireless networks," in *Proc. IEEE Workshop on Mobile Computing Systems and Applications (WMCSA'99)*, 1999, pp. 51–60.

[24] F. Gu, L. M. Ni, and A. H. Esfahanian, "HOPOVER: a new handoff protocol for overlay networks," in *Proc. IEEE International Conference on Communications (ICC 2002)*, vol. 5, 2002, pp. 3234–3239.

[25] C. E. Perkins, "Mobile IP," *IEEE Communication Magazine*, pp. 84–99, May 1997.

[26] R. Caceres and V. N. Padmanabhan, "Fast and scalable handoffs for wireless internetworks," in *Proc. ACM Mobicom 96*, 1996, pp. 56–66.

[27] C. Castelluccia, "Extending mobile IP with adaptive individual paging: a performance analysis," in *Proc. IEEE Symposium on Computer and Communications*, 2000, pp. 113–118.

[28] H. Omar, T. Saadawi, and M. Lee, "Supporting reduced location management overhead and fault tolerance in mobile-IP systems," in *Proc. IEEE Symp. Computer and Comm*, 1999, pp. 347–353.

[29] H. Xie, S. Tabbane, and D. J. Goodman, "Dynamic location area management and performance analysis," in *Proc. 43rd IEEE Vehicular Technology Conference,*, 1993, pp. 536–539.

[30] W. S. Wong and C. M. Leung, "An adaptive distance-based location update algorithm for next-generation PCS networks," *IEEE J. Selected Areas in Comm. (JSAC)*, vol. 19, no. 10, pp. 1942–1952, Oct 2001.

[31] I. F. Akyildiz and W. Wang, "A dynamic location management scheme for next-generation multitier PCS systems," *IEEE Trans. Wireless Communications*, vol. 1, no. 1, pp. 178–189, January 2002.

[32] I. F. Akyildiz, Y.-B. Lin, W.-R. Lai, and R.-J. Chen, "A new random walk model for PCS networks," *IEEE Journal on Selected Areas in Communications (JSAC) Wireless Series*, vol. 18, no. 7, July 2000.

[33] J. S. Ho and I. F. Akyildiz, "Mobile user location update and paging under delay constraints," *ACM Journal of Wireless Networks (WINET)*, vol. 1, no. 4, pp. 413–425, December 1995.

[34] J. Xie and I. F. Akyildiz, "An optimal location management scheme for minimizing signaling cost in mobile ip," in *Proc. IEEE International Conference on Communications (ICC 2002)*, April 2002, pp. 3313–3317.

[35] J. Xie and I. Akyildiz, "A distributed dynamic regional location management scheme for mobile IP," in *Proc. IEEE INFOCOM 2002*, vol. 2, 2002, pp. 1069–1078.

[36] J. Xie and I. F. Akyildiz, "A novel distributed dynamic location management scheme for minimizing signaling costs in Mobile IP," *IEEE Transactions on Mobile Computing*, vol. 1, no. 3, pp. 163–176, July-September 2002.

[37] Y. Wang, W. Chen, and J. S. Ho, "Performance analysis of mobile IP extended with routing agentst," Southern Methodist University, Tech. Rep. 97-CSE-13, 1997.

[38] A. Bar-Noy, I. Kessler, and M. Sidi, "Mobile users: To update or not to update?" *ACM Journal of Wireless Networks (WINET)*, vol. 1, no. 2, pp. 175–185, July 1995.

[39] I. F. Akyildiz and J. S. M. Ho, "On location management for personal communications networks," *IEEE Communication Magazine*, pp. 138–145, September 1996.

[40] B. Lampson, V. Srinivasan, and G. Varghese, "IP lookups using multiway and multicolumn search," *IEEE/ACM Transactions on Networking*, vol. 7, no. 3, pp. 324–334, June 1999.

[41] H.-Y. Tzeng and T. Przygienda, "On fast address-lookup algorithms," *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 17, no. 6, pp. 1067–1082, June 1999.

[42] W. R. Stevens, *TCP/IP Illustrated, Volume 1: The Protocols.* Addison Wesley Longman, Inc., 1994.

[43] J. S. M. Ho and I. F. Akyildiz, "Local anchor scheme for reducing signaling costs in personal communications networks," *IEEE/ACM Trans. Networking*, vol. 4, no. 5, pp. 709–725, October 1996.

[44] ——, "Dynamic hierarchical database architecture for location management in PCS networks," *IEEE/ACM Trans. Networking*, vol. 5, no. 5, pp. 646–660, October 1997.

[45] A. Abutaleb and V. O. K. Li, "Paging strategy optimization in personal communication systems," *ACM Journal of Wireless Networks (WINET)*, vol. 3, pp. 195–204, August 1997.

[46] I. F. Akyildiz, J. S. M. Ho, and Y.-B. Lin, "Movement-based location update and selective paging for PCS networks," *IEEE/ACM Trans. Networking*, vol. 4, no. 4, pp. 629–638, August 1996.

[47] C. Rose and R. Yates, "Minimizing the average cost of paging under delay constraints," *ACM Journal of Wireless Networks (WINET)*, vol. 1, pp. 211–219, February 1995.

[48] A. Abutaleb and V. O. K. Li, "Location update optimization in personal communication systems," *ACM Journal of Wireless Networks (WINET)*, vol. 3, pp. 205–216, August 1997.

[49] W. Wang and I. F. Akyildiz, "A new signaling protocol for intersystem roaming in next-generation wireless systems,," *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 19, no. 10, pp. 2040–2052, October 2001.

[50] A. Misra and *et al.*, "IDMP: An intra-domain mobility management protocol using mobility agents (work in progress)," Internet Engineering Task Force, draft-mobileip-misra-idmp-00.txt, July 2000.

[51] J. Xie, "User independent paging scheme for Mobile IP," *to appear in ACM Wireless Networks (WINET)*, 2004.

[52] K. van der Wal, M. Mandjes, and H. Bastiaansen, "Delay performance analysis of the new internet services with guaranteed QoS," *Proceedings of the IEEE*, vol. 85, no. 12, pp. 1947–1957, December 1997.

[53] T. Liu, P. Bahl, and I. Chlamtac, "Mobility modeling, location tracking, and trajectory prediction in wireless ATM networks," *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 16, no. 6, pp. 922–936, August 1998.

[54] G. P. Pollini and C.-L. I, "A profile-based location strategy and its performance," *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 15, no. 8, pp. 1415–1424, October 1997.

[55] S. Tabbane, "Location management methods for third-generation mobile systems," *IEEE Communications Magazine*, pp. 72–84, August 1997.

[56] E. Cayirci and I. F. Akyildiz, "User mobility pattern scheme for location update and paging in wireless systems," *IEEE Trans. Mobile Computing*, vol. 1, no. 3, pp. 236–247, July-September 2002.

[57] W. Wang and I. F. Akyildiz, "On the estimation of user mobility pattern for location tracking in wireless networks," in *Proc. IEEE GlOBECOM 2002*, vol. 1, 2002, pp. 610–614.

[58] A. Papoulis, *Probability, Random Variables, and Stochastic Processes.* New York: McGraw-Hill, 1965.

[59] T. Simunic, L. Benini, P. Glynn, and G. D. Micheli, "Dynamic power management for portable systems," in *Proc. ACM Mobicom 2000*, 2000, pp. 11–19.

[60] S.-U. Yoon, J.-H. Lee, K.-S. Lee, and C.-H. Kang, "QoS support in mobile/wireless IP networks using differentiated services and fast handoffmethod," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC 2000)*, vol. 1, 2000, pp. 266–270.

[61] A. Terzis, M. Srivastava, and L. Zhang, "A simple QoS signaling protocol for mobile hosts in the integrated services Internet," in *Proc. IEEE INFOCOM 99*, vol. 3, 1999, pp. 1011–1018.

[62] G. L. Grand, J. Ben-Othman, and E. Horlait, "Providing quality of service in mobile environments with MIR (Mobile IP Reservation Protocol)," in *Proc. IEEE International Conference on Networks (ICON2000)*, 2000, pp. 24–29.

[63] C.-C. Tseng, G.-C. Lee, and R.-S. Liu, "HMRSVP: A hierarchical mobile RSVP protocol," in *Proc. 2001 International Conference on Distributed Computing Systems Workshop*, 2001, pp. 467–472.

[64] K.-I. Kim and S.-H. Kim, "Domain based approach for QoS provisioning in Mobile IP," in *Proc. IEEE GLOBECOM 2001*, vol. 4, 2001, pp. 2230–2234.

[65] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation protocol (RSVP), version 1 functional specification," Request for Comments (RFC) 2205, September 1997.

[66] A. K. Talukdar, B. R. Badrinath, and A. Acharya, "Integrated services packet networks with mobile hosts: architecture and performance," *ACM Wireless Networks (WINET)*, vol. 5, no. 2, pp. 111–124, 1999.

[67] J. Xie, "Paging-aided connection setup for real-time communication in Mobile Internet," in *Proc. IEEE International Conference on Communications (ICC 2003)*, 2003, pp. 1858–1862.

[68] C. Morris and J. Nelson, "Mobility support in evolving third generation mobile systems," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM'98)*, vol. 5, 1998, pp. 2580–2585.

[69] R. Beaubrun, S. Pierre, P. Flocchini, and J. Conan, "Global roaming management in next-generation wireless systems," in *Proc. IEEE International Conference on Communications (ICC 2002)*, vol. 4, 2002, pp. 2070–2074.

[70] Y.-B. Lin and I. Chlamtac, "Heterogeneous personal communications services: integration of PCS systems," *IEEE Communication Magazine*, pp. 106–113, September 1996.

[71] T. B. Zahariadis, K. G. Vaxevanakis, C. P. Tsantilas, N. A. Zervos, and N. A. Nikolaou, "Global roaming in next-generation networks," *IEEE Communication Magazine*, pp. 145–151, February 2002.

[72] A. Bertrand, "Jambala mobility gateway — convergence and intersystem roaming," in *Proc. IEEE INFOCOM 99*, no. 2, 1999, pp. 89–93.

[73] Y. Black, *Second generation mobile and wireless networks.* Prentice Hall PTR, New Jersey, 1998.

[74] F. D. Priscoli, "Interworking of a satellite system for mobile multimedia applications with the terrestrial networks," *IEEE Journal on Selected Areas in Communicaions (JSAC)*, vol. 17, no. 2, pp. 385–394, February 1999.

[75] M. Buddhikot and *et al.*, "Integration of 802.11 and third-generation wireless data networks," in *Proc. IEEE INFOCOM 2003*, 2003.

[76] Q. Tian and D. C. Cox, "Location management in a heterogeneous network environment," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC 2000)*, vol. 2, 2000, pp. 753–758.

[77] K. I. Park and Y.-B. Lin, "Reducing registration traffic for multitier personal communications services," *IEEE Trans. Vehicular Technology*, vol. 46, no. 3, pp. 597–602, August 1997.

[78] L. F. Chang, A. R. Noerpel, and I. Park, *Private communications*, 1995.

[79] Y.-B. Lin, L. F. Chang, and A. R. Noerpel, "Performance modeling of multi-tier PCS system," *International Journal of Wireless Information Networks*, vol. 3, pp. 67–78, 1996.

[80] Y.-B. Lin, "A comparison study of the two-tier and the single-tier personal communications services systems," *ACM Mobile Networks and Applications (MONET)*, vol. 1, pp. 29–38, 1996.

[81] S. Mohanty, J. Xie, and I. F. Akyildiz, "AMC: An architecture for mobile computing in next generation heterogeneous wireless systems," *submitted to ACM MobiQuitous*, 2004.

[82] J. Xie and I. F. Akyildiz, "Location management in next generation heterogeneous overlay networks," *submitted to IEEE Trans. Mobile Computing*, 2004.

[83] A.-C. Pang, Y.-B. Lin, and Y. Fang, "Implicit deregistration with forced registration for PCS mobility management," *ACM Wireless Networks (WINET)*, vol. 7, no. 1, pp. 99–104, 2001.

[84] N. Efthymiou, Y. F. Hu, and R. Sheriff, "Performance of intersegment handover protocols in an integrated space/terrestrial-UMTS environment," *IEEE Trans. Vehicular Technology*, vol. 47, no. 4, pp. 1179–1199, November 1998.

[85] K. Ushiki and M. Fukazawa, "A new handover method for next generation mobile communication systems," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM 98)*, 1998, pp. 1118–1123.

[86] J. McNair, I. F. Akyildiz, and M. D. Bender, "An inter-system handoff technique for the IMT-2000 system," in *Proc. IEEE INFOCOM 2000*, vol. 1, 2000, pp. 208–216.

[87] M. E. Kounavis, A. T. Campbell, G. Ito, and G. Bianchi, "Design, implementation and evaluation of programmable handoff in mobile networks," *ACM Mobile Networks and Applications (MONET)*, vol. 6, pp. 443–461, 2001.

[88] ETSI, "Requirements and architectures for interworking between HIPERLAN/3 and 3rd generation cellular systems," ETSI, Tech. Rep. ETSI TR 101 957, 2001.

[89] H. Honkasalo and *et al.*, "WCDMA and WLAN for 3G and beyond," *IEEE Wireless Communications*, pp. 14–18, April 2002.

[90] V. K. Varma, S. Ramesh, and *et al.*, "Mobility management in integrated 3G/WLAN networks," in *Proc. IEEE International Conference on Communications (ICC 2003)*, May 2003.

[91] S. Tsao and C. Lin, "VGSN : A gateway approach to interconnect UMTS/WLAN networks," in *Proceedings of IEEE PIMRC 2002*, September 2002.

[92] D. Wisely and E. Mitjana, "Paving the road to systems beyond 3G — the IST BRAIN and MIND projects," *Journal of Communications and Networks*, vol. 4, no. 4, pp. 292–301, December 2002.

[93] A. Jamalipour and S. Tekinay, "Fourth generation wireless networks and interconnecting," *IEEE Personal Communications*, vol. 8, no. 5, October 2001.

[94] A. T. Campbell, J. Gomez, S. Kim, C.-Y. Wan, Z. R. Turanyi, and A. G. Valko, "Comparison of IP micromobility protocols," *IEEE Wireless Communications*, pp. 72–82, February 2002.

VITA

Jiang (Linda) Xie was born in Beijing, P. R. China in January 1974. She received the Bachelor of Engineering degree from Tsinghua University, Beijing, China, in 1997 and the Master of Philosophy degree from the Hong Kong University of Science and Technology in 1999, both in Electrical Engineering. From September 1997 to January 1999, she was a Teaching Assistant in the Department of Electrical and Electronic Engineering at the Hong Kong University of Science and Technology. From September 1999 to April 2004, she was a research assistant in the Broadband and Wireless Networking Laboratory (BWN-LAB) at the Georgia Institute of Technology. She received the Master of Science degree and the Doctor of Philosophy degree in May 2002 and May 2004, respectively, in Electrical and Computer Engineering from the Georgia Institute of Technology.