Design and Analysis of Algorithms for Efficient Location and Service Management in Mobile Wireless Systems

Baoshan Gu

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Committee:
Dr. Ing-Ray Chen, Chairman
Dr. Luiz A. DaSilva
Dr. Denis Gracanin
Dr. Chang-Tien Lu
Dr. Scott F. Midkiff

Date: September 30, 2005
Falls Church, Virginia

Keywords: Location management, service management, integrated location and service management, service handoff, mobile wireless networks, performance evaluation

Design and Analysis of Algorithms for Efficient Location and Service Management in Mobile Wireless Systems

Baoshan Gu

ABSTRACT

Mobile wireless environments present new challenges to the design and validation of system supports for facilitating development of mobile applications. This dissertation concerns two major system-support mechanisms in mobile wireless networks, namely, location management and service management. We address this research issue by considering three topics: location management, service management, and integrated location and service management.

A location management scheme must effectively and efficiently handle both user location-update and location-search operations. We first quantitatively analyze a class of location management algorithms and identify conditions under which one algorithm may perform better than others. From insight gained from the quantitative analysis, we design and analyze a hybrid replication with forwarding algorithm that outperforms individual algorithms and show that such a hybrid algorithm can be uniformly applied to mobile users with distinct call and mobility characteristics to simplify the system design without sacrificing performance.

For service management, we explore the notion of *location-aware* personal proxies that cooperate with the underlying location management system with the goal to minimize the network communication cost caused by service management operations. We show that for cellular wireless networks that provide packet services, when given a set of model parameters characterizing the network and workload conditions, there exists an optimal proxy service area size such that the overall network communication cost for service operations is minimized. These proxy-based mobile service management schemes are shown to outperform non-proxy-based schemes over a wide range of identified conditions.

We investigate a class of integrated location and service management schemes by which service proxies are tightly integrated with location databases to further reduce the overall network

signaling and communication cost. We show analytically and by simulation that when given a user's mobility and service characteristics, there exists an optimal integrated location and service management scheme that would minimize the overall network communication cost for servicing location and service operations. We demonstrate that the best integrated location and service scheme identified always performs better than the best decoupled scheme that considers location and service managements separately.

ACKNOWLEDGEMENTS

First of all, I'd like to thank my advisor Dr. Ing-Ray Chen, for his continuous guidance and support. He was always there to listen and to give advice. His expertise in mobile wireless computing and performance evaluation improved my research skills and prepared me for future challenges. Thanks also go to my other committee members, Dr. Scott Midkiff, Dr. Chang-Tien Lu, Dr. Luiz DaSilva and Dr. Denis Gracanin. Their friendly comments and constructive criticism on my work as well as insightful ideas for the research subjects were invaluable.

I want to thank Dr. Bouguettaya for his help during my Ph.D. study. Thanks are also due to Ms. Marija Telbis-Forster, Ms. Diane Wang and Ms. Charmaine Carter for their general support. I also thank all the graduate students in northern Virginia center. Thank you for all the good time and all the help you've given me over years.

Last, but not least, I'd like to thank my parents Changen Gu and Xiaohua Zhang, my wife Daqi and my son Han, all of whom patiently stood by me throughout my Ph.D. study.

Table of Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Over the last few years, we have seen rapid progress in wireless and mobile network communication. The evolution of wireless and mobile network technologies has enabled the development of ubiquitous personal communications services (PCS), providing mobile users with voice, data and multimedia personalized services at any time, any place, and in any format [35].

Mobile wireless environments also present several challenges to provide these services. This dissertation concerns two major system-support mechanisms in mobile wireless networks, namely, location management and service management.

Location management addresses the issues of how to track and locate a mobile user efficiently. Service management addresses the issues of how to efficiently deliver services to mobile user through limited wired and wireless network resources. This dissertation aims to design and analyze location and service management schemes that are efficient for cellular personal communication service (PCS) systems. We propose to address this research issue by considering three topics: location management, service management, and integrated location and service management.

1.1 Mobile Wireless Environment

Mobile wireless environments present several specific challenges. First of all, the location of a mobile host (MH) is not static. An MH's point of attachment to the fixed network changes as the mobile user moves. A central issue in a mobile wireless system is locating mobile hosts. There are two primitive operations [46], namely, a location update operation to update the new location of a user when the user moves to a new location, and a lookup (or search) operation to find the current location of a user when a call or a data packet to that user is to be delivered. A location management scheme must handle both operations effectively. One consequence of mobility is heterogeneity. The connectivity an MH experiences may be highly variant due to high variability in bandwidth and reliability in wireless networks. For instance, an MH may move from low bandwidth Global System for Mobiles/General Packet Radio Services (GSM/GPRS) to high speed wireless local area networks (LANs), or vice versa. Even in the same network, due to the

change of network load in a base station, the availability of bandwidth and resources may change dramatically. Moreover, there may be areas with inadequate coverage, thus resulting in disconnections as the mobile user moves.

Secondly, wireless networks are expensive, offer less bandwidth, and are less reliable than wired networks in general. Consequently, bandwidth conservative algorithms are needed, taking into consideration of the weak and intermittent connections.

Lastly, mobile hosts are considered resource-poor when compared to static hosts. In general, these devices have limited memory, limited computational power, limited battery life, and small display size. These limitations will inevitably affect the design of service applications and communication protocols.

1.2 Research Statement and Contribution

This dissertation aims to design location and service management schemes that are efficient for PCS cellular systems. We propose to address this research issue by considering three topics: location management, service management, and integrated location and service management. For the first topic, we propose to quantitatively analyze existing location management algorithms, identify conditions under which one algorithm can perform better than all others, and then design a hybrid scheme that can perform better than all individual algorithms and can be uniformly applied to all users with different mobility and call characteristics. For service management, we explore the notion of *location-aware* personal proxies. For the last topic, we propose to co-locate location databases and personal proxies to minimize the overall network signaling and communication cost caused by location and service management operations.

1.2.1 Location Management

In mobile wireless systems, one of the most important issues is locating MHs. In cellular systems, a well known basic and simple scheme is to update the location of each MH at its home location register (HLR) whenever it moves to a new visitor location register (VLR) area. This location management scheme exists in IS-41 [16] in the United States and GSM [39] in Europe, commonly known as the basic HLR/VLR two-tier scheme.

The basic HLR/VLR two-tier scheme works well for a relatively small number of mobile users. For future PCS networks, the population of MHs will increase dramatically. Also many new services like multimedia applications are emerging. In order to meet the quality of service requirements, the registration area size is expected to reduce for saving the power of transmission and greater frequency reuse [62]. However, smaller registration areas adversely cause more boundary crossings by MH, which in turn leads to a higher cost to location management.

An important research issue for location management is minimizing the network signaling cost associated with location update and search operations under location management strategies as these operations need to be performed frequently. For voice communication in a cellular network, search operations are related to how often an MH is called, i.e., the call arrival rate, while update operations are related to the MH's mobility rate. Thus, in general, the call to mobility ratio (*CMR*) parameter, defined as the ratio of an MH's call arrival rate to the MH's mobility rate, captures the MH's call and mobility patterns.

In recent years, various location management strategies have been proposed in the literature. It has been separately reported that when call to mobility ratio (*CMR*) is high, the location caching/replication scheme [25][47][55] is effective, while when *CMR* is low, the forwarding and resetting algorithm (FRA) [9][26], the paging and location update algorithm (PLA) [2] and the local anchor algorithm (LAA) [23] are effective. Thus, under the notion of per-user-based location management, the best algorithm among all can be selected for execution by the system based on the user's *CMR* value. In addition to lacking a comprehensive comparative study of existing algorithms to identify the best algorithm when given a *CMR* value in a system environment, this also introduces undesirably high complexity in managing and maintaining location-related information stored in the system as different algorithms may be applied to different users. The correlations among the locations of mobile users are explored in [19] to reduce the paging delay and increase the throughput.

We are motivated to develop *generic* analytical models to analyze and compare various location management schemes. For two-tier architecture based HLR-VLR algorithms, the dissertation research investigates a two-level hierarchical performance model as a uniform framework for assessing and comparing the performance characteristics of these algorithms. At the top level comes a model that calculates the total cost incurred to the PCS network as a result of location-update and call delivery operations during the period between two consecutive calls. At the low

level model comes a stochastic model that estimates the values of high-level model parameters. We show that by utilizing simple Markov models at the low level, we can easily assess and compare the performance characteristics of two-tier based location management algorithms.

The dissertation research then proposes and analyzes a hybrid algorithm that can be applied to all users with different *CMR* values without sacrificing the optimality of individual algorithms. Specially, the hybrid scheme combines "replication" and "forwarding" techniques. Replication is known to be most effective in reducing user search and update costs when *CMR* is high. The reason is that when the call arrival rate to the user is much higher than the mobility rate, the communication cost would be dominated by search operations. By replicating the user's location information at selected VLRs, the cost of querying the HLR (the only copy when replication is not used) is avoided and, thus, the search cost is reduced. Conversely, forwarding is known to be most effective when *CMR* is low. When the mobility rate is much higher than the call arrival rate, the communication cost would be dominated by location update operations. The update cost can be reduced by simply forming a forwarding chain of VLRs to locate the MH from the HLR, thus avoiding updating the HLR upon every update operation. The HLR is updated only periodically when the forwarding chain becomes too long (say after K moves). Our hybrid scheme takes advantages of the benefits of replication and forwarding techniques at high and low *CMR* values, respectively.

In addition to having the advantage of uniformity, this thesis shows that the hybrid scheme performs better than not only either scheme under all CMR values, but also a binary "CMR threshold-based" scheme that applies the replication technique when a mobile user's CMR is higher than a threshold and applies the forwarding technique otherwise. For this binary CMR threshold-based scheme, we identify the optimal threshold value and show the hybrid scheme outperforms the threshold-based scheme under optimal threshold values.

1.2.2 Service Management

Though the cellular systems have traditionally provided wireless access for voice communication, it is expected that there will be a surge in demand for wireless data services. The Yankee Group projected that the number of wireless data users will soar from 22.3 million in 2002 to some 96 million by 2006 [63]. Mobile data communication will be pervasive in cellular systems, such as 3G, and in wireless LANs, such as 802.11. It is becoming clear that the existing data services will

be accessible to the wireless mobile users. Such services include personalized news and financial, sports information, banking, sales inventory, travel information and the like. At the same time, new advanced services that provide personalized, context-aware and multimedia services will be emerging. These emerging services will impose new requirements on the underlying mobile wireless communication networks.

The characteristics of mobile wireless environments, combined with emerging location-aware and context-aware services for mobile users, demand personalized services. For example, an MH who subscribes to a wireless service provider may provide user preferences for the type of news it would like to receive. The service provider then provides information based on the user's interest, and nothing else. It is quite different from our current desktop web-based model where a user can "surf the web" until an interest area is found. The traditional client/server model is not sufficient in a mobile wireless environment. Extended client/server models place proxies (agents or gateways) between the client and server to reduce the communication overhead by performing data service functions, such as filtering data contents, caching, and tracking client locations [21]. In 3GPP, Virtual Home Environment (VHE) [68] is introduced as a concept for personal service environment portability across network boundaries and between terminals. The VHE consistently presents users with the same personalized feature, user interface customization and services no matter what network, what terminal (within the capabilities of the terminal and network), and where the user may be located. In the Wireless Application Protocol (WAP) 1.0 [55][61] architecture, a proxy/gateway is a required component to handle the protocol conversion between a mobile host and an origin server. In WAP 2.0 architecture, the proxy is needed to provide feature/performance-enhancing services, such as pushing and context-aware services.

This research investigates the use of a personal proxy-based architecture for service managements. Our objective is to design service management schemes that can take advantage of the underlying location management schemes such that the overall network signaling and communication cost for service management may be minimized. We consider two possible designs. One design is to use a personal "mobile" service proxy that tracks the location of the mobile user such that the proxy can determine when and how often it should move with the user as the mobile user moves from one location area to another. Another design, which we will discuss later in more detail, is to use a personal mobile proxy that is tightly coupled with the per-user location management scheme, i.e., the service proxy is co-located with the location database in an integrated fashion.

The personal “mobile” service proxy in our proposed service management schemes resides in the wired network on behalf of a mobile wireless client. All messages exchanged between the client and the server will go through the personal service proxy. The personal proxy performs tasks such as tracking the location of the user, accepting communication messages on the user's behalf, converting the communication data into different application formats, and forwarding communication data to the mobile user. Since all communications to the mobile user must go through the personal proxy, inefficient routes will result in increased communication costs, so we like to place the proxy to be as close to the MH as possible. However, moving the proxy frequently with the MH as the MH moves from one location to another incurs a service context migration cost as well as a reconnection cost. An important design consideration is *when* and *how often* one should move the personal proxy with the mobile user in order to minimize the overall network signaling and communication cost for service management.

When a service proxy moves, a “service handoff” occurs. The service handoff considered in this research refers to the process of migrating the MH’s personal proxy as a user moves into another “service area” (whose optimal sizes are to be studied in the dissertation) so as to move the proxy closer to the mobile. The migrating process involves a cost of transferring the service context from one service area to another, so the proxy can continue with the service, and a cost of reestablishing the connection between the proxy and the application server. The proxy holds the service context information, which must be migrated with the proxy during a service handoff. The context information can include both static context information such as user profile, device profile, and authentication data, as well as dynamic context information such as files opened, objects updated, locks and timestamps, etc. Since the service handoff cost is due to the migration of the proxy as the MH moves, it is considered as part of the proxy maintenance overhead. There is a tradeoff between minimizing the service handoff overhead (e.g., by moving the proxy with the MH less often) versus minimizing the service management overhead (by moving the proxy with the MH more often and thus placing the proxy closer to the MH). This dissertation aims to explore this tradeoff with the goal to identify conditions under which the proxy can operate optimally to minimize the network signaling and communication cost for mobile services.

To provide personalized services and further reduce network traffic, this dissertation also investigates the notion of per-service personal proxy. That is, a proxy is created for each client-server application running on the MH. The advantage of using distinct per-service proxies instead

of a single proxy to interface with multiple server applications is that a per-service proxy is application specific, thus fully understanding the message exchange protocols between the client and server. This dissertation will analyze performance characteristics of service management due to the use of per-service proxies.

1.2.3 Integrated Location and Service Management

As location and service management operations are two major sources of communication cost in mobile wireless systems, this dissertation proposes proxy-based *integrated* location and service management schemes to further reduce the network signaling and communication cost.

Jain and Krishnakumar [24] suggested that location and service handoffs should be integrated to reduce the overall communication cost, but no follow-up analysis or research has been done to investigate the potential benefit of the integrated approach. Their notion of service handoff is based on the assumption of fully replicated servers in service areas, such that whenever an MH crosses a service area, its ongoing service can be handed-off from one server to another by means of service context information transfer to allow the service to be continued at the new server. An example would be a video on-demand application with replicated servers such that the context information could include the video title, minutes played, and the current frames being buffered at the server and played at the MH. While it is possible to have integrated location and service management for these applications, the difficulty of integrating location handoffs (due to movements of the MH crossing VLR boundaries) with service handoffs (due to movement of the MH crossing service area boundaries) in the PCS network lies in the very large scale deployment of a large number of replicated servers in VLRs.

We investigate integrated location and service management for minimizing network cost without making the assumption of fully replicated servers in VLRs in the PCS network. Under our design of integrated location and service management, a per-user service proxy[1] is used as a gateway between the MH and all client-server applications engaged by the MH concurrently. The proxy keeps track of service context information such as the current state of the execution for maintaining service continuity. All user requests and server replies would pass through the proxy.

[1] A single per-user proxy is created to interface with all server applications in integrated location and service management because the proxy will co-locate with location database, making per-service personal proxies unfeasible.

Furthermore, the proxy is co-located with the user's location database all the time, so whenever there is a move of the user's location database, the proxy also moves so as to co-locate with the location database. The proxy thus knows the location of the MH all the time. The server application does not need to be informed of the location change of the MH but only needs to be informed of the location change of the proxy, thus effectively reducing the cost of location and service management.

Four integrated location and service management schemes are proposed and investigated in the dissertation with the goal to identify conditions under which a particular scheme should be adopted by an MH based on its mobility and service characteristics for network cost minimization. These four schemes are derived from the basic HLR/VLR and LA schemes for location management, and the personal service proxy scheme for service management in the PCS network. We are motivated to investigate and identify the best integrated location and service management scheme that can be applied on an individual user basis to minimize the overall cost incurred to the PCS network per time unit for servicing location and service operations of all users.

Our contribution is that we propose and analyze new integrated location and service management schemes not considered before and show that integrated location and service management is a viable concept applicable to the PCS network on a per-user basis for general server applications. We show that, when given an MH's mobility and service characteristics through a set of parameters identified in the dissertation, there exists an optimal integrated location and service management scheme that would minimize the overall network communication cost as a result of executing the MH's location and service operations. We also demonstrate that the best integrated location and service scheme identified always performs better than the best decoupled scheme that considers location and service managements separately.

1.3 Thesis Organization

The rest of the thesis is organized as follows. In Chapter 2 we present the system model for location and service management in wireless networks. In Chapters 3 and 4, we present our research results on location and service management, respectively. In Chapter 5, we investigate integrated location and service management schemes with the goal to provide personalized

services while minimizing the *overall* communication cost for servicing both location and service management operations. Chapter 6 summarizes the dissertation and describes future work.

Chapter 2

SYSTEM MODEL

Our system model is based on cellular PCS networks. A cellular PCS network can be modeled as hierarchical or flat, depending on the size of the network and the structure adopted by the network service provider. Figure 2-1 shows a hierarchical PCS network architecture in a two-tier HLR-VLR structure as in IS-41 [16] and GSM [39]. A Registration Area (RA) can cover a single Base Station (BS) or a group of base stations. A mobile switching center (MSC) is used to connect all the cells in one RA. A two level hierarchical database are used: the Home Location Register (HLR) and the Visitor Location Register (VLR) [1]. An HLR is responsible for keeping track of a mobile host's (MH) current location as well as its profile such as services subscribed. Each MH is permanently associated with an HLR. Conceptually, the HLR of an MH is at a higher level, while VLRs that the MH wanders into from time to time are at the lower level. Each VLR stores the information downloaded from the HLR and the location information of the MHs visiting its service area. There may be some network switches connecting the HLR to VLRs in the mobile network. Each VLR or HLR is connected to the rest of the signaling network through a local signal transfer point (LSTP); one or more LSTPs belonging to one region may be connected to a regional signal transfer Point (RSTP). Separate RSTPs may be connected by a public switched telephone network (PSTN). We assume that each VLR corresponds to one RA. When a mobile user moves to a new RA, the mobile user sends the registration information to the new VLR, which in turn can perform appropriate update actions, depending on the location management scheme used.

Please note that this is a structure from the point of view of an MH. A VLR can be an HLR for some MHs permanently and at the same time acts as a VLR for other MHs roaming into its RA area temporarily. We assume the average communication cost between a VLR and the HLR is equal to the communication cost between any two randomly placed VLRs, represented by T. The average communication cost between two neighboring VLRs is represented by τ. In general, τ is less than T and their values can be calculated by means of a network coverage model (e.g., hexagonal) characterizing the underlying wireless network as in [9]. The time that a particular MH stays in a VLR before moving to another one is characterized by an exponential distribution with mean value $1/\sigma$. Such a parameter can be estimated using the approach described in [26] on a

per-user basis. The inter-arrival time between two consecutive calls to a particular MH is also assumed to be exponentially distributed with an average rate of λ. Similarly, a service rate γ is introduced for service management. For analysis purposes, the assumption of exponentially distributed times can be relaxed by using Stochastic Petri Net (SPN) tools that support specifications of general time distributions such as SPNP version 6 [59] and TimeNET version 3 [71]. The assumption of the inter-arrival time of calls being exponentially distributed is widely accepted in the literature while the assumption of the residence time being exponentially distributed is questionable. For the latter case, we perform a sensitivity analysis to test the sensitivity of performance evaluation results with respect to different probability distributions for the residence time.

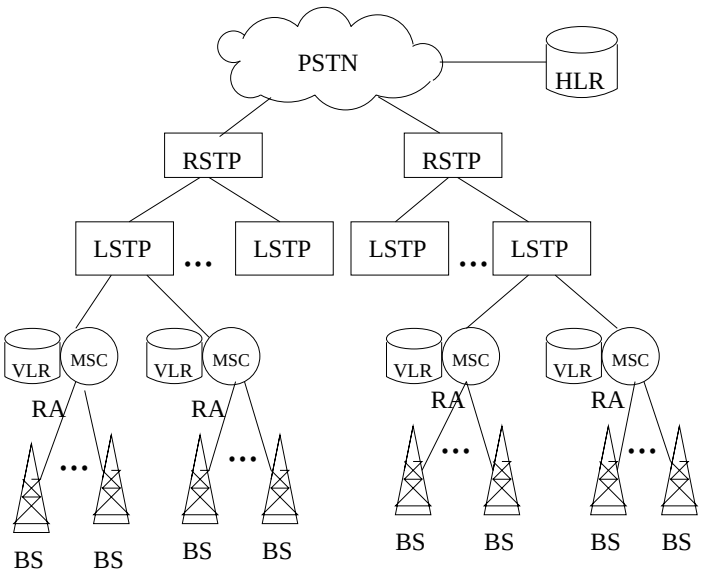

Figure 2-1: Hierarchical PCS Cellular Architecture.

The general model parameters are separated into two classes, i.e. the *per-mobile-user* parameter class, which is inherently associated with a mobile user and the service accessed by a mobile user, and the *network* parameter class, which depends on the mobile network structure.

Table 2-1 and Table 2-2 show these two classes of parameters.

Table 2-1: Per-Mobile-User Parameters.

Symbol	Meaning
σ	the average rate at which the mobile user moves across VLR boundaries
λ	the average rate at which the mobile user is being called
γ	the average rate at which the MH requests services
CMR	λ/σ, the call to mobility ratio of an MH
SMR	γ/σ, the service to mobility ratio of an MH

Table 2-2: Mobile Network Parameters.

Symbol	Meaning
T	the average VLR-HLR single-trip communication cost
τ	the average neighboring VLR-VLR single-trip communication cost or the average VLR-VLR single-trip communication cost under one switch

A PCS cellular network can also be flat as shown in Figure 2-2, in which the PCS system is described by a hexagonal network coverage model (such that a cell has a hexagonal shape) and an n-layer area contains $3n^2-3n+1$ cells. For example, when n=2, an area will contain 7 cells and when n=3, it will contain 19 cells.

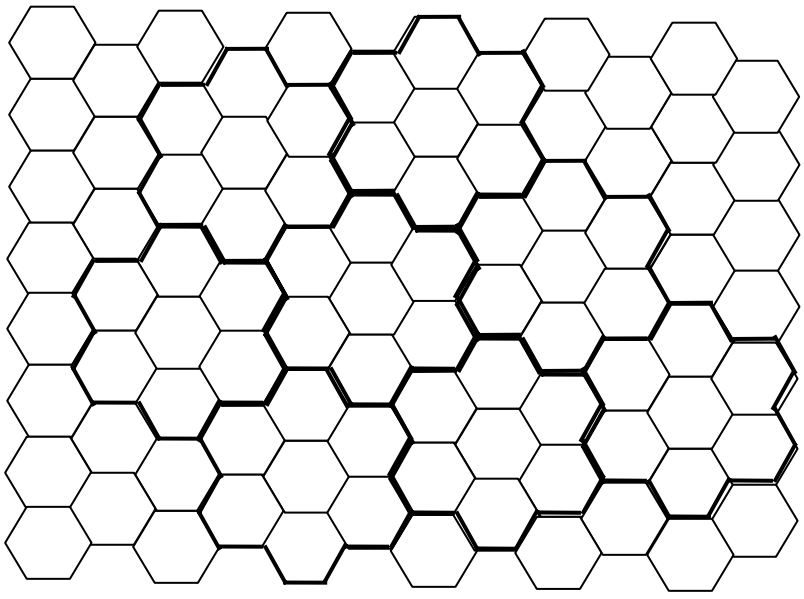

Figure 2-2: Flat PCS Cellular Architecture.

Chapter 3

LOCATION MANAGEMENT

To address location management issues, our strategy is to quantitatively analyze existing location management algorithms to understand conditions under which one algorithm can perform better than others. Then, based on the insight gained, we design hybrid algorithms that can combine the benefits to perform better than individual algorithms. In this chapter, we utilize a two-level hierarchical performance model as a uniform framework for quantitatively assessing and comparing the performance characteristics of a number of existing location management algorithms. We then develop and analyze a hybrid scheme that combines replication and forwarding techniques, known to be effective in reducing user search and update costs, respectively. We show that the hybrid scheme not only can be uniformly applied to all users with different *CMR* ratios, but also outperforms both replication (being most effective when *CMR* is high) and forwarding (being most effective when *CMR* is low) as well as an algorithm that switches between replication and forwarding based on the *CMR* value of the mobile user.

3.1 Location Management Algorithms

In this section, we present some existing location management algorithms. In the next section, analytical models based on hierarchical modeling are developed to assess the performance behaviors of these algorithms.

3.1.1 Basic HLR/VLR

Under the basic HLR/VLR scheme [16], a mobile user is permanently registered under a location register called the home location register (HLR). When the mobile user enters a new VLR area, it reports to the new VLR, which in turn informs the HLR by means of a location update operation. The location update operation under IS-41 scheme proceeds as follows:

- When an MH moves into a new RA, it sends a location update message to the current base station, which forwards this message to the current serving VLR.
- The current serving VLR forwards the message to the MH's HLR.
- The HLR updates the location information of the MH and sends an acknowledgment message together with a copy of the MH's profile to the current serving VLR.

The entries stored in the old VLR can be implicitly removed after a timeout period or explicitly removed via a location cancellation request from the HLR to the old serving VLR.

When there is a call asking for the mobile user, the PCS network checks with the HLR of the mobile user to know the current VLR of the mobile user and then the call is delivered to the current VLR. A call delivery under IS-41 scheme proceeds as follows:

- The calling MH sends a call initiation message through its base station to the currently serving VLR.
- The VLR determinates the associated HLR serving the called MH and sends a location request message to the HLR.
- The HLR determines the callee VLR and sends a route request message to this VLR/MSC.
- The callee VLR sends the route information to the HLR.
- The HLR forwards the route information to the calling VLR. Now the calling VLR can set up a connection to the callee VLR via the SS7 signaling network using the usual call setup protocol.

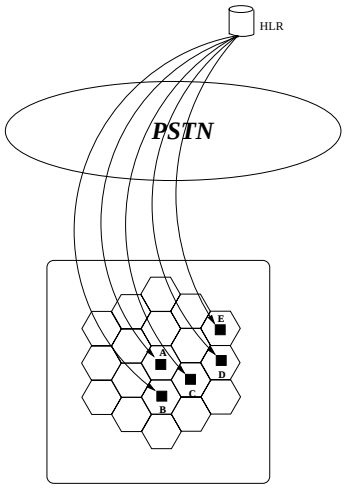

Figure 3-1: Update Operations Performed under IS-41.

In Figure 3-1, when a mobile user moves from VLR A to VLR B, the HLR is informed to point to VLR B. All subsequent moves to C, D and E behave similarly. That is, the HLR is updated to point to C, D and E, respectively, in these subsequent moves.

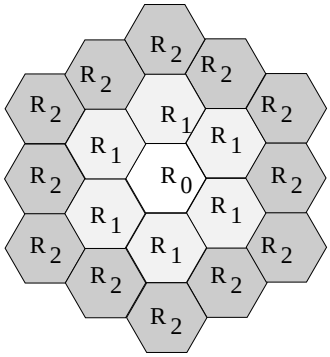

Figure 3-2: Hexagonal Network Coverage Model.

3.1.2 The Local Anchor Algorithm (LAA)

In [23], Ho and Akyildiz proposed a scheme called local anchoring. The basic idea is that location registration operations should be as localized as possible so as to reduce the number of registration messages to the HLR. The VLR that performs the last registration operation with the HLR is called the local anchor (LA) of the mobile user. There is one LA per region where the size of the region is a parameter to be determined. With the hexagonal network coverage model illustrated in Figure 3-2, the number of VLRs covered by the LA in the local region under LAA is $3n^2 - 3n + 1$, where n is a design parameter, e.g., $n=2$ means that 7 VLRs are covered by a LA in the local region and $n=3$ means that 19 VLRs are covered instead. Note that an n-layer region contains layers *0* through n-*1*. When the mobile user crosses a VLR boundary, the following procedure is followed:

- If the MH is still within the same local region, the new VLR only sends a location update message to the LA instead of informing the HLR;
- If the MH makes a regional move, the new VLR sends a location update message to the HLR and it becomes the new LA of the mobile user.

The HLR at all times knows only the address of the LA. A search operation proceeds as follows:

- The calling MH sends a call initiation message through its base station to the currently serving VLR.
- The VLR sends a location request message to the HLR, which determines the callee LA and sends a route request message to this LA.
- The LA forwards the message to the callee VLR which sends the route information to the HLR.

- The HLR forwards the route information to the calling VLR. Now the calling VLR can set up a connection to the callee VLR via the SS7 signaling network using the usual call setup protocol.

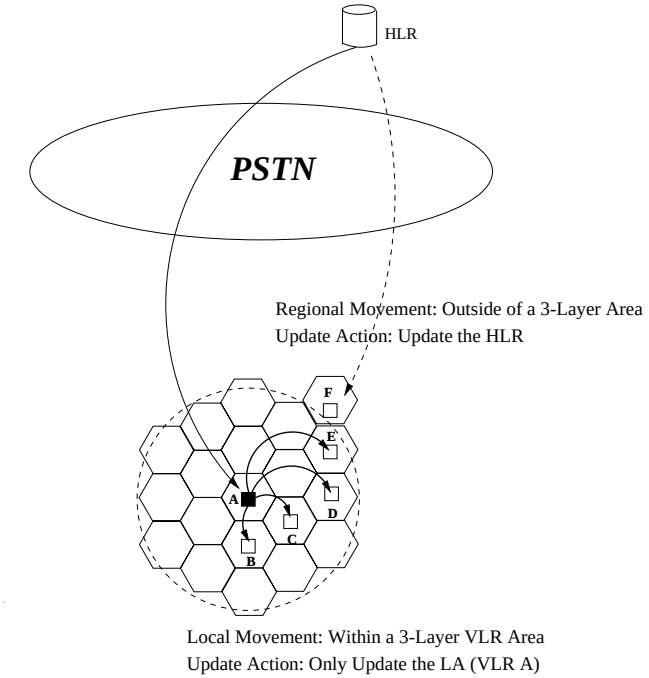

Figure 3-3: Location Updates Performed under LAA.

In Figure 3-3, assume that n=3, at which a network switch covers 19 VLRs, VLR A is the LA initially. When the mobile user makes a local movement from VLR A to VLR B, a location update operation to the LA is performed without updating the HLR. Similarly, when the mobile user subsequently moves to VLR C, the LA's database is updated to point to VLR C. If there is a call looking for the mobile user, the search path starts from the HLR, then VLR A (the LA), and finally to VLR C. The LA is changed only when the mobile user makes a regional movement, in which case a location update operation to the HLR is required. In Figure 3-3, this happens when the mobile user moves to VLR F, which is outside of the 19 VLRs covered by the LA.

3.1.3 Forwarding and Resetting Algorithm (FRA)

A per-user forwarding strategy was proposed in [26] to reduce the total signaling network cost. It was shown that the forwarding strategy could result in 20-60% cost reduction when the user's call-to-mobility ratio is low. In the forwarding strategy, when a user moves to a new RA, it updates the HLR only when the current forwarding chain length reaches a predefined constant K.

Otherwise, a forwarding pointer is set up from the old VLR to the new VLR. The update procedure is described as follows:

- When an MH moves into a new RA, it sends a location update message to the current base station, which then forwards this message to its associated VLR.
- If the current forwarding length is less than the maximum forwarding chain length K, then:
    - The new VLR de-registers the MH at the old VLR, but asks the old VLR to keep a pointer to point to the new VLR.
    - The old VLR sends an acknowledgment to the new VLR.
    - The forwarding length is increased by one.
- Otherwise, the IS-41 basic strategy is followed and the forwarding chain is reset, after which the HLR points to the new VLR directly.

When serving a call to an MH, the HLR is queried first to determine the first VLR at which the MH was registered, and then a chain of forwarding pointers is followed to reach the MH's current VLR. A call delivery under the forwarding strategy proceeds as follows:

- The first VLR is obtained from the callee's HLR as in the IS-41 scheme.
- The callee's current VLR is reached by following the forwarding chain.
- The callee's current MSC/VLR sends user route information to the HLR.
- The HLR forwards route information to the calling MSC. Now the calling MSC can set up a connection to the callee MSC via the SS7 signaling network using the usual call setup protocol.

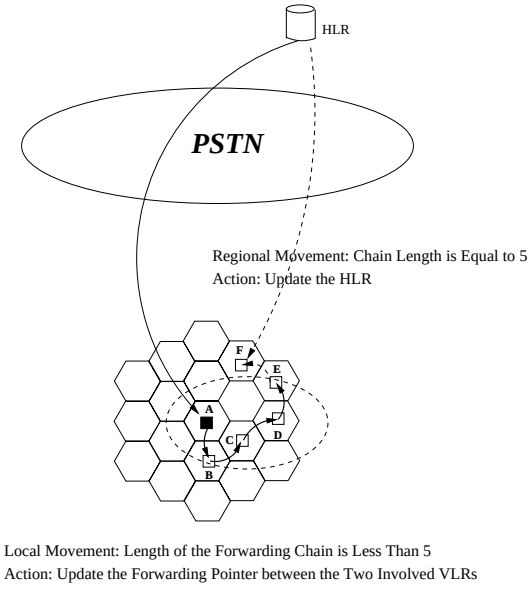

Figure 3-4: Location Updates Performed under FRA.

In Figure 3-4, assume that VLR A is at the beginning of the forwarding chain, thus behaving as the local agent of the mobile user. When the mobile user makes a local movement, the length of the forwarding chain is still smaller than a predefined value *K*, so only forwarding pointers are set up. In Figure 3-4, A-B-C-D-E is the forwarding chain with a length of 4. Suppose *K=5*, then when the mobile user goes from VLR E to VLR F, a regional movement occurs, thus triggering a reset operation to be performed to the HLR, after which VLR F becomes the local agent.

3.1.4 Paging and Location Updating Algorithm (PLA)

PLA (*n*) is a movement-based location update scheme discussed in [2]. Under this scheme, a mobile user performs a location update to the HLR only when the distance between the agent and the current VLR is greater than or equal to a predefined distance value *n*. The VLR that performs the last update operation to the HLR is called the *agent* of the mobile user. When the mobile user makes a movement, the following operations follow:

- If the distance away from the agent is still smaller than the specified distance value *n*, no location update operation to the agent or the HLR is required. Such a movement is called a local movement.
- Otherwise, a location update message is sent to the HLR and the new VLR becomes the agent. A movement that causes the HLR to be updated is called a regional movement.

When a call arrives, the search operation proceeds as follows:

- On receiving the location request message from the caller VLR, the HLR sends a route request message to the callee agent.
- The callee agent searches for the mobile user using an outward paging method, i.e., starting from layer 0 (the agent itself), layer 1 and so on until the mobile user is found.
- The callee's current MSC/VLR sends user route information to the HLR.
- The HLR forwards route information to the calling MSC. Now the calling MSC can set up a connection to the callee MSC via the SS7 signaling network using the usual call setup protocol.

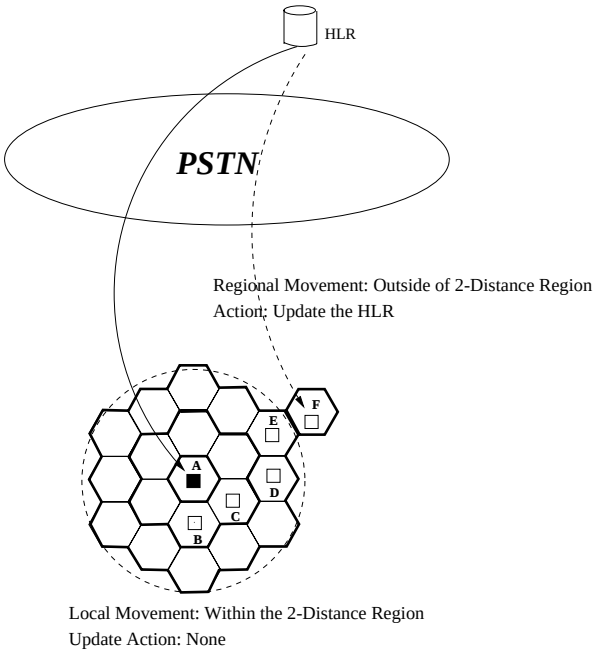

Figure 3-5: Update Operations Performed under PLA.

In Figure 3-5, assume that $n=3$, VLR A is the agent initially, and the shortest-distance-first (SDF) partitioning scheme is used to divide the registration areas into layered areas such that VLR A is in layer 0, VLRs B and C are in layer 1, and VLRs D and E are in layer 2. When the mobile user makes a local movement with a distance away from A less than n=3, e.g. to B, C, D or E, no location update operation is performed and the agent remains unchanged. However, when the mobile user makes a regional move, e.g., to VLR F, the distance from A (the agent) to F now is equal to n=3. An update operation must be performed to the HLR in this case, after which VLR F becomes the new agent. Note the basic difference between LAA and PLA: in LAA, an update operation is performed to the local anchor upon a local movement. In PLA, a local movement does not trigger any location update operation to the agent and thus does not incur any update cost at all.

3.1.5 Replication Algorithm

It has been observed that even for the case when the number of possible communication areas is very large, the set of communication areas from which calls are made to an MH is relatively static and confined [48][55]. For example, a study on e-mail traffic patterns [48] shows that about 80% of mails are from 3 most frequently calling communication sites. If we place location replicas in these 3 sites, we can reduce the search cost by 80%. Another study tracing actual calls over a six-

month period in Stanford [55] shows that more than 70% of the calls made by callers in a week are to their top 5 callees. Also 80% of the calls made by callers in one day are to their top 3 callees. If we store the top 3 callees' locations in the VLR of the caller, then 80% of the remote search cost can be eliminated. The results showed that through replication, a significant call latency and network cost could be eliminated when the user's call-to-mobility ratio is high. Under the replication strategy, the location of a specific MH is replicated at selected sites. A location update operation due to mobility proceeds as follows:

- The MH sends a location update message to the serving MSC/VLR, which forwards it to the MH's HLR as in IS-41.
- The HLR updates the location of the MH. In addition, it also sends an update message to all VLRs where a replica of the MH's profile is stored. All other steps are the same as in IS-41.

A call delivery operation under the replication strategy proceeds as follows:

- The calling MH sends a call initiation message to its current serving MSC/VLR through a base station.
- If the MSC/VLR stores a location replica, it contacts the callee's current MSC directly and gets the routing information. Otherwise the IS-41 procedure is followed.

3.2 Comparative Analysis of Location Management Algorithms in PCS Cellular Networks

In this section, we develop analytical models to describe the behavior of a mobile user under various two-tier based location management algorithms in PCS cellular networks. Our intent is to assess and compare location management algorithms and be able to classify them based on the location update cost required. Our approach is based on hierarchical modeling. At the high level comes a cost model for defining the cost of the PCS network in servicing “location update” and “location search” operations for a mobile user. We assume that all times are exponentially distributed, thus allowing us to use a Markov model defined at the low level for parameterizing the cost defined at the high level. We note that this assumption may not be justified for the residence time in a registration area [62] but it can be relaxed by using modeling tools such as SPNP version 6 [59] and TimeNET version 3 [71] at the low level in which times are generally distributed. The modeling approaches proposed in defining Markov models and in calculating location update and search costs then can be similarly applied.

3.2.1 High-level Model

The high-level model adopts the cost model in [34] and includes two components: (a) update cost, the cost of updating the location of the mobile user due to user movements; and (b) search cost, the cost of searching the user in response to a call. Assume the time that a particular MH stays in a VLR before moving to another one is characterized by an exponential distribution with an average rate of σ. The inter-arrival time between two consecutive calls to a particular MH is also assumed to be exponentially distributed with an average rate of λ. An MH is thus characterized by its *CMR* value, defined as λ/σ. For a location management scheme X, let X_{update} be the average cost of the PCS network in servicing a location update operation due to a user movement crossing VLR registration boundaries. Note that for some location management algorithms, a user movement may not cause any update cost at all, e.g., a local movement under the PLA scheme causes zero update cost. Therefore, X_{update} here stands for the **average** cost of a user movement over the lifetime of the mobile user, covering both the local and regional moves. Similarly, let X_{search} be the **average** cost in locating the mobile user in a location search operation. Furthermore, let X_{cost} be the average cost of the PCS network in servicing the above two types of operations between two consecutive calls. Then,

$$X_{cost} = X_{update} \times \sigma/\lambda + X_{search} \quad (1)$$

The equation is obtained above because between two consecutive calls, the number of movements across VLR registration boundaries by the mobile user is equal to σ/λ on average. One can imagine that a mobile user moves across registration boundaries for a number of times (σ/λ on average) before receiving a call and then the same pattern repeats again. The X_{cost} parameter above gives the total cost incurred to the PCS network in each such repeated period accounting for both the location update and search costs, thus providing a uniform cost measurement for fairly comparing all location management schemes.

3.2.2 Low-level Model

We now develop separate Markov models for PLA, FRA and LAA. The objective is to parameterize Equation (1). To make our presentation concrete, we consider the hexagonal network coverage model. Note that T and τ defined in Table 2-2 represent the **average**

communication costs. Their values can be calculated by means of a network coverage model (e.g., hexagonal) characterizing the underlying wireless network as in [9].

3.2.2.1 Modeling PCS Network Operating under PLA

For notational convenience, we introduce the following additional parameters as we model PLA. Note that σ_i, β_i, μ_r and μ_i can each be expressed as a function of the per-mobile-user and network parameters introduced in Chapter 2. We will explain how these functions are obtained later.

Table 3-1: Additional Parameter in PLA.

Symbol	Meaning
n	the n parameter used by PLA to specify an n-layer region within which a user move causes no update cost
σ_i	the mobility rate of the mobile user moving from layer i to layer $i+1$
β_i	the mobility rate of the mobile user moving from layer $i+1$ to layer i
μ_r	the execution rate to perform a location update to the HLR
μ_i	the execution rate to locate the mobile user currently located in layer i
$P_{(i,j)}$	the probability that the system is in a particular state in equilibrium

For a hexagonal model as shown in Figure 3-2, it can be shown [2] that the probability of a mobile user moves from layer i to layer $i+1$, $i \geq 0$, given that a random move has been made by the mobile user, is given by:

$$\begin{cases} \dfrac{2i+1}{6i} & if\ i \geq 1 \\ 1 & Otherwise \end{cases}$$

The special case is that the probability is 1 for moving from layer 0 (containing only the agent itself) to layer 1. Similarly, the probability of moving from layer $i+1$ to layer i, $i \geq 0$, given that a random move had been made by the mobile user can be derived as

$$\frac{2(i+1)-1}{6(i+1)}$$

Hence, the mobility rate of the mobile user moving from layer i to layer $i+1$, σ_i, is given by

$$\sigma_i = \begin{cases} \dfrac{(2i+1)\sigma}{6i} & if\ i \geq 1 \\ \sigma & Otherwise \end{cases}$$

and the mobility rate of the mobile user moving from layer $i+1$ to layer i, β_i, is given by

$$\beta_i = \frac{2(i+1)-1}{6(i+1)}\sigma \quad i \geq 0$$

Furthermore, since it takes an average of *2T* time to do a round-trip VLR-HLR communication, the execution rate to perform a location update from a new VLR agent to the HLR, μ_r, can be parameterized as

$$\mu_r = \frac{1}{2T}$$

Note that the above includes a one-way communication cost from the new VLR agent to the HLR to update the HLR's database and another one-way communication time from the HLR to the new agent to acknowledge the request. The update action is triggered by the new VLR agent outside of the *(n-1)*-distance region.

The time to locate a mobile user in layer *i* incurs the following communication costs: (a) from the caller VLR to the callee HLR; (b) from the HLR to the agent; (c) from the agent (in layer 0) to the current VLR (in layer *i*); (d) from the current VLR back to the HLR and (e) the HLR forwards the routing information to caller VLR. The step (d) also updates the HLR's database such that the current VLR becomes the new agent. This is so because the search event is triggered by the HLR which expects to receive the user's location information from the current VLR. Please note that paging is used from the agent to locate the current VLR. Hence, the execution rate μ_i to locate the mobile user in layer *i* (note the value of *i* starts from 0), can be parameterized as

$$\mu_i = 1/(4T + (3i^2 + 3i) \times (1/2) \times 2\tau)$$

Here, the second term in the denominator accounts for the fact that on average the agent will have to query one half of the VLRs in the *i*-distance region to find the current VLR in layer *i*.

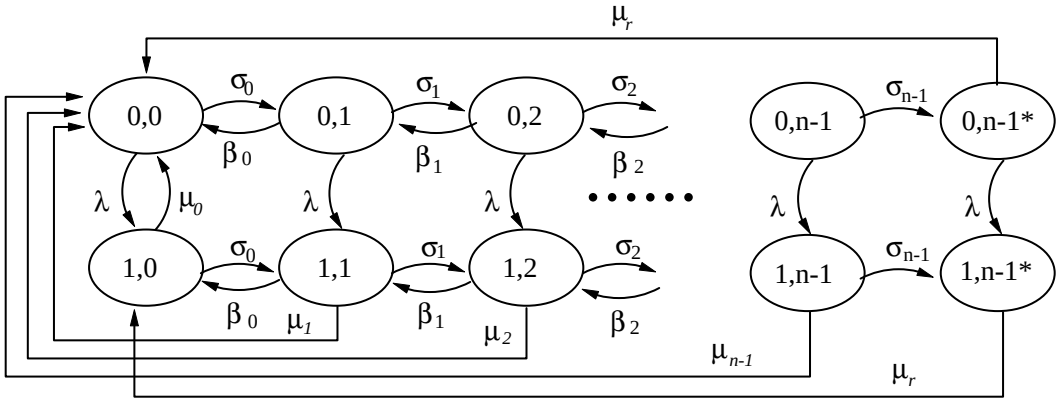

Figure 3-6: The Markov model for the PCS network under PLA.

Figure 3-6 shows a low level Markov model for describing the behavior of a mobile user under PLA. Here a state is represented by *(a,b)* where *a* is either *0* (standing for IDLE) or *1* (standing for CALLED), and *b* indicates the current distance between the mobile user and the local agent. Of course, $0 \le b \le n-1$, where *n* is the distance value. Initially, the mobile user is in the state of *(0, 0)*, meaning that it is not being called and the mobile user resides in the area of the current VLR agent. Below, we explain briefly how we construct the Markov model.

- If the mobile user is in the state of *(0,i)* and a call arrives, then the new state is *(1,i)*, i.e., the mobile user is now in the state of being called. This behavior is modeled by the (downward) transition from state *(0,i)* to state *(1,i)*, $0 \le i \le n-1$, with a transition rate of λ.
- If the mobile user is in the state of *(1,i)* and another call arrives, then the mobile user will remain in the same state, since the mobile user remains in the state of being called. This behavior is described by a hidden transition from state *(1,i)* back to itself with a transition rate of λ. This self transition is not shown in the model since it does not need to be considered as we solve the Markov model for the steady state probability $P_{(i,j)}$.
- If the mobile user is in layer *i* while it is being called, the network serves the call with a service rate μ_i whose magnitude depends on *i*. This is modeled by a transition from state *(1,i)* to *(0,0)* with rate μ_i. Note that in the above transition, the new state is *(0,0)* because all calls can be serviced at once and after the HLR is informed of the location update, the new VLR which the mobile user currently resides under becomes the new agent.
- Regardless of whether the mobile user is in the state of being called or not, if the mobile user moves from layer *j* to layer *j+1*, $0 \le j \le n-2$, the distance between the mobile user and the current local agent is increased by 1. This is described by a transition from state *(i,j)* to *(i,j+1)* with rate σ_j. This also describes the behavior of the user in migrating from an inner layer to an outer layer.
- If the mobile user moves from an outer layer *j+1* to an inner layer *j*, the distance between the mobile user and the current local agent is reduced by 1. This is described by a transition from state *(i,j)* to state *(i,j-1)* with rate β_{j-1}, $1 \le j \le n-1$. This also describes the behavior of the user in migrating from an outer layer to an inner layer.
- If the mobile user is in the state of *(i,n-1)*, where *n* is the distance value, and it makes a move to a new outer layer *n*, then a regional movement occurs and a reset operation must be invoked to update the new local agent to the HLR. This behavior is described first by a transition from *(i,n-1)* to *(i,n-1)** with rate σ_{n-1}, after which a transition occurs from *(i,n-1)** to *(i,0)* with rate μ_r, representing the time it takes for the PCS network to execute the reset operation and update the HLR with the new agent information.

If the call arrival rate λ is much higher than the mobility rate σ, then the probability that the system is found to stay in state *(1, i)* would be much greater than in state *(0,i)*. Let pla_{update} denote the average cost of the PCS network for executing a location update operation and let pla_{search} denote the average cost for executing a location query operation. Then,

$$pla_{update} = \sum_{i=0}^{1} (P_{(i,n-1)} + P_{(i,n-1)^*}) \times \frac{1}{\mu_r}$$

$$pla_{search} = \sum_{i=0}^{1} \sum_{j=0}^{n-1} P_{(i,j)} \times \frac{1}{\mu_j} + (P_{(0,n-1)^*} + P_{(1,n-1)^*}) \times \frac{1}{\mu_0}$$

where $P_{(i,j)}$ is the percentage of time the system is found to be staying at state *(i,j)* in equilibrium. Therefore, based on Equation (1),

$$pla_{cost} = pla_{update} \times \sigma/\lambda + pla_{search} \quad (2)$$

3.2.2.2 Modeling PCS Network Operating under FRA

Here we develop a Markov model to account for the following behaviors in order to compare FRA with other algorithms fairly: (a) the forwarding chain will be reset after a location query operation is performed; (b) when the mobile user moves back to the previously visited VLR in the chain, i.e., from V_i to V_{i-1}, the length of the forwarding chain is reduced by 1 and no pointer deletion operation is required between V_i and V_{i-1}. The second point is based on the assumption that obsolete pointers will be purged automatically after a period of time much greater than the average reset period. The Markov model will account for the looping behavior involving two neighbor VLRs in the forwarding chain. Note that in the above scenario if the mobile user subsequently moves from V_{i-1} to V_i again, a pointer set-up operation is still required. We assume that when the mobile user moves across a VLR, the forwarding pointer information will be updated before it crosses another VLR. This assumption has the implication that the dwell time of the mobile user in a VLR is much larger than the forwarding pointer update operation time.

For FRA, we also introduce a set of parameters as follows to facilitate discussion:

Table 3-2: Additional Parameters in FRA.

Symbol	Meaning
k	the length of the forwarding chain at which a reset operation is performed
σ_n	the mobility rate of the mobile user moving to a new VLR
σ_b	the mobility rate of the mobile user moving to the previous VLR
μ_r	the execution rate to reset a forwarding chain, i.e., to update the HLR
μ_f	the execution rate to set-up or travel a pointer between two VLRs
μ_i	$0 \le i \le k-1$, the execution rate to locate the mobile when the length of forwarding pointer is i

Again based on the hexagonal structure and random movement, parameterized as follows:

$$\sigma_n = \sigma \times (5/6)$$

$$\sigma_b = \sigma \times (1/6)$$

The time to update the HLR from a VLR is $2T$. Hence,

$$\mu_r = 1/(2T)$$

Similarly, the time to set-up a pointer between two VLRs is 2τ. Hence,

$$\mu_f = 1/(2\tau)$$

Finally, the time to locate the mobile user when the length of the forwarding chain is i includes the time from the caller VLR to the callee HLR, the time from the callee's HLR to the first VLR on the chain, the time to traverse the chain from the first VLR to the ith VLR, and finally the time to go from the ith VLR back to the HLR, and the time to forward the message from the HLR to caller VLR. Hence,

$$\mu_i = 1/(4T+2\,i\,\tau)$$

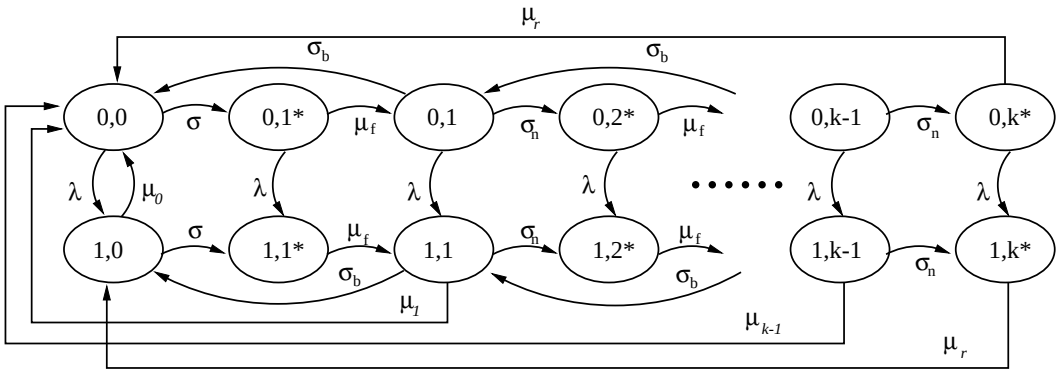

Figure 3-7: The Markov Model for the PCS Network under FRA.

Figure 3-7 shows a Markov model describing the behavior of a mobile user wherein a state is represented by (s_1, s_2) where s_1 is either *0* (standing for IDLE) or *1* (standing for CALLED), while the other component s_2 indicates the current length of the forwarding chain. Initially, the mobile user is in the state of *(0, 0)*, meaning that it is not being called and the number of forwarding steps is zero. A symbol "*" is put in a state if the mobile user just enters a new VLR but the forwarding pointer operation is not yet performed. For example, in state *(0, 1)** the mobile user has just crossed V_1 from V_0 but the forwarding operation between V_0 and V_1 is not yet performed. Of course, after the forwarding pointer operation is performed, the length of the forwarding chain will be 1 in this case.

We briefly discuss the meaning of the Markov chain as follows. First, if the mobile user is in the state of *(0,i)* or *(0,i)** and a call arrives, then the new state is *(1,i)* or *(1,i)** in which the number of forwarding steps remains at *i* but the mobile user is now in the state of being called. This behavior is modeled by the (downward) transition from state *(0,i)* to state *(1,i)* or from state *(0,i)** to state *(1,i)** , $0 \le i \le k$, with a transition rate of λ. Second, if the mobile user is in the state of *(1,i)* or *(1,i)** and another call arrives, then the mobile user will remain in the same state. Third, if the mobile user is in the state of *(1,i)*, the signaling network can service all pending calls simultaneously with a service rate of μ_i. After the service, the new state is *(0,0)* since all calls are serviced and the reset operation is also performed. This behavior is described by the state transition from state *(1,i)* to state *(0,0)* with a transition rate of μ_i. Note that this rate depends on the length of the forwarding chain. Fourth, if the mobile user moves back to the previously visited VLR on the forwarding chain, a transition from state *(i,j)* to state *(i,j-1)* will take place, where $1 \le j \le k-1$. Note that obsolete pointers will be deleted implicitly, so there is no need to take time to perform the pointer delete operation. Lastly, regardless of whether the mobile user is in the state of being idle or having been called, if the mobile user moves across a new VLR boundary, a pointer connection or a reset operation must be performed in response to the move event. This behavior is first modeled by a transition from state *(i,j)* to state *(i,j+1)** with a mobility rate of σ_n where $0 \le i \le 1$ and $0 \le j \le k-1$, after which one of the following two cases may occur:

- If $0 \le j \le k-2$, then the new VLR simply sets up a forwarding pointer connection. This behavior is modeled by a transition from state *(i,j+1)** to state *(i,j+1)* with rate μ_f.
- If $j=k-1$, however, then the length of the forwarding chain has reached *k*, so the new VLR must perform a reset operation. This latter behavior is modeled by a transition from state *(i,k)** to state *(i,0)* with rate μ_r.

Now let $P_{(i,j)}$ represent the probability of the system is found to be staying at state *(i,j)* in equilibrium. Let fra_{update} be the average cost to perform a location update operation. Let fra_{search} be the average cost to perform a location search operation.

Then,

$$fra_{update} = (P_{(0,k-1)} + P_{(1,k-1)} + P_{(0,k)^*} + P_{(1,k)^*}) \times \frac{1}{\mu_r} + \left(\sum_{i=0}^{k-2} (P_{(0,i)} + P_{(1,i)}) + \sum_{i=1}^{k-1} (P_{(0,i)^*} + P_{(1,i)^*}) \right) \times \frac{1}{\mu_f}$$

$$fra_{search} = \sum_{i=0}^{k-1} (P_{(0,i)} + P_{(1,i)} + P_{(0,i+1)^*} + P_{(1,i+1)^*}) \times \frac{1}{\mu_i}$$

Hence, based on Equation (1),

$$fra_{cost} = fra_{update} \times \sigma/\lambda + fra_{search} \quad (3)$$

Equation (3) above yields the average cost of the signaling network as a function of *K*. For a given set of parameter values, we can first compute the values of $P_{(i,j)}$ for all states and then use Equation (3) to compute the average cost. The optimal *K* value is the one that minimizes the cost measure defined in Equation (3). Once the optimal *K* value is determined, it can be used to compute fra_{update} so as to determine where the optimal FRA algorithm will fall in the spectrum of degradable location management algorithms for a given set of per-mobile-user and network conditions.

3.2.2.3 Modeling PCS Network Operating under LAA

Under the LAA scheme, if a mobile user makes a move under the same network switch, i.e., a local movement, then the new VLR only informs the local anchor without updating the HLR. However, if the mobile user crosses a network switch boundary, a registration operation must be initiated to update the HLR and the new VLR will become the local anchor.

We introduce some more parameters below to ease our discussion:

Table 3-3: Additional Parameters in LAA.

Symbol	Meaning
n	the n parameter to specify the n-layer VLR region which covers $3n^2-3n+1$ VLRs
P_1	the probability that when the mobile user moves it remains under the same network switch
σ_1	the mobility rate of the mobile user moving under the same network switch, i.e., $\sigma_1 = P_1 \sigma$
σ_r	the mobility rate of the mobile user crossing a network switch boundary, i.e., $\sigma_r = (1- P_1) \sigma$
μ_g	the search execution rate when the mobile user is directly covered by the anchor, i.e., located in the anchor's VLR
μ_a	the search execution rate when the mobile user is not directly covered by the anchor, i.e., the mobile user is located in other VLRs in the local anchor area
δ_1	the execution rate to update the anchor, i.e., to set-up a pointer between the new VLR and the anchor
δ	the execution rate to update the HLR

Here we also consider a hexagonal network coverage model, in which each network switch covers an n-layer VLR region where n can be either 2 or 3. In this case, it can be derived easily (see [9]) that the probability of the MH staying under the same switch when making a move, i.e., P_1, is equal to

$$P_1 = (3n^2 - 5n + 2) / (3n^2 - 3n + 1)$$

Hence,

$$\sigma_1 = \sigma \times (3n^2 - 5n + 2) / (3n^2 - 3n + 1)$$

and

$$\sigma_r = \sigma \times (2n-1) / (3n^2 - 3n + 1)$$

When the mobile user is under the area directly covered by the anchor, the average time needed to find the mobile user is the same as that in basic IS-41 scheme. Hence,

$$\mu_g = 1 / (4T)$$

Otherwise, the anchor must communicate with the mobile user's current VLR via a pointer to locate the mobile user. Thus,

$$\mu_a = 1 / (4T+2\tau)$$

where τ is the average single-trip communication cost between the anchor and a non-anchor VLR within the anchor area (under the same switch).

The average time to update the anchor from a new VLR within the same switch is the VLR-VLR round-trip communication time. Hence,

$$\delta_l = 1 / (2\tau)$$

On the other hand, the average time to update the HLR from a new VLR when a new switch area is entered is the VLR-HLR round-trip communication time, Hence,

$$\delta = 1 / (2T)$$

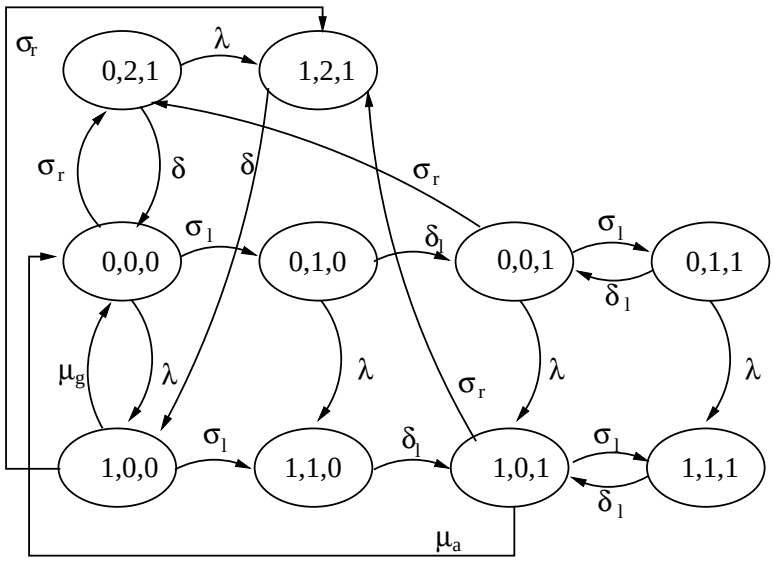

Figure 3-8: The Markov Model for the PCS Network under LAA.

Figure 3-8 shows a Markov model describing the behavior of a mobile user with the state representation *(a,b,c)*. The first component *a* indicates whether or not the mobile user is in the state of being called, with *0* standing for idle and *1* standing for busy. The second component *b* indicates if the mobile user makes a move, with 0 meaning that it does not, 1 meaning that it just makes a local move, and 2 meaning that it just makes a regional move. The third component indicates if the anchor directly covers the mobile user, with *0* meaning yes and *1* meaning no. Initially, the mobile user is in the state of *(0,0,0)*, meaning that the mobile user is not being called,

has not yet made any move, and is directly covered by the anchor. Below, we explain briefly how we construct the Markov model.

- First, if the mobile user is in the state of *(0,i,j)* and a call arrives, the new state will be *(1,i,j)*, i.e., the mobile user is now in the state of being called. This behavior is modeled by the transition from state *(0,i,j)* to state *(1,i,j)*, $0 \le i \le 1$, or from state *(0,2,1)* to state *(1,2,1)*, with a transition rate of λ.
- Calls will be serviced with a rate of μ_g when the system is in state *(1,0,0)* since in this case the HLR database points to the anchor which directly covers the mobile user. This is modeled by a transition from state *(1,0,0)* to *(0,0,0)* with a transition rate of μ_g. On the other hand, calls will be serviced with a rate of μ_a when the system is in state *(1,0,1)* since in this case the HLR database points to the anchor which does not cover the mobile user directly; therefore, we must follow the pointer stored at the anchor's database to locate the mobile user's current VLR, after which the VLR also becomes the new anchor. This latter case is modeled by a transition from state *(1,0,1)* to *(0,0,0)* with a transition rate of μ_a.
- If the mobile user moves across a VLR boundary, a location update operation will be performed either to the anchor or to the HLR, depending on whether a network switch is crossed. We therefore distinguish the following two cases:
    - If the mobile user has moved across a network switch, a reset operation must be performed to update the HLR. This behavior is modeled first by a transition from state (i,0,j) to state (i,2,1), $0 \le i,j \le 1$, with a transition rate σ_r, after which a transition occurs from (i,2,1) to (i,0,0), $0 \le i \le 1$, with a transition rate δ to account for the HLR update time. Here, after an update is done to the HLR, the HLR database points to the new anchor which covers the mobile user directly.
    - If the mobile user just makes a local move, then only a pointer set up between the local anchor and the new VLR is required. This behavior is modeled first by a transition from state (i,0,j) to state (i,1,j), $0 \le i,j \le 1$, with a transition rate $\sigma 1$, after which a transition occurs from (i,1,j) to (i,0,1), $0 \le i,j \le 1$, with a execution rate $\sigma 1$ to account for the update time to the anchor.

Now, from the Markov chain,

$$laa_{update} = \sum_{i=0}^{1}\sum_{j=0}^{1} P_{(i,0,j)} \times \left[P_1 \times \frac{1}{\delta_1} + (1-P_1) \times \frac{1}{\delta} \right] + \sum_{i=0}^{1}\sum_{j=0}^{1} P_{(i,1,j)} \times \frac{1}{\delta_1} + (P_{(0,2,1)} + P_{(1,2,1)}) \times \frac{1}{\delta}$$

$$laa_{search} = \sum_{i=0}^{1}(P_{(i,0,0)} + P_{(i,2,1)}) \times \frac{1}{\mu_g} + \sum_{i=0}^{1}(P_{(i,1,0)} + P_{(i,0,1)} + P_{(i,1,1)}) \times \frac{1}{\mu_a}$$

Therefore, based on Equation (1),

$$laa_{cost} = laa_{update} \times \sigma/\lambda + laa_{search} \quad (4)$$

Again, Equation laa_{update} above can be used to classify LAA for a given set of per-mobile-user and network conditions.

3.2.3 Analysis and Comparison

All the data reported here were obtained by solving the Markov models using the SHARPE software package [52] to obtain $P_{(i,j)}$'s for all states and then by computing the location update, search or total cost based on the equations derived. We report a case in which the ratio of the VLR-to-VLR cost to the VLR-to-HLR cost is equal to 0.3, i.e., $T=1$ and $\tau=0.3$. This selection reflects a reasonable ratio between T and τ. The exact ratio of T to τ depends on the wireless network employed and can be computed using the approach described in [9] by means of a network coverage model. Here we report only the result for this case since the main objective is to demonstrate how our two-level hierarchical modeling method can be effectively used to compare degradable location management algorithms when given a set of per-mobile-user and network parameters that characterize a wireless network environment. For completeness, we will first show individual performance characteristics of these algorithms compared with the basic HLR/VLR IS-41 algorithm. Then we will compare these algorithms head to head under identical per-mobile-user and network conditions.

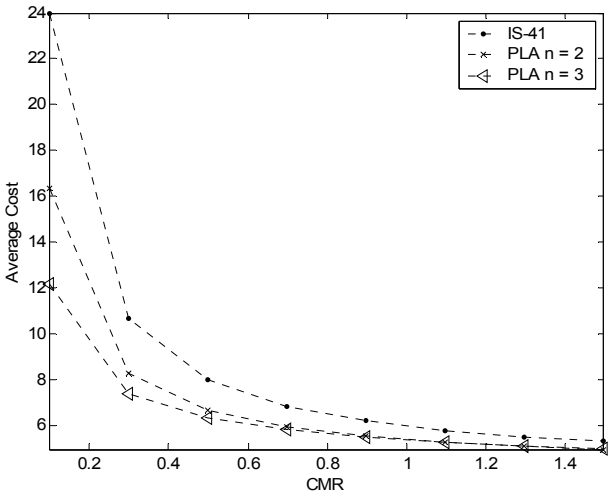

Figure 3-9: Comparison of PLA under Different *n*-distance Values.

The average cost of the PCS network for location management under IS-41 is given by

$$IS\text{-}41_{cost} = IS\text{-}41_{update} \times \sigma/\lambda + IS\text{-}41_{search}$$

where $IS\text{-}41_{\text{update}} = 2T$ and $IS\text{-}41_{\text{search}} = 4T$.

The average cost of PLA, FRA and LAA, as defined by Equations (2) (3) and (4) are shown in Figure 3-9, Figure 3-10 and Figure 3-11, respectively, along with that of IS-41 for comparison purposes. These figures show that for the case when $\tau = 0.3\ T$ and when call-to-mobility (*CMR*) is small, PLA, FRA and LAA all can significantly outperform IS-41.

Figure 3-9 shows a plot of the PCS network cost under PLA(*n*) for various *n* values. When the *CMR* value is small, the performance of PLA with *n=3* is better than PLA with *n=2*. This behavior can be explained as follows. Recall that a larger *n* value means that a local agent can cover a larger area and thus there is a smaller probability for the mobile user to cross a regional boundary. Consequently, the number of update operations to the HLR is reduced as *n* increases. This is reflected in Figure 3-9 where we see that the performance of PLA with *n=3* is better than PLA with *n=2* at low *CMR* values where the cost of location updates dominates that of location queries. As the *CMR* value increases, however, the larger location query cost attributed by the larger cover area starts to dominate the reduced location update cost. Therefore, after the *CMR* value exceeds a threshold (CMR=2, not shown in Figure 3-9), PLA with *n=2* becomes better than PLA with *n=3*.

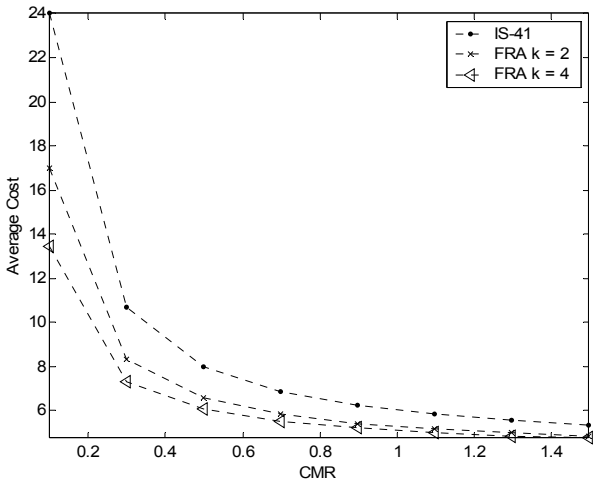

Figure 3-10: Comparison of FRA under Different Lengths in the Forwarding Chain.

Figure 3-10 demonstrates that FRA has a better performance with a long forwarding chain when *CMR* is small, again due to the fact that at a low *CMR* value, the location update cost dominates the location query cost, so a longer chain is favored at low *CMR* since it reduces the location update cost. Again, as *CMR* increases, the higher cost associated with location search operations which happen frequently starts to offset the benefit of lower cost associated with location update operations which do not happen as frequent. In general, for any combination of mobile user and network conditions, there exists an optimal *K* value for which the network cost is minimized.

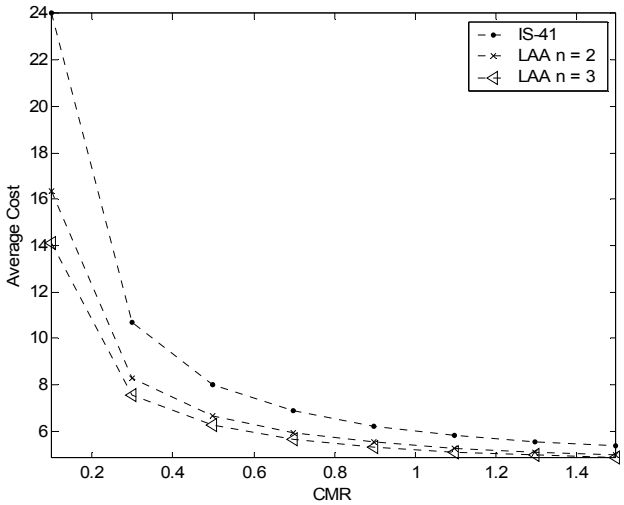

Figure 3-11: Comparison of LAA under Different Distance Values.

Figure 3-11 shows that LAA with *n=3* is better than LAA with *n=2* for *CMR*∈[0.1,1.5]. The reason is that we assume the communication cost under one switch that covers an *n*-layer VLR region is the same. This indicates that the network switch with *n=3* is more powerful than that with *n=2*. The effect of the network switch difference is not included in the average cost.

We compare these PLA, FRA and LAA algorithms head to head under the selected set of network conditions in Figure 3-12, Figure 3-13 and Figure 3-14.

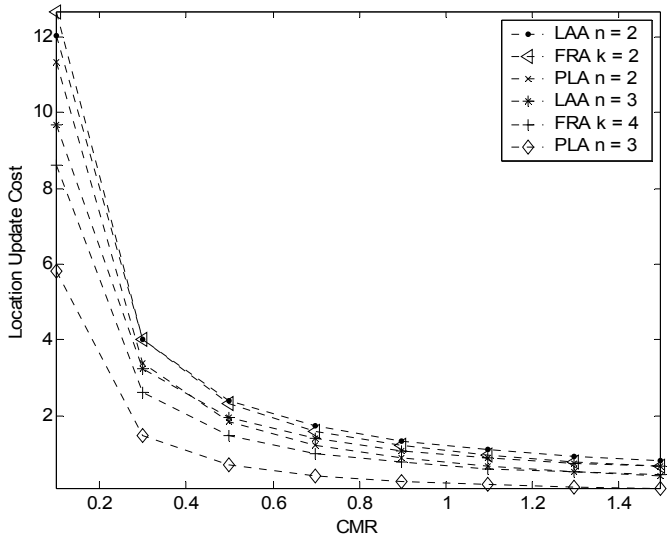

Figure 3-12: Comparison of the Location Update Cost Only

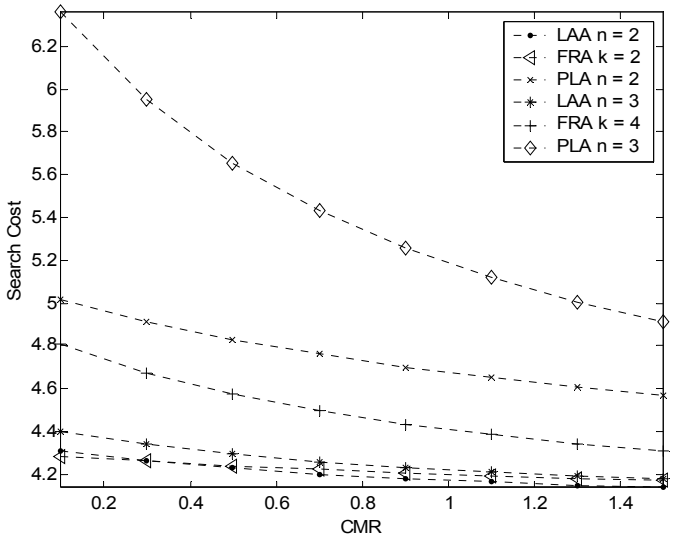

Figure 3-13: Comparison of the Search Cost Only.

Figure 3-12 shows only the location update cost part, that is, the $X_{update} \times \sigma/\lambda$ part (or X_{update}/CMR) in Equation (1). It provides us a basis for classifying existing degradable location management algorithms based on the update cost per move relative to IS-41 for a wide range of *CMR* values. Specifically, the IS-41 HLR/VLR algorithm can be considered as an algorithm which keeps the system in the strong state all the time. FRA(2) and LAA(2) are the next two among the 6 algorithms listed in terms of maintaining the location information in a good state. From Figure 3-12, we see that PLA(3) incurs the least amount of update overhead among the 6 algorithms listed since under PLA(3) it is likely that the mobile user makes only local movements because the size of the local region is large. Consequently, it hardly does any update at all. We therefore expect that PLA(3) will have to spend more time to search for the mobile user when a call arrives. This is confirmed in Figure 3-13 which displays only the location query cost part, that is, the X_{search} part in Equation (1), in which it shows indeed PLA(3) incurs the most overhead to deliver a call. Figure 3-12 and Figure 3-13 thus clearly demonstrate the trade-off between the location update and search operations.

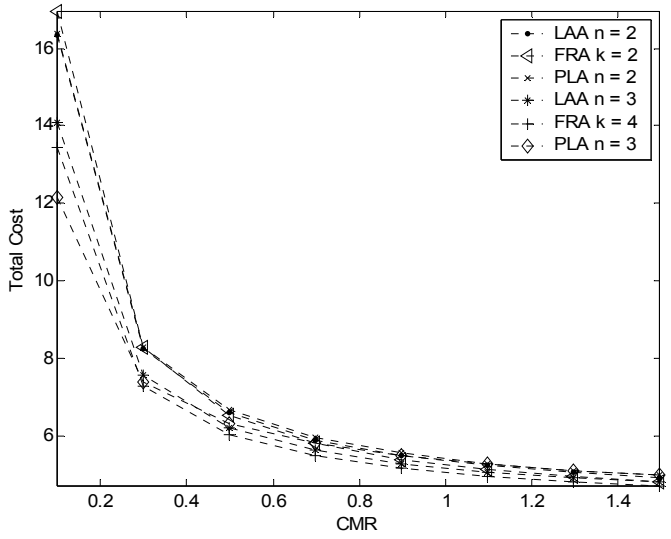

Figure 3-14: Comparison of the Total Communication Cost for Location Management.

We also observe that the increasing order in location search cost under these degradable location management algorithms shown in Figure 3-13 is not exactly the same as the decreasing order in location update cost shown in Figure 3-12, suggesting that certain algorithms can actually perform better under the same network condition. Figure 3-14 compares all the 6 algorithms listed by combining both parts of the network cost, i.e., based on Equation (1). For the 6 algorithms listed under this selected network condition, we can see that PLA(3) is the best when

CMR is 0.1, FRA(4), PLA(3) and LAA(3) are better than FRA(2), PLA(2) and LAA(2) when *CMR* is 0.3, and all six algorithms performances are very close when *CMR* is larger than 0.5.

3.3 Hybrid Location Management Algorithm

From the analysis in Section 3.2, we see different location management algorithms could be applied to different users having different *CMR* to reduce the system signaling cost. However, this will introduce undesirable high complexity in managing and maintaining location-related information. Motivated by providing a uniform algorithm that can be generally applied to all users with different *CMR* values without sacrificing the optimality of individual algorithm, we investigate and analyze hybrid schemes that can combine the benefits of two or more existing schemes.

Specifically, we develop and analyze a hybrid scheme that combines "replication" and "forwarding" techniques. Replication algorithm is most effective in reducing search and update costs when *CMR* is high. The reason is that when the call arrival rate to the user is much higher than the mobility rate, the communication cost would be dominated by search operations and the search cost can be reduced by replicating the location of a frequently called MH at selective VLRs from which most calls to the mobile user originate, thus avoiding the cost of querying the HLR (the only copy when replication is not used) for the location of the called MH. Conversely, forwarding algorithm is known to be most effective when *CMR* is low because when the mobility rate is much higher than the call arrival rate, the communication cost would be dominated by location update operations and the update cost can be reduced by simply forming a forwarding chain of VLRs through which the location of the MH can be found, thus avoiding updating the HLR upon every update operation. The HLR is updated only periodically when the forwarding chain becomes too long (say after *K* moves). Our hybrid scheme takes advantages of the benefits of replication and forwarding techniques at high and low *CMR* values, respectively.

A key concept of the hybrid scheme design is that it will degenerate into the forwarding scheme when the *CMR* value is sufficiently low and, on the other hand, into the replication scheme when the *CMR* value is sufficiently high. Under the hybrid scheme, the system applies the optimal number of replicas and the optimal forwarding chain length in order to minimize the total signaling cost, when given the user profile as input characterizing the user's calling and moving

patterns. A lookup table is built at static time and applied to all users at run time to identify the optimal number of replicas and the optimal forwarding chain length to be used for each user.

We develop a Stochastic Petri Net (SPN) model to help built the table. Because of the uniformity property associated with the hybrid scheme, the internal data structure used by the system to manipulate and maintain the location information for all users is the same, thus greatly easing the system maintenance task. Moreover, since the per-user profile is typically kept at the HLR, we can keep the lookup table at the HLR. The run time overhead of determining the optimal number of replicas and forwarding chain length involves only a table lookup operation which can be executed efficiently.

Later the hybrid scheme is shown to perform better than either scheme under all *CMR* values, as well as a binary "*CMR* threshold-based" scheme that applies the replication technique when a mobile user's *CMR* is higher than a threshold and applies the forwarding technique otherwise. For this binary *CMR* threshold-based scheme, we identify the optimal threshold value and show the hybrid scheme outperforms the threshold-based scheme under optimal threshold values.

3.3.1 Hybrid Replication with Forwarding Strategy

We first observe that the forwarding strategy is most beneficial when the user's *CMR* is below a threshold. The effect is especially pronounced when τ/T is small. On the other hand, the replication strategy attempts to exploit locality in calling patterns to reduce the call delivery cost, and is most beneficial when user's *CMR* value is above a threshold.

We analyze a hybrid strategy that combines per-user replication and forwarding. The basic idea is that under a low *CMR*, the hybrid strategy attempts to reduce the total cost by replacing expensive HLR update operations with adjacent VLR forwarding pointer operations, thus behaving like the forwarding strategy. Under a modest *CMR*, the hybrid strategy combines the benefits from both forwarding and replication strategies. Finally, under a high *CMR*, the hybrid strategy behaves like the replication strategy by exploiting call locality to reduce the total signaling cost.

Under the hybrid scheme, the system maintains N replicas and a maximum forwarding chain length of K for each user. How many replicas to be used depends on a user's *CMR* value and its call arrival profile, e.g., if 50% of the total calls to the user originate from a single VLR (so N=1),

70% of the total calls originate from 2 VLRs (N=2), and 80% may come from 3 VLRs (N=3), then the call arrival profile can be represented by a series of *(N,P)* values, e.g., (1,50%), (2,70%) and (3,80%). This per-user call arrival profile information is kept at the MH's HLR as part of the user profile.

When given this per-user profile information and by applying the model and methodology described in Sections 3.3.2 and 3.3.3, the HLR determines the user's best (N,K) combination by performing a table lookup operation at run time with the goal to minimize the overall cost associated with location search and update operations for the user. Each of the N replicas stores the identical content, i.e., a pointer to the first VLR of the forwarding chain (say v_0). Note that a location update operation would not change this content stored in replicas unless the current forwarding chain length reaches K. Also, once the best (N, K) is determined for each MH, the HLR knows exactly which calling VLRs keep a replica by consulting the MH's call arrival profile. For example, if the best N value is 2, then only the two VLRs with the highest calling probability to the MH will each keep a replica; all other VLRs will not keep a replica. Recall that the call arrival profile regarding which VLRs have the highest calling probability to the MH is part of the user profile kept by the HLR. The HLR then is responsible for updating v_0 to all N replicas when a reset operation on the forwarding chain is performed at the HLR.

A location update under our hybrid strategy proceeds as follows:

- When an MH moves into a new registration area, it sends a location update message to the new VLR/MSC via its current base station.
- The new VLR/MSC examines the current forwarding chain length. If it reaches K, the HLR is updated to point to the new VLR as in IS-41, the forward chain is reset as in the forwarding scheme, and all N replicas are updated to store the location of the new VLR as in the replication scheme. This is the only condition under which replicas are being updated upon a location update operation since the first VLR of the forwarding chain has been changed as a result of the reset operation being executed when the maximum forwarding chain length K is reached.
- Otherwise:
    - The MH registers at the new VLR and the new VLR de-registers MH at the old VLR;
    - The old VLR setups a pointer to the new VLR, and then sends an ACK and MH's location profile to the new VLR;
    - The current forwarding chain length is increased by one.

- The replicas are not changed.

When a call is placed to an MH, the caller's local VLR is searched first to see if it stores a location replica. If a replica is found (i.e., a replica hit), the caller will use the local information to get the first VLR and follow the forwarding chain to contact the callee's current VLR instead of contacting the callee's HLR. If a location replica is not found (i.e., a replica miss), the HLR is queried to determine the first VLR at which the callee was registered, and then the chain of forwarding pointers is followed to reach the current VLR.

A call delivery in the hybrid strategy proceeds as follows:

- The calling MH sends a call initiation message to its current serving VLR/MSC through its base station.
- The VLR/MSC checks if it has a location replica of the callee. If a replica is found, the caller's VLR sends a routing request message directly to reach the first VLR in the forwarding chain. Otherwise, the callee's HLR is contacted which sends a routing request message to the first VLR.
- The query message reaches the callee's current serving VLR/MSC by following the forwarding chain.
- The callee's current serving VLR/MSC sends routing information to the calling VLR/MSC (if it is a replica hit) or its HLR (if it is a replica miss), which forwards it to the calling VLR/MSC. Now the calling MSC can set up a connection to the called MSC via the SS7 signaling network using the usual call setup protocol.

3.3.2 Model

In this section we develop a Stochastic Petri Net (SPN) model to study the performance of the hybrid strategy compared with both replication forwarding schemes, as well as the binary *CMR* threshold-based scheme, for users with different call and mobility patterns. We use the average cost of the PCS signaling network between two consecutive calls as the performance metric. That is, let $Hybrid_{update}$ be the average cost of the signaling network in serving a location update operation and $Hybrid_{search}$ be the average cost of the signaling network in locating an MH. Then, $Hybrid_{cost} = Hybrid_{search} + Hybrid_{update} \times (\sigma/\lambda)$ where σ/λ is the average number of update operations issued by the user between two consecutive location search operations.

The SPN model of the hybrid strategy is shown in Figure 3-15 with the MH's incoming call pattern being modeled in the upper part and its mobility pattern being modeled in the lower part. These two parts interconnect with each other by several inhibitor arcs to model the fact that the system will perform location updates before location queries if they come simultaneously.

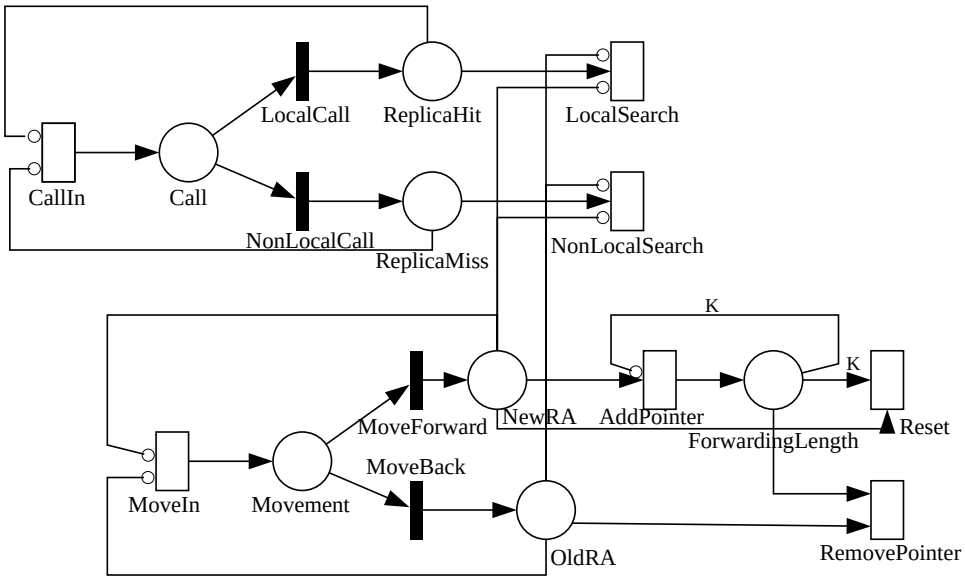

Figure 3-15: SPN Model for Hybrid Strategy.

Table 3-4 gives the notation for model parameters used. Table 3-5 gives the meaning of places defined in the SPN model. Table 3-6 shows transition rates (for timed transitions) or probabilities (for immediate transitions) assigned to transitions defined in the SPN model.

Table 3-4: Notation Used in Hybrid Strategy.

Symbol	Meaning
λ	arrival rate of calls to a particular MH
σ	mobility rate of a particular MH
CMR	call-to-mobility ratio of a particular MH, i.e., λ / σ
P_f	probability that a particular MH moves to a new VLR
P_b	probability that a particular MH returns to the last-visited VLR; $P_b = 1\text{-}P_f$
P_h	probability that a call is serviced by the local replica, i.e., the hit probability
P_m	Probability that a local replica is not found to service a call, i.e., the miss probability; $P_m = 1\text{-}P_h$

μ_p	execution rate to set up, delete, or travel a pointer between two adjacent VLRs
$\mu_{h,i}$	execution rate to find the current VLR where the MH resides for a forwarding chain of length *i* when there is a replica hit
$\mu_{m,i}$	execution rate to find the current VLR where the MH resides for a forwarding chain of length *i* when there is a replica miss
K	maximum number of forwarding steps after which a reset operation is performed
N	number of replicas
$\mu_{r,N}$	execution rate to reset a forwarding chain of length K so that after the reset operation is performed, all N replicas and the HLR are updated to point to the new VLR, and the old forwarding chain is released
$Hybrid_{update}$	average cost of the signaling network in serving a location update operation
$Hybrid_{search}$	average cost of the signaling network in locating an MH
$Hybrid_{cost}$	average cost of the signaling network in serving location update and MH locating between two consecutive calls, i.e. $Hybrid_{search} + Hybrid_{update} / CMR$

Table 3-5: Meaning of Places.

Place	Meaning
Call	*mark*(**Call**)[2]=1 indicates that a call has just arrived
ReplicaHit	*mark*(**ReplicaHit**)=1 indicates that the call can be serviced by the local replica
ReplicaMiss	*mark*(**ReplicaMiss**)=1 indicates that the call cannot be serviced locally because a replica does not exist
Movement	*mark*(**Movement**)=1 indicates that the MH just makes a move across the VLR boundary
NewRA	*mark*(**NewRA**)=1 indicates a forward movement has just been made
OldRA	*mark*(**OldRA**)=1 indicates a backward (return to the previously visited RA) movement has just been made
ForwardingLength	*mark*(**ForwardingLength**) indicates the current forwarding chain length

[2] *mark(***p***)* returns the number of tokens held in place **p**.

Table 3-6: Transition Rates or Probabilities.

Transition	*Symbol*
CallIn	λ
LocalCall	P_h
NonLocalCall	P_m
LocalSearch	$\mu_{h,i}$
NonLocalSearch	$\mu_{m,i}$
MoveIn	σ
MoveForward	P_f
MoveBack	P_b
AddPointer	μ_p
RemovePointer	μ_p
Reset	$\mu_{r,N}$

The SPN model is constructed as follows:

- When a call arrives, a token is placed in place **Call**. The immediate transition **LocalCall** and **NonLocalCall** are enabled with the probabilities of P_h and P_m, respectively.
  - If a location replica exists (a replica hit with probability of P_h) in the caller's VLR, the token will flow through transition **LocalCall** to **ReplicaHit** which in turn will be serviced by a **LocalSearch** procedure in the hybrid strategy. The **LocalSearch** procedure will directly find the first VLR via the local replica and subsequently find the current VLR (where the MH currently resides) by following the forwarding chain whose length is indicated by the number of tokens stored in place **ForwardingLength**, i.e., *mark*(**ForwardingLength**). This is modeled by making the execution rate of **LocalSearch** marking-dependent, that is, *rate*(**LocalSearch**) = $\mu_{h,i}$ where i=*mark*(**ForwardingLength**).
  - If a location replica does not exist (a replica miss with probability P_m) in the caller's VLR, the token will flow through transition **NonLocalCall** to **ReplicaMiss** which in turn will be serviced by a **NonLocalSearch** procedure in the hybrid strategy. The **NonLocalSearch** procedure will go to the HLR to locate the first VLR and subsequently find the current VLR (where the MH currently resides) by following the forwarding chain whose length is indicated by the number of tokens stored in place **ForwardingLength**, i.e., *mark*(**ForwardingLength**). This is also modeled by

making the execution rate of **NonLocalSearch** marking-dependent, that is, *rate*(**NonLocalSearch**) = $\mu_{m,i}$ where i = *mark*(**ForwardingLength**).

- When the MH moves across a RA/VLR boundary, a token is placed in place **Movement**.
  - If the MH moves to a new RA/VLR, the immediate transition **MoveForward** will consume the token, after which a token will be placed in **NewRA**. In this case, transitions **LocalSearch** and **NonLocalSearch** will subsequently be disabled. Either **AddPointer** or **Reset** is enabled based on the current length of the forwarding chain indicated by *mark*(**ForwardingLength**) as follows:
    - If the current forwarding length is less than the predefined maximum forwarding length K, Transition **AddPointer** is enabled. A forwarding pointer between the two adjacent VLRs is set up and a token is added to **ForwardingLength**.
    - If the current forwarding length is equal to K, transition **AddPointer** is disabled and transition **Reset** fires. A reset operation is performed which resets the forwarding chain and updates all replicas. Transition **Reset** consumes the token stored in **NewRA**, as well as the K tokens stored in **ForwardingLength**. After the reset operation, the number of token in **ForwardingLength** is zero, i.e., *mark*(**ForwardingLength**) = 0.
  - If the MH moves back to the previous RA, the immediate transition **MoveBack** will consume the token, after which one token will be placed in **OldRA**. At this time, there should exist at least one token in **ForwardingLength**. Therefore, **RemovePointer** fires and consumes one token from **ForwardingLength** and one token from **OldRA**, thus reducing the current forwarding length by one. This models the fact that the forwarding chain length can be reduced by one if the MH moves back to the last visited VLR.

The SPN model shown in Figure 3-15 will generate a semi-Markov model that contains a number of states with each state being represented by a 5-component tuple (**ReplicaHit**, **ReplicaMiss**, **NewRA**, **OldRA**, **ForwardingLength**). The **ReplicaHit** component will take on the value of either 1 or 0, with 1 meaning a call just arrives and there is a replica hit because of the existence of a local replica, and 0 meaning otherwise. The **ReplicaMiss** component will also take on the value of either 1 or 0, with 1 meaning a call just arrives and there is a replica misses because of the nonexistence of a local replica, and 0 meaning otherwise. Note that a value of 1 can appear in either of these two components, but not in both at the same time. When both components contain the value of 0, it means that there is no call arrival in the state. The **NewRA** component will take on the value of 1 or 0, with 1 meaning a movement crossing a VLR boundary just occurs and it is

a forward movement, and 0 meaning otherwise. The **OldRA** component will take on the value of 1 or 0, with 1 meaning a movement crossing a VLR boundary just occurs and it is a backward movement, and 0 meaning otherwise. When both **NewRA** and **OldRA** contain the value of 0, it means that there is no movement. The **ForwardingLength** component will take the value in the range of 0 to *K* indicating the length of the forwarding length in the state. For example, (0,1,0,0,3) represents a state in which a call just arrives with a replica miss and the forwarding length is 3. Thus the system needs to go to the HLR to find the first VLR of the forwarding chain and then follows the forwarding chain of length 3 to find the current serving VLR of he MH to deliver the call in that state. It should be noted that the number of replicas *N*, replica hit/miss probability, and length of the forwarding chain *K* will affect the transition rates of the SPN model, thus affecting the probability that the system is found in a particular state in the steady state. Moreover, the value of *K* will affect the total number of states that exist in the underlying state model. For the SPN model shown in Figure 3-15, the total number of states is less than 200 for *K* in the range of 0 to 10 (maximum forwarding length). Using an evaluation tool such as SPNP [59] designed to solve thousands of states, we can solve the SPN model very efficiently.

3.3.3 Methodology

In this section, we describe how to use the SPN model to evaluate the hybrid scheme proposed. Our goal is to determine the best *K* value (the maximum length of the forwarding chain) and the best *N* value (the replica number) to minimize the average cost of the PCS network between two consecutive calls, when given the profile of an MH.

3.3.3.1 $Hybrid_{cost}$ Calculation

Suppose that there are altogether N_s states in the underlying semi-Markov model of the SPN. Let P_i be the steady state probability that the system is found in state *i*, as solved by SPNP. The average cost of the PCS signaling network in serving location update and call delivery operations between two consecutive calls can be obtained by assigning "cost" values to states of the system. Let $Hybrid_{i,search}$ be the search cost assigned to state *i* given that a search operation is being serviced in state *i*. Then the average search cost is given by

$$Hybrid_{search} = \sum_{i=1}^{N_s} P_i \times Hybrid_{i,search}$$

Similarly, let $Hybrid_{i,update}$ be the update cost assigned to state *i* given a location update operation is being serviced in state *i*. Then the average location update cost is given by

$$Hybrid_{update} = \sum_{i=1}^{N_s} P_i \times Hybrid_{i,update}$$

For the search operation, if the calling VLR contains a location replica, then the local search cost applies; otherwise, the non-local search cost applies. Thus,

$$Hybrid_{i,search} = \begin{cases} \dfrac{1}{rate(\text{LocalSearch})} & if\ enabled(\text{LocalSearch}) \\ \dfrac{1}{rate(\text{NonLocalSearch})} & if\ enabled(\text{NonLocalSearch}) \\ \dfrac{P_h}{rate(\text{LocalSearch})} + \dfrac{P_m}{rate(\text{NonLocalSearch})} & Otherwise \end{cases}$$

where *enabled*(**T**) means that transition **T** is enabled; *rate*(**T**) stands for the rate at which transition **T** fires in the SPN (see Table 3-6 for rates associated with transitions). When *enabled*(**T**) is true it means that the system is in a state in which the event associated with transition ***T*** is occurring. Thus, when *enabled*(**LocalSearch**) is true, it means that the system is in a state in which a local replica exists and consequently the rate at which the system serves the call is *rate*(**LocalSearch**). Conversely, when *enabled*(**LocalSearch**) is false, it means that the system is in a state in which a local replica does not exist and consequently the rate at which the system serves the call is *rate*(**NonLocalSearch**). That is, the system needs to go to the HLR to find the location of the called user.

For the update operation, if the maximum length *K* is reached then the cost of **Reset** applies; else if the movement is a forward movement then the cost of **AddPointer** applies; else if the movement is a backward movement then the cost of **RemovePointer** applies. Thus,

$$Hybrid_{i,update} = \begin{cases} \dfrac{1}{rate(\text{AddPointer})} & if\ enabled(\text{AddPointer}) \\ \dfrac{1}{rate(\text{RemovePoin ter})} & if\ enabled(\text{RemovePoin ter}) \\ \dfrac{1}{rate(\text{Reset})} & if\ enabled(\text{Reset}) \\ \dfrac{P_f}{mark(\text{Forwarding Length}) = K\ ?\ rate(\text{Reset}) : rate(\text{AddPointer})} + \dfrac{P_b}{rate(\text{RemovePoin ter})} & Otherwise \end{cases}$$

Here the syntax **condition?yes-value:no-value** as in programming language C is used.

Finally, we obtain $Hybrid_{cost}$ as follows:

$$Hybrid_{cost} = Hybrid_{search} + Hybrid_{update} / CMR \quad (5)$$

3.3.3.2 $Hybrid_{\text{cost}}^{\min}(N, P, K_{opt})$ Calculation

An MH's profile is characterized by its *CMR* value and (*N,P*) value sets describing the call arrival profile to the MH. For example, if out of the total number of calls received by an MH, 50% comes from VLR 1, 70% comes from VLRs 1 and 2, and 80% comes from VLRs 1, 2 and 3, then it means that one replica (resided in VLR 1) can provide 50% local replica hit, two replicas (resided in VLRs 1 and 2) can provide 70% local replica hit, and three replicas (resided in VLRs 1, 2 and 3) can provide 80% local replica hit. Obviously, the local replica hit ratio will be 0% when the number of replicas *N* is zero under which the hybrid strategy becomes a pure forwarding strategy. In the scenario described above, the *(N, P)* value sets characterizing the call arrival profile would be (1, 50%), (2, 70%) and (3, 80%).

Here we note that a different combination of (*N,P,K*) will result in a different $Hybrid_{cost}$ being calculated based on Equation (5) because $Hybrid_{search}$ and $Hybrid_{update}$ depend on the state probabilities P_i's calculated and P_i's themselves in turn depend on the (*N, P, K*) combination considered. To get the optimal *N* and *K*, we first consider all possible combinations of replica hit ratio *P* and replica number *N*, e.g., with *P* in the range of [0%,100%] in 5% increments, and *N* in the range of [0,10] in increment of 1. Of course *P*=0% when *N=0*. Use the SPN model developed, we can statically obtain the minimum communication costs for possible combinations of *(N,P)* as shown in Table 3-7.

Table 3-7: Minimum Communication Cost.

	0%	P=5%	…	P=100%
N=0	$C_{total}^{\min}(0,0\%, K_{opt})$	-	-	-
N=1	-	$C_{total}^{\min}(1,5\%, K_{opt})$	…	$C_{total}^{\min}(1,100\%, K_{opt})$
…	-	…	…	…
N=10	-	$C_{total}^{\min}(10,5\%, K_{opt})$	…	$C_{total}^{\min}(10,100\%, K_{opt})$

In Table 3-7, $Hybrid_{\text{cost}}^{\min}(N, P, K_{opt})$ represents the minimum communication cost under a specified *CMR* value. The "not applicable" cases are marked with "-." As mentioned, when *N=0* there is no replica and the hybrid strategy proposed degenerates into a pure forwarding strategy.

To obtain $Hybrid_{cost}^{\min}(N,P,K_{opt})$, we utilize the SPN model developed to find the optimal forwarding length K when given N and P. This is achieved by calculating $Hybrid_{cost}$ based on Equation 5 under each possible K value in the range of $[0,n_K]$ as shown in Table 3-8 with n_K representing the maximum allowable forwarding chain length by the system for all MHs.

Table 3-8: Finding $Hybrid_{cost}^{\min}(N,P,K_{opt})$.

K	$Hybrid_{cost}(N,P,K)$
0	$Hybrid_{cost}(N,P,0)$
1	$Hybrid_{cost}(N,P,1)$
…	…
n_K	$Hybrid_{cost}(N,P,n_K)$

We then obtain $Hybrid_{cost}^{\min}(N,P,K_{opt})$ by taking the minimum $Hybrid_{cost}$ among all, i.e., $Hybrid_{cost}^{\min}(N,P,K_{opt}) = \{\min Hybrid_{cost}(N,P,K) \mid K \in [0,n_K]\}$. When $K=0$, the maximum forwarding length is zero and the hybrid strategy degenerates to the pure replication strategy. If both N and K are equal to zero, the hybrid strategy becomes the IS-41 HLR/VLR basic scheme.

Table 3-9: Finding $Hybrid_{cost}^{\min}(1,50\%,K_{opt})$.

K	$Hybrid_{cost}(1,50\%,K)$
0	7.000
1	5.601
2	5.246
3	5.249
4	5.380
5	5.572
6	5.797
7	6.041

As an example, consider (*N,P*)=(1,50%) under *CMR*=1.0. Table 3-9 shows $Hybrid_{cost}$ (1,50%,*K*) values at different forwarding chain length *K*'s. We see in this case, the optimal *K* value is equal to 2 since $Hybrid^{\min}_{\cos t}(1,50\%,K_{opt})$ = $Hybrid_{cost}$ (1,50%,2) = 5.246 is the lowest among all.

3.3.3.3 Using Minimum Communication Cost Table

For a given MH, we can obtain its *CMR* value and call patterns from its profile. Here we present an example of how to use the minimum communication cost table obtained statically to determine the optimal *K* and *N*. Suppose that one MH has *CMR*=1 and its incoming call pattern is as follows:

- *P*=0% when *N=0* (trivial condition)
- *P*=50% when *N*=1.

The MH in this case has only two possible (*N,P*) combinations, so we only need to compare $Hybrid^{\min}_{\cos t}(0,0\%,K_{opt})$ with $Hybrid^{\min}_{\cos t}(1,50\%,K_{opt})$ to determine this MH's optimal *K* and *N*. Follow our last scenario $Hybrid^{\min}_{\cos t}(1,50\%,K_{opt})$ = $Hybrid_{cost}$ (1,50%,2) = 5.246. Suppose that $Hybrid^{\min}_{\cos t}(0,0\%,K_{opt})$ = $Hybrid$ (0,0%,1) = 5.600 and this information is also stored in the table. Now we simply compare the values of $Hybrid^{\min}_{\cos t}(0,0\%,K_{opt})$ and $Hybrid^{\min}_{\cos t}(1,50\%,K_{opt})$. Since $Hybrid^{\min}_{\cos t}(0,0\%,K_{opt}) > Hybrid^{\min}_{\cos t}(1,50\%,K_{opt})$, the optimal *K* is equal to the K_{opt} in $Hybrid^{\min}_{\cos t}(1,50\%,K_{opt})$ which is 2 and the optimal *N* is equal to 1. In other words, for this particular MH it is better that we use hybrid scheme with *N=1* and *K=2* to minimize the cost associated with location update and search operations.

3.3.4 Analysis

In this section, we first discuss the parameterization process, i.e., how to estimate values for the parameters of the SPN model in Figure 3-15. Then we present the analysis results with physical interpretations given.

3.3.4.1 Parameterization

Let the average communication time (single trip) between the HLR and a VLR or between two random VLRs be *T* and the average communication time (single trip) between two neighboring

VLRs be τ. These two parameters can be estimated by considering a network coverage (e.g., hexagonal) model characterizing the underlying wireless network [9]. As a case study, we consider τ/T = 0.3. Also consider the case that the call arrival rate λ is 1.4/hr/MH as in [35]. Therefore, the mobility rate $\sigma = 1.4/CMR$ where *CMR* is given from the MH's profile. Also, assume that we can obtain the local replica hit probability P_h from the MH's profile. The local replica miss probability $P_m = 1 - P_h$. For example, if we consider the combination *(N, P)* = (3, 80%) then $P_h = 0.8$.

The move forward probability P_f and move backward probability P_b are also network structure dependent and their values will be different depending on the network coverage model considered. Suppose that the network structure is again modeled by the hexagonal coverage model and that the MH moves randomly to one of its neighbors with equal probability, i.e., 1/6 for the hexagonal network coverage model. Then, we can calculate these two probabilities as:

$$P_f = 5/6$$

$$P_b = 1/6$$

The communication cost to set up, delete or travel a pointer connection between two neighboring VLRs is:

$$Hybrid_p = cost(src_{VLR} \rightarrow dest_{VLR}) + cost(dest_{VLR} \rightarrow src_{VLR}) = \tau + \tau = 2\tau$$

Therefore, the execution rate μ_p to set up or delete a pointer connection between two neighboring VLRs is:

$$\mu_p = 1/2\tau$$

The communication cost for a call delivery under a replica hit condition is:

$$Hybrid_h = cost(caller_{VLR} \rightarrow first_{VLR}) + cost(travel\_a\_pointer\_connection) \times current\_forwarding\_length + cost(current_{VLR} \rightarrow caller_{VLR})$$
$$= T + 2i\tau + T$$
$$= 2T + 2i\tau$$

Here i is the current forwarding chain length corresponding to the number of tokens contained in place **ForwardingLength**, that is, *mark*(**ForwardingLength**). Therefore, the execution rate for a call delivery operation under a replica hit condition is:

$$\mu_{h,i} = 1/(2T + 2i\tau)$$

The communication cost for a call delivery operation under a replica miss condition is:

$$Hybrid_m = cost(caller_{VLR} \rightarrow callee_{HLR}) + cost(callee_{HLR} \rightarrow first_{VLR}) +$$
$$cost(travel\_a\_pointer\_connection) \times current\_forwarding\_length + cost(current_{VLR} \rightarrow callee_{HLR})$$
$$+ cost(callee_{HLR} \rightarrow caller_{VLR})$$
$$= T + T + 2i\tau + T + T$$
$$= 4T + 2i\tau$$

Here i again is the current forwarding chain length corresponding to *mark*(**ForwardingLength**). Therefore, the execution rate for a call delivery operation under a replica miss condition is:

$$\mu_{m,i} = 1/(4T + 2i\tau)$$

The communication cost for a reset operation is:

$$Hybrid_r = cost(current_{VLR} \rightarrow HLR) + cost(HLR \rightarrow current_{VLR}) + (cost(HLR \rightarrow all\_replicas) +$$
$$cost(all\_replicas \rightarrow HLR))$$
$$= T + T + 2NT$$
$$= 2T + 2NT$$

Here N is the number of replicas. We assume that obsolete pointers will be deleted implicitly, so there is no need to take time to perform the pointer delete operation. Consequently, the execution rate for a reset operation is:

$$\mu_{r,N} = 1/(2T + 2NT)$$

3.3.4.2 Example

In this section, we present a detailed analysis and numerical data obtained for a case study to demonstrate the effectiveness of our approach. The input is the *CMR* value and a series of *(N,P)* value sets characterizing the call arrival profile of an MH. The output is the optimal values of N and K identified by our hybrid scheme. We show that not only we can easily determine the best number of replicas N and forwarding chain length K for minimizing $Hybrid_{cost}$, but also the $Hybrid_{cost}$ value obtained is better than that obtained under either replication or forwarding, as well as that obtained under the *CMR* threshold-based scheme.

Consider an MH's call pattern given as follows:

- P=0% when N=0 (trivial condition)
- P=50% when N=1
- P=70% when N=2
- P=80% when N=3

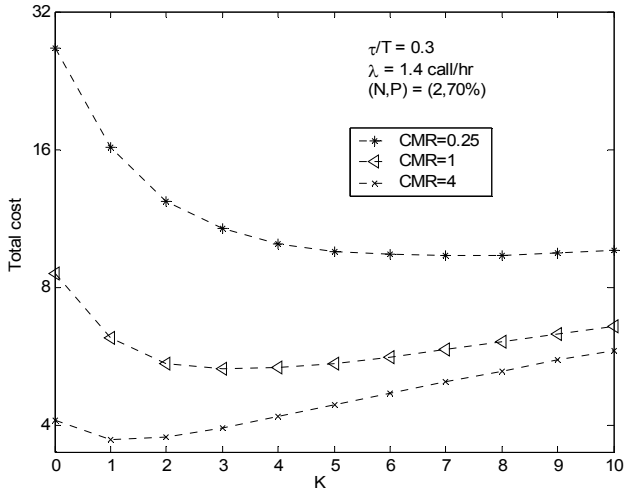

Figure 3-16 : Finding the Optimal *K* under a Constant Number of Replicas *N*.

We first use the SPN model developed to obtain the optimal forwarding length *K* under these 4 given (*N,P*) combinations. Specifically, we use SPNP [59] to solve the SPN model in Figure 3-15 based on Equation (5). Figure 3-16 shows the case in which (*N,P*) = (2,70%) with the *K* value ranging from 0 to 10 under different *CMR* ratios. We normalize the cost in the Y-coordinate with respect to *T*. Thus a ratio of *τ/T=0.3* means that *τ* is set to 0.3*T*. The result shows that the optimal *K* is 7 when *CMR*=0.25, 3 when *CMR*=1 and 2 when *CMR*=4.

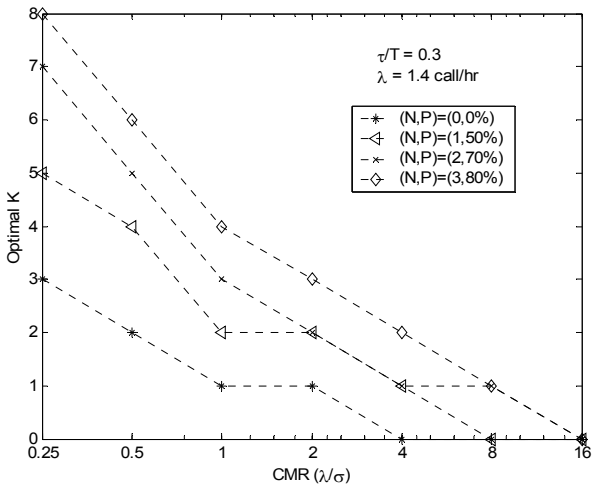

Figure 3-17: Optimal *K* under Different *N* and *CMR* Values.

Figure 3-17 summarizes the effects of *N* and *CMR* on *K*. We first discuss the effect of *CMR* on *K*. As we can see from Figure 3-17, when the *CMR* value becomes larger or, equivalently stated,

when the mobile user is called more often than it crosses VLR boundaries, the system would incur a lower communication cost with a smaller *K* value. On the other hand, with a smaller *CMR* value, the system would perform better with a larger *K* value. For example, with *N* fixed at 2, the optimal *K* is 0 (no forwarding at all) when *CMR* = 16 and then becomes 7 when *CMR* = 0.25. The interpretation of this result is clear: when *CMR* is low, the MH is not called very often relative to its mobility, so it is not judicious to minimize the call delivery cost by performing very costly resetting operations frequently.

Figure 3-17 also shows the effect of *N*. Specifically, when *N* is large, it is better that *K* is a large value; otherwise, it is better that *K* is a small value. For example, with *CMR* fixed at 1 in Figure 3-17, the optimal *K* is 4 when *N* is 3, and then drops to 1 when *N* is 0 (e.g., no replica at all). This result is attributed to the fact that the resetting operation cost depends on *N*. The larger the replica number *N*, the larger the resetting cost. In this case, the forwarding length tends to be large to avoid the high resetting cost.

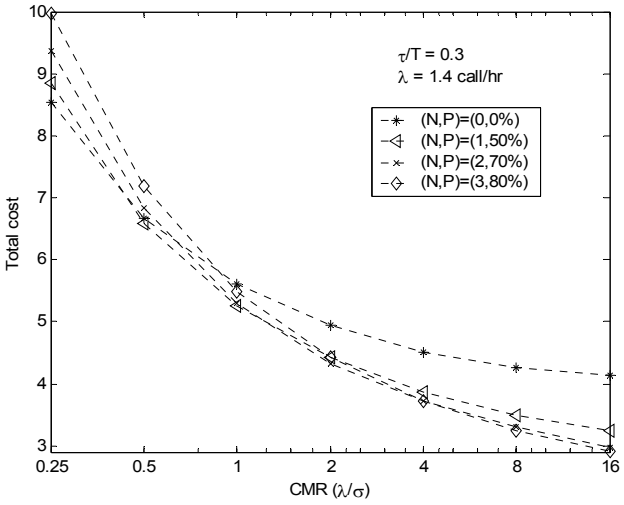

Figure 3-18: Cost versus Number of Replica *N* at Optimal *K*.

Figure 3-18 summarizes the total communication cost incurred due to location update and search operations under different replica numbers *N* at their respective optimal *K* values as determined from Figure 3-17. From Figure 3-18 and with the reference to Figure 3-17, we can easily obtain the optimal *N* and *K* values for the given MH profile. For example, when *CMR*=0.25, the lowest communication cost can be obtained at *N*=0 and *K*=3; when *CMR*=16, the optimal *N* increases to 3 while the optimal *K* decreases to 0. Figure 3-18 also demonstrates the superiority of the hybrid strategy because it encompasses the advantages of both the forwarding and replication strategies.

When the call arrival rate is low compared with the mobility rate (i.e. when *CMR* is low), it behaves like a pure forwarding strategy with N=0. For example, when *CMR*=0.25, (N,P,K_{opt}) = (0,0%,3) for this MH. When the call arrival rate is high compared with the mobility rate (i.e. when *CMR* is high), it behaves like a pure replication strategy with K=0. For example, when *CMR*=16, (N,P,K_{opt})=(3,80%,0) for this MH.

To further demonstrate the performance gain compared with both pure forwarding and replication schemes. We consider a "threshold-based" scheme such that if the user's *CMR* is less than a threshold *CMR* value then the pure forwarding scheme is applied to the user since it is known that the pure forwarding scheme performs excellent under low *CMR* values. On the other hand, if the user's *CMR* value is higher than or equal to the threshold *CMR* value, then the pure replication scheme is used since it is known that the pure replication scheme performs excellent under high *CMR* values. Such a threshold-based scheme has the advantage of simplicity compared with the hybrid scheme at the expense of uniformity.

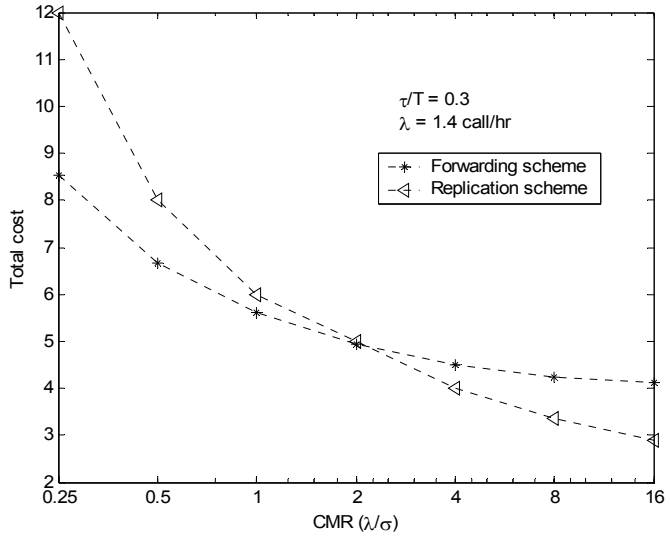

Figure 3-19: Pure Forwarding Scheme versus Pure Replication Scheme.

Figure 3-19 identifies the cross-over threshold *CMR* value (at 2.8) below which the pure forwarding scheme (at *N=0* and the optimal K value) performs better than the pure replication scheme (at *K=0* and the optimal N value) and vice versa for the same example case considered. Such a threshold *CMR* value for the threshold-based scheme again can be determined statically using the methodology described earlier since both pure forwarding and replication schemes are encompassed by the more general hybrid scheme. Figure 3-20 displays the cost difference

between the cost obtained under the threshold-based scheme just described and that obtained under the hybrid replication with forwarding scheme. The curve for the threshold-based scheme in Figure 3-20 is obtained by combining the lower-cost portions of the two curves in Figure 3-19. We see that the hybrid scheme outperforms the threshold-based scheme over the *CMR* range in [0.5, 8] because the hybrid scheme allows the best *N* and *K* values to be identified to minimize the total communication cost due to location management operations. The cost difference is close to zero when the user's *CMR* value is very small or very large because in these cases the hybrid scheme degenerates to a pure forwarding scheme and a pure replication scheme, respectively.

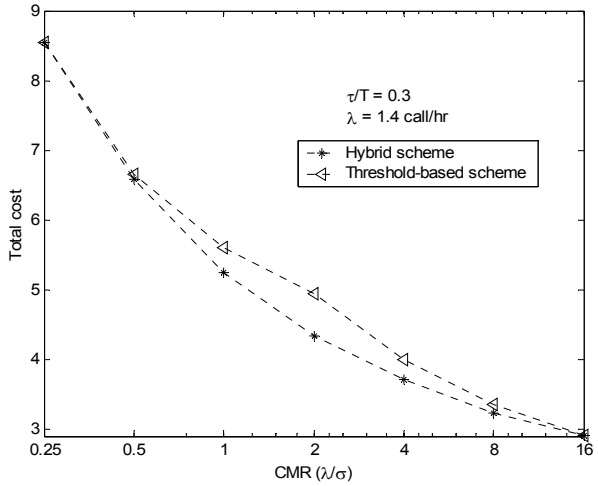

Figure 3-20: Comparing Hybrid versus Threshold-based Schemes.

3.4 Summary

In this chapter, we analyze a class of location management algorithms by quantitatively evaluating the network signaling overhead for each of these algorithms and identifying conditions under which one algorithm may perform better than others. From insight gained from the quantitative analysis, we design and analyze a hybrid replication with forwarding algorithm that outperforms individual algorithms and show that such a hybrid algorithm can be uniformly applied to all mobile users with distinct call and mobility characteristics to simplify the system design without sacrificing performance.

Chapter 4

SERVICE MANAGEMENT

It's envisioned that future personal communications services (PCS) networks will provide a wide range of personalized services, such as personal banking, personalized stock services, location-aware travel advisory [17][24], etc. with most of these services based on the client-server computing paradigm. In this chapter, we investigate the notion of *location-aware*, *mobile* service proxies to interact with services on behalf of the mobile user to reduce the network signaling and communication cost.

4.1 Proxy-based Architecture

In our proxy-based architecture, a mobile host communicates with the application server via a proxy or agent. A proxy or agent is a surrogate of the client residing in the fixed network. The proxy performs tasks such as location tracking, accepting communications on behalf of the user, converting the communications into different application formats, and forwarding the resulting communication to the mobile user. This alleviates the impact of limited bandwidth and poor reliability of the wireless link by continuously maintaining the client's presence in the fixed network via the proxy.

The placement of the proxy in the fixed network is crucial. Placing the proxy at the edge of the fixed network, i.e., at the base station, has some advantages [67] since it is easier to gather information regarding wireless network characteristics so proper adaptation can be applied. Also personalized information about the mobile host is available locally, thus allowing efficient data packet delivery from the application server to the proxy and then from the proxy to the mobile host. However, since all communications to the mobile user must go through the personal proxy, a "static" proxy may utilize inefficient routes between the service and the mobile user once the user moves, thus incurring a high network signaling and communication cost to the PCS system. Our proxy-based architecture advocates mobile proxies to address this problem.

4.2 Personal Proxy-based Location-aware Service Management

We investigate service management schemes based on location-aware "mobile" proxies with the objective to reduce the network cost for client-server personalized services in future PCS networks. Our approach is based on the notion of personal proxy, that is, the proxy is created on a per-user basis; however, our personal proxy performs location tracking by cooperating with the underlying location management system with the objective to process service management operations efficiently.

To remedy the problem of inefficient routes, we consider the design of moving the personal proxy with the mobile user during location handoffs to minimize the network signaling cost. How often we move the personal proxy, that is, how often we perform *service handoffs*, depends on the user profile. We investigate the notion of "personal proxy service area." A fast-moving mobile user with low packet rate may require a large proxy area, while a slow–moving user with high packet rate may require a small proxy area.

We also differentiate an *aggregate* proxy from a *per-service* proxy. Both types of proxies are created on a per-user basis. However, the former is an aggregate proxy that interfaces with all mobile services that the mobile user concurrently engages, while the latter only interfaces with a specific mobile service, that is, a proxy is created for each service accessed by the mobile user. For the case in which the user only engages with one mobile service, the aggregate proxy degenerates into the per-service proxy. Our per-service proxy performs *service handoff* optimally based on the specific characteristics of the mobile service involved in order to minimize the network signaling cost.

The *service handoff* refers to the process of moving the personal proxy (aggregate or per-service) as a user moves from one *service area* into another *service area* so as to move the proxy closer to the mobile user to reduce the network communication cost. The cost associated with a service handoff includes a reconnection cost to setup a new connection from the new proxy to the server and a context transfer cost to transfer service context from the old proxy to the new proxy. We aim to design and validate location-aware mobile service management schemes based on intelligent proxies that can optimally determine if a service handoff should occur during a location handoff as the user moves such that the overall signaling and communication cost incurred to the network is minimized.

We follow the assumption in [24] for the overhead involved in performing a service handoff, namely, a reconnection cost and a service context transfer cost. The physical reconnection cost refers to the communication cost for the proxy to inform the server of the new network address (and session reestablishment for connection-oriented services such as those based on TCP), while the service context transfer cost refers to the communication cost to move the service context with the moving proxy. The amount of service context information is application-dependent and may include both static context information (such as user profile and authentication information) and dynamic context information (such as files opened, objects updated, locks and time-stamps, and status of execution). In addition to the service handoff cost incurred when the proxy moves across a service area, there is also a cost of informing the proxy of the location of the mobile user when the mobile user crosses a cell regardless of whether the proxy performs a service handoff or not. The sum of these costs will be called the "proxy maintenance cost" in the dissertation. Our scheme aims to find the optimal proxy service area, when given a set of parameter values characterizing the network and workload conditions, such that the overall communication cost (including the proxy maintenance cost and packet delivery cost) is minimized.

Table 4-1: Parameters in Personal Proxy-based Scheme.

Parameter	Meaning
γ	The aggregated service request rate, i.e. the data packet delivery rate for all services currently accessed by a mobile user.
σ	The average rate at which the mobile user moves across cell boundaries.
SMR	Service request to mobility ratio, e.g., γ/σ.
T	The average communication cost between a proxy and a server per packet.
τ	The average communication cost between two neighboring cells per packet.
α	The reconnect parameter, i.e., the communication cost parameter to setup a new connection with the server when the personal proxy moves. For example, for an application on TCP, α is the number of messages to tear down the old TCP connection, setup a new TCP connection and any application specific messages in a service handoff.
β	The context transfer parameter, i.e., communication cost parameter to transfer context when the personal proxy moves.
C_{pt}	Service handoff cost, including the connection setup cost and context transfer

	cost, i.e. $\alpha T + \beta N \tau$, where N is the distance factor between two proxies.
γ_i	The service request rate for a particular service, i.e., the data packet delivery rate.
α_i	The service-specific communication cost parameter related with the physical connection with the server when the service proxy moves.
β_i	The service-specific communication cost parameter related with context transfer when a service proxy moves.

System parameters that characterize the network and user workload condition of a PCS system are summarized in Table 4-1. Here we note that three sets of parameters are considered, namely, user parameters (e.g. σ), application-specific parameters (e.g., γ, α, β, γ_i, α_i, β_i) and network parameters (e.g. T, τ).

4.3 Operation of Personal Proxy Schemes

In this section, we first describe a service management scheme based on the notion of aggregate personal proxy, that is, a single personal proxy is used for all services engaged by the mobile user. Then we extend it to the case of per-service personal proxy for performance optimization. Later, we will develop analytical model to analyze their performance characteristics.

4.3.1 Aggregate Personal Proxy Scheme

Under the aggregate personal proxy scheme, each mobile user on power up creates a client-side personal proxy that acts on behalf of the mobile user. Initially the aggregate proxy will reside in the base station of the cell in which the mobile user resides. All messages exchanged between the client and any service will go through the personal proxy. The personal proxy performs tasks such as location tracking (through event notification services from the underlying location management system), accepting user requests to access services, converting communication data in different application formats, and forwarding data packets to the mobile user. There is only a single proxy regardless of the number of services engaged by the user. All servers at all times only know the personal proxy. The personal proxy may move when the user moves across a cell boundary if justified, in which case the proxy will move from the base station it had resided to the base station which the mobile user just entered. When a proxy moves, a service handoff cost

incurs for reestablishing the connection and transferring service context. In return, the proxy is moved closer to the mobile user so the packet delivery cost from the proxy to the mobile user is reduced. Thus, there exists a tradeoff between the cost incurred due to moving the proxy versus the cost saved due to close proximity between the proxy and the mobile user.

The aggregate personal proxy scheme has its root derived from the notion of "local anchor" (LA) proposed by Ho and Akyildiz [23] in the context of location management. The basic idea is that within a personal proxy service area, we use the user's personal proxy to keep track of the location of the mobile user within the area. The underlying location management system informs the proxy whenever the mobile user crosses a cell boundary, so the proxy at all times knows the current cell of the user. As a personal proxy area normally covers a large geographic region spanning several cells, so when a mobile user crosses a cell boundary it may be still within the same service area. In this case, the personal proxy stays in the same location without moving with the mobile user. On the other hand, if the mobile user moves out of the current service area into another service area upon a cell boundary crossing, then the proxy will move with the mobile user into the new area. In this latter case, in addition to a cost incurred for the underlying location management system to inform the proxy of the location change of the mobile user, there is also a service handoff cost to inform the server of the network address of the new proxy and to transfer the service context to the new proxy.

In addition to keeping service context information for each service accessed by the mobile user, a mobile user's personal proxy also keeps the mobile user's statistics information, such as the mobility rate, the packet rate for each service, and characteristics of services currently accessed by the mobile user to determine the optimal service area. Upon being informed of the new location of the mobile user when the mobile user moves into a new cell, the proxy will check if the service area is crossed. If yes, after the proxy moves into the new service area, a new optimal personal proxy service area size can be determined by executing a computational procedure developed based on the up-to-date statistics information maintained.

It should be noted that in the aggregate personal proxy scheme, there is only a single user proxy that acts on behalf of the user in the fixed network serving as the client-side agent for all services engaged by the mobile user. As a result, the optimal personal proxy service area determined by the proxy to minimize the network signaling and communication cost will be based on the

aggregate service characteristics exhibited by all services, e.g., an aggregate packet rate, as the service area determined by the proxy will apply to all services engaged by the mobile user.

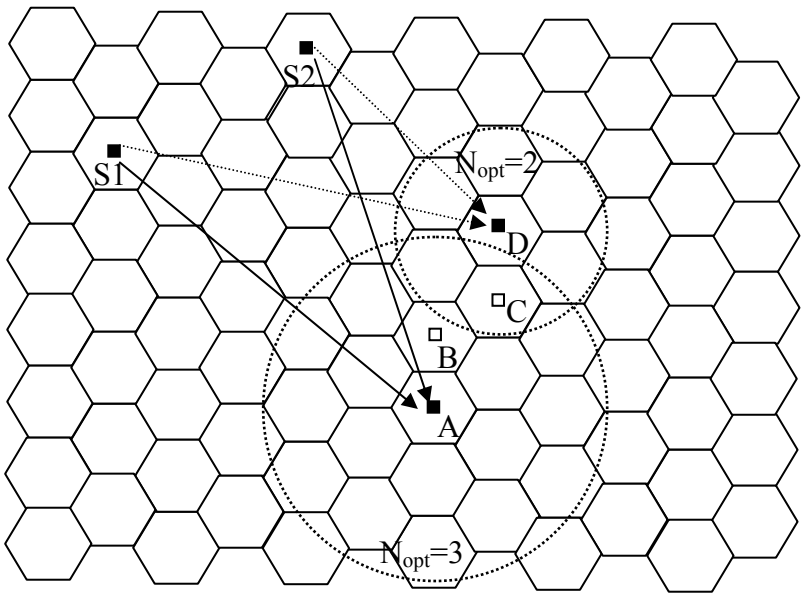

Figure 4-1: Aggregate Personal Proxy Scheme.

Figure 4-1 illustrates a scenario in which a mobile user moves under the aggregate personal proxy scheme in a flat PCS network. Initially the optimal proxy area size is determined to be $N_{opt} = 3$ and a mobile user resides in cell A together with the proxy who always resides in the center cell of the service area. When the mobile user makes a move from cell A to B, the proxy in cell A is notified. A similar location management operation is performed when the mobile user subsequently moves from cell B to C. In the meantime, all packets from S1 and S2 to the mobile user will be delivered to the proxy in A first, and then forwarded to the mobile user by the proxy. When the mobile user moves to D, which is outside of the proxy's service area, the proxy, along with the services context, is moved to D, triggering a service handoff to inform all services (S1 and S2) of the network address of the proxy (now in D) and to transfer context information from cell A to cell D. Depending on the current state information, the new proxy service area may or may not be the same as before. It will be determined dynamically by the proxy after a service handoff. Figure 4-1 shows that the new proxy service area size is $N_{opt} = 2$ after the proxy moves to D.

4.3.2 Per-Service Personal Proxy Scheme

Unlike the aggregate proxy scheme where the optimal proxy service area is determined based on aggregate characteristics of services being accessed by the mobile user, the per-service personal proxy scheme creates a separate proxy for each client-server application engaged by the mobile user. Each proxy created is application-specific and, since it knows specific service characteristics of the application, can optimally determine the best service area for the application. For example, a multimedia application with a small service handoff cost may dictate a different optimal proxy area from the one having a low speed data service with a high service handoff cost. The disadvantage of the scheme is a processing overhead added to the mobile user since each mobile user needs to keep a list of proxies for multiple services that it is currently accessing. The advantage in return is that each service can have its own service-tailored optimal proxy service area, thus collectively reducing the overall network signaling and communication cost compared with the aggregate personal proxy scheme.

Each personal proxy behaves the same as the one in the aggregate scheme except that it only maintains its own service-specific context and statistics information. Note that it is possible that different proxies may have different optimal proxy service areas since in general different services exhibit different service characteristics.

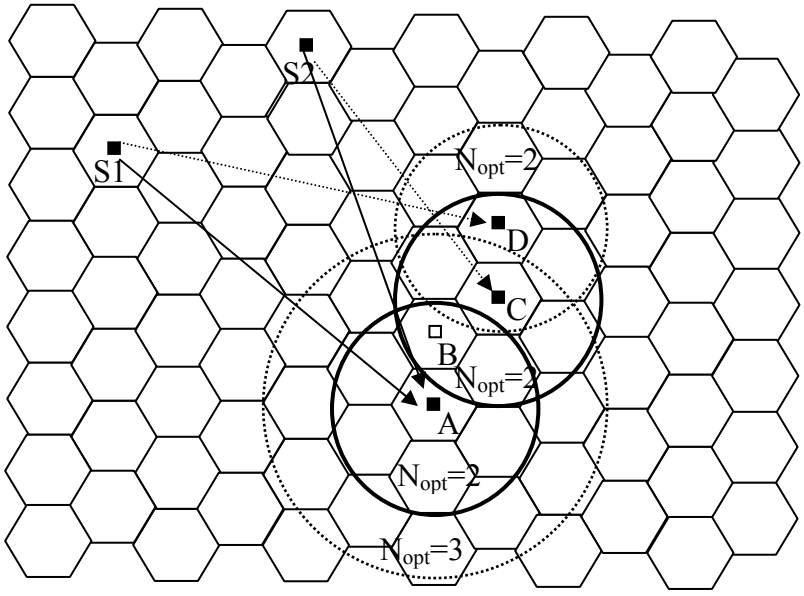

Figure 4-2: Per-Service Personal Proxy Scheme.

We illustrate a user movement scenario under the per-service personal proxy scheme in Figure 4-2. Initially the optimal proxy service area sizes for S1 and S2 are N_{opt} =3 and 2, respectively. The service area for S1 is marked dashed, while that for S2 is marked solid. The mobile user initially resides in cell A together with the two proxies of S1 and S2. When the mobile user makes a move from A to B, both proxies are notified of the movement. When the mobile user subsequently moves to C, it is still inside S1's proxy service area but outside of S2's proxy area. Thus, a service handoff for S2 is triggered, after which the proxy of S2 moves to cell C and a new proxy area size is determined (still 2 in the diagram). When the mobile user moves to D, a service handoff for S1 is triggered, after which the proxy of S1 moves to D and a new optimal proxy service area size is calculated (2 in the diagram). The proxy for S2 remains in C since the current location of the mobile user, namely, D, is still within S2's service area. All packets from S1 and S2 to the mobile user are delivered to their respective proxies who in turn forward them to the mobile user.

4.4 Performance Model

In this Section, we develop analytical models for evaluating the aggregate and per-service personal proxy schemes introduced in previous section. We first define the performance metric used as the basis for evaluation. Then, we show how the performance metric can be assessed through our analytical model.

4.4.1 Performance Metrics

Our performance metric used for evaluating location-aware mobile-proxy-based service management schemes is based on the *total communication cost per time unit* for the network to serve service management operations. Specifically, our performance metric consists of two cost parameters:

- Proxy maintenance cost C_{move} – this includes the cost for tracking the location of the mobile user and the service handoff cost for moving the proxy to stay closer to the mobile user if necessary as a mobile user moves across a cell boundary.
- Service packet delivery cost $C_{service}$ – this is the cost to deliver data packets to the mobile user.

Note that the above two cost parameters refer to the average cost. Let C_{total} be the average cost of the PCS network in servicing the above two types of operations per time unit. Then, our

performance metric C_{total}, defined as the cost incurred to the PCS network per time unit for servicing service management operations, is given by:

$$C_{total} = C_{move} \times \sigma + C_{service} \times \gamma$$

Here σ and γ are the mobile user's cell boundary crossing rate and service request rate, respectively, as described in Table 4-1.

The communication cost between a mobile user and its proxy is assumed proportional to the separating distance. The proxy is always located at the center of the layered structure as shown in Figure 2-2, so the distance between a mobile user in layer *i* and the proxy (located in layer *0*) is exactly *i* cells apart. Note that an *n*-layer structure contains layers *0* to *n-1*.

4.4.2 Model for Aggregate Personal Proxy Scheme

For the aggregate personal proxy scheme, a Stochastic Petri Net (SPN) model as shown in Figure 4-3 is developed to analyze its behavior.

Table 4-2 gives the meanings of places and transitions defined in the SPN model. Here *mark(***p***)* returns the number of tokens held in place **p**.

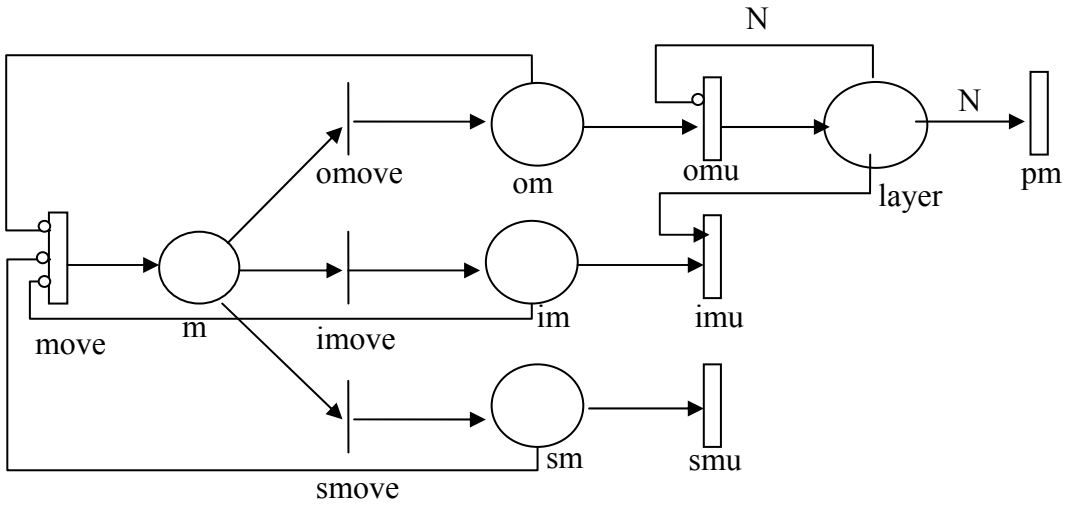

Figure 4-3: SPN Model for the Aggregate Personal Proxy Scheme.

Table 4-2: Meaning of Places and Transitions in the SPN Model.

Symbol	Meaning
m	*mark*(**m**) indicates that the mobile user has just moved across a cell boundary.
om	*mark*(**om**) indicates that the mobile user has just moved outwards, i.e., from layer i to layer i+1.
im	*mark*(**om**) indicates that the mobile user has just moved inwards, i.e., from layer i to layer i-1.
sm	*mark*(**om**) indicates that the mobile user has just moved, but still remains in the same layer.
layer	*mark*(**layer**) indicates the layer number at which the mobile user currently resides; it also is the distance between the mobile user and its proxy located at layer 0.
move	A timed transition representing a cell boundary crossing with a rate of σ.
omove	An immediate transition following a move event with P_{omove} representing the probability that the movement is an outward movement.
imove	An immediate transition following a move event with P_{imove} representing the probability that a movement is an inward movement.
smove	An immediate transition following a move event with P_{smove} representing the probability that a movement is an inside-the-same-layer movement.
omu	A timed transition to service an outward movement.
imu	A timed transition to service an inward movement.
smu	A timed transition to service an inside-the-same-layer movement.
pm	A timed transition to service a proxy transfer.

The SPN model describes the behavior of the mobile user in a PCS system operating under the aggregate personal proxy scheme for which the personal proxy area is an n-layer structure in the PCS network. It is constructed as follows:

- When a mobile user moves across a cell boundary, a token is placed in place **m**.
- If the movement is an "outward movement" (i.e. the user moves from layer *i* to layer *i+1*) with probability P_{omove}, transition **omove** will consume the token immediately, after which a token will be placed in **om,** which subsequently disables transition **m** and enables transition **omu.** This indicates that a local proxy update operation is being performed by the underlying location management system to inform the proxy of the current location of the mobile user. After that, a token is placed in place **layer**.
- If the token number in place **layer** is equal to *N,* it means that the mobile user has just moved out of the personal proxy area, in which case a service handoff occurs and the personal proxy along with the service context will move to a new proxy area centered at the cell which the mobile user just entered into.

- If the movement is an "inward movement" (i.e. the user moves from layer *i* to layer *i-1*) with probability P_{imove}, then transition **imove** will consume the token immediately, after which a token will be placed in **im** which subsequently disables transition **m** and enables transition **omu,** thus triggering a local proxy update operation to be performed. After that, one token in place **layer** will be consumed, meaning that the layer number has been reduced by 1 as a result of the inward movement. Note that there must exist at least a token in place **layer** when an inward movement occurs.
- If the movement is an "inside-the-same-layer movement" (i.e. the user moves from one cell to another cell in the same layer i) with probability P_{smove}, then transition **smove** will consume the token immediately, after which a token will be placed in **sm** which subsequently disables transition **m** and enables transition **smu**, representing that a local proxy update operation is being performed. After that, the token in place **sm** is consumed while the number of tokens in place **layer** remains the same, meaning that the layer number is not changed (since the mobile user stays at the same layer) as a result of this movement.

Note that there is no service request being modeled in the SPN. The reason is that place **layer** keeps track of the current status, i.e. the current layer number a mobile user locates, and a packet delivery cost is only dependent on this status. Thus we are able to calculate the service request cost without having to model the service request behavior explicitly.

Suppose the personal proxy area size is *N* (from layer *0* to layer *N-1*). Let $\boldsymbol{P_i}$ be the steady state probability that the system is found to contain $\boldsymbol{i}$ tokens in place **layer.** Let $\boldsymbol{\omega}$ be the steady state average number of tokens found in place **layer**. Then the service packet delivery cost $C_{service}$ per request can be calculated by:

$$C_{service} = \sum_{i=0}^{N} P_i \times C_{i,service} = \sum_{i=0}^{N} P_i \times (T + \tau \times i) = \sum_{i=0}^{N} P_i \times T + \sum_{i=0}^{N} P_i \times (\tau \times i) = T + \tau \times \omega \quad (6)$$

where $C_{i,service}$ is the service packet delivery cost per service request when the mobile user is in layer *i* and is equal to the communication cost between the proxy and server (T) plus the communication cost ($\tau \times i$) between the proxy and the mobile user which are *i* cells apart in distance. Similarly, let C_{move} be the proxy maintenance cost per move, including the cost to inform the proxy of the current location of the mobile user and possibly a service handoff if the proxy moves. We have:

$$
\begin{aligned}
C_{move} &= \sum_{i=0}^{N} P_i \times C_{i,move} \\
&= P_0 \times \tau + \sum_{i=1}^{N-2} P_i \times (\tau \times (i+1) \times P_{omove} + \tau \times i \times P_{smove} + \tau \times (i-1) \times P_{imove}) \\
&+ P_{N-1} \times (C_{pt} \times P_{omove} + \tau \times (N-1) \times P_{smove} + \tau \times (N-2) \times P_{imove}) + P_N \times C_{pt}
\end{aligned} \tag{7}
$$

where C_{pt} is the service handoff cost, including the context transfer cost and connection re-establishment cost with remote servers. We assume it has the form $\alpha T + \beta N\tau$ as described in Table 4-1, with α, β being parameters dependent on services characteristics. The total cost per time unit incurred to PCS network under personal proxy scheme, C_{total}, then can be calculated by:

$$C_{total} = C_{service} \times \gamma + C_{move} \times \sigma \tag{8}$$

4.4.3 Model for the Per-Service Personal Proxy Scheme

In the per-service personal proxy scheme, a proxy is used for each service accessed by a mobile user. We can use the same performance model in Figure 4-3 to analyze the scheme with some adjustment made on service parameters values to account for the fact that each service has its set of service parameters. These parameters include service-specific request rate γ_i and context transfer cost parameters α_i and β_i (see Table 4-1). Specifically, for each service accessed by the mobile user we will analyze its behavior separately utilizing the performance model shown in Figure 4-3. Thus, the performance model will be utilized as many times as the number of services concurrently accessed by the mobile user to obtain the cost incurred due to service management activities.

Recall that the advantage of the per-service personal proxy scheme over the aggregate personal proxy scheme is that optimizing conditions in terms of the best personal proxy areas to reduce the per-service communication cost can be separately determined for different services concurrently accessed by the mobile user. Thus in calculating the per-service cost, the parameters in equation (6), (7) and (8) will be service-specific, e.g., replacing γ, α and β by γ_i, α_i and β_i respectively. Suppose we have M services concurrently being accessed by the mobile user. Then the overall cost incurred will be:

$$C_{total} = \sum_{i=1}^{M} C_{total}(service_i)$$

where C_{total} $(service_i)$ is calculated from Equation (8) above for each separate service with γ, α and β being replaced by γ_i, α_i and β_i respectively.

4.5 Analysis

In this section, we use SPNP [59] as a tool to define and evaluate the SPN models developed to yield numerical results with physical interpretations given.

4.5.1 Computational Procedure for Calculating C_{total}

To calculate the total communication cost C_{total} based on Equations (6), (7) and (8), we need to obtain the steady state probability P_i, that i tokens are found in place **layer,** and the steady-state average number of tokens in place **layer**, ω. SPNP [59] was used to help obtaining these when given a set of parameter values characterizing the network and workload conditions. Specifically, we used the following reward assignment to calculate P_i:

$$r_i = \begin{cases} 1 & if\ mark(ring) = i \\ 0 & otherwise \end{cases}$$

In effect, this will calculate the *average reward* weighted by the state probabilities, which in this case, is exactly the probability that i tokens are found in place **layer**.
To calculate ω, we used the following reward assignment: $r = mark($**layer**$)$.

4.5.2 Numerical Data

We report numerical data to (a) show that there exists an optimal personal proxy area in our proposed service management schemes based on location-aware personal proxies to minimize the overall network signaling and communication cost, (b) compare our proposed schemes with non-proxy-based schemes in the PCS system and (c) study the effects of certain model parameters, including the *SMR* and context transfer parameters, on the optimal personal proxy area size. The

numerical data are obtained by using SPNP as a tool to define and evaluate the SPN models developed following the reward assignment process explained in Section 4.5.1.

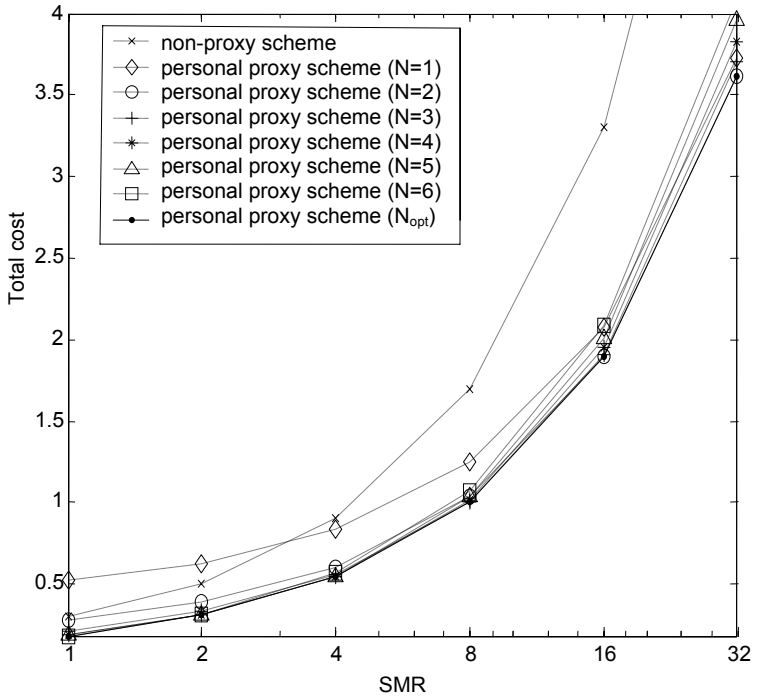

Figure 4-4: Comparison of Proxy-Based vs. Non-Proxy Service Management.

We first compare location-aware personal proxy schemes with non-proxy service management schemes as a function of model parameters to analyze conditions under which, if any, non-proxy can perform better than proxy-based schemes. Figure 4-4 compares the cost of non-proxy scheme and personal proxy-based schemes under varying *SMR* (i.e., γ/σ) values with mobility rate $\sigma=0.1$ and proxy-move parameters $\alpha=4$ and $\beta=2$. The effect of other proxy-move parameters will be shown later. The top curve shows the total cost obtained under the non-proxy scheme. The bottom solid-line curve shows the total cost obtained under the location-aware proxy scheme operating at optimizing proxy service areas (that is, at N_{opt}) as identified from our model. There are several middle curves in between these two curves showing the total cost obtained at various proxy service areas. Of particular interest is the case when N=1 for which the proxy always moves with the user whenever the user moves across a cell boundary.

We observe that the non-proxy scheme possibly could perform better than the proxy-based scheme under low *SMR* ratios and large proxy areas (that is, large N). However, if the proxy service area is optimally selected at N_{opt}, the proxy-based scheme always performs better than the

non-proxy scheme. Further, the advantage of the proxy-based scheme becomes more and more pronounced with the increase of *SMR*. The reason is that when *SMR* is low, the packet arrival rate is low compared with the user mobility rate; thus, the service packet delivery cost incurred in the non-proxy scheme due to triangular routing in servicing packets is minimal. This factor, when coupled with a large service area which incurs a large service handoff cost in the proxy-based scheme, can make the non-proxy scheme perform better than the proxy-based scheme in terms of the service management cost incurred to the network. On the other hand, as *SMR* increases the high service packet delivery cost due to the triangular routing in the non-proxy scheme dominates the proxy maintenance cost, making non-proxy schemes perform worse than proxy-based schemes, regardless of the service area in the proxy-based scheme in this case.

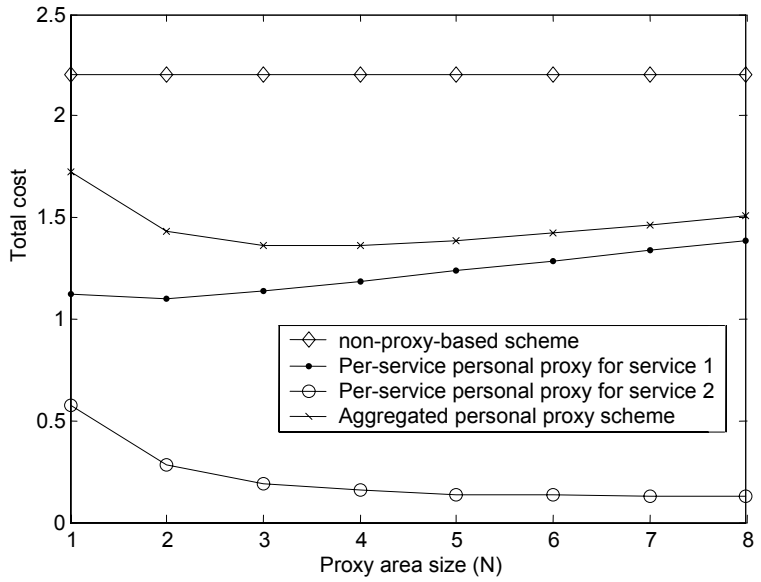

Figure 4-5: Total Cost under Different Proxy Area Sizes.

The above analysis is based on the aggregate proxy-based service management schemes where there is a single service or there are multiple services but only one proxy is used to interface with all services taking the aggregate characteristics into consideration. Below we construct a case study to compare aggregate vs. per-service proxy-based service management schemes.

Figure 4-5 shows the effect of the size of the proxy service area (an *n*-layer area) on the cost incurred to the system due to service management operations under the proxy schemes, with $\tau = 0.1T$ and $\sigma=0.1$. We consider the case in which there are two services being accessed by a mobile user concurrently. One service is a UDP-like application with packet delivery rate $\gamma_1 = 1$

packet/second, and proxy-move cost parameters $\alpha_1 = 1$, $\beta_1 = 1$. Another is a TCP-like application with packet delivery rate $\gamma_2 = 0.05$ packet/second, and proxy-move cost parameters $\alpha_2 = 5$, $\beta_2 = 2$. Thus the aggregate $\gamma = \gamma_1 + \gamma_2 = 1.05$, $\alpha = \alpha_1 + \alpha_2 = 6$ and $\beta = \beta_1 + \beta_2 = 3$.

We also consider the cost of a non-proxy-based scheme for which the HLR is being informed of the new network address of the mobile user whenever the mobile user moves across a cell boundary (with a cost of T per move), and a triangular routing is incurred for packet delivery as in Cellular Digital Packet Data (CDPD) systems, that is, each packet to the mobile host will travel from the server to the HLR (with cost T) and then from the HLR to the current address (with another cost T). Consequently for non-proxy-based scheme, the total cost is $T \times \sigma + (T+T) \times \gamma = 2.2T$ and remains a constant with the change of N.

From Figure 4-5 we first see that there exists an optimal proxy service area size that highlights the tradeoff between proxy maintenance cost and service packet delivery cost. On the one hand, with the increase of N, and thus a larger service area, the service packet delivery cost is higher due to the higher communication cost from the proxy to the current location of mobile user. On the other hand, with a larger service area, the proxy maintenance cost is reduced more because a user movement crossing a cell boundary is more likely to be within the same service area and thus the proxy needs not to be moved, thus resulting in a lower proxy maintenance cost. The optimal value is reached when N=4 for the aggregate personal proxy scheme, at which the network signaling and communication cost incurred to the PCS network is minimized while maintaining the required service and location functionality. For the per-service personal proxy scheme, service1 and service2 have the optimal N values at 2 and 7, respectively.

Figure 4-5 also demonstrates the superiority of the per-service personal proxy scheme over the aggregate personal proxy scheme. The total costs incurred for service1, service2 and the aggregate service under optimal proxy area sizes are 1.0998, 0.1294 and 1.3624, respectively. Thus, the cost is reduced by (1.3624-1.0998-0.1294)/ 1.3624=9.8% in the per-service personal proxy scheme compared to the aggregate personal proxy scheme.

Lastly, we observed that the improvement of the per-service personal proxy scheme over the aggregate personal proxy scheme is more pronounced when the services characteristics (e.g. packet rate, context transfer cost, etc.) of multiple services accessed by the mobile user are dramatically distinct, because otherwise the aggregate service characteristics would be close to

those of individual ones and the optimal service area found by the aggregate scheme would be close to those individually found by separate services, making the performance behavior virtually the same between these two schemes.

4.6 Summary

In this chapter, we propose and analyze mobile service management schemes based on location-aware proxies with the objective to reduce the network signaling and communication cost in future personal communication systems (PCS). Under these schemes, a mobile user uses personal proxies as intelligent client-side agents to communication with services engaged by the mobile user. A personal proxy cooperates with the underlying location management system so that it is location-aware and can optimally decide when and how often it should move with the roaming user. We show that, when given a set of model parameters characterizing the network and workload conditions, there exists an optimal proxy service area size for service handoffs such that the overall network signaling and communication cost for servicing location and service operations is minimized. We demonstrate via Petri net models that our proposed proxy-based mobile service management schemes outperform non-proxy-based schemes over a wide range of identified conditions. Further, when the mobile user is concurrently engaged in multiple services, the per-service proxy scheme that uses a separate proxy for each service outperforms the aggregate proxy scheme that uses a single proxy to interface with multiple services taking their aggregate service characteristics into consideration.

Chapter 5

INTEGRATED LOCATION AND SERVICE MANAGEMENT

In this chapter we investigate the notion of integrated location and service management in personal communication service (PCS) networks with the objective to reduce the *overall* network communication cost of servicing mobility-related and service-related operations. Here the *overall* cost includes the location management cost for explicitly maintaining the user's location database to service location lookup/update operations, as well as the service management cost for maintaining the service proxy and delivering service packets. Earlier in Chapter 4, we considered the use of personal proxies to interact with services with the objective to minimize the cost associated with service delivery without considering the cost for location management. In this chapter we explore the design concept of co-locating an MH's location database (for location management) with the MH's service proxy (for service management) to tightly integrate location management with service management.

5.1 Co-Locating Service Proxy with Location Database

Our approach follows the per-user design concept taking into consideration of per-user specific mobility and service characteristics. Each user has a location database to track its location. For PCS cellular systems, this location database can be in the form of a single entry stored at the HLR directly pointing to the current VLR under which the user resides (as in the basic HLR/VLR scheme discussed in Chapter 3), a forwarding chain stored at VLRs on the chain (as in FRA) eventually pointing to the current VLR, or a double-pointer with the HLR pointing to a local anchor which in turn points to the current VLR (as in LAA). Whenever there is a call looking for the mobile user, the database is queried to locate the user to establish the connection. The cost associated with maintaining the database and servicing location update/search operations is referred to as the location management cost.

Our notion of integrated location and service management is based on the concept of using a per-user service proxy as a gateway between the MH and all client-server applications engaged by the MH concurrently. All user requests and server replies would pass through the proxy. If the

backend server is replicated, e.g., for multimedia streaming applications [57][60], the server may change its location for load balance and performance reasons, in which case the server would inform the proxy of its location change without involving the MH.

A feature of our integrated location and service management scheme is that we always co-locate the MH's service proxy with the MH's location database that stores the current location of the MH. The service proxy knows the current location of the MH all the time so as to eliminate the cost associated with tracking the user location on behalf of the services for data delivery. In the PCS network, whenever the MH moves across a registration area boundary, a location handoff occurs for the location management system to update the location database. In our integrated location and service management scheme, associated with a location handoff is a service handoff[3] to update the service proxy. If a location handoff results in moving the MH's current location database to stay closer to the MH (e.g., as in the LA scheme), then the associated service handoff will also move the service proxy to the same location. Whether the MH's service proxy should move with the MH as the MH crosses VLRs in the PCS network depends on the specific integrated location and service management scheme employed. An integrated scheme that frequently moves the proxy would have the advantage of low-cost service and call management because of the proximity of the service proxy with the MH at the expense of high-cost location management, and vice versa.

5.2 Integrated Location and Service Management Schemes

Four integrated location and service management schemes are investigated and analyzed, i.e., centralized, fully distributed, dynamic anchor, and static anchor. We describe the operational procedures used to handle location update, call delivery, and service requests in these four schemes. One should note that the best integrated scheme is selected on a per-user basis for network cost minimization, not to be affected by other users in the system.

[3] A service handoff refers to the event that the MH crosses a service area, which in our case would coincide with a location registration area. Note that we do not consider fully replicated servers in service areas, which we consider would be difficult to deploy due to its large economical scale. Therefore a service handoff in our schemes actually involves moving the MH's service proxy from one service area to another area.

5.2.1 Centralized Scheme

Under the centralized scheme, the location management operations are handled as the basic HLR/VLR scheme and the service proxy is centralized and “co-located" with the HLR, so as to avoid extra costs to locate the MH by the service proxy when forwarding server responses to the MH. A location update operation to the HLR is performed when an MH moves across a VLR boundary. A search operation at the HLR database is performed when a call is placed to the MH. A service operation involves a request/reply communication cost between the MH and the server through the service proxy.

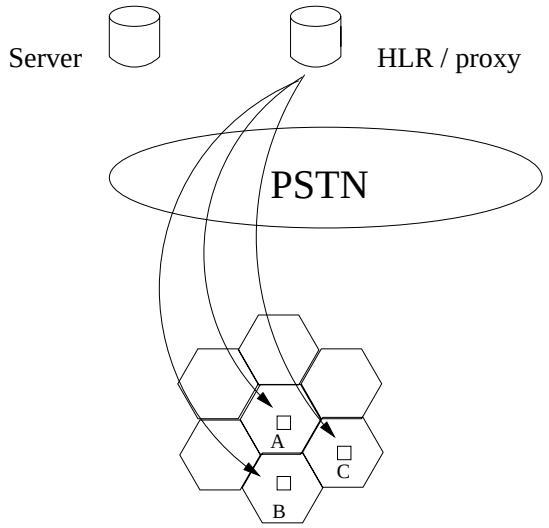

Figure 5-1: Centralized Scheme.

We illustrate the centralized scheme in Figure 5-1. As the MH moves from VLR A to VLR B and subsequently to VLR C, the HLR and the service proxy are updated to point to VLR B and then to VLR C sequentially.

5.2.2 Fully Distributed Scheme

Under the fully distributed scheme, both the location and service handoffs occur whenever the MH moves into a new VLR. The location handoff behaves the same as the basic HLR/VLR scheme. The service handoff migrates the service proxy along with the service context to the new serving VLR that the MH just enters into. Thus, the service proxy is always located at the MH’s current VLR pointed to by the HLR.

We illustrate the fully distributed scheme in Figure 5-2. When the MH moves from VLR A to VLR B, the service proxy migrates from VLR A to VLR B, and the HLR and the server are updated to point to VLR B. The subsequent move to C behaves similarly. To service a location search request (not initiated from the current VLR), the HLR database is accessed first to know the current VLR (A, B, or C) and then the MH is found within the current VLR. When the service proxy needs to forward packets to the MH, no extra search cost is required to find the current VLR, since the service proxy is located in the current VLR.

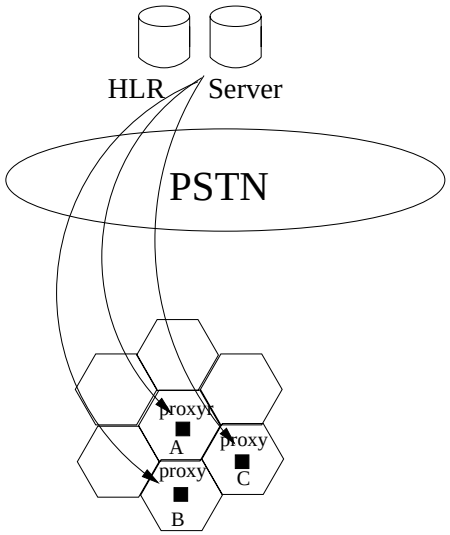

Figure 5-2: Fully Distributed Scheme.

5.2.3 Dynamic Anchor Scheme

Under the dynamic anchor scheme, a location anchor is used for location management such that the anchor changes whenever the MH crosses an anchor boundary. In addition, the anchor may also change its location within an anchor area when a call delivery operation is serviced. The service proxy dynamically moves with the anchor and is always co-located with the anchor. Below we give an algorithmic description of the dynamic anchor scheme for processing location update, call delivery, and service requests.

Location Update:

If (this is an anchor boundary crossing movement)

A location update message is sent to the HLR through the base station

The service context is moved to the new VLR which now serves as the new anchor

A location update message is sent to all application servers that the MH is engaged in

Else

The new VLR sends location update message to the anchor

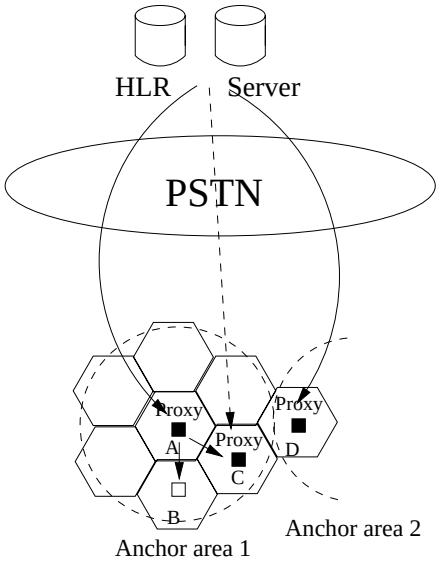

Figure 5-3: Dynamic Anchor Scheme.

Call Delivery:

A location request message is sent to the HLR to know the anchor of the called user

If (the local anchor is the current serving VLR)

The anchor sends the route information to the HLR, which in turn forwards the route information to the calling VLR

Else

The local anchor forwards the request to the current serving VLR

The current VLR sends the route information to the HLR

The HLR updates its record such that the current VLR becomes the new anchor

The service context is moved to the current VLR (which is the new anchor)

A location update message is sent to all application servers that the MH engages with

Service Request:

A request is sent from the MH to its current VLR

If (the current VLR is the local anchor)

The request is sent to the server and then a response is sent back to the MH

Else

The current VLR forwards the request to the anchor

The anchor forwards the service request/response to the server/MH

In Figure 5-3, when an MH moves within anchor area 1 from VLR A to VLR B, only the local anchor in VLR A is updated to point to the current location. Thus, the location update to the HLR and application servers is avoided. Suppose that a call arrives after the MH moves into VLR C. The call will invoke a search operation in the HLR database and a subsequent search operation in the anchor. Once the call is serviced, the HLR database will be updated to point to VLR C; the anchor and the service context are moved from VLR A to VLR C; and the application servers are informed of the address change. Later, if the MH subsequently moves from VLR C to VLR D due to an inter-anchor movement, the HLR database will be updated to point to VLR D who subsequently will become the new anchor after the service context is transferred to it. Data delivery from the server will pass through the service proxy co-located with the anchor to reach the MH.

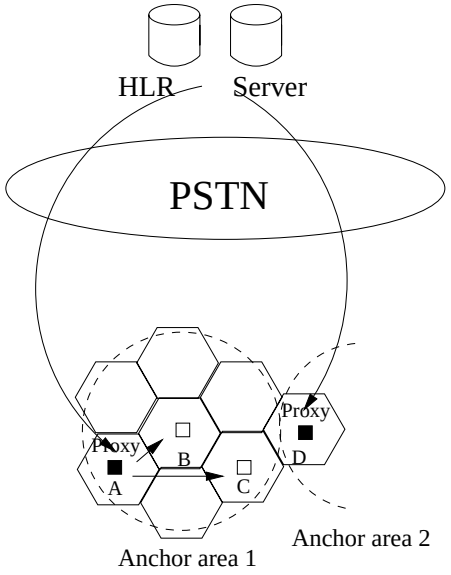

Figure 5-4: Static Anchor Scheme.

5.2.4 Static Anchor Scheme

Under the static anchor scheme, the service proxy is again co-located with the anchor. However, the anchor will remain at a fixed location as long as the MH stays in the same anchor area. The only condition to move the anchor (along with the service context transferred) is when the MH moves across an anchor boundary. The procedures for processing location update, call delivery,

and service requests are the same as in the dynamic anchor scheme except that upon a successful call delivery, the anchor's location remains unchanged. Thus, there is no need to migrate the service proxy to the current serving VLR (if they are not the same) after serving a call delivery operation.

We illustrate static anchor in Figure 5-4. When the MH moves within anchor area 1 from VLR A to VLR B and then to VLR C, the local anchor in VLR A is updated to point to the current VLR without updating the HLR. An incoming call will invoke a search operation at the HLR database to first find the anchor and then from the anchor to find the current VLR. The location of the anchor (where the service proxy is co-located) remains unchanged after a call is serviced. The anchor moves only when the MH moves out of the current anchor area (from VLR C to VLR D in this case). For each service request issued from the MH, it is serviced by the service proxy co-located with the anchor. As in dynamic anchor, there is no extra cost for the service proxy to find the MH, since the service proxy is co-located with the anchor.

5.3 Model

In this Section, we develop analytical models for evaluating and comparing various integrated schemes introduced in the previous section. We first define the performance metric as the basis for evaluation. Then, we show how the performance metric can be assessed for various schemes. In particular, we develop Stochastic Petri Net (SPN) models for analyzing the static and dynamic anchor schemes.

Table 5-1: Parameters for Integrated Schemes.

Symbol	Meaning
λ	the average rate at which the MH is being called
σ	the average rate at which the MH moves across VLR boundaries
γ	the average rate at which the MH requests services
CMR	call to mobility ratio, e.g., λ/σ
SMR	service request to mobility ratio, e.g., γ/σ
T	the average communication cost between a VLR and the HLR (or between a VLR and the server) per message

τ_1	the average communication cost between the anchor and a VLR in the anchor area per message
τ_2	the average communication cost between two neighboring anchor areas per message
τ_3	the average communication cost between two neighboring VLRs per message
M_{cs}	the number of packets required to transfer the service context
N_s	the number of applications concurrently engaged by the MH
P_{InA}	the probability that an MH moves within the same anchor area when a VLR boundary crossing movement occurs
P_{OutA}	the probability that an MH moves out of the current anchor area when a VLR boundary crossing movement occurs, $P_{OutA} = 1 - P_{InA}$

5.3.1 Cost Model

Our performance metric used for evaluating various integrated schemes is based on the total communication cost per time unit for handling three basic operations, namely, location update, call delivery, and service delivery. To be more specific, our cost model consists of three cost components: (a) update cost C_{update} - the cost for updating the locations of the MH and service proxy and transferring the service context if needed when a user moves across a VLR boundary; (b) search cost C_{search} - the cost for locating the MH to deliver a call; and (c) service delivery cost $C_{service}$ - the cost for the MH to communicate with the server through the proxy. Note that the cost here stands for the **average** cost. Let C_{total} be the average cost of the PCS network in servicing the above three types of basic operations per time unit. Then, our performance metric C_{total}, defined as the total cost incurred to the PCS network per time unit for servicing location and service operations of the MH, is given by:

$$C_{total} = C_{update} \times \sigma + C_{search} \times \lambda + C_{service} \times \gamma$$

where σ, λ and γ are the MH's VLR boundary crossing rate, call arrival rate and service request rate, respectively, as described in Table 5-1. Note that the paging cost for locating the location of the MH within the current VLR is not considered in the cost model because the paging cost is the same for all schemes.

5.3.2 Centralized Scheme

For the centralized scheme, each operation incurs a communication cost between the user's current VLR and the HLR co-located with the centralized service proxy. Thus, we have,

$$C_{update} = T;$$

$$C_{search} = T;$$
$$C_{service} = T + T.$$

where in the last equation the first T accounts for the communication cost from the MH to the service proxy while the second T accounts for the cost from the proxy to the server. Thus,

$$C_{total}^{centralized} = T \times \sigma + T \times \lambda + 2T \times \gamma$$

5.3.3 Fully Distributed Scheme

In the fully distributed scheme, each time the MH moves across a VLR boundary, three costs occur, i.e., a cost of T is required to update the HLR database for keeping track of the MH, a cost of $M_{cs} \times \tau_3$ is required to transfer the service context to the new VLR to provide continuous services where τ_3 stands for the communication cost between two neighboring VLRs, and finally a cost of $N_s \times T$ is required to inform N_s application servers of the address change of the service proxy. Each time a call is placed for the mobile user, the HLR consults the current VLR to get the location information with the communication cost T. For each service request, since the service proxy is always co-located with the current VLR of the MH, the only communication cost is from the proxy to the server. Summarizing above,

$$C_{update} = T + M_{cs} \times \tau_3 + N_s \times T;$$
$$C_{search} = T;$$
$$C_{service} = T.$$

Therefore,

$$C_{total}^{distributed} = (T + M_{cs} \times \tau_3 + N_s \times T) \times \sigma + T \times \lambda + T \times \gamma$$

5.3.4 Dynamic Anchor

For the dynamic anchor scheme, a SPN model as shown in Figure 5-5 is developed to analyze its behavior[4]. Table 5-2 gives the meanings of places and transitions defined in the SPN model. Here *mark(***p***)* returns the number of tokens in place **p**. The SPN model is constructed as follows:

[4] We could have directly used a finite-state continuous-time Markov chain for performance analysis except that the number of states would be large and the state diagram would be unwieldy. Instead of using a Markov model, we have used a Stochastic Petri Net (SPN) model to provide a concise definition of the corresponding finite-state continuous Markov chain. Tools such as SPNP [59] allow us to automatically generate the underlying finite-state Markov chain corresponding to an SPN model defined for stochastic analysis.

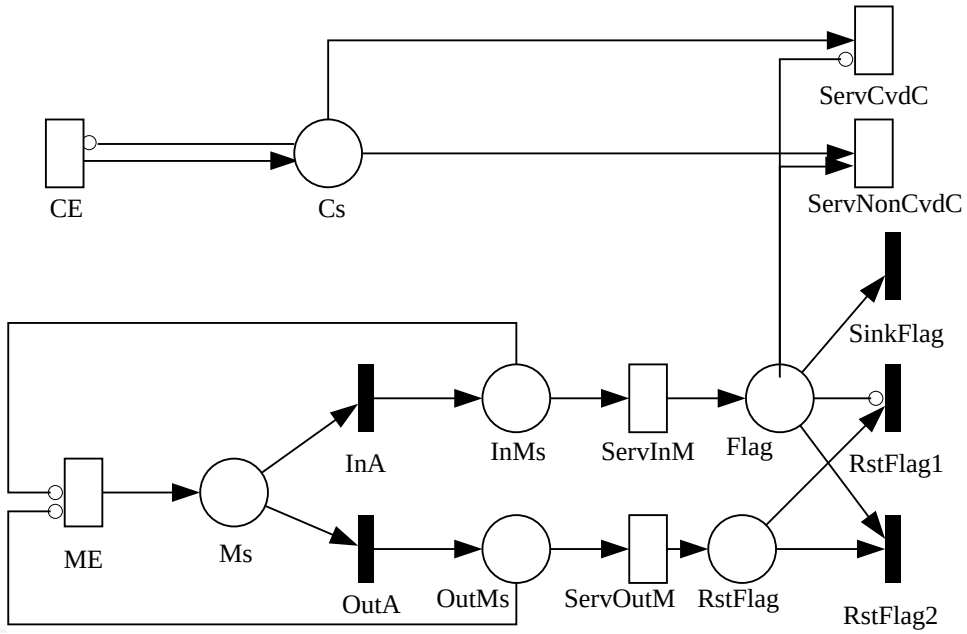

Figure 5-5: SPN Model for the Dynamic Anchor Scheme.

- When a call arrives, a token is placed in place **Cs**. The system serves the call based on the current status stored in place **Flag**:
    - If place **Flag** contains a token, or, *mark(***Flag***)*>0, then transition **ServNonCvdC** is enabled, which means that the current VLR is not the same as the anchor VLR. To delivery the call, the HLR is first queried to locate the anchor which in turn queries the current serving VLR to return the MH's current location. Note that after the search operation is performed, the anchor is moved to the current VLR as modeled by resetting *mark(***Flag***)*=0.
    - If *mark(***Flag***)*=0, transition **ServCvdC** is enabled. It means that the user resides in the same VLR with the anchor, so the search request is sent to HLR which in turn forwards the request to the local anchor. The local anchor returns the MH's location immediately.
- When an MH moves across a VLR boundary, a token is placed in place **Ms**.
    - If this is an intra-anchor movement with probability P_{InA}, transition **InA** will consume the token immediately, after which a token will be placed in **InMs** which subsequently disables transition **ME** and enables transition **ServInM**, representing that a local anchor update operation is being performed. After that, a token is placed in **Flag** to indicate that the current VLR is not the anchor VLR. If multiple tokens exist in **Flag**, **SinkFlag** is enabled and only one token remains in place **Flag**.

- If it is an inter-anchor movement with probability P_{OutA}, transition **OutA** will consume the token immediately, after which a token will be generated in **OutM** which subsequently disables transition **ME** and enables transition **ServOutM**. To serve the inter-anchor movement, the HLR is updated to point to the current VLR (e.g. the new anchor), the service context is transferred from the old anchor to the new anchor, and the application servers are updated with the new address of the proxy, after which a token is placed in **RstFlag** to reset the token of **Flag** to 0 using immediate transitions **RstFlag1** and **RstFlag2**. This models the fact that current VLR is the anchor VLR.

Note that there is no service request being modeled in the SPN. The reason is that the cost involved in a service request depends on the system status **Flag**, which we have already modeled in the SPN. Thus we would be able to calculate the service request cost without having to model the service request behavior explicitly.

Table 5-2: Places and Transitions for the SPN Model shown in Figure 5-5.

Place/Transition	Meaning
Cs	*mark(***Cs***)*[5] = 1 indicates that a call has just arrived
Ms	*mark(***Ms***)* = 1 indicates that the MH has just moved across a VLR boundary
InMs	*mark(***InMs***)* = 1 indicates that an intra-anchor movement has just been made
OutMs	*mark(***OutMs***)* = 1 indicates that an inter-anchor movement has just been made
Flag	*mark(***Flag***)* > 0 indicates that the current VLR is different from the anchor, i.e., the MH is located in a VLR that is not the same as the anchor VLR
RstFlag	*mark(***RstFlag***)* = 1 indicates that an inter-anchor movement has just been serviced. The Flag should be reset.
CE	call arrival transition with a rate of λ
ME	VLR boundary crossing transition with a rate of σ
ServCvdC	transition to service a call when the current VLR is the same as the anchor

[5] *mark*(**P**) returns the number of tokens in place **P**.

ServNonCvdC	transition to service a call when the current VLR is different from the anchor
InA	immediate transition with probability P_{InA} that the movement is an intra-anchor movement
OutA	immediate transition with probability P_{OutA} that the movement is an inter-anchor movement
ServInM	transition to service an intra-anchor movement
ServOutM	transition to service an inter-anchor movement
SinkFlag	immediate transition to consume one token from **Flag** if multiple tokens exist in **Flag**. It is used to ensure that at most one token exists in **Flag**. The enabling function[6] is *mark(***Flag***)*>1
RstFlag1	immediate transition to consume the token generated after an inter-anchor movement is serviced
RstFlag2	immediate transition to reset the **Flag** after an inter-anchor movement is serviced

To calculate C_{total} of the dynamic anchor scheme, we introduce additional cost parameters in Table 5-3 for ease of presentation.

Table 5-3: Additional Parameters for Dynamic Anchor.

Parameters	**Meaning**
$C_{ServInM}$	the average cost of performing an intra-anchor location update operation when the MH changes its VLR within the same anchor area
$C_{ServOutM}$	the average cost of performing an inter-anchor location update operation when the MH moves out of the current anchor area
$C_{ServCvdC}$	the cost to handle a call delivery operation when the current VLR is the same as the anchor VLR, i.e., the cost from the HLR to the anchor, *T*

[6] Enabling function is an advanced feature of SPNP [59] For a transition to fire, not only the general SPN marking conditions muse be met, but also the associated enabling function must be evaluated true. In SPNP, Each transition **t** can be associated with a boolean enabling function **e**. The function is evaluated in marking **M** when "there is a possibility that **t** is enabled", that is, when (1) no transition with priority higher than **t** is enabled in **M**; (2) the number of tokens in each of its input places is larger than or equal to the (variable) cardinality of the corresponding input arc; (3) the number of tokens in each of its inhibitor places is less than the (variable) cardinality of the corresponding inhibitor arc. Only then **e**(**M**) is evaluated; **t** is declared enabled in **M** iff **e**(**M**) = TRUE. The default for **e** is the constant function TRUE.

$C_{ServNonCvdC}$	the cost for handling a call delivery operation when the current VLR is different from the anchor VLR, including a cost from the HLR to the anchor (i.e, *T*), a communication cost (τ_1) from the anchor to the current VLR, a service context transfer cost (i.e. $M_{cs} \times \tau_1$) to migrate the anchor to the current VLR, and a cost ($N_s \times$T) to inform all N_s application servers of the address change of the proxy.
$C_{ServCvdS}$	the cost to handle a service request when the anchor resides in the current serving VLR; it is *T* under the dynamic scheme
$C_{ServNonCvdS}$	the cost to handle a service request when the anchor is different from the current serving VLR; it is $\tau_1 + T$ under the dynamic scheme

These cost parameters can be calculated as follows:

$$C_{ServInM} = \tau_1;$$

$$C_{ServOutM} = T + M_{cs} \times \tau_2 + N_s\, T;$$

$$C_{ServCvdC} = T;$$

$$C_{ServNonCvdC} = T + \tau_1 + M_{cs} \times \tau_1 + N_s\, T;$$

$$C_{ServCvdS} = T;$$

$$C_{ServNonCvdS} = \tau_1 + T.$$

Suppose *N* states exist in the underlying Markov model of the SPN. Let P_i be the steady state probability that the system is found in state *i*. The average cost to serve location update, call delivery and service requests can be obtained by assigning “cost" values to these *N* system states. Specifically, let $C_{i,call}^{da}$ be the search cost assigned to state *i* given that a search operation is being serviced in state *i* under the dynamic anchor scheme. Then, the average search cost under dynamic anchor, C_{seach}^{da}, can be calculated as the expected value of $C_{i,call}^{da}$ weighted by the state probability, i.e.,

$$C_{seach}^{da} = \sum_{i=1}^{N} P_i \times C_{i,call}^{da}$$

where,

$$C_{i,call}^{da} = \begin{cases} C_{ServNonCvdC} & if\ mark(\mathbf{Flag}) > 0 \\ C_{ServCvdC} & Otherwise \end{cases}$$

Here, $C_{i,call}^{da}$ is $C_{ServNonCvdC}$ if in state i, the current VLR is different from the anchor, i.e. *mark(***Flag***)* > 0. Otherwise, $C_{i,call}^{da}$ is assigned the value of $C_{ServCvdC}$ to account for the fact that the current VLR is the same as the anchor in state i.

Similarly, let $C_{i,update}^{da}$ and $C_{i,service}^{da}$ be the costs for serving location update and service requests in state i, respectively. We have:

$$C_{update}^{da} = \sum_{i=1}^{N} P_i \times C_{i,update}^{da}$$

$$C_{service}^{da} = \sum_{i=1}^{N} P_i \times C_{i,service}^{da}$$

where,

$$C_{i,update}^{da} = \begin{cases} C_{ServInM} & if\ enabled(\textbf{ServInM}) \\ C_{ServOutM} & if\ enabled(\textbf{ServOutM}) \\ P_{inA} \times C_{ServInM} + P_{outA} \times C_{ServOutM} & Otherwise \end{cases}$$

$$C_{i,service}^{da} = \begin{cases} C_{ServNonCvdS} & if\ mark(\textbf{Flag}) > 0 \\ C_{ServCvdS} & Otherwise \end{cases}$$

Here *enabled(***T***)* means that transition **T** is enabled. In the first equation above, $C_{i,update}^{da}$ is assigned a value that reflects if the movement is intra-anchor or inter-anchor. If the MH has just made an intra-anchor movement in state i, transition **ServInM** would be enabled. Thus the location update cost in state i would be $C_{ServInM}$. If the MH has just made an inter-anchor movement, transition **ServOutM** would be enabled instead. Thus, the location update cost would be $C_{ServOutM}$. If in state i, the MH has not yet made a move, then the location update cost in state i is the average cost weighted on the probability of whether the user's next move is inter- or intra-anchor, i.e., P_{InA} × $C_{ServInM}$ + P_{OutA} × $C_{ServOutM}$. In the second equation above, $C_{i,service}^{da}$'s value depends on if the current VLR is different from the anchor VLR in state i. If yes (modeled by *mark(***Flag***)* > 0 in the SPN), then the service request cost in state i is $C_{ServNonCvdS}$. Otherwise, the cost is $C_{ServCvdS}$.

The total cost per time unit incurred to PCS network under dynamic anchor, C_{total}^{da}, can be calculated by

$$C_{total}^{da} = C_{update}^{da} \times \sigma + C_{seach}^{da} \times \lambda + C_{service}^{da} \times \gamma$$

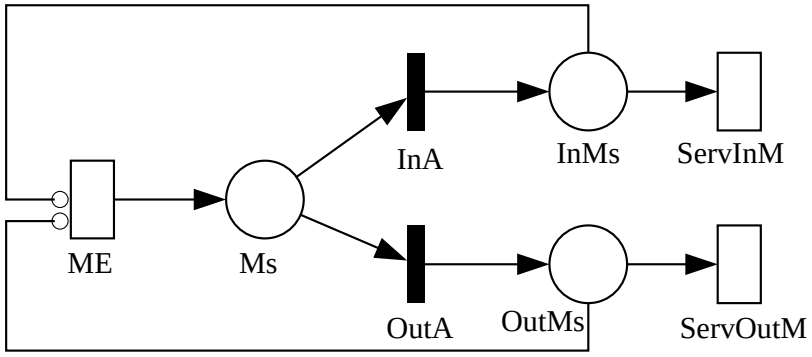

Figure 5-6: SPN Model for the Static Anchor Scheme.

5.3.5 Static Anchor

In the static anchor scheme, the local anchor and the service proxy remain static in one VLR as long as the MH resides in an anchor area. Its behavior is modeled by an SPN model as shown in Figure 5-6.

Table **5-4** lists the meanings of transitions and places in the SPN. Table 5-5 lists the cost parameters for the static anchor scheme. The major difference between the static anchor model and the dynamic anchor model is that there is no **Flag** to indicate whether the anchor VLR is located in the current serving VLR because unlike in the dynamic anchor scheme, the anchor is at a fixed location upon entry to a new anchor area and remains there until the MH departs the anchor area. Therefore, we only need to consider the *average* cost of accessing the anchor from any VLR in the anchor area without having to track if the current VLR is the same as the anchor VLR. Let τ_1 be this average communication cost between the anchor and a VLR in the anchor area as described in Table 5-1. Then, the cost parameters listed in Table 5-5 can be calculated as:

$$C_{ServInM} = \tau_1;$$
$$C_{ServOutM} = T + M_{cs} \times \tau_2 + N_s \times T;$$
$$C_{ServC} = T + \tau_1;$$
$$C_{ServS} = T + \tau_1.$$

Table 5-4: Places and Transitions for the SPN Model shown in Figure 5-6.

Place/Transition	Meaning
Ms	*mark(***Ms***)* = 1 indicates that the MH has just moved across a VLR boundary
InMs	*mark (***InMs***)* = 1 indicates that an intra-anchor movement has just been made
OutMs	*mark(***OutMs***)* = 1 indicates that an inter-anchor movement has just been made
ME	VLR boundary crossing transition with a rate of σ
ServC	transition to service an incoming call
InA	immediate transition with probability P_{InA} that the movement is an intra-anchor movement
OutA	immediate transition with probability P_{OutA} that the movement is an inter-anchor movement
ServInM	transition to service an intra-anchor movement
ServOutM	transition to service an inter-anchor movement

Table 5-5: Additional Parameters for Static Anchor.

Parameter	Meaning
$C_{ServInM}$	the average cost of performing an intra-anchor location update operation when the MH changes its VLR within the same anchor area
$C_{ServOutM}$	the average cost of performing an inter-anchor location update operation when the MH moves out of the current anchor area
C_{ServC}	the cost to handle a call delivery
C_{ServS}	the cost to handle a service request

By following a similar approach performed for the dynamic anchor scheme, the costs incurred to the PCS system per time unit under the static anchor scheme for serving location update, call delivery and service requests can be calculated, respectively, as:

$$C_{seach}^{sa} = \sum_{i=1}^{N} P_i \times C_{i,call}^{sa} = \sum_{i=1}^{N} P_i \times C_{ServC} = C_{ServC}$$

$$C_{update}^{sa} = \sum_{i=1}^{N} P_i \times C_{i,update}^{sa}$$

$$C_{service}^{sa} = \sum_{i=1}^{N} P_i \times C_{i,service}^{sa} = \sum_{i=1}^{N} P_i \times C_{ServS} = C_{ServS}$$

where,

$$C_{i,update}^{sa} = \begin{cases} C_{ServInM} & if\ enabled(ServInM) \\ C_{ServOutM} & if\ enabled(ServOutM) \\ P_{inA} \times C_{ServInM} + P_{outA} \times C_{ServOutM} & Otherwise \end{cases}$$

Therefore, the total cost per time unit incurred to PCS network under static anchor, C_{total}^{sa}, is calculated as:

$$C_{total}^{sa} = C_{update}^{sa} \times \sigma + C_{seach}^{sa} \times \lambda + C_{service}^{sa} \times \gamma$$

5.4 Evaluation

In this Section, we first parameterize the performance models developed by means of a hexagonal network coverage model for describing a PCS network to evaluate the performance of the four integrated location and service management schemes proposed so as to identify conditions under which one scheme could perform the best when given a set of parameters characterizing an MH's mobility and service behaviors. Then we present analytical results with physical interpretation given. We compare integrated vs. decoupled location and service management and show that the best integrated scheme outperforms the best decoupled scheme, as well as management schemes that do not use any service proxy. Lastly, a simulation study is conducted to do validation and sensitivity analysis for our analytical results.

5.4.1 Parameterization

We use a hexagonal network coverage model to describe a PCS network as shown earlier in Figure 2-1 where cells are assumed to be hexagon shape, with each cell having six neighbors. At the lowest level of Figure 2-1, an n-layer VLR covers $3n^2$ -3n+1 cells where n is equal to either two or three [32]. Going into the second lowest level of Figure 2-1, we can again view each hexagon-shaped cell as corresponding to a VLR and therefore an n-level LSTP will contain $3n^2$-3n+1 VLRs. This view continues as we recursively go up to the higher levels of the PCS network

until the RSTP level is reached. For the dynamic and static anchor schemes, we consider an anchor area corresponding to one LSTP area.

For a PCS system described by the hexagonal network coverage model as such, it can be shown that [9] with random movements, the probability that an MH moves within the same anchor area (e.g., same LSTP area), that is, the probability of an intra-anchor movement, as the MH moves across a VLR boundary, is given by:

$$P_{InA} = \frac{3n^2 - 5n + 2}{3n^2 - 3n + 1}$$

Thus, the probability of an inter-anchor movement, when the MH moves across a VLR boundary, is given by:

$$P_{OutA} = 1 - P_{InA} = \frac{2n - 1}{3n^2 - 3n + 1}$$

In the evaluation, we consider $n = 2$ for n-layer VLRs, LSTPs and RSTPs composing the PCS. Then, the probability P_{InR} that an MH moves within the same RSTP, that is, the probability of an intra-RSTP movement, when the MH moves across a VLR boundary, is given by:

$$P_{InR} = \frac{21n^2 - 27n + 10}{7(3n^2 - 3n + 1)}$$

Let C_{vl} be the cost of transmitting a message between a VLR and its LSTP. Let C_{lr} be the cost of transmitting a message between an LSTP and its RSTP. Let C_{pstn} be the communication cost to pass through a PSTN. The communication between a VLR and the HLR will traverse through a VLR-LSTP-RSTP-PSTN path sequence. Therefore,

$$T = C_{vl} + C_{lr} + C_{pstn}$$

For the centralized scheme, there are no additional parameters to parameterize. For the fully distributed scheme, we need to parameterize τ_3 standing for the average communication cost between two neighboring VLRs. With reference to the PCS network shown in Figure 2-1, the communication cost between two VLRs within the same LSTP (with probability P_{InA}) is $2\ C_{vl}$; the communication cost between two VLRs out of the same LSTP but within the same RSTP (with probability $P_{InR} - P_{InA}$) is $2(C_{vl} + C_{lr})$; the communication cost between two VLRs out of the same RSTP (with probability $1-P_{InR}$) is $2\ C_{vl} + 2\ C_{lr} + C_{pstn}$. Therefore, τ_3 can be parameterized as:

$$\tau_3 = 2\ C_{vl} \times P_{InA} + 2(C_{vl} + C_{lr}) \times (P_{InR} - P_{InA}) + (2\ C_{vl} + 2\ C_{lr} + C_{pstn}) \times (1 - P_{InR})$$

For the dynamic anchor scheme, we need to parameterize τ_1 for the average communication cost between the anchor VLR and another VLR (other than the anchor VLR itself) in an anchor area, as well as τ_2 for the average signaling communication cost between two neighboring LSTP areas. τ_1 is equal to the communication cost between two VLRs within the same LSTP. To calculate τ_2, two scenarios are considered: the communication between two VLRs within the same RSTP with cost $2(C_{vl} + C_{lr})$ and the communication between two VLRs out of the same RSTP with cost $2\ C_{vl} + 2\ C_{lr} + C_{pstn}$. Thus,

$$\tau_1 = 2\ C_{vl}$$

$$\tau_2 = 2(C_{vl} + C_{lr}) \times \frac{P_{InR} - P_{InA}}{1 - P_{InA}} + (2\ C_{vl} + 2\ C_{lr} + C_{pstn}) \times \frac{1 - P_{InR}}{1 - P_{InA}}$$

For the static anchor scheme, we need to parameterize τ_1 for the average communication cost between the anchor VLR and any VLR (including possibly the static anchor VLR itself) in an anchor area, as well as τ_2 for the average signaling communication cost between two neighboring LSTP areas. Since the static anchor scheme does not track the location of the MH within an anchor area, the MH can reside in each VLR with equal probability. Thus, for a PCS network with $n = 2$ where each LSTP has 7 VLRs, we have:

$$\tau_1 = 2\ C_{vl} \times \frac{6}{7} + 0 \times \frac{1}{7} = C_{vl} \times \frac{12}{7}$$

$$\tau_2 = 2(C_{vl} + C_{lr}) \times \frac{P_{InR} - P_{InA}}{1 - P_{InA}} + (2\ C_{vl} + 2\ C_{lr} + C_{pstn}) \times \frac{1 - P_{InR}}{1 - P_{InA}}$$

Note that costs above are expressed in terms of the delay required to process or transfer a signaling message. For example, C_{vl} and C_{lr} represent the communication delay for sending a signaling message through particular network node and link. However, it doesn't mean a cost considered here must have a time unit. In general, other measurements for the cost parameters are possible. The network administrator can assign a relative high cost to C_{lr} to discourage the use of the link from *LSTP* to *RSTP*, which in turn may favor certain schemes. In IP networks, the cost parameters may be proportional to hops. We do not intend to introduce a method for determining

these costs. Instead, we perform the analysis of our proposed schemes assuming cost parameters are available.

5.4.2 Results

In the section, we present numerical data obtained based on our analysis for a PCS network consisting of 2-layer VLRs, LSTPs, RSTPs and HLR as shown earlier in Figure 2-1. All costs are normalized with respect to the cost of transmitting a message between a VLR and its LSTP, i.e., $C_{vl} = 1$, such that $C_{lr} = 0.5$ and $C_{pstn} = 6$. We also assume there is a single server, i.e. $N_s = 1$. For the dynamic anchor and static anchor schemes, we used SPNP [59] as a tool to evaluate their respective SPN models defined in Figure 5-5 and Figure 5-6 to obtain the data.

Figure 5-7 shows the cost incurred to the PCS network per second as a function of the MH's *CMR* for the four integrated schemes. The *X* coordinate represents the *CMR* value in the range of [0.125, 16] with the mobility rate σ fixed at 10/hour while changing the call arrival rate λ. To isolate the effect of *CMR*, we let *SMR*=1 and M_{cs} =1 such that the service request rate γ is the same as the mobility rate σ and the average number of packets to transfer the service context is 1. The *Y* coordinate is the cost rate, i.e., the total cost incurred per second (normalized with respect to the cost of transmitting a message between a VLR and its LSTP) to the network.

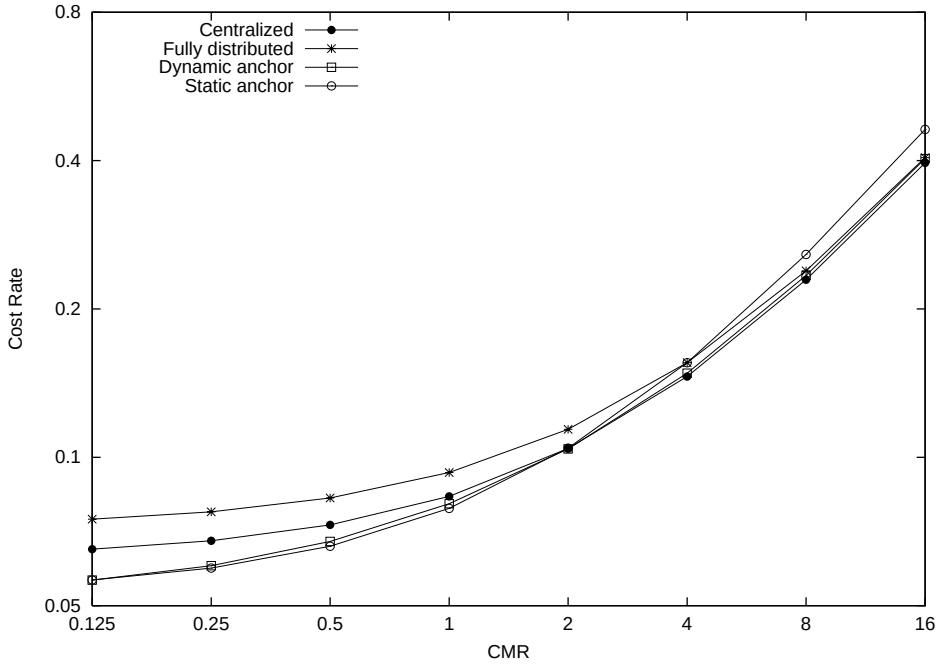

Figure 5-7: Cost Rate under Different Call to Mobility Ratio (CMR) Values.

When the *CMR* value is low, both the centralized and fully distributed schemes perform worse than the dynamic and static anchor schemes. This is attributed to the fact that the total cost rate is dominated by mobility-related cost factors at low *CMR* at which the mobility rate is much higher than the call arrival rate. Specifically, the centralized scheme performs badly in this condition because of the high cost of servicing location update operations as these operations need to access the HLR in the centralized scheme. The fully distributed scheme performs badly at low *CMR* because with a high mobility rate, the location update cost and the context transfer cost are high in the fully distributed scheme. On the other hand, the dynamic and static anchor schemes employ an anchor to reduce the location update cost and the context transfer cost when the user's mobility rate is high.

As the *CMR* value increases, the performance of both centralized and fully distributed improves. At very high *CMR*, the static anchor scheme is the worst and the performances of the other three schemes are very close. Dynamic anchor performs better than static anchor in this extreme case because in the dynamic anchor scheme the anchor co-located with the service proxy is close to the MH. Thus, the cost for service requests and location updates due to movements within an anchor area is low. Another reason is that when a call arrives and the anchor VLR is not the current serving VLR, the dynamic anchor scheme will update the HLR after the call is serviced and move the anchor to the current VLR. This keeps the HLR database up-to-date and keeps the anchor close to the MH. As a result, it reduces the call delivery cost since the system is able to find the MH quickly on subsequent calls, the effect of which is especially pronounced when *CMR* is high.

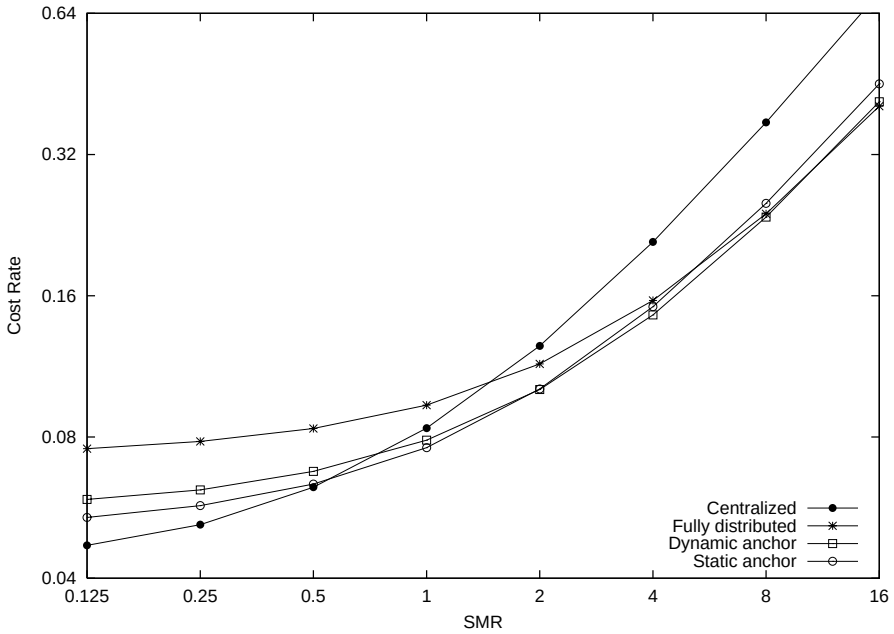

Figure 5-8: Cost Rate under Different Service to Mobility Ratio (*SMR*) Values.

Figure 5-8 shows the cost rate as a function of the service request to mobility ratio (*SMR*) to analyze the effect of the service request rate. Again we isolate the effect of SMR by fixing *CMR* = 1 and M_{cs} =1. Here by setting *CMR*=1, we set the calling rate to be the same as the mobility rate fixed at 10/hour. Figure 5-8 shows that as *SMR* increases, the cost rate under all four schemes increases because when the mobility rate σ is fixed, increasing *SMR* increases the service request rate, which in turn incurs more service-related costs for all four schemes. At very high *SMR*, however, fully distributed and dynamic anchor schemes perform better over static anchor and centralized schemes because in the fully distributed scheme the MH's service requests can be serviced quickly by the local service proxy located in the current VLR database, although each service request still unavoidably incurs a communication cost from the service proxy to the server. As the service rate increases while keeping other rates constant, we see that the service request cost dominates other costs, thus making the fully distributed scheme outperform both static and centralized schemes at high *SMR*.

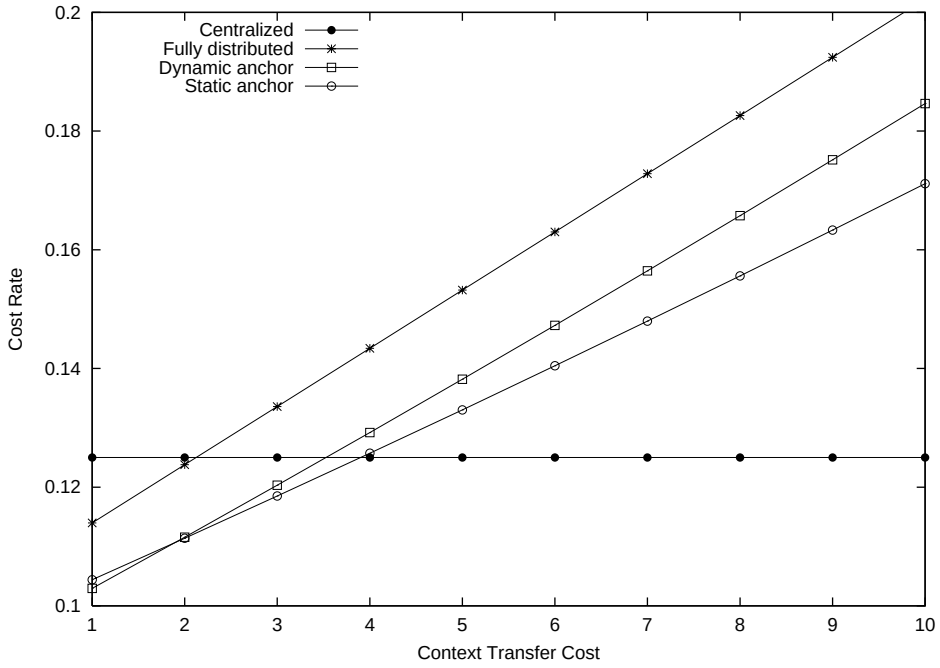

Figure 5-9: Cost Rate under Different Context Transfer Cost Values.

Figure 5-9 summarizes the effect of the service context transfer cost on the cost rate. As expected, as the context transfer cost increases, the cost rate under the fully distributed, dynamic anchor, or static anchor scheme all increases, while that for the centralized scheme remains unchanged because there is no service context transfer cost in the centralized scheme. The fully distributed scheme is most sensitive to the increase of the context transfer cost in terms of the increase of the cost rate, followed by dynamic anchor and static anchor. This order corresponds to the context transfer frequency under various schemes. At one end of the spectrum, the fully distributed scheme must transfer the service context with the migrated service proxy whenever the MH moves across a VLR boundary. The dynamic anchor scheme transfers the service context when the MH moves across an anchor boundary, or after a call delivery operation is serviced if the anchor VLR is not the same as the current VLR. In the static anchor scheme, the service context is transferred only when the user moves across an anchor boundary. At the other end of the spectrum, the centralized scheme in entirely insensitive to the increase of the service context transfer cost because the service proxy is co-located with the HLR which requires no service context transfer.

5.4.3 Integrated vs. Decoupled Location and Service Management

To demonstrate the viability of the integrated location and management scheme, we have conducted a performance study to compare integrated against decoupled location and service management for which location management is decoupled from service management. By decoupling, the MH's service proxy is not co-located with the MH's location database, and the MH's location registration areas are decoupled from the MH's service areas. Three location management schemes are feasible, namely, fully distributed (corresponding to basic HLR/VLR), dynamic anchor and static anchor. The centralized scheme is not feasible because it is meaningless to put a regional location database co-located with the HLR also pointing to the current VLR. For service management, again fully distributed, static anchor and dynamic scheme are feasible. The centralized scheme is not feasible because it would place the service proxy at a fixed location (not at the HLR) so the communication cost for servicing user requests would be excessively high. The dynamic anchor scheme for service management here refers to the feature that the anchor can change its location to the current serving VLR within the anchor area after serving a user request. Therefore, there are nine possible combinations (i.e., three for location management and three for service management) through which decoupled location and service management can be applied. For fair comparison, we only compare the best cost rate achievable by both schemes, i.e., for the decoupled scheme the best combination out of the nine selections is used, and for the integrated scheme the best out of four is used, when given an MH's mobility and service characteristics.

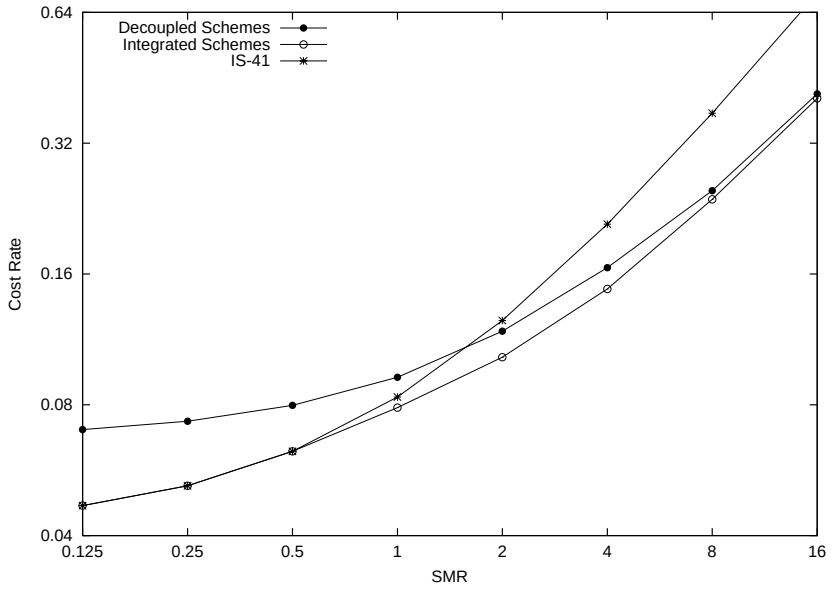

Figure 5-10: Integrated vs. Decoupled Location and Service Management: Best Cost Rate under Different *SMR* Values.

Figure 5-10 (corresponding to Figure 5-8) compares integrated vs. decoupled at various SMR values. As a baseline, Figure 5-10 also shows a cost curve for the basic HLR/VLR scheme that does not use a proxy for location and service management as in IS-41 and GSM. Figure 5-10 demonstrates the superiority of integrated over decoupled schemes and the basic HLR/VLR scheme. We attribute the superiority of the integrated scheme to the fact that the service proxy knows the MH's location at all times through integration of location and service management. The superiority of the integrated scheme over the decoupled scheme is especially pronounced when SMR is low at which the service proxy in the decoupled scheme has to explicitly track the MH's location which incurs extra costs. On the other hand, the integrated scheme outperforms the basic scheme significantly especially at high SMR at which the server in the basic scheme must go through the HLR for data delivery.

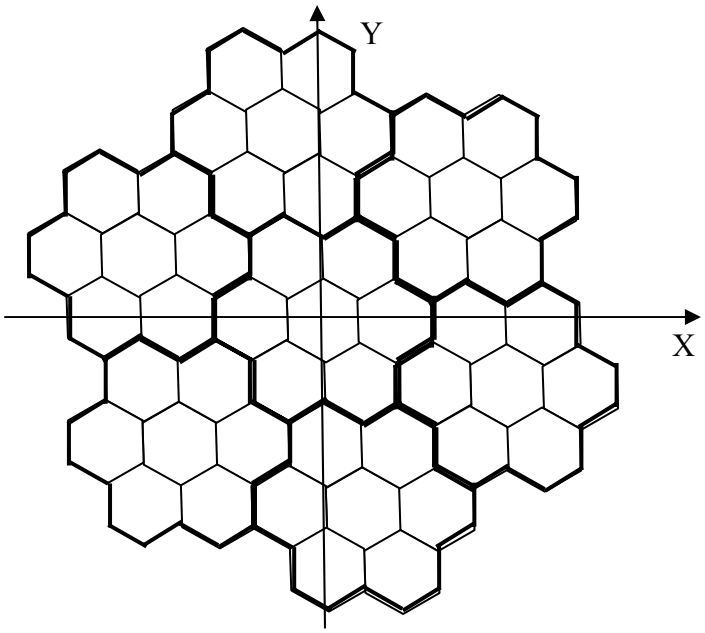

Figure 5-11: Simulation Environment.

5.4.4 Simulation Validation

We have conducted a simulation study using a discrete event simulation language called SMPL (Simulation Model Programming Language) [37] to validate the analytical results. The simulation environment (Figure 5-11) consists of a large 2-layer RSTP area covering 7 LSTP's each corresponding to an anchor area (for the dynamic and static anchor schemes) that in turn covers 7 hexagonally-shaped VLRs. The center of the RSTP is at *(0, 0)*. Each hexagonal VLR area is represented by its center location *(x, y)*. An MH is characterized by its own mobility and service

behaviors with the mobility rate σ and service rate γ. The MH can move from the current VLR to one of the 6 neighbor VLRs randomly. If the MH moves out of the simulated RSTP area, its location will be circled to the other side of the simulated area, i.e., its location will be changed from *(x, y)* to *(-x,-y)*, thus allowing the simulated RSTP area to be reused. At all times, the location of the MH is known. The service proxy moves according to the specific integrated scheme considered. As the simulation program knows the locations of the MH and its service proxy all the time, whenever a location or service management event occurs, such as a call, a move to another VLR, or a service request, it knows exactly the cost incurred in response to the event. These per-event costs are then accumulated to the overall cost during the course of the simulation. At the end of each simulation batch run, the time-average cost is computed by dividing the cumulative cost over the simulation period.

To ensure statistical significance of simulation results, a batch mean analysis technique has been adopted by which the simulation period is divided into batch runs with each batch consisting of 2000 time-average cost observations for computing an average value. A minimum of 10 batches was run to compute a grand mean of the cost rate value. Additional batches were added if necessary until the mean cost rate value is within 95% confidence level and 10% accuracy from the true mean.

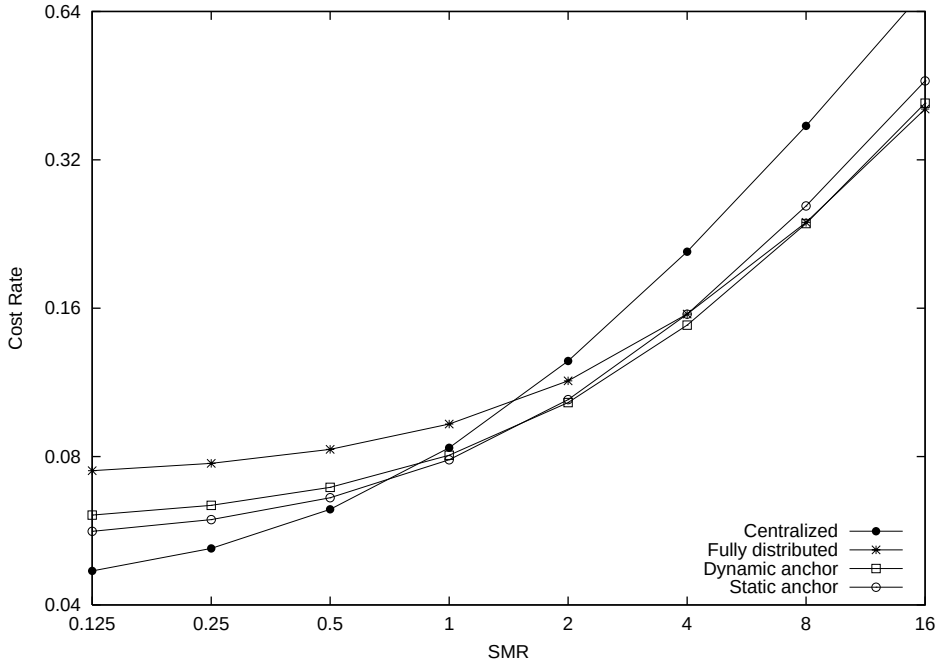

Figure 5-12: Simulation Results: Cost Rate under Different SMR Values.

The simulation results show very good correlations with analytical results. As an example Figure 5-12 shows the simulation results for the cost rate as a function of the MH's SMR, corresponding to Figure 5-8 for analytical results. We see that Figure 5-12 and Figure 5-8 are virtually identical despite the fact that simulation results are obtained based on the cumulative cost over the simulation period in response to mobility, call and service events divided by the simulation period, while analytical results are obtained based on the average cost rate as calculated above. We conclude that the analytical results are valid and there exists an optimal integrated scheme for integrated location and service management on a per-user basis.

5.4.5 Random Waypoint Mobility Model

Our analytical and simulation results obtained thus far are based on the random movement mobility model. Below we test the sensitivity of the results with respect to a different mobility model. The Random Waypoint (RWP) model [40] is commonly used in simulation of user mobility. Using the environment depicted in Figure 5-11, we conducted a simulation based on a RWP model. In this model, the MH is assigned an initial location, a movement direction, a distance and a speed. The MH's initial position is (0, 0). The direction is chosen from six possible directions to reach the MH's neighbor cells. The distance is chosen uniformly in the range of [1, 10] cells, with each cell spanning 2 miles. The speed is chosen uniformly in the range of [2, 38] mph, independently of both the initial location and direction. After reaching the next destination, a new direction, distance and speed are chosen from their respective uniform distribution functions, independently of all previous values.

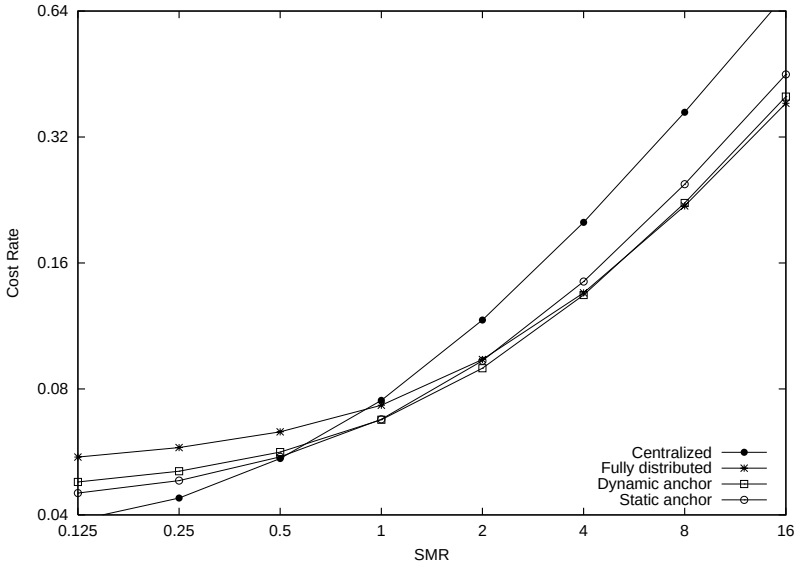

Figure 5-13: Random Waypoint Model Results: Cost Rate under Different SMR Values.

The simulation results based on this RWP model are shown in Figure 5-13. Clearly, the cost rate values are smaller than the corresponding values based on the random movement model as shown in Figure 5-8. The reason is that the distribution of network nodes in the RWP model is non-uniform in general [6] and the highest movement density tends to be near the center of the simulation area. In our simulation, this means that more movements occur within the same RSTP/LSTP, resulting in a lower cost rate compared with the random movement model. Nevertheless, the relative positioning of the four integrated schemes virtually remains the same, with the same cross-over trend being observed. A small discrepancy from Figure 5-8 is that the cost rate curve for the full distributed scheme rises less sharply, causing its cross-over point with the centralized scheme to happen earlier (at SMR=1 in Figure 5-13) than it used to be under the random movement model (at SMR=2 in Figure 5-8). The reason is that while we fix the average mobility rate to be the same (σ=10/hr) for both random movement and RWP mobility models for fair comparison, the RWP model tends to generate a larger residence time than the random movement model. A large residence time especially favors the fully distributed scheme because for the fully distributed scheme the cost of location updated due to boundary crossings is reduced while it can service data delivery efficiently since the proxy is always located in the current VLR. We conclude from this study that, while the RWP model affects the relative ordering of the four schemes over a small range of cross-over areas (e.g., around SMR=1 in the study) because of its peculiar movement density and residence time characteristics, the cost-rate curves obtained for the four schemes remain virtually the same, with the same cross-over trend observed. As a result, the analytical and simulation results obtained based on the random movement model are valid.

5.4.6 Sensitivity Analysis

In our analytical modeling, we have assumed the residence time is exponentially distributed. In this section, we study the sensitivity of the results obtained with respect to the residence time distributions including Erlang[7], hyper-exponential[8], normal, and uniform. Table 5-6 reports the cost rate obtained under different distributions with *CMR* = 1, M_{cs}= 1 and *SMR* ranging from 0.125 to 16.

[7] The Erlang distribution models a r-stage exponential center connected in a series structure such that it has the same mean residence time x=1/lambda as that under the exponential distribution with the standard deviation s (square root of the variance) less than x. The number of stages is equal to (floor(x/s))^2.

[8] The Hyper-exponential distribution models a 2-stage exponential center in a parallel structure such that the mean(x) is the same as that under the exponential distribution but the standard deviation (s) is higher than the mean.

Although different distributions or different deviations in the normal distribution may affect the exact results, we can see that the relative costs of these four integrated schemes are quite stable, i.e., for a given network and a user profile, the choice of the best scheme (in some cases the best two schemes are very close) among the four integrated schemes remains the same. Thus we conclude that the analysis results reported are valid and are insensitive to the residence time distribution used.

Table 5-6: Cost Rates under Various Residence Time Distributions.

SMR	Analysis				Exponential (Tr)			
	Centralized	Fully Distributed	Dynamic Anchor	Static Anchor	Centralized	Fully Distributed	Dynamic Anchor	Static Anchor
0.125	0.0469	0.0749	0.0609	0.0564	0.0468	0.0748	0.0595	0.0556
0.25	0.0521	0.0775	0.0637	0.0596	0.0521	0.0775	0.0624	0.0589
0.5	0.0625	0.0827	0.0693	0.0660	0.0625	0.0828	0.0680	0.0653
1	0.0833	0.0931	0.0805	0.0788	0.0832	0.0933	0.0791	0.0779
2	0.1250	0.1140	0.1029	0.1044	0.1250	0.1141	0.1013	0.1037
4	0.2083	0.1556	0.1478	0.1556	0.2084	0.1557	0.1461	0.1548
8	0.3750	0.2390	0.2375	0.2580	0.3751	0.2391	0.2352	0.2570
16	0.7083	0.4056	0.4170	0.4627	0.7079	0.4058	0.4134	0.4622

SMR	Hyper (Tr, s = 1.5Tr)				Erlang (Tr,s = 0.5Tr)			
	Centralized	Fully Distributed	Dynamic Anchor	Static Anchor	Centralized	Fully Distributed	Dynamic Anchor	Static Anchor
0.125	0.0468	0.0750	0.0586	0.0557	0.0468	0.0749	0.0613	0.0556
0.25	0.0522	0.0776	0.0615	0.0590	0.0520	0.0775	0.0641	0.0588
0.5	0.0624	0.0828	0.0669	0.0651	0.0626	0.0828	0.0697	0.0651
1	0.0833	0.0933	0.0779	0.0780	0.0833	0.0931	0.0809	0.0782
2	0.1250	0.1138	0.1001	0.1037	0.1249	0.1141	0.1036	0.1035
4	0.2081	0.1559	0.1444	0.1550	0.2084	0.1555	0.1490	0.1546
8	0.3752	0.2386	0.2330	0.2573	0.3750	0.2389	0.2394	0.2575
16	0.7079	0.4055	0.4101	0.4618	0.7080	0.4053	0.4199	0.4621

SMR	Normal (Tr,s = Tr)				Uniform (a = 0.5Tr, b = 1.5Tr)			
	Centralized	Fully Distributed	Dynamic Anchor	Static Anchor	Centralized	Fully Distributed	Dynamic Anchor	Static Anchor
0.125	0.0422	0.0634	0.0536	0.0496	0.0469	0.0749	0.0618	0.0557
0.25	0.0475	0.0659	0.0564	0.0528	0.0521	0.0774	0.0647	0.0590
0.5	0.0579	0.0712	0.0620	0.0591	0.0626	0.0828	0.0702	0.0653
1	0.0788	0.0816	0.0731	0.0721	0.0834	0.0932	0.0815	0.0779
2	0.1204	0.1024	0.0955	0.0978	0.1250	0.1139	0.1043	0.1038
4	0.2036	0.1442	0.1400	0.1488	0.2086	0.1556	0.1498	0.1547
8	0.3705	0.2276	0.2296	0.2511	0.3751	0.2391	0.2404	0.2568
16	0.7029	0.3940	0.4081	0.4560	0.7081	0.4058	0.4218	0.4622

Due to the nature of certain distributions and limitation of the simulation package, it is not possible to make all distributions have the same standard deviation. We take normal distribution as an example to study the effect of standard deviation. The standard deviation of a random variable indicates how far the variable's value is from its expected value. From **Table 5-7**, we observe that the higher the standard deviation, the higher the difference between simulation results based on normal distribution and analytical results. Nevertheless, the trend regarding the relative costs of the four schemes remains the same and valid.

Table 5-7: Cost Rates under Normal Distributions with Different Variances.

SMR	Normal (Tr, s=0.5Tr)				Normal (Tr, s=1.5Tr)			
	Centralized	Fully Distributed	Dynamic Anchor	Static Anchor	Centralized	Fully Distributed	Dynamic Anchor	Static Anchor
0.125	0.046373	0.073530	0.060536	0.054911	0.038699	0.054789	0.047782	0.045264
0.25	0.051525	0.076244	0.063261	0.058025	0.043957	0.057375	0.050494	0.048461
0.5	0.062009	0.081448	0.068946	0.064486	0.054292	0.062647	0.056134	0.054846
1	0.082879	0.091752	0.080145	0.077339	0.075265	0.073131	0.067075	0.067439
2	0.124540	0.112520	0.102843	0.102921	0.116807	0.093919	0.089202	0.093028
4	0.207462	0.154182	0.147870	0.154036	0.200223	0.135573	0.133538	0.144446
8	0.374327	0.237789	0.238393	0.256461	0.366763	0.218817	0.221991	0.246556
16	0.707615	0.404279	0.419000	0.461400	0.699629	0.385453	0.398678	0.451627

5.5 Summary

In this chapter, we investigated and analyzed a class of integrated location and service management schemes by means of SPN models and identified conditions under which one scheme may perform better than others. The analysis results are useful for identifying the most efficient scheme to be adopted to provide personalized services to individual users based on their profiles. Our analysis and simulation results show that the dynamic anchor scheme performs very well in most conditions except when the context transfer cost is high (when the server is heavy). The centralized scheme performs the best at low SMR and high CMR. The fully distributed scheme performs the best at high SMR and high CMR. The static anchor scheme performs reasonably well under a wide range of parameter values, especially when CMR is low, SMR is low, or the context transfer cost is low. However its performance degrades more rapidly with the increase of CMR or SMR values, when compared with fully distributed scheme or dynamic anchor scheme. These results suggest that different users with vastly different mobility patterns should adopt different integrated location and service management methods to optimize system performance. We have tested the sensitivity of these results with respect to a random waypoint mobility model and the simulation results indicate that the cost-rate curves obtained for the four

schemes remain virtually the same compared with those obtained based on the random movement model, with the same cross-over trend observed. We have further tested the sensitivity of the analytical results with respect to the probability distribution function of the residence time. The sensitivity analysis shows our analytical results are valid and are not sensitive to the residence time distribution used.

Chapter 6

CONCLUSION AND FUTURE WORK

This dissertation proposes and analyzes effective and efficient algorithms for location management and service management in mobile wireless networks, focusing on personal communication service (PCS) cellular networks. We tackle this issue by considering three topics: location management, service management, and integrated location and service management.

6.1 Summary of Contributions

We have developed a two-level hierarchical performance model as a uniform framework to quantitatively assess and compare the performance characteristics of a number of existing location management algorithms. This work lays the foundation for designing and analyzing efficient algorithms for location management. Based on the insight gained, we have developed a hybrid scheme that combines replication and forwarding techniques, known to be effective in reducing user search and update costs at high *CMR* and low *CMR* values, respectively. We demonstrated that the hybrid scheme not only can be uniformly applied to all users with different *CMR* ratios, but also outperforms both replication (being most effective when *CMR* is high) and forwarding (being most effective when *CMR* is low), as well as an algorithm that switches between replication and forwarding based on the *CMR* value of the mobile user.

For service management, we have proposed and analyzed per-user proxy-based service management schemes and identified the best service area for service handoff to minimize the signaling and network cost based on user and service characteristics. Our results showed that the best proxy-based mobile service management scheme outperforms non-proxy-based schemes over a wide range of identified conditions. Further, we have demonstrated that when the mobile user is concurrently engaged in multiple services, the per-service proxy scheme that uses a separate proxy for each service outperforms the aggregate proxy scheme that uses a single proxy to interface with multiple services taking their aggregate service characteristics into consideration.

To further reduce the network signaling and communication cost, we have investigated a class of integrated location and service management schemes by co-locating an MH's service proxy with the MH's location database. We showed that integrated location and service management is a

viable concept applicable to the PCS network on a per-user basis for general server applications. Our analytical and simulation results showed that, when given an MH's mobility, service characteristics and network conditions through a set of parameters identified in the dissertation, there exists an optimal integrated location and service management scheme (in some cases the best two schemes are very close) that would minimize the overall network communication cost as a result of executing the MH's location and service operations.

6.2 Publications

The research work has resulted in the following publications:

- I.R. Chen and B. Gu, "A comparative cost analysis of degradable location management algorithms in wireless networks," *The Computer Journal,* Vol. 45, No. 3, 2002, pp. 304-319.

- I.R. Chen and B. Gu, "Quantitative analysis of a hybrid replication with forwarding strategy for efficient and uniform location management in mobile wireless networks," *IEEE Transactions on Mobile Computing*, Vol. 2, No. 1, 2003, pp. 3-15.

- B. Gu and I.R. Chen, "Performance analysis of location-aware mobile service proxies for reducing network cost in personal communication systems," *ACM/Kluwer Journal on Mobile Networks and Applications (MONET),* Vol. 10, No. 4, 2005, pp. 453-463.

- I.R. Chen, B. Gu and S.T. Cheng, "On integrated location and service handoff schemes for reducing network cost in personal communication systems," *IEEE Transactions on Mobile Computing,* accepted to appear, 2005.

The first and second publications are based on the concept of analyzing existing location management algorithms and designing hybrid schemes to combine the benefits of individual algorithms to minimize the network signaling cost for location management in PCS cellular networks. The content of Chapter 3 in the dissertation is largely based on these two papers. The third publication is based on the concept of using location-aware mobile proxy on a per-user basis to balance the cost between proxy maintenance and service packet delivery by determining the optimal service area per service so as to minimize the network cost for service management. The content of Chapter 4 is largely based on this paper. The fourth paper discusses integrated location and service management. Specifically, we investigate and analyze four integrated location and

service management schemes to explore this cost tradeoff with the goal to identify conditions under which a particular scheme should be adopted by an MH based on each MH's own mobility and service characteristics for network cost minimization. The content of Chapter 5 is based on this paper augmented with simulation validation and sensitivity analysis.

6.3 Potential Future Research

Our work has focused on PCS cellular networks. Here we outline future possible research extensions from this work. While next generation mobile wireless systems still lack a clear definition, a wide recognition is that they will likely be all IP-based [31][40]. The IETF solution for enabling the IP mobility is Mobile IP [3][43]. While Mobile IP provides the basic mobility management in IP networks, there are still many challenges, including handoff latency and inefficiency for micro-mobility. A future work deriving from the dissertation is to extend the integrated scheme to all IP-based wireless environments. For example, a recent study [53] shows that the strict layered architecture is not well suited for seamless inter-domain handoffs. High performance handoffs require high interaction between layers as well as new functionality for optimally handling these interactions. An integrated scheme utilizing underlying IP-based location management with upper-layer service characteristics represented by a personal service proxy can be considered. On the one hand, the proxy takes advantages of the visibility of user profile and network/device profiles and utilizes the location and presence information from the underlying network to provide location-aware and personalized services for value-added services providers or mobile users. On the other hand, the proxy is service-based and thus can take specific service characteristics of mobile applications into consideration to make right choices about when and how often to move to minimize the network signaling cost and data traffic. The work on location management, service management and possibly integrated location and service management for future all IP-based wireless networks can leverage the modeling and analysis methodologies developed in the dissertation, possibly tailoring them to all IP-based networking environments.

Another future research extension is to consider future mobile applications involving smart terminals [58]. Smart terminals are capable of reporting their locations which may necessitate new location and service management schemes to be used (e.g., paging and letting smart terminals inform ongoing services of their location changes) for efficiency reasons and providing better QoS support.

BIBLIOGRAPHY

[1] I.F. Akyildiz, J. McNair, J. Ho, I. Uzunalioglu and W. Wang, "Mobility management in next generation wireless systems," *Proceedings of the IEEE*, Vol. 87, No. 8, Aug. 1999, pp. 1347-1384.

[2] I.F. Akyildiz, J.S.M. Ho. and Y.-B. Lin, "Movement-based location update and selective paging for PCS networks," *IEEE/ACM Transactions on Networking*, Vol. 4, No. 4, Aug. 1996, pp. 629-638.

[3] I. F. Akyildiz, J. Xie, and S. Mohanty, "A Survey on Mobility Management in Next Generation All-IP Based Wireless Systems," *IEEE Wireless Communications*, Vol. 11, No. 4, Aug. 2004, pp. 16-28.

[4] A. Bar-Noy, I. Kessler and M. Sidi, "Mobile users: to update or not to update?," *ACM/Baltzer Wireless Networks*, Vol. 1, No. 2, July 1995, pp. 175-185.

[5] P. Bellavista, A. Corradi and C. Stefanelli, "The ubiquitous provisioning of Internet services to portable devices," *IEEE Pervasive Computing*, Vol. 1 No. 3, 2002. pp 81 –87.

[6] C. Bettstetter, G. Resta and P. Santi, "The node distribution of the random waypoint mobility model for wireless ad hoc networks," *IEEE Transactions on Mobile Computing,* Vol. 2, No. 3, July-Sept. 2003, pp. 257-269.

[7] A. Bhattacharya and S.K. Das, "LeZi-Update: an information theoretical approach to track mobile users in PCS networks," *ACM/IEEE MobiCom'99*, Seattle, WA, Aug. 1999, pp. 1-12.

[8] T. Campbell et al., "Comparison of IP Micromobility Protocols," *IEEE Personal Communications*, Vol. 9, No. 1, Feb. 2002, pp 72-82.

[9] I.R. Chen, T.M. Chen and C. Lee, "Performance evaluation of forwarding strategies for location management in mobile networks," *The Computer Journal*, Vol. 41, No. 4, 1998, pp. 243-253.

[10] I.R. Chen, T.M. Chen and C. Lee, "Agent-based forwarding strategies for reducing location management cost in mobile networks," *ACM/Baltzer Journal on Mobile Networks and Applications (MONET)*, Vol. 6, No. 2, 2001, pp. 105-116.

[11] I.R. Chen and B. Gu, "A comparative cost analysis of degradable location management algorithms in wireless networks," *The Computer Journal*, Vol. 45, No. 3, 2002, pp. 304-319.

[12] G. Cho and L.F. Marchall, "An Efficient Location and Routing Scheme for Mobile Computing Environments," *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 5, 1995, pp. 868-879.

[13] H. Choi, V.G. Kulkarni, and K.S. Trivedi, "Markov regenerative stochastic Petri nets," *Performance Evaluation*, Vol. 20, No. 1-3, 1994, pp. 337-357.

[14] Y.N. Doganata, T.X. Brown and E.C. Posner, "Call setup strategy trade-off for universal digital portable communication," *Computer Networks and ISDN system,* Vol. 20, Netherlands, 1990, pp. 455-464.

[15] M. H. Dunham and V. Kumar, "Impact of mobility on transaction management," *Proceedings of the International Workshop on Data Engineering for Wireless and Mobile Access* (MobiDE '99), Seattle, WA, USA, Aug. 1999, pp.14-21.

[16] EIA/TIA, *Cellular Radio Telecommunication Intersystem Operations*, Technical Report IS-41 (Revision B), EIA/TIA, July 1991.

[17] I. Elsen et al., "Streaming Technology in 3G Mobile Communication Systems," *Computer*, Vol. 34, No. 9, Sep., 2001, pp. 46-52.

[18] M. Endler, D. M. Silva, K. Okuda, "RDP: A result delivery protocol for mobile computing," *Proc. of the Int. Workshop on Wireless Networks and Mobile Computing* (WNMC), Taiwan, 2000, pp D36-D43.

[19] R.H. Gau and C.W. Lin, "Location Management of Correlated Mobile Users in the UMTS," *IEEE Transactions on Mobile Computing*, Vol. 4, No. 6, Nov./Dec., 2005, pp. 641-651.

[20] C. Gourraud, *Services and System Aspects; Virtual Home Environment/Open Service Access,* V5.1.0, 3GPP TS 23.127, Mar. 2002.

[21] S. Hadjiefthymiades and L. Merakos, "Using proxy cache relocation to accelerate Web browsing in wireless/mobile communications," *Proc. of the 10th international Conference on World Wide Web* (WWW'01), Hong Kong, May 2001, pp. 26-35.

[22] J. Hjelm, *Creating Location Service for the Wireless Web,* Wiley Computer Publishing, 2002.

[23] J.S.M. Ho, and I.F. Akyilidiz, "Local anchor scheme for reducing signaling costs in personal communications networks," *IEEE/ACM Transactions on Networking*, Vol. 4, No. 5, Oct. 1996, pp. 709-725.

[24] R. Jain and N. Krishnakumar, "Network support for personal information services to PCS users," *IEEE Conference on Networks for Personal Communications*, 1994, pp. 1-7.

[25] R. Jain, Y.B. Lin, C. Lo and S. Mohan, "A caching strategy to reduce network impacts of PCS," *IEEE Journal on Selected Areas in Communications*, Vol. 12, No. 8, Oct. 1994, pp. 1434-1444.

[26] R. Jain, Y.B. Lin, C. Lo and S. Mohan, "A forwarding strategy to reduce network impacts of PCS," *14th Annual Joint Conference of the IEEE Computer and Communications Societies, (IEEE INFOCOM '95)*, Vol. 2, 1995, pp. 481-489.

[27] L. Kleinrock, *Queuing Systems, Volume 1: Theory*, New York: John Wiley and Sons, 1975.

[28] P. Krishna, N.H. Vaidya and D.K. Pradhan, "Location management in distributed mobile environment," *3rd Inter. Conf. Parallel and Distributed Information Systems*, 1994, pp. 81-88.

[29] N. Krishnakumar and R. Jain, "Escrow techniques for mobile sales and inventory applications," *ACM Wireless Networks,* Vol. 3, No. 3, 1997, pp. 235-246.

[30] G. Krishnamurthi, M. Azizoglu and A. Somani, "Optimal location management algorithms for mobile networks," *ACM/IEEE MobiCom'98*, Dallas, TX, Oct. 1998, pp. 223-232.

[31] T. T. Kwon et al., "Mobility Management for VoIP Service: Mobile IP vs. SIP," IEEE Wireless Communications, Vol.9, No.5, Oct. 2002, pp. 2-11.

[32] W.R. Lai and Y.B. Lin, "Mobility database planning for PCS," *1996 Workshop Distributed System Technologies and Applications*, Tainan, Taiwan, 1996, pp. 263-269.

[33] J. Li and Y. Pan, "A dynamic HLR location management scheme for PCS networks," *Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2004)*, Vol. 1, Mar. 2004, pp 276-279.

[34] Y.B. Lin, "Reducing location update cost in a PCS network," *IEEE/ACM Transactions on Networking*, Vol. 5, No. 1, Feb. 1997, pp. 25-33.

[35] Y.B. Lin and I. Chlamtac, *Wireless and Mobile Network Architecture,* John Wiley & Sons, Inc. 2001.

[36] Y.B. Lin, L.F. Chang and A. Noerpel, "Modeling Hierarchical Microcell and Macrocell PCS Architecture," *Proc. of IEEE ICC'95*, June 1995, pp 405-409.

[37] M. H. MacDougall, *Simulating Computer Systems*, MIT Press, 1987.

[38] S. Mohan, and R. Jain, "Two user location strategies for personal communications services," *IEEE Personal Communications*, Vol. 1, No. 1, 1994, pp. 42-50.

[39] M. Mouly and M.B. Pautet, *The GSM System for Mobile Communications,* 49 rue Louise Bruneau, Palaiseau, France, 1992.

[40] K.Murakami et al., "Mobility Management Alternatives for Migration to Mobile Internet Session-Based Services," *IEEE Journal on Selected Areas in Communications*, Vol. 22, No. 5, June 2004, pp. 818-833.

[41] W. Navid and T. Camp, "Stationary Distributions for the Random Waypoint Model," *IEEE Transactions on Mobile Computing,* Vol. 3, No. 1, Jan-Feb 2004, pp. 99-108.

[42] S. Panagiotakis and A. Alonistioti, "Intelligent Service Mediation for Supporting Advanced Location and Mobility-aware Service Provisioning in Reconfigurable Mobile Networks," *IEEE Wireless Communications*, Vol. 9, No. 5, Oct. 2002, pp 28-38.

[43] C. Perkins, *IP Mobility Support,* IETF RFC 2002, Oct. 1996; http://www.rfc-editor.org/rfc/rfc2002.txt.

[44] C. Perkins, *IP Mobility Support for IP v4*, IETF RFC 3344, Aug. 2002; http://www.rfc-editor.org/rfc/rfc3344.txt.

[45] C. Perkins and D. Johnson, *Route Optimization in Mobile IP*, Internet draft, work in progress, Feb. 2000.

[46] E. Pitoura and G. Samaras, *Data Management for Mobile Computing,* Kluwer Academic Publishers, 1998.

[47] A Quintero, "A User Pattern Learning Strategy for Managing Users' Mobility in UMTS Networks," *IEEE Transactions on Mobile Computing*, Vol. 4, No. 6, Nov./Dec., 2005, pp. 552-566.

[48] S. Rajagopalan and B.R. Badrinath, "An adaptive location management strategy for Mobile IP," *Proceedings of the first annual international conference on Mobile computing and networking*, Berkeley, CA, USA, 1995, pp. 170-180.

[49] S. Rao, B. Gopinath and D. Kurshan, "Optimizing call management of mobile units," *3rd IEEE Inter. Symp. Personal, Indoor and Mobile Communications*, 1992, pp. 225-229.

[50] C. Rose and R. Yates, "Ensemble polling strategies for increased paging capacity in mobile communication networks," *ACM/Baltzer Wireless Networks*, Vol. 3, No. 2, May 1997, pp. 159-167.

[51] M. Roussopoulos et al., "Personal-level routing in the mobile people architecture," *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, Oct. 1999.

[52] R. Sahner, K. S. Trivedi and A. Puliafito, *Performance and Reliability Analysis of Computer Systems: An Example-Based Approach Using the SHARPE Software Package,* Kluwer Academic, 1996.

[53] A. Sanmateu et al., "Using Mobile IP for provision of seamless handoff between heterogeneous access networks, or how a network can support the always-on concept," *EURESCOM Summit2001,* Nov. 2001, Heidelberg.

[54] S.K. Sen, A. Bhattacharya and S.K. Das, "A selective location update strategy for PCS users," *ACM/Baltzer Wireless Networks*, Vol. 5, No. 5, Sept. 1999, pp. 313-326.

[55] G. Shih and S. Shim, "A service management framework for M-commerce applications," *Mobile Networks and Applications*, Vol. 7, No. 3, 2002, pp. 199-212.

[56] N. Shivakumar, J. Jannink, and J. Widom, "Per-user profile replication in mobile environments: algorithms, analysis and simulation results," *ACM-Baltzer Journal of Mobile Networks and Nomadic Applications (MONET)*, Vol. 2, No. 2, 1997, pp. 129-140.

[57] R. Subrata and A.Y. Zomaya, "Evolving Cellular Automata for Location Management in Mobile Computing Networks," *IEEE Transactions on Parallel and Distributed Systems*, Vol. 14, No. 1, Jan. 2003, pp. 13-26.

[58] M. Tariq and A. Takeshita, "Management of cacheable streaming multimedia content in networks with mobile hosts," *IEEE 2002 Global Telecommunication Conference*, Vol. 3, Taipei, Taiwan, Nov. 2002, pp. 2245-2249.

[59] K.S. Trivedi, G. Ciardo, and J. Muppala, *User Manual SPNP Version 6*, Dept. of Electrical Engineering, Duke University, Durham, N.C., 1999.

[60] E. Wedlund and H. Schulzrinne, "Mobility Support using SIP," *ACM/IEEE Internationl Conference on Wireless and Multimedia*, (WOWMOM) Aug. 1999, pp 76-82.

[61] Wireless Application Protocol Forum, *WAP Architecture: Wireless Application Protocol Architecture Specification,* WAP-210-WAPArch-20010712, July 2001.

[62] V.W.S. Wong and V.C.M. Leung, "Location management for next-generation personal communications networks," *IEEE Network*, Sept. /Oct. 2000, pp. 18-24.

[63] J. Wrolstad, "Wireless Content Flood Ahead," *Wireless NewsFactor,* Jan. 28, 2002; http://www.wirelessnewsfactor.com/perl/story/16014.html.

[64] J. Wu, H.P. Lin and L.S. Lan, "A New Analytic Framework for Dynamic Mobility Management of PCS Networks," *IEEE Transactions on Mobile Computing*, Vol. 1, No. 3, 2002, pp 208-220.

[65] T. Yoshimura, Y. Yonemoto, T. Ohya, M. Etoh, and S. Wee, "Mobile streaming media CDN enabled by dynamic SMIL," *11th International World Wide Web Conference*, Honolulu, Hawaii, May 2002, pp. 651-661.

[66] F. Yu, V.W.S. Wong, V.C.M. Leung, "Performance Enhancement of Combining QoS Provisioning and Location Management in Wireless Cellular Networks," *IEEE Transactions on Wireless Communications*, Vol. 4, No. 3, May 2005, pp 943- 953.

[67] H. Yumiba, K. Imai and M. Yabusaki, "IP-Based IMT Network Platform," *IEEE Personal Communication*, Vol. 8, No. 5, Oct. 2001, pp 18-23.

[68] M. Zarri, *Service Aspects; The Virtual Home Environment,* V5.3.0, 3GPP TS 22.121, Mar. 2003.

[69] B. Zenel and D. Duchamp, "General Purpose Proxies: Solved and Unsolved Problems," *Proceedings of the Sixth Workshop on Hot Topics in Operating Systems*, May 1997, pp 87-92.

[70] X. Zeng, R. Bagrodia and M. Gerla, "GloMoSim: A Library for Parallel Simulation of Large-scale Wireless Networks," *Proc. of the 12th Workshop on Parallel and Distributed Simulations*, May 1998, pp 154-161.

[71] A. Zimmermann, *TimeNET: A Software Tool for the Performability Evaluation with Stochastic Petri Nets*, TU Berlin, 2001.

Appendix A: Acronyms and Abbreviations

BS	base station
CDPD	cellular digital packet data
CMR	call to mobility ratio
FRA	forwarding and resetting algorithm
GPRS	General Packet Radio Services
GSM	Global System for Mobiles
HLR	home location register
LA	local anchor
LAA	local anchor algorithm
LAN	local area network
LSTP	local signal transfer point
MH	mobile host
MONET	mobile networks and applications
MSC	mobile switching center
PCS	personal communications services
PLA	paging and location updating algorithm
PSTN	public switched telephone network
RA	registration area
RSTP	regional signal transfer point
RWP	random waypoint
SDF	shortest-distance-first
SMPL	simulation model programming language
SMR	service to mobility ratio
SPN	Stochastic Petri Net
VHE	virtual home environment
VLR	visitor location register
WAP	wireless application protocol

Appendix B: Vita

Baoshan Gu received the BS degree from University of Science and Technology of China, Hefei, China, in 1992 and the MS degree in computer architecture from Institute of Computing Technology, Chinese Academia of Sciences, Beijing, China, in 1995. From 1995 to 2000, he was a research and development engineer in Institute of Computing Technology, Chinese Academia of Sciences. From 2003, he has been a software engineer at QueTel Corp. in Virginia where he developed mobile applications. He has been a PhD student since 2000 in the Department of Computer Science, Virginia Tech, where he is a research assistant in the Systems and Software Engineering Laboratory and is completing his PhD degree in the fall of 2005. His research interests include next-generation wireless system architectures, design and evaluation of location and service management schemes in mobile computing environments, and mobile multimedia systems.