# MOBILITY AND SERVICE MANAGEMENT FOR FUTURE ALL-IP BASED WIRELESS NETWORKS

Weiping He

Preliminary Research Document 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, Chair
Dr. Csaba Egyhazy
Dr. Mohamed Eltoweissy
Dr. Chang-Tien Lu
Dr. Gregory Kulczycki

December, 2006
Falls Church, Virginia

Keywords: Mobile IP, proxy, network protocol, regional registration, integrated mobility and service management, data consistency, performance analysis, optimization, Petri net.

©Copyright 2006, Weiping He

# Abstract

The next generation wireless network will provide not only voice but also data services. With the success of the Internet, it is widely believed that IP will become the foundation of next generation wireless networks. With the help of IETF standardization, IP-based wireless networks can benefit from existing and emerging IP related technologies and services. One key issue is how to provide uninterrupted, reliable and efficient data services to a mobile node (MN) in wireless networks. This dissertation concerns two major system-support mechanisms in future all-IP based wireless networks, namely, mobility management and service management.

Mobility management addresses the issues of how to track and locate a mobile node efficiently. Service management addresses the issues of how to efficiently deliver services to mobile nodes. This dissertation aims to design and analyze integrated mobility and service management schemes for future all-IP based wireless systems. We propose and analyze per-user regional registration schemes for integrated mobility and service management with the goal to minimize the network signaling and packet delivery cost in future all-IP based wireless networks. We show that, when given a set of parameters characterizing the operational and workload conditions of a MN, there exists an optimal regional area size for the MN such that the network communication cost is minimized for serving mobility and service management operations of the MN.

If access routers in future all-IP based wireless networks are restricted to perform network layer functions only, we investigate the design of intelligent routers, called dynamic mobility anchor points (DMAPs), to implement per-user regional management in IP wireless networks. These DMAPs are access routers (ARs) chosen by individual MNs to act as a regional router to reduce the signaling overhead for intra-regional movements. The DMAP domain size, i.e., the number of subnets covered by a DMAP, is based on a MN's mobility and service characteristics. A MN optimally determines when and where to launch a DMAP to minimize the network cost in serving the user's mobility and service management operations. We show that there exists an optimal DMAP domain size for each individual MN.

If access routers are powerful and flexible in future all-IP based to perform network-layer and application-layer functions, we propose the use of per-user proxies that can run on access routers. The user proxies can carry service context information such as cached data items and Web processing objects, and perform context-aware functions such as content adaptation for services engaged by the MN to help application executions. Under the proxy-based regional management scheme, a client-side proxy is created on a per-user basis to serve as a gateway between a MN and all services engaged by the MN. Leveraging Mobile IP with route optimization, the proxy runs on a foreign agent/access router and cooperates with the home agent and foreign agent/access router of the MN to maintain the location information of the MN, in order to facilitate data delivery by services engaged by the MN. Further, the proxy optimally determines when to move with the MN so as to minimize the network cost associated with the user's mobility and service management operations.

The proxy-based scheme supports query processing mobile applications. To improve query performance, the MN stores frequently used data in its cache. The MN's proxy receives invalidation reports or updated data objects from application servers, i.e., corresponding nodes (CN) for cached data objects stored in the MN. If the MN is connected, the proxy will forward invalidation reports or fresh data objects to the MN. If the MN is disconnected, the proxy will store the invalidation reports or fresh data objects, and, once the MN is reconnected, the proxy will forward the latest cache invalidation report or data objects to the MN. We show that there is an optimal "service area" under which the overall cost including query processing cost and location management cost is minimized.

We demonstrate that our proposed per-user regional management scheme outperforms basic Mobile IPv6, Mobile IPv6 Regional Registration, and Hierarchical Mobile IPv6 that do not consider integrated mobility and service management and that use static regional routers to server all MNs in the system. We will develop a simulation model based on ns2 to validate analytical results. We will also investigate mobile applications to which the proposed integrated mobility and service management scheme can be applied in Mobile IP systems.

# Acknowledgments

I would like to give my special and sincere thanks to my advisor, Dr. Ing-Ray Chen, who has constantly provided valuable guidance and patience to my research work.

I am indebted to my friend and big-brother Baoshan Gu for his support and advice. He has always been keen to help even in most difficult moments.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

The next generation wireless networks will provide not only voice but also data services. With the success of Internet, it is widely believed that IP will become the foundation of next generation wireless network. With the help of IETF standardization, IP-based wireless networks can benefit from existing and emerging IP related technologies and services. They can provide data, voice and multimedia services over an IP bearer. Such network architectures are referred as all-IP based wireless networks. Future all-IP based wireless networks will allow users to maintain service continuity while moving through wireless networks.

IP stands for "Internet Protocol". It specifies the packet format and address scheme. Each packet has a header that specifies the two end points of a packet transfer. IP is a connectionless packet delivery protocol. One of the design principles of IP is *IP over everything and everything over IP*. It means that IP can run on top of any link layer technology and any service can run on top of IP [56]. IP has well defined interfaces between layers. IP has been the ubiquitous protocol for fixed networks. IP-based equipment will become cheaper for wireless networks in the very near future. Using IP for both fixed and wireless networks enables a single unified infrastructure for all service delivery. This can reduce capital and maintenance costs. In an all-IP network, the same application can be available for mobile users as well as for fixed users. Existing applications will not need to be rewritten specially for wireless networks. The flexibility of an all-IP based wireless networks will liberate application developers from having to understand network details so they can concentrate on application development. Furthermore, new applications can benefit from mobility such as location based services. Wireless networks operators can also increase the range of services and gain extra revenues [56].

A mobile node (MN) in all-IP based wireless systems is a computing device that can change its point of attachment from one network or subnet to another [1]. It can continue communicating with other nodes when changing its location. A MN experiences many constraints that distinguish it from the traditional fixed hosts in distributed computing environments, such as relatively poor resources and highly variable wireless connectivity. These constraints require us to rethink about how to design mobile applications and systems. Furthermore, traditional wireless networks are based on circuit-switch technology, while IP is based on packet technology. The migration from circuit switched networks to packet based IP networks enables many more flexible services. However, it also brings out a number of design issues. One of the key issues is how to provide uninterrupted, reliable and cost efficient data services to mobile nodes in wireless networks.

In this dissertation, we aim to design and analyze integrated mobility and service management schemes for future all-IP based wireless networks to reduce network signaling and communication cost. The goal is to minimize not only the mobility management overhead, but also the service overhead associates with data delivery.

## 1.1 Mobility Management

Mobility is the ability of a node to change its point-of-attachment from one link to another while maintaining all existing communications [48]. Mobility is an important feature in any wireless networks. Mobility management is a key design element and an integrated part of any wireless network architectures [39]. Mobility management enables wireless networks to locate the roaming mobile node for service delivery and to maintain active connections as the MN is moving and changing its access point to the network.

To ensure that a mobile node $A$ can communicate with some node $B$ in a wireless network, the network infrastructure needs to ensure that (a) $A$'s location can be determined so that a route can be established between $A$ *and* $B$; (b) when $A$ moves out of the range of the current access point (AP), it establishes a connection with another AP; and (c) data packets are routed correctly to the new AP. Mobility Management includes two parts: location Management and handoff Management. Points (a) above deals with location management, while points (b) and (c) deal with handoff management.

- Location Management

  As in PCS wireless networks [20], location management keeps track the location of a MN for service delivery. It informs the sender of MN's new address so that future packets can be routed to the MN's current address. It is a two-stage process. The first stage is location registration. In this stage, the MN notifies the network of its new access point when it moves across location boundaries. The second stage is packet delivery. The network is queried for the MN's location and a connection to the MN is initiated. Important research issues with location management include how to reduce signaling costs and packet delivery latency.

- Handoff Management

  Handoff management is the process by which an MN keeps ongoing connections when it moves from one subnet to another. Handoff management ensures that the MN always remains reachable to receive/send packets. A handoff process can be divided into four tasks: (1) deciding when a handoff should occur; (2) selecting a new AP from several APs nearby; (3) acquiring resources such as channels; (4) informing the old AP that it can reroute packets for the MN and transfer state information to the new AP. Important research issues in handoff management include how to reduce power consumption, how to eliminate packet loss and service disruption, and how to improve reliability and scalability.

## 1.2 Service Management

Service management ensures that mobile nodes get data services reliably, correctly and efficiently. Service management consists of service request management and service handoff management.

Service request management includes request handling; request delivery; request accounting, authentication and authorization (AAA). Request handling is to accept service requests from an MN and transform requests into proper form. Request delivery is to forward server replies to the MN.

Service handoff management is the process by which an MN keeps its services connection when it moves from one access point to another one. Service context transfer is a key issue. For some real-time services, the context transfer cost many be high if there is a large number of users with frequent movements. For example, for a video on-demand application, the context information may include the video title, minutes played, and current frames being buffered at the server and played at MN.

## 1.3 Thesis Organization

This thesis is organized as follows. In Chapter 2, we present the motivation, methods and plan of the research. In Chapter 3, we give a survey of related work. In Chapters 4 and 5, we describe integrated mobility and service management for MIPv6. In Chapter 4 we consider the case in which access routers are restricted to perform network layer functions only. We discuss the design notion of dynamic mobility anchor points (DMAPs) by which A MN can optimally determine when and where to launch a DMAP to minimize the network cost in serving the MN's mobility and service management operations. In chapter 5 we consider the case in which access routers are powerful and flexible so that user proxies can be dynamically executed on access routers to perform network layer and application layer functions on behalf of users and applications. We propose and analyze a proxy-based regional registration scheme for integrated mobility and service management with the goal to minimize the network signaling and packet delivery cost in future all-IP based wireless networks. In Chapter 6, we demonstrate the applicability of our proxy-based regional registration scheme with a mobile query processing application for cache consistency and mobility management. Finally, Chapter 7 summarizes work completed and work to be done, along with the work schedule.

# Chapter 2

# Research Statement and Methods

In this chapter, we discuss our research statements, plan and methods for conducting the dissertation research. We first present the network architecture. By illustrating the challenges in IP-based wireless networks, we explain why future all-IP wireless networks need new mobility and service management schemes. We recognize that mobility and service management is fundamental to supporting ubiquitous mobile applications in all-IP based wireless networks. Therefore, the objective of this research is to develop new mobility and service management schemes to minimize the overall network cost.

## 2.1 Future All-IP based Wireless Network Architecture

Future all-IP based wireless networks provide network services based on the ubiquitous communication protocol: IP.

Figure 2.1 shows the basic architecture of future all-IP based networks. A permanent IP address is assigned to each MN. Each MN has a home agent (HA) that at all times knows the location of the MN. The HA accepts registration requests and updates the current point of attachment of the MN. While the MN is away from home, the HA intercepts packets on the home link destined to the MN's home address, encapsulates them, and tunnels them to the MN's registered care of address (CoA). A correspondent node (CN) communicates with the MN to provide services to MN via IP packets. Access routers (ARs) in IP networks provide routing services to MNs. An AR resides on the edge of an IP subnet and connects to one or more access points (APs). The APs may run different technologies. An AR offers IP connectivity to MNs, acting as a default router to the MNs it is currently serving. The AR may include intelligence beyond a simple forwarding service offered by ordinary IP routers. An AP is a layer 2 device which is connected to one or more ARs and offers the wireless link connection to the MN. APs are sometimes called base stations or access point transceivers. An AP may be a separate entity or co-located with an AR [30].

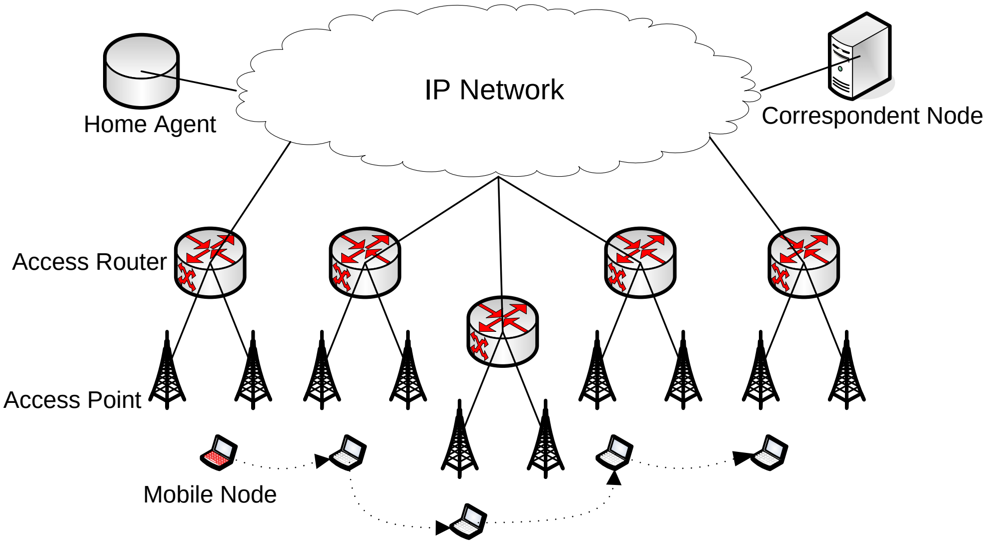

Figure 2.1: Future All-IP based Wireless Network Architecture

Future ARs are expected to be quite powerful and flexible. Some research have been done on deploying programmable agents in wireless IP systems [31, 41]. Assuming that future ARs are able to host agents or proxies, proxies can be dynamically downloaded and roam among ARs to perform network layer and application layer functions on behalf of users and applications to satisfy user QoS needs and reduce network overheads. This dissertation will investigate such proxy-based designs for mobility and service management.

## 2.2 Research Challenges

We recognize the following challenges facing mobility and service management in IP-based wireless networks [15] [42]:

1. Mobile connectivity is highly variable – Some subnets may offer reliable, high-bandwidth wireless connectivity, while others may only offer low-bandwidth. Wireless networks are expensive, offering low bandwidth and low reliability due to interference and noise than wireline networks. Therefore, MNs will encounter wide variation in connectivity ranging from adequate bandwidth to total lack of connectivity.

2. Mobile nodes are resource-poor relative to static elements – power and computational resource (CPU speed, memory size, disk capacity) are relatively limited for MNs. Furthermore, resources available to a MN vary, for example, a notebook vs. a palm. Power consumption must be taken into consideration in hardware and software design.

3. Workload to ARs is highly variable – The attachment of MNs to the wireless network changes as they move. This causes a highly variable workload to ARs in the system. Efficient mobility management schemes are needed to spread workload to ARs in the system and to handle rapidly changing location information of MNs.

4. Mobility and service characteristics of MNs are highly variable – If the mobility of a MN is high and the HA is far away, the handoff signaling cost would be high. On the other hand, if the packet exchange rate between the MN and CNs is high, the packet service cost would be high. To best balance the handoff signaling cost vs. packet delivery cost, each individual MN's mobility and service characteristics must be taken into account in the design. A management protocol without considering individual MN's mobility and service characteristics will not optimize the system performance.

It is predicted that vast majority of terminals will be mobile in a few years. The vast majority of traffic will originate from IP-based applications. The challenge is to deliver

IP-based applications to MNs [56]. In this research, we aim to design and analyze efficient integrated mobility and service management schemes for future all-IP based wireless systems. The goal is to minimize not only the mobility handoff overhead, but also the service overhead associated with data delivery.

## 2.3 Research Methods

The primary methods for achieving the goals and objectives of this research consist of the following steps:

- Extensive literature survey on existing mobility and service management schemes;
- Investigation of new techniques for efficient mobility and service management scheme based on existing protocols and methods;
- Assessment of the performance of proposed integrated mobility and service management schemes via modeling and analysis;
- Demonstration of the applicability of proposed mobility and service management schemes with a set of mobile applications;
- Development of a simulation model based on ns2 to validate analytical results.

## 2.4 Contribution

In this dissertation we will propose and analyze per-user regional registration schemes for integrated mobility and service management with the goal to minimize the network signaling and packet delivery cost in future all-IP based wireless networks. We will show that, when given a set of parameters characterizing the operational and workload conditions of a MN, there exists an optimal regional area size for the MN such that the network communication cost is minimized for serving mobility and service management operations of the MN.

If access routers in future all-IP based wireless networks are restricted to perform network layer functions only, we investigate the design of intelligent routers, called dynamic mobility anchor points (DMAPs), to implement per-user regional management in IP wireless networks. These DMAPs are access routers (ARs) chosen by individual MNs to act as a regional router to reduce the signaling overhead for intra-regional movements. The DMAP domain size, i.e., the number of subnets covered by a DMAP, is based on a MN's mobility and service characteristics. A MN optimally determines when and where to launch a DMAP to minimize the network cost in serving the user's mobility and service management operations. We show that there exists an optimal DMAP domain size for each individual MN.

If access routers are powerful and flexible in future all-IP based to perform network-layer and application-layer functions, we propose the use of per-user proxies that can run on access routers. The user proxies can carry service context information such as cached data items and Web processing objects, and perform context-aware functions such as content adaptation for services engaged by the MN to help application executions. Under the proxy-based regional management scheme, a client-side proxy is created on a per-user basis to serve as a gateway between a MN and all services engaged by the MN. Leveraging Mobile IP with route optimization, the proxy runs on a foreign agent/access router and cooperates with the home agent and foreign agent/access router of the MN to maintain the location information of the MN, in order to facilitate data delivery by services engaged by the MN. Further, the proxy optimally determines when to move with the MN so as to minimize the network cost associated with the user's mobility and service management operations.

The proxy-based scheme supports query processing mobile applications. To improve query performance, the MN stores frequently used data in its cache. The MN's proxy receives invalidation reports or updated data objects from application servers, or CNs for cached data objects stored in the MN. If the MN is connected, the proxy will forward invalidation reports or fresh data objects to the MN. If the MN is disconnected, the proxy will store the invalidation reports or fresh data objects, and, once the MN is reconnected, the proxy will forward the latest cache invalidation report or data objects to the MN. We show that there is an optimal "service area" under which the overall cost including query processing cost and location management cost is minimized.

We demonstrate that our proposed per-user regional management scheme outperforms basic Mobile IPv6, Mobile IPv6 Regional Registration, and Hierarchical Mobile IPv6 that do not consider integrated mobility and service management and that use static regional routers to server all MNs in the system. We plan to develop a simulation model based on ns2 to validate analytical results. We plan to investigate mobile applications to which the proposed integrated mobility and service management scheme can be applied in Mobile IP systems to demonstrate the applicability of our proposed per-user management schemes.

# Chapter 3

# Related Work

## 3.1 Efforts in PCS networks

The evolution of wireless technology has proliferated Personal Communications Service (PCS) networks [7]. Substantial research work in mobility management has been done in the PCS domain. Some of them can be extended to the IP based wireless networks. Currently there are two standards for PCS networks: Interim Standard 41(IS-41) of Electronic/Telecommunications Industry Association (EIA/TIA) and the Global System for Mobile Communications (GSM) [2]. IS-41 is widely used in North America while GSM exists mainly in Europe and Asia. Both are based on the two-level database hierarchy architecture, consisting of the Home Location Register (HLR) at the top level and Visitor Location Registers (VLRs) at the bottom level. IS-41 and GSM are very similar in mobility management. For simplicity, we would use IS-41 in this section.

In IS-41, the network coverage area is divided into cells. Each cell has a base station. Mobile hosts in a cell communicate with the network through the base station. A few cells can be grouped together to form a larger area called a Registration Area (RA). All base stations in one RA are connected to a mobile switching center (MSC). The MSC serves as an interface between the wireless network and the public switching telephone network (PSTN). Each MSC has a unique `MSCID` assigned to it. Each MH also has a unique identification `MHID`.

The HLR is a centralized database. It contains the user profiles of its assigned subscribers.

These user profiles record information such as the type of subscribed services, quality of service (QoS) requirement, billing information, and current location of the Mobile Host (MH) [21]. The VLRs are distributed throughout the PCS network. Each one stores the information of current MHs residing in its location. In the basic HLR/VLR scheme, the HLR keeps the location of the MH by knowing the VLR that the MH currently resides.

A review of mobility management schemes in PCS networks can be found in [2] [42] [57] . These schemes fall into two categories. The first category includes improvements on the current VLR/HLR scheme based on centralized database architectures. To reduce the lookup cost, the location of a mobile host can be maintained at additional sites. It may be cached at its caller's site or replicated at the sites from where the majority of calls originate [2] [22]. In [6], forwarding pointers are proposed to reduce the location update cost. If a MH moves, a new forwarding pointer between two VLRs is created. To search for a MH, the HLR is queried to know the first VLR in the forwarding chain. Then the system follows a chain of databases (VLRs) to locate the MH. This method reduces location updating cost when the call-to-mobility ratio is low. However, a number of VLR queries have to be performed in order to locate the mobile host. This introduces additional overhead in the call delivery process.

The profile based location strategy takes advantage of the user's mobility pattern. The network maintains a profile for each user. It includes a sequential list of registration areas (RAs) that the user is most likely to be located at in different time periods [27]. When a call arrives, the RAs on the list are paged sequentially. Location update is performed only when the mobile host moves to a new RA which is not on the list. In [32] a user profile replication with caching strategy is proposed. User profiles based on user mobility patterns are replicated and combined with caching. This reduces the communication cost (signaling message) and computational (database access) cost. In [46] a stochastic mobility model based on daily activity patterns is discussed. It provides a balance between deterministic and random mobility models.

The second category is based on a hierarchical database organization for locating mobile hosts. Ho et al. [21] discuss a dynamic hierarchical database architecture to allow the user location information distribution to be adjusted based on the mobility and calling patterns of MHs. It also adopts location pointers to indicate the current location of MHs. In [4], every level of the hierarchy represents a partition of geographic regions. The system records the location of every mobile user with a certain degree of accuracy in each level of the hierarchy. The degree of accuracy is increased as we go down the levels. The advantage of this scheme is that the upper bound of a search operation does not depend on the network size. Wang et al. [53] integrates two mechanism: distributed temporary location caches and distributed home location registers to reduce database access delay and to decrease network signaling traffic in updating and paging.

In [43], a load balancing protocol is proposed. Location updates and queries for a MH are multi-casted to subsets of location servers. These subsets change with time depending on the locations of MH/querying MSC and loads on the servers. They are chosen such that any two subsets have at least one common database. Upon receiving a location update request, the MSC uses the hash function with the MH's `MHID` and its `MSCID` as parameters in selecting a subset for update. Upon receiving a call delivery request, the MSC uses the hash function with its own `MSCID` and the called mobile's `MHID` as parameters to choose a subset for the query. The construction of the subsets guarantees that the MSC is able to find a database with the desired location information. In [20] this idea is extended by using caching to reduce the cost of call delivery. Furthermore, we use hashing to choose only one location server. This process would simplify the load balancing protocol and reduce the database operation overhead.

## 3.2 Link Layer Solutions

Link layer solutions provide mobility in the underlying radio systems. They ensure uninterrupted communications when a MN changes its position within the scope of an access router. For example, IEEE 802.11b allows a device to move within the scope of a single broadcast medium without breaking ongoing connections. Mobility support by link layer is also called access mobility [12]. However, it is possible for a MN to move between different links that are connected to different routers. Furthermore, the MN can move between different link layers such as IEEE 802.11b and UMTS. Link layer solutions are tightly coupled with specific wireless technologies. A link layer solution may not be enough and cannot be used as a general solution [49].

One example is the Inter-Access Point Protocol (IAPP) [23]. It is an extension of the IEEE 802.11 standard to support interoperability, mobility, handover and coordination among Access Points (AP) of wireless local area networks (WLAN) as shown in Figure 3.1. IAPP ensures all relevant information is delivered to the correct AP to which the station is associated. It provides a means to make APs communicate with each other. However, the IAPP does not scale to large number of MNs. It is specific for a particular layer-2 wireless local area networks, so inter-technology handoffs will be impossible.

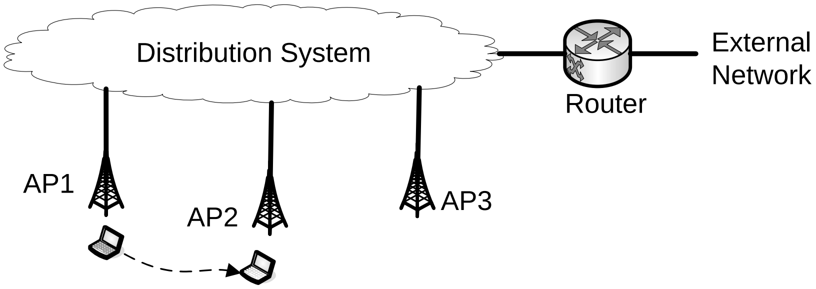

Figure 3.1: Inter-Access Point Protocol

## 3.3 Network Layer Solutions

Most of the mobile data services are based on a client-server computing paradigm [17]. in IP-based wireless networks. a packet delivered to a MN needs to know the current IP address of the MN. a MN can connect to the network via its network interface and be assigned with an IP address. An IP address consists of a network address and a host address. All nodes in the same subnet have the same network address but different host addresses. A router will only need one entry in its routing table for all hosts that have the same network address.

A data packet can be sent from a source node to a destination node through routers by its IP address. However, when a MN moves from one network to another, its IP address may become invalid in the new network. No IP packets can be routed the MN unless all routers on the path from the source node to the destination node are updated to include an entry for this IP address. It is undesirable because many routers will need to be updated whenever the MN moves.

Since the IP layer is about packet delivering, according to the design principle "obey the layer model," the IP layer is the ideal place to handle user mobility. In order to deliver IP packets properly, the system needs fixed identifiers to decide which node a packet is to be delivered. It also requires a variable locator to indicate where the MN is in the network. To deal with the problem of mobility, the system must have some kind of dynamic mapping between the fixed identifier and the variable locator.

### 3.3.1 MIPv4

Mobile IPv4 (MIPv4) [40] is an IETF standard and also the first network layer protocol to provide transparent dynamical mapping between a MN's fixed identifier and its variable locator. The basic infrastructure of Mobile IPv4 is showed in Figure 3.2. A MN is identified by its home address. If the MN is not in its home area, it has another address called its Care of Address (CoA) associated with its current foreign location. The Home Agent (HA) maintains a dynamic mapping between the home address and the CoA of the current foreign agent (FA) that the MN currently resides under. A corresponding node (CN) always sends packets to the MN by the MN's home address. The HA intercepts and tunnels data packets to the current FA which forwards to the MN. The Mobile IP protocol is transparent to mobile applications such that an application can reach the MN by the same home IP address while the MN is roaming. Standard IP routing protocols can be used with Mobile IP without modification. Only the MN, HA and FA need to know the internal working of Mobile IP.

When the MN crosses a foreign agent boundary, a location handoff occurs by which the MN must send its change of the CoA to the HA. If the mobility of the MN is high, and

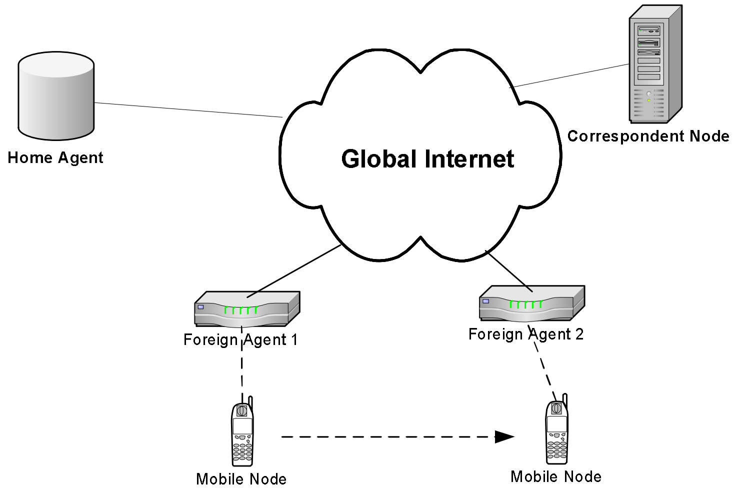

Figure 3.2: Basic Infrastructure of Mobile IP Systems.

the HA is far away from the current foreign agent, the overhead of informing the HA of the address change is high.

The mobile IP protocol incurs some problems: (1) it causes a high transmission and processing overhead because of the triangle HA-FA-MN routing issue. (2) handover may be slow. The MN must send its change of CoA to the HA. This may take a long time if the HA is far away. The latency involved in this handoff can exceed the threshold required for supporting real-time services.

### 3.3.2 Mobile IP Regional Registration (MIP-RR)

Mobile IP Regional Registration (MIP-RR) [18] has been proposed to reduce the location handoff overhead. As illustrated in Figure 3.3, when a MN arrives at a regional registration area govern by gateway foreign agent (GFA), it registers the address of the GFA as the CoA to the HA. When the MN moves among the subnets within the regional registration area, i.e., moving from one foreign agent (FA) to another, it only registers the address of the FA as the CoA to the GFA. When the MN moves to another regional area, it will perform a home registration to the HA again. If a CN wants to send packets to a MN, the CN will send them to the MN's home address. These packets are tunneled from the HA to the GFA of the MN. The GFA then forwards packets to the current FA under which the MN resides and then the FA finally forwards data packets to the MN. The MIP-RR protocol effectively reduces the overhead of location handoffs [37]. However, if a GFA covers too many FAs, then it tends to be away from the current FA so it introduces another layer of routing delay for data delivery, viz., data packets will take a HA-GFA-FA-MN route instead of a HA-FA-MN route as in basic Mobile IP. How to decide the number of FAs under one GFA is an open issue [5]. This dissertation research aims to solve this problem by determining the optimal service area under a GFA to minimize the network signaling cost and packet delivery cost.

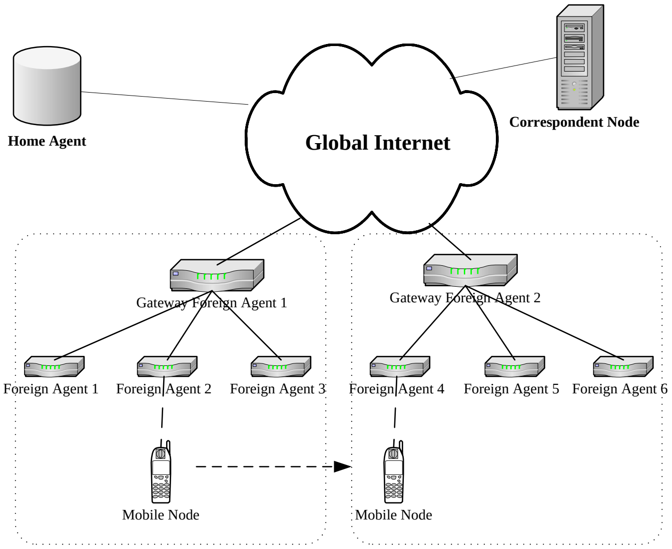

Figure 3.3: Basic Infrastructure of Mobile IP Regional Registration Systems.

### 3.3.3 MIPv6

MIPv6 is Mobile IP for IPv6, extending from MIPv4 for IPv4. The current state of mobility management under MIPv6 is summarized in Figure 3.4 [48, 49]. First an MN determines its current location using the IPv6 router discovery protocol. Then the MN acts as a fixed host to connect to its home link. In the case of a foreign link, the MN uses the IPv6 address autoconfiguration mechanism to acquire a care of address (CoA) on the foreign link. The MN then notifies its HA of its CoA. The MN also reports its CoA to CNs (i.e., application servers) with which it currently engages. If a CN knows the MN's CoA, then data packets

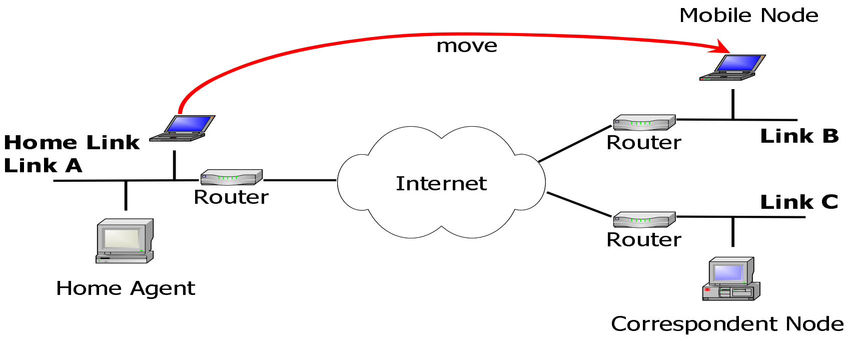

Figure 3.4: Infrastructure of Mobile IPv6 Systems.

from the CN can be directly sent to the MN with an IPv6 routing header specifying the CoA as the forwarding address.

A comparison of Mobile IPv6 and IPv4 is showed in Table 3.1 [48]. In particular, in MIPv6 there is no need to use FAs. A MN will acquire a CoA and will act as a FA in MIPv6. The route optimization is integrated in MIPv6. While in MIPv4, a change to the stack protocol in the CN is required to implement route optimization, in MIPv6 MN can inform CNs of CoA change directly as part of the protocol.

No IETF standard has been made for regional registration under MIPv6. The dissertation research proposes a regional registration protocol for MIPv6 with the goal to determine the optimal regional registration area to minimize the network signaling cost and packet delivery cost.

### 3.3.4 Hierarchical MIPv6

Hierarchical MIPv6 (HMIPv6) [50] is designed to reduce the network signaling cost for mobility management based on the observation that statistically local mobility accounts for more than 60% of movements made by a MN. Thus, instead of using a CoA, HMIPv6 also employs a Regional CoA (RCoA), which is the IP address allocated to the MN in the subnet of a mobility anchor point (MAP) that covers a number of subnets in a region. Whenever a MN moves from one subnet to another but is still within a region covered by the MAP,

| **Mobile IPv4** | **Mobile IPv6** |
|---|---|
| Foreign Agent | A “plain” IPv6 router on the foreign link (Foreign Agents no long exist) |
| Foreign Agent care-of address vs. colocated care-of address | All care-of addresses are collocated |
| Care-of address obtained via Agent Discovery, DHCP, or manually | Care-of address obtained via stateless address autoconfiguration, DHCP, or manually |
| Agent Discovery | Router Discovery |
| Authenticated registration with home agent | Authenticated registration with home agent and correspondents |
| Routing to mobile node via tunneling | Routing to mobile node via tunneling and source routing |
| Route optimization via separate protocol specification | Integrated support for route optimization |

Table 3.1: Comparison between Mobile IPv4 and Mobile IPv6.

the CoA change is only propagated to the MAP instead of to the HA and CNs, thus saving the signaling cost for mobility management. The HA and CNs ideally only know the MN's RCoA, so whenever the MN moves across a regional area and triggers a RCoA address change, the new RCoA address needs to be propagated to the HA and CNs. This concept can be applied at multiple levels in a hierarchical manner [38, 50] and can be combined with the use of Fast Handovers [14] that use forwarding pointers between the current and next subnets in a hybrid manner. MAPs in HMIPv6 are statically configured and shared by all MNs in the system. Access routers (ARs) are responsible for announcing their MAP's identity by means of router advertisement packets so that roaming MNs would know if they have crossed a MAP domain and need to perform a RCoA update to the HA and CNs.

In this dissertation, we adopt the MAP design concept in HMIPv6 for regional registration to reduce the network signaling cost of mobility management. Moreover, we propose and analyze a dynamic DMAP (DMAP) design so that the optimal service area of a MAP is *dynamically* determined on a per-MN basis to minimize the combined mobility and service management cost.

### 3.3.5 Intra-Domain Mobility Management Protocol

The IETF work-in-progress draft Intra-Domain Mobility Management Protocol (IDMP) [34] introduces the concept of domain mobility with a domain corresponding to a region in MIP-RR and HMIPv6, and a domain agent corresponding to a MAP in HMIPv6 to keep track of CoA of a MN as the MN roams within a domain. IDMP can be combined with fast handoff mechanisms utilizing multicasting [33] to reduce handoff latency and paging mechanisms to reduce the network signaling cost for intra-domain movements.

### 3.3.6 Cellular IP

Cellular IP uses mobile-originated data packets to maintain reverse path routes. Nodes in a cellular IP access network monitor mobile originated packets and maintain a distributed, hop-by-hop location database that is used to route packets to mobile hosts. Cellular IP uses IP addresses to identify mobile hosts [16]. Cellular IP incorporates a number of important cellular system design principles such as paging in support of passive connectivity.

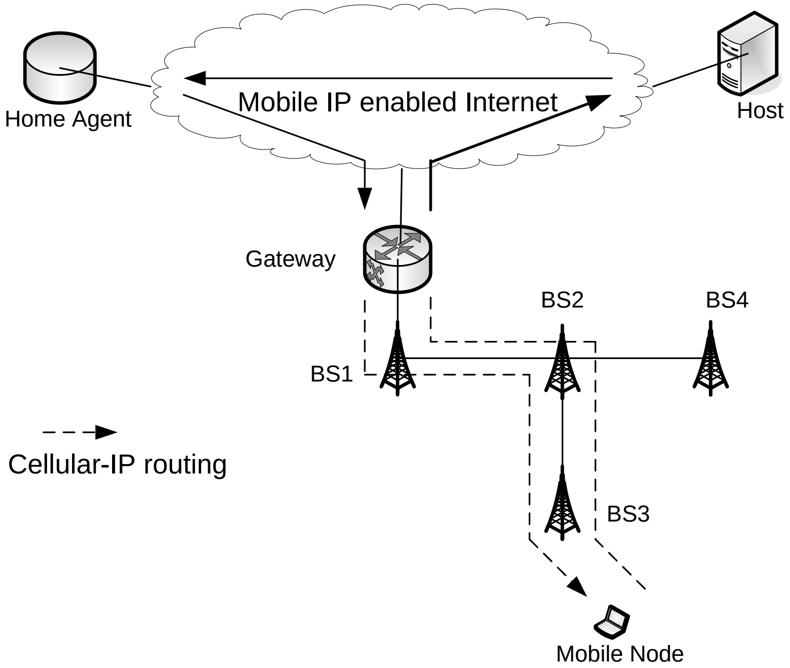

Figure 3.5: Cellular IP

As shown in Figure 3.5, the mobile node is identified by its home address. The gateway's address is used as its CoA. Base stations each have an AR. A route update packet travels from the MN all the way to the gateway, updating all the ARs on the way. Old routing entries in ARs are purged by timeout. Two types of handoffs are supported in Cellular IP:

- Hard Handoff: A MN listens to beacons transmitted by base stations and initiates a hard handoff based on signal-strength measurements. To perform a hard handoff, a MN tunes its radio to a new base station and sends a route-update packet. The route update message creates a routing path to the gateway, which configures a downlink route cache entry to point to the new base station.

- Soft Handoff: A MN notifies the new AR before actual handoff. The MN will also listen to transmissions from the old AR. The MN sets a flag in the route update packet, enabling cross-over routers to forward downstream packets to both the old and new ARs.

### 3.3.7 Handoff-Aware Wireless Access Internet Infrastructure (HAWAII)

Handoff-Aware Wireless Access Internet Infrastructure (HAWAII) is a domain-based approach for supporting mobility [44]. The combination of HAWAII for micromobility within a domain and Mobile IP for macromobility across domains provides for mobility across all levels. HAWAII has the following features:

- The MN obtains a co-located care-of address using DHCP when it is not in its home domain. If the MN moves within the foreign domain, the mobile host retains its care-of address unchanged.

- Forwarding entries for MNs are created and maintained using explicit signaling messages initiated by the MNs. When a MN transmits such a message on power-up or change of location, it is relayed, along the optimal path, to the domain root in the form of a Hawaii signaling message. All routers receiving this message establish and update mobile-specific entries for the reverse path packet forwarding.

- The domain root maintains a flat-address lookup table with forwarding metrics for all active mobile hosts in its domain. Each routing node is also required to maintain part of this table.

- Route updates only travel as far as the cross-over router.

- There are two approaches for handling handoffs. If the MN can connect to only one AR, packets will be forwarded from the old AR to the new AR. If the MN can connect to both ARs simultaneously, packets are diverted at the cross-over router as soon as the patch update message reaches it, and there will be no packet being forwarded from the old AR.

The integrated mobility and service management scheme proposed in the dissertation is generic and potentially can take a micromobility management scheme such as IDMP, Cellular IP or HAWAII for the intra-domain mobility management, and Mobile IP for the inter-domain management.

### 3.3.8 Multicasting-based approach

IP multicasting provides a mechanism for location independent addressing and packet delivery to a group of hosts that belong to a multicast group. It also provides efficient mechanisms for hosts to join and leave multicast groups. The principle of supporting host mobility using IP multicasting is to assign a unique multicast IP address to the MN. When the MN moves, the new AR will be added to the multicast group; the old AR will be removed from the multicast group. All packets from a correspondent node to a MN are treated as multicast packets, while packets from a MN to a static host are treated as standard unicast packets.

Mysore and Bharghavan proposed an approach called MSM-IP [36] based on multicast protocols. Each MN is assigned with a unique location independent (multicast) address. Packets from the correspondent node to the mobile node will be tunneled through a sequence of multicast routers in the multicast distribution tree and reach the mobile node, rather than go through a home network as in Mobile IP. To make handoff seamless, MNs will perform advance registration and have packets delivered to the next cell in advance. A comparison between multicast based approach and Mobile IP based approaches is given in the following table [36]:

| **Functionality** | **Mobile IP based** | **Multicast based** |
|---|---|---|
| Registration | A MN must register with a local agent when entering a new network | A MN that wishes to receive multicast datagrams must register with the local multicast router |
| Connectivity | Connectivity to the rest of the internet is provided by the foreign agent | A multicast router provides connectivity to the rest of the virtual multicast network to a host |
| Data forwarding | The foreign agents forward datagrams to the mobile host | The multicast router multicasts datagrams to the members in its subnet |
| Address translation | The HA translates the home address to the CoA | Multicast messages need no address translation |
| Routing | The HA tunnels messages to MN through the current FA | Messages sent to a destination multicast address are forwarded to multicast routers |

Table 3.2: A comparison between multicast based approach and Mobile IP based approaches

IP multicasting based addressing is designed mainly for multicast applications and represents a totally different approach from Mobile IP. This dissertation only deals with Mobile IP based applications.

## 3.4 Transport Layer Solutions

The solutions in the transport layers must be able to modify sockets and inform a peer's transport layer of such modification when a MN moves and changes its IP address.

One example is MSOCKS [28] for TCP connections proposed by Maltz and Bhagwat. The architecture is built around a split-connection proxy that is inserted into the communication path between a MN and its CN. As show in Figure 3.6, the MSOCKS architecture consists of three pieces: a user level MSOCKS proxy process running on a proxy machine; an in-kernel modification on the proxy machine to provide the TCP Splice service; and a shim MSOCKS library that runs under the application on the mobile node. No modifications are needed in the server, the mobile node, or the client application. Each logical communication session between MN and CN is split into two separate TCP connections. By using a technique called TCP splice, the proxy is able to maintain one stable data stream to the MN while allowing connection to be redirected. The proxy can simultaneously make and break connections to the MN as needed to migrate data streams between networks.

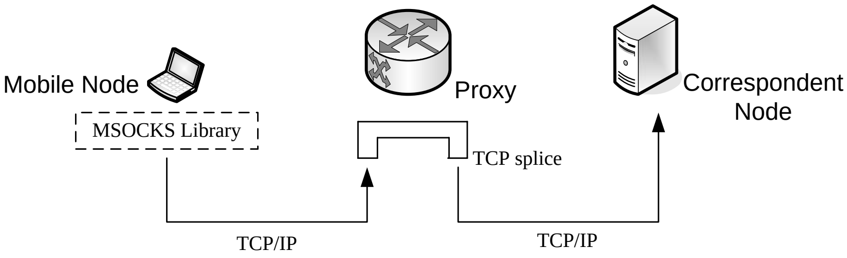

Figure 3.6: MSOCKS Architecture

However, this solution is for TCP only. Applications that do not use TCP will lose their connections when MN moves. There are different transport layer protocols such as TCP and UDP, so modifying every transport layer protocol to support mobility is a heavy task [49]. The integrated mobility and service management solution proposed in this dissertation deals with Mobile IP which is intrinsically a network layer protocol.

## 3.5 Application Layer Solutions

A couple of application layer solutions have been proposed for mobility management in the past few years. One approach is the SIP protocol [19], allowing an application to initiate, modify and terminate network service sessions. There are four basic elements in SIP: users agents, registrars, proxy servers and redirect servers. User agents listen for incoming messages and send SIP messages upon user actions or incoming messages. Registrars keep track of users within the assigned network domain. Proxy servers are application-layer gateways that forward SIP requests and responses. The redirect server returns the location of another SIP user agent or server where the user might be found [45]. Since a MN can register with the SIP server independent of its location, the MN can be reached even if it changes its location. Elin Wedlund et al. [55] proposed a scheme in which a SIP server is used in each MN's home network to receive registrations from the MN whenever the MN changes its location. When the CN sends an INVITE SIP message to the MN, the redirect server knows the current location of the MN and then forwards the INVITE message to the MN. If a MN moves during an active session, it must send a new INVITE message to the CN using the same call ID as in the original call setup. The new IP address is put in the contact field of the SIP messages. The CN will send future SIP messages to the new address. Thus the MN can keep the connection with the CN after handoffs. However, if the MN is far away from the home network, every time it moves, it will send a new registration to the home SIP server. This may incur a high load on the network and to the SIP server.

Another approach is the "Mobile People Architecture" (MPA) [29]. The main idea is that a user can be reached at any time no matter if the user is using a pager, computer, or cellular phone. This goal is achieved through the use of personal proxies to maintain personal mobility, i.e. a person moving to different terminals can remain in contact. The drawback of MPA is that it does not provide continued connectivity of ongoing applications.

The dissertation does not use application protocols for mobility management. However, for service management, we also use personal proxies to maintain service context to facilitate uninterrupted context-aware services when a MN moves from one subnet to another in MIP environments.

## 3.6 End to End Solutions

Snoeren and Balakrishnan presented an end-to-end host mobility support solution in [47]. It uses secure updates to the Domain Name System (DNS) upon an address change to track mobile node locations. It provides a set of connection migration options to securely and efficiently negotiate a change in the IP address of a peer without breaking the end-to-end connection. When the MN moves, currently open TCP connections are suspended and reactivated from the new IP address. This is transparent to an application that expects uninterrupted reliable communication with the peer.

This approach leaves the decision of whether to support transparent connectivity across network address changes to applications, which increases the complexity of application development. Some issues exist with DNS servers such as update frequencies to DNS servers and consistency between DNS servers. For these reasons this dissertation does not consider this approach for mobility management.

## 3.7 Summary of Mobility Support Approaches

Figure 3.7 summarizes various mobility support approaches that have been proposed in different ISO-OSI layers.

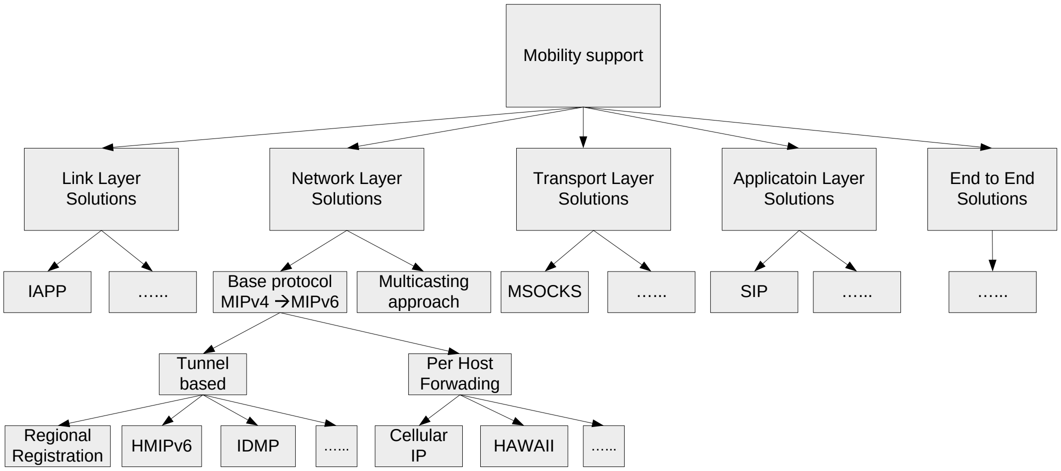

Figure 3.7: Protocol Family Tree for Mobility Support

All existing approaches try to reduce the cost associate with mobility management, including, most noticeably, IETF work-in-progress MIPv4 Regional Registration (MIP-RR) [18], Hierarchical MIPv6 [25] and IDMP [34]. However, these schemes deal with mobility management only without considering service characteristic of individual MNs.

Non-multicasting based addressing schemes under network layer solutions can be further categorized into two classes. In tunnel-based approaches, the CN always sends packets to the home address of MN. The HA uses IP-in-IP to tunnel packets to the CoA of the MN. The HA will create new packets with the header containing the destination IP address and the data payoff holding the original packet. At the other end of the tunnel, the original packet can be extracted by removing the outer IP header. MIPv4, MIPv6, MIP-RR, IDMP, and HMIPv6 are in this category.

In the per-host forwarding based approaches, the location information of the MN is stored in routers in the network. The location information stores in each router only indicates the next router to forward a packet to, not its final destination [56]. Cellular IP and HAWAII are in this category.

This dissertation deals with MIP based systems, so the solutions we investigate for mobility management are inherently network layer solutions. For the case in which ARs are not capable of executing user proxies to carry service context information for users, our DMAP solution (discussed in Chapter 4) is a pure network layer solution providing integrated mobility and service management transparent to user applications. For the case in which ARs are capable of executing user proxies, on the other hand, our IMSA solution (Chapter 5) is a cross-layer (network and application) solution allowing uninterrupted context-aware services (see Chapter 6) in the presence of user mobility in MIP systems.

## 3.8 Service Management

The concept of service management in the context of PCS networks has been investigated in [15, 24, 26]. Furthermore, the concept of integrated mobility and service management has been discussed in [7, 17].

Jain et al. [24] first suggested that mobility and service handoff be integrated to reduce the overall cost in distributed service environments. However the service architecture was based on fully replicated servers. Endler et al. [15] proposed a service delivery protocol named Result Delivery Protocol (RDP) by using a service proxy to provide reliable message delivery to MNs. It is based on the notion of proxies. A proxy is created on behalf of a MN that wishes to interact with servers within the wired network. It is created when a MN initiates a new series of service requests. The main purpose of the proxy is to provide a fixed location for the reception of server replies, keep track of pending requests, store the request results, and forward the results to the MSS responsible for the cell in which the MN located. However, this protocol is suitable only for connectionless request-reply communications. It runs on the application layer. Further, the proxy moves whenever the MN moves across a location boundary, so it may incur a high communication cost. MARCH [3] runs at the application layer for adapting media content for devices with various resource and computational capability. MARCH leverages the use of service proxies for content adaptation in client-server environments so that service contents could match network capabilities of the mobile device and the user preferences.

Gu and Chen [17] proposed and analyzed mobile service management schemes based on location-aware proxies with the objective to reduce the network signaling and communication cost in PCS networks. Under these schemes, a mobile user uses personal proxies as intelligent client- side agents to communicate 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. They showed 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.

Chen, Gu and Cheng [7] very recently investigated the notion of per-user integrated location and service management in PCS networks by which a per-user service proxy is created to serve as a gateway between the mobile user and all client-server applications engaged by the mobile user. The service proxy is colocated with the mobile user's location database such that whenever the MU's location database moves during a location handoff, a service handoff also ensues to colocate the service proxy with the location database. This allows the proxy to know the location of the mobile user all the time to reduce the network communication cost for service delivery. They investigated four integrated location and service management schemes and identified operational conditions under which one scheme may perform better than others.

The above two studies, however, are in the context of HLR/VLR based PCS networks. This dissertation addresses integrated mobility and service management in MIP based systems. The addressing scheme used for routing packets is very different in these two systems and the use of proxies for integrated mobility and service management in IP-based networks is vastly different from that in PCS networks. Unlike PCS networks where base stations and VLRs can be assumed to have regular shapes, IP subnets are shapeless. In PCS networks, distance can be used to measure network cost. However, in IP networks, the network cost is normally measured by hops, which do not correspond to distances. Network routers in IP networks normally are specific routing devices. However, MSS, VLR and HLR in PCS networks are powerful devices capable of performing both routing and computational functions. This dissertation work is the first study that investigates integrated mobility and service management in Mobile IP networks.

# Chapter 4

# DMAP: Dynamic Mobility Anchor Points

In this chapter, we consider the case that access routers are restricted to perform network layer functions. We investigate an integrated mobility and service management scheme based on MIPv6 with the goal to minimize the overall network signaling cost in MIPv6 systems for serving mobility and service management related operations. Our design extends Hierarchical Mobile IPv6 (HMIPv6) with the notion of dynamic mobility anchor points (DMAPs) for each mobile node (MN) instead of static ones for all MNs. These DMAPs are access routers (ARs) chosen by individual MNs to act as a regional router to reduce the signaling overhead for intra-regional movements. The DMAPs only perform network layer functions. The DMAP domain size, i.e., the number of subnets covered by a DMAP, is based on the MN's mobility and service characteristics. Under our DMAP protocol, a MN optimally determines when and where to launch a DMAP to minimize the network cost in serving the user's mobility and service management operations.

## 4.1 Algorithm Description

We describe a "DMAP table lookup" below. This design has the advantage of simplicity, scalability and efficiency and is HMIPv6-compliant. When a MN crosses a service area, it makes the AR of the subnet just crossed as the DMAP as in HMIPv6. The MN also determines the size of the new service area (or MAP domain). Concurrently, it acquires a RCoA as well as a CoA from the current subnet and registers the address pair (RCoA, CoA) to the current DMAP (the AR of the current subnet) in a binding request message. Note that the RCoA could be the same as the CoA upon the MN's entry into a new DMAP domain. The MN also informs the HA and CNs of the new RCoA address change in another binding message so that the HA and CNs would know the MN by its new RCoA address. When the HA and CNs subsequently send packets to the MN, they would use the RCoA as the MN's address.

A packet destined for RCoA will first be intercepted by the DMAP. By inspecting the address pair (RCoA, CoA) stored in the internal table, the DMAP knows that the MN's address is actually the CoA and will forward the packet to the MN through tunneling. If the RCoA and CoA are in the same subnet, the DMAP can directly forward the packet to the MN without using tunneling. When the MN subsequently crosses a subnet but is still located within the service area, it would inform the MAP of the CoA address change without informing the HA and CNs to reduce the network signaling cost. This "DMAP table lookup" design maps RCoA to CoA by having the current DMAP maintain an internal table, so the DMAP can intercept a packet destined for RCoA and forward it to the MN's CoA. It is efficient since the RCoA-CoA routing function can be performed efficiently by DMAPs (which are routers) through simple table lookup operations. It is scalable because the design is scalable to a large number of MNs by having all ARs in MIPv6 networks DMAP-enabled and randomly spreading the routing and table lookup functions to all ARs in the network. In terms of security and fault tolerance, it can also leverage existing solutions in HMIPv6 because this design is HMIPv6-compliant except that a MN dynamically selects ARs to be MAPs.

The idea of determining a dynamic "service area" to minimize the network signaling cost is similar in concept to determining a gateway foreign agent (GFA) coverage area in MIP Regional Registration [58]. However, MIP Regional Registration only deals with mobility management while our scheme deals with both mobility and service management.

A MN's service area can be modeled as consisting of $K$ IP subnets. We develop a computational procedure to determine the optimal service area size in terms of $K$. The optimal $K$ value would be determined at runtime dynamically to minimize the network cost. For the special case in which $K$ is constant for all MNs, our scheme degenerates to HMIPv6 with a two-level MAP-AR structure.

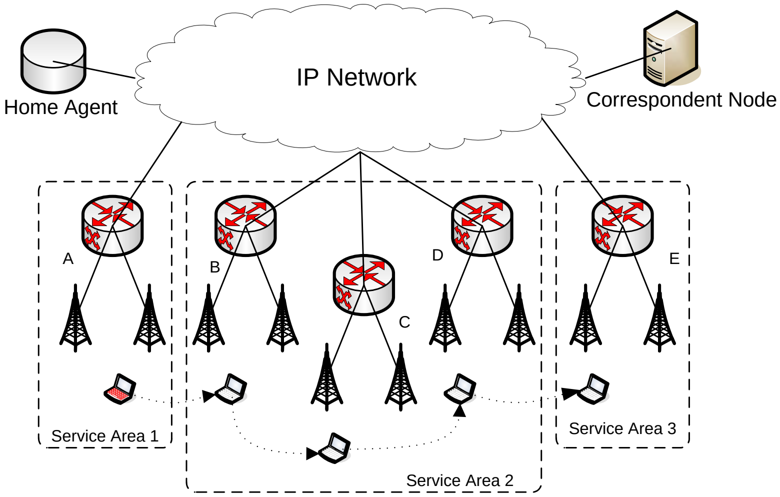

Figure 4.1: DMAP: Integrated Mobility and Service Management in MIPv6.

Figure 4.1 illustrates DMAP in MIPv6 environments. When moving from one subnet and another within service area 1, the MN only informs its CoA change to the DMAP without informing the HA or CNs. When the MN exits service area 1, the AR of subnet B becomes the DMAP. The MN obtains a CoA and a RCoA from subnet B and an entry (RCoA, CoA) is recorded in the routing table of the AR of subnet B. Subnet B's AR now acts as the DMAP of the MN. Both the HA and CNs are informed of the RCoA address change by the MN.

In our DMAP scheme, the MN appoints a new DMAP only when it crosses a service area whose size is determined based on the mobility and service characteristics of the MN in the new service area. One should note that the service area size of the DMAP is not necessarily uniform. In the above scenario although subnet B appears to be at the edge of service area 2, it is actually at the center of the new service area since our service area is defined by the number of subnets (or the number of moves) starting from the first subnet at which the MN enters into a new service area. Within service area 2 (the new service area that the MN just moves into), if the MN moves from subnet B to subnet D through C (with 2 location handoffs), then the DMAP will be informed of the CoA change by the MN but will remain at the same location (subnet B).

A large service area size means that the DMAP will not change often. The consequence of not changing the DMAP often is that the service delivery cost would be high because of the triangular routing path CN-DMAP-MN for data communication between the CN and MN. On the other hand, a small service area size means that the DMAP will be changed often so it will stay close to the MN. The consequence is that the communication cost for service data delivery would be low because of the short CN-DMAP-MN route. However, a DMAP change involves the cost of informing the HA and CNs of the RCoA address change. Therefore, there is a trade-off between these two cost factors and an optimal service area exists.

The service and mobility characteristics of a MN are summarized by two parameters. The first parameter is the residence time that the MN stays in a subnet. This parameter can be collected by each MN based on statistical analysis [10]. We expect that future MNs are reasonably powerful for collecting data and doing statistical analysis. The residence time in general would be characterized by a general distribution. Loosely, we use the MN's mobility rate ($\sigma$) to represent this parameter. The second parameter is the service traffic between the MN and server applications. The MN can also collect data statistically to parameterize this. Loosely, we use the data packet rate ($\lambda$) between the MN and CNs to represent this parameter. Both of these parameters are to be determined by the MN. For efficiency, the MN could build a table to lookup its mobility and service rates as a function of its location, time of the day, and day of the week, based on statistical analysis. The ratio of $\lambda/\mu$ is called the service to mobility ratio (SMR) of the MN.

When a MN moves across a service area boundary, the DMAP changes, thus incurring a "service handoff." This service handoff cost includes the signaling cost to inform the HA and CNs of the new RCoA of the MN.

Table 4.1 lists a set of identified system parameters that characterize the mobility and service characteristics of a MN in a MIPv6 system. Our DMAP scheme is per-user based.

| Symbol | Meaning |
|---|---|
| $\lambda$ | data packet rate between the MN and CNs |
| $\sigma$ | mobility rate at which the MN moves<br>across subnet boundaries |
| SMR | service to mobility ratio ($\lambda/\sigma$) |
| $N$ | number of server engaged by the MN |
| $K$ | number of subnets in one service area |
| $\tau$ | 1-hop communication delay per packet in wired networks |
| $\alpha$ | average distance between HA and MAP |
| $\beta$ | average distance between CN and MAP |
| $\gamma$ | cost ratio between wireless vs. wired network |

Table 4.1: Parameters for DMAP.

Data packets from a server application are sent to a MN's DMAP first (which is just an AR selected to be the current DMAP) and then forwarded to the MN. The DMAP will receive packets addressed to the RCoA from the HA and CNs. Packets will be tunneled from the DMAP to the MN.

**Intra-Regional Move:** When a MN performs a location handoff within a service area, it acquires a CoA from the subnet and informs the DMAP of the CoA address change. The DMAP is not changed. Also, the HA and CN are not informed of the CoA address change of the MN since they only know the MN by the RCoA address which is not changed in this case.

**Inter-Regional Move:** When a MN moves across a service area thus incurring a service handoff, the MN acquires a CoA and a RCoA from the AR that now becomes the DMAP and an entry (RCoA, CoA) is recorded in the lookup table of the AR. The MN then informs the HA and CNs of its new RCoA to complete the service handoff. The implementation of the proposed DMAP scheme based on the "DMAP table lookup" design is totally transparent to the HA and CNs. The HA and CNs are informed of the RCoA as part of the service handoff process. Packets from the HA or CNs will use the RCoA as the destination address, which will be intercepted by the AR that serves as the MN's DMAP who will then forward them to the MN. The packet routing and forwarding will be done entirely in the network layer. In effect the system behaves as if a two-level HMIPv6 structure has been used to do both mobility and service management, except that the service area is to be determined by the MN. Instead of having the MN discover the presence of a MAP through announcement packets and initiating the MAP migration process, the MN will dynamically determine which AR will act as a DMAP to minimize the network cost.

## 4.2 Modeling and Analysis

We devise a computational procedure to determine the optimal service area size utilizing stochastic Petri net (SPN) techniques. The intent to find the optimal service area based on the MN's mobility and service behaviors. The designer would utilize the computational procedure developed here to build a table at static time listing the optimal service area as a function of these parameters each covering a reasonable value range. Such a table is then loaded into the MN. The actual values of these parameters are dynamically collected by the MN at runtime. Based on the values of these parameters at the time a service area is crossed, the MN performs a simple table lookup to determine the optimal service area. The metric that we aim to minimize is the "communication cost" incurred per time unit due to mobility and service operations. Our SPN model is shown in Figure 4.2. Table 4.2 gives the meaning of places and transitions defined in the model.

We choose SPN because of its ability to deal with general time distributions for events, its concise representation of the underlying state machine to deal with a large number of states, and its expressiveness to reason about a MN's behavior as it migrates from one state to another in response to events occurring in the system. Moreover SPN models allow the residence time of a MN in a subnet or the packet interarrival time between a MN and its CNs to be generally distributed. Once the parameters of the SPN model are given proper values, numerical analysis methods for solving SPN models based on SOR or Gauss Seidel [52] are readily available to compute the optimal service area size.

A token in the SPN model represents a subnet crossing event by the MN. The function

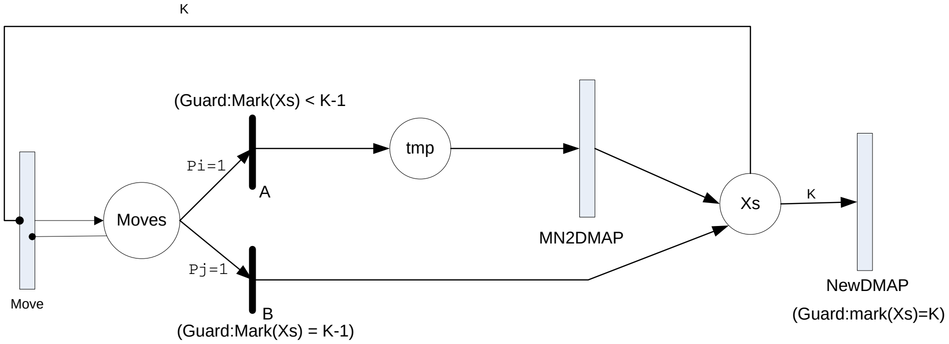

Figure 4.2: Stochastic Petri Net Model for DMAP.

| Symbol | Meaning |
|---|---|
| `Move` | a timed transition for the MN to move across subnet areas |
| `Moves` | Mark(`Moves`)=1 meaning that the MN just moves across a subnet |
| `MN2DMAP` | a timed transition for the MN to inform the DMAP of the CoA change |
| `Xs` | Mark(`Xs`) holds the number of subnets crossed in a service area |
| `NewDMAP` | a timed transition to inform the HA and CNs of the RCoA change |
| `K` | number of subnets under a service area |
| Guard:Mark(`Xs`) $<K-1$ | a guard for transition `A` that is enabled if a move will not cross a service area |
| Guard:Mark(`Xs`) $=K-1$ | a guard for transition `B` that is enabled if a move will cross a service area |
| Guard:Mark(`Xs`) $=K$ | a guard that is enabled if the MN just moves across a service area |
| `A` | an immediate transition when Mark(`Xs`)$<K-1$ |
| `B` | an immediate transition when Mark(`Xs`)$=K-1$ |
| `tmp` | a temporary place that holds tokens from transition A |

Table 4.2: Meaning of Places and Transitions in SPN Model for DMAP.

Mark(P) returns the number of tokens in place P. The number of tokens accumulated in place Xs, that is, Mark(Xs), represents the number of subnets already crossed by the MN since it enters a new service area. The SPN model describes the behavior of a MN operating under the DMAP scheme:

- When a MN moves across a subnet area, thus incurring a location handoff, a token is put in place Moves. The mobility rate at which location handoffs occur is $\sigma$ which is the transition rate assigned to Move.

- If the current move is an intra-regional move such that the guard for transition A will return *true*, then the MN will only inform the DMAP of the CoA change. This is modeled by firing immediate transition A, allowing the token in place Moves to move to place tmp. Subsequently, once the MN obtains a CoA from the subnet it just enters, it will communicate with the DMAP of the new CoA change. This is modeled by enabling and firing transition MN2DMAP. After MN2DMAP is fired, a token in place tmp flows to place Xs, representing that a location handoff has been completed and the DMAP has been informed of the CoA change of the MN.

- If the current move results in the total number of moves being $K$ such that the guard for transition B will return *true*, then the move will make the MN cross a service area. This is modeled by enabling and thus firing immediate transition B, allowing the token in place Moves to move to place Xs in preparation for a service handoff event. Note that in an SPN, firing an immediate transition does not take any time.

- If the number of moves, including the current one, in place Xs has accumulated to $K$, a threshold determined by the DMAP representing the size of a service area, then it means that the MN has just moved into a new service area and a service handoff ensues. This is modeled by assigning an enabling function that will enable transition NewDMAP when $K$ tokens have been accumulated in place Xs. After transition NewDMAP is fired, all $K$ tokens are consumed and place Xs contains no token, representing that the AR of the subnet that the MN just enters has been appointed as the DMAP by the MN in the new service area.

Below we show an example of parameterizing transition rates of `MN2DMAP` and `NewDMAP` based on the set of base parameters defined in Table 4.1. The firing time of transition `MN2DMAP` stands for the communication time of the MN informing the DMAP of the new CoA through the wireless network. This time depends on the number of hops separating the MN and its DMAP. Thus, the transition rate of transition `MN2DMAP` is calculated as:

$$\frac{1}{\gamma\tau + F(Mark(\mathrm{Xs}) + 1) \times \tau}$$

where $\tau$ stands for the one-hop communication delay per packet in the wired network and $\gamma$ is a proportionality constant representing the ratio of the communication delay in the wireless network to the communication delay in the wired network. $F(Mark(\texttt{Xs}) + 1)$ returns the number of hops between the current subnet and the DMAP separated by $Mark(\texttt{Xs}) + 1$ subnets. The argument of the $F(x)$ function is added by 1 to satisfy the initial condition that $Mark(\texttt{Xs}) = 0$ in which the DMAP has just moved into a new service area, so at the first subnet crossing event, the distance between the DMAP and the subnet is one subnet apart. Note that this transition rate is state-dependent because the number of tokens in place `Xs` changes dynamically over time.

When transition `NewDMAP` fires, the AR of the subnet that the MN moves into will be selected as the DMAP. The communication cost includes that for the MN to inform the HA and CNs of the new RCoA address change, i.e., $(\alpha + N\beta)\tau$, where $\alpha$ is the average distance in hops between the MN and the HA, $\beta$ is the average distance in hops between the MN and a CN, and $N$ is the number of CNs that the MN concurrently engages. Thus, the transition rate of transition `NewDMAP` is calculated as:

$$\frac{1}{(\alpha + N\beta)\tau}$$

The stochastic model underlying the SPN model is a continuous-time semi-Markov chain with the state representation of $(a, b)$ where $a$ is the number of tokens in place `Moves`, $b$ is the number of tokens in place Xs. Let $P_i$ be the steady state probability that the system is found to contain $i$ tokens in place Xs such that Mark(Xs) $= i$. Let $C_{i,service}$ be the communication overhead for the network to service a data packet when the MN is in the $i^{th}$ subnet in the service area.

Let $C_{service}$ be the average communication overhead to service a data packet weighted by the respective $P_i$ probabilities. The communication overhead includes a communication delay between the DMAP and a CN in the fixed network, a delay from DMAP to the AR of the MN's current subnet in the fixed network, and a delay in the wireless link from the AR to the MN. Thus, $C_{service}$ is calculated as follows:

$$\begin{aligned} C_{service} &= \sum_{i=0}^{K} (P_i \times C_{i,service}) \\ &= \gamma\tau + \beta\tau + \sum_{i=0}^{K} (P_i \times F(i)\tau) \qquad (4.1) \end{aligned}$$

Let $C_{i,location}$ be the network signaling overhead to service a location handoff operation given that the MN is in the $i^{th}$ subnet in the service area. If $i < K$, only a minimum signaling cost will incurred for the MN to inform the DMAP of the CoA address change. On the other hand, if $i = K$, then the location handoff also triggers a service handoff. A service handoff will incur a higher communication signaling cost to inform the HA and $N$ CNs (or application servers) of the RCoA address change. Let $C_{location}$ be the average communication cost to service a move operation by the MN weighted by the respective $P_i$ probabilities. Then, $C_{location}$ is calculated as follows:

$$\begin{aligned} C_{location} &= \sum_{i=0}^{K} (P_i \times C_{i,location}) \qquad (4.2) \\ &= P_K(\gamma\tau + \alpha\tau + N\beta\tau) + \sum_{i=0}^{K-1} \{P_i(\gamma\tau + F(K)\tau)\} \end{aligned}$$

Summarizing above, the total communication cost *per time unit* for the Mobile IP network operating under our DMAP scheme to service operations associated with mobility and service management of the MN is calculated as:

$$C_{DMAP} \quad = \quad C_{service} \times \lambda + C_{location} \times \sigma \tag{4.3}$$

where $\lambda$ is the data packet rate between the MN and CNs, and $\sigma$ is the MN's mobility rate. The steady-state probability $P_i$, $1 \leq i \leq K$, needed in Equations 1 and 2 can be solved easily utilizing numerical method solution techniques such as SOR or Gauss Seidel [52].

## 4.3 Numerical Results

Here we apply Equations 4.1, 4.2 and 4.3 to calculate $C_{DMAP}$ as a function of $K$ and determine the optimal $K$ representing the optimal "service area" size that will minimize the network cost, when given a set of parameter values characterizing the MN's mobility and service behaviors. Below we present results to show that there exists an optimal service area for systems operating under DMAP for network cost minimization, and demonstrate the benefit of DMAP over basic MIPv6 and HMIPv6.

For basic MIPv6, there is no DMAP. Thus, the communication cost $C^{MIPv6}_{service}$ for servicing a packet delivery in basic MIPv6 includes a communication delay from the CN to the AR of the current subnet, and a delay in the wireless link from the AR to the MN as follows:

$$C^{MIPv6}_{service} = \gamma\tau + \beta\tau \tag{4.4}$$

Under basic MIPv6, when a MN crosses a subnet boundary, thus incurring a location handoff, the MN informs the HA and CNs of its CoA change. The cost $C^{MIPv6}_{location}$ for servicing a location handoff under basic MIPv6 consists of a delay in the wireless link from the MN to the AR of the subnet that it just enters into, a delay from that AR to the CNs, and a delay from that AR to the HA as follows:

$$C^{MIPv6}_{location} = \gamma\tau + \alpha\tau + N\beta\tau \tag{4.5}$$

The total cost per time unit for servicing data delivery and mobility management operations under MIPv6 is given by:

$$C_{MIPv6} = C_{service}^{MIPv6} \times \lambda + C_{location}^{MIPv6} \times \sigma \tag{4.6}$$

For HMIPv6, the placement of MAPs is pre-determined. We compare DMAP with an implementation of HMIPv6 in which each MAP covers a fixed number of subnets, say, $K_H = 4$. Table 4.3 compares the communication cost incurred per time unit by DMAP vs. that by basic MIPv6 and HMIPv6 head-to-head as a function of SMR, from the perspective of $K_{opt}$ that minimizes the overall communication cost. The example is based on $F(K) = \sqrt{K}$ [59], $\alpha = \beta = 30$, and $\gamma = 10$. The cost metric is normalized with respect to $\tau$=1. First we observe that $K_{opt}$ exists under DMAP. Furthermore, as SMR increases, $K_{opt}$ decreases because when SMR is large, the service rate is high compared with the mobility rate, so the DMAP likes to stay as close to the MN as possible to minimize the communication overhead through the CN-DMAP-MN route. This table clearly exhibits the behavior of DMAP. We see that DMAP dominates basic MIPv6 when SMR is low. As SMR increases exceeding a threshold (e.g., 64), $K_{opt}$ approaches 1 under which DMAP degenerates to basic MIPv6. The reason is that when SMR is sufficiently high, the MN's packet arrival rate is much higher than the mobility rate, so the data delivery cost dominates the mobility management cost. Therefore the DMAP will stay close to the MN to lower the data delivery cost, thus making $K_{opt} = 1$ in our DMAP scheme in order to reduce the CN-DMAP-MN (or CN-MAP-MN) triangular routing cost for packet delivery.

Comparing DMAP with HMIPv6 from the perspective of $K_{opt}$, we observe from Table 4.3 that DMAP degenerates to HMIPv6 (with a fixed $K_H$) when SMR is moderate. When SMR is either high or low, DMAP performs substantially better than HMIPv6. The trend would be true with other choices of $K_H$ for HMIPv6 as well (say $K_H = 8$). When SMR is low, $K_{opt}$ is high for DMAP compared with a fixed $K_H$ for HMIPv6, in which case the cost saving is due to mobility cost reduction. On the other hand, when SMR is high, $K_{opt}$ is low for DMAP compared with a fixed $K_H$ for HMIPv6, in which case the cost saving is due to service cost reduction. Overall, there is a wide range of SMR values under which

| **SMR** | $C_{MIPv6}$ | $C_{HMIPv6}$<br>**(**$K_H$**=4)** | $C_{DMAP}$<br>**(**$K_{opt}$**)** | $K_{opt}$ |
|---|---|---|---|---|
| 0.1250 | 1.8750 | 0.7897 | 0.5522 | 34 |
| 0.2500 | 2.0000 | 0.9187 | 0.6892 | 32 |
| 0.5000 | 2.2500 | 1.1766 | 0.9619 | 28 |
| 1.0000 | 2.7500 | 1.6925 | 1.5034 | 22 |
| 2.0000 | 3.7500 | 2.7242 | 2.5758 | 16 |
| 4.0000 | 5.7500 | 4.7876 | 4.6960 | 11 |
| 8.0000 | 9.7500 | 8.9144 | 8.8859 | 7 |
| **16.0000** | **17.7500** | **17.1681** | **17.1681** | **4** |
| 32.0000 | 33.7500 | 33.6754 | 33.5475 | 2 |
| 64.0000 | 65.7500 | 66.6901 | 65.7500 | 1 |

Table 4.3: Comparing DMAP with Basic MIPv6 and HMIPv6 Head-to-Head from the Perspective of $K_{opt}$.

DMAP dominates HMIPv6 because of the choice of $K_{opt}$ for DMAP vs. a fixed $K_H$ for HMIPv6. Here we should note that a small difference of 0.1 is considered significant because the metric is expressed in terms of the network cost per time unit (normalized to the per-hop cost $\tau = 1$) so the cumulative effect would be significant over the lifetime of a single MN. It is also worth noting that the cumulative effect of cost saving for a large number of MN would be even more significant.

Figure 4.3 summarizes the cost difference between basic MIPv6 and DMAP vs. HMIPv6 and DMAP as a function of SMR. The Y coordinate is the cost difference incurred per time unit. There are two curves shown in Figure 4.3. The first curve shows the cost difference between basic MIPv6 and DMAP ($C_{MIPv6} - C_{DMAP}$), and the second curve shows the cost difference between HMIPv6 and DMAP ($C_{HMIPv6} - C_{DMAP}$), as a function of SMR. We first observe that the cost difference between basic MIPv6 and DMAP (the first curve) decreases as SMR increases. The reason is that DMAP degenerates to MIPv6 when SMR becomes sufficiently large. We conclude that DMAP performs significantly better than basic MIPv6 especially when SMR is low. Next we observe that the cost difference between HMIPv6 and DMAP (the second curve) initially decreases as SMR increases until $K_{opt}$ coincides with $K_H$ at which point DMAP degenerates to HMIPv6, and then the cost difference increases sharply as SMR continues to increase. We conclude that DMAP performs significantly better than HMIPv6 when SMR is either low and high.

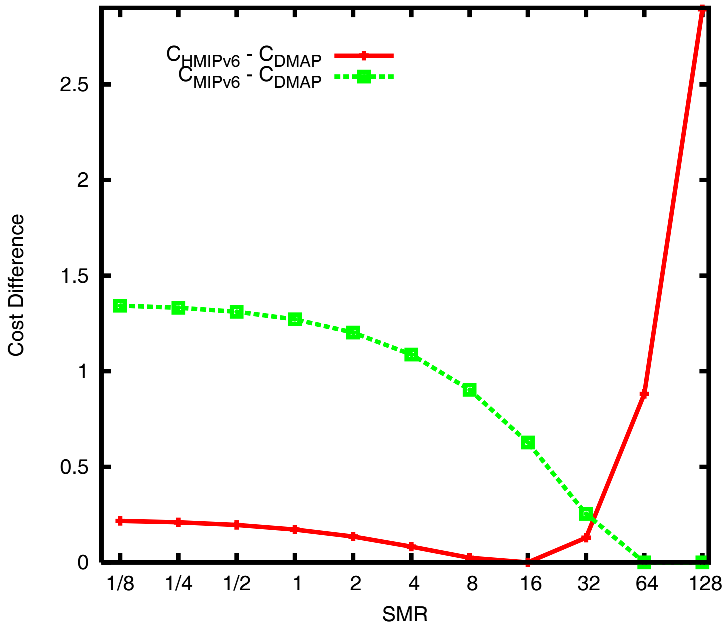

Figure 4.3: Cost Difference between Basic MIPv6, HMIPv6 and DMAP.

Figure 4.4 tests the sensitivity of the results with respect to $\alpha$ and $\beta$ representing the hop distances between the MN and the HA and CNs. We see that as the hop distance between the HA (and CNs) increases, the cost difference between HMIPv6 and DMAP ($C_{HMIPv6} - C_{DMAP}$) becomes more pronounced, especially when SMR is small. The reason is that when SMR is small, the mobility management cost dominates the data delivery cost, so $K_{opt}$ tends to be large to reduce the mobility management cost. In this case DMAP dictates a high $K_{opt}$ value to be used to reduce the mobility management cost as opposed to a fixed $K_H = 4$ used by HMIPv6, thereby resulting in a more substantial cost difference between HMIPv6 and DMAP as SMR decreases. Recall that the cost is normalized with respect the per-hop communication cost $\tau = 1$. By observing the 4 curves in Figure 4.4, nevertheless, we see that the trend remains the same. That is, the cost difference ($C_{HMIPv6} - C_{DMAP}$) decreases with the increase of SMR and then increases sharply after a threshold point at which DMAP degenerates to HMIPv6. Figure 4.4 shows that under a wide range of $\alpha$ and

$\beta$ values, DMAP always incurs less network overheads than HMIPv6, the effect of which is pronounced when SMR is relatively low or high.

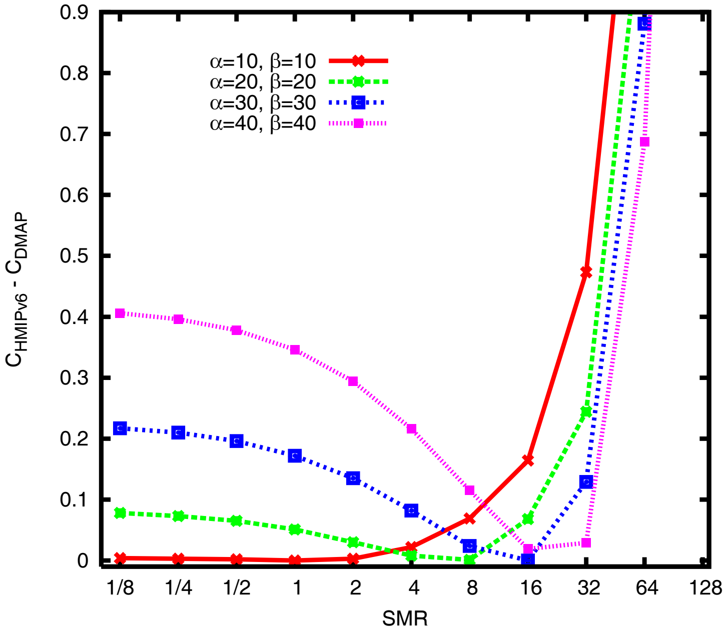

Figure 4.4: Effect of $\alpha$ and $\beta$ on $C_{HMIPv6} - C_{DMAP}$.

All the above results are based on the assumption that the average number of hops between the DMAP and MN separated by $k$ subnets is given by $F(k)= \sqrt{k}$, which are based on the fluid flow model [59]. We have tested the sensitivity to the form of $F(k)$. Figure 4.5 shows the effect of $F(k)$ on the cost difference $C_{HMIPv6} - C_{DMAP}$ for $F(k)= \sqrt{k}$ and $F(k) = k$. The cost difference curves exhibited are very similar in shape and are not sensitive to the form of $F(k)$. We conclude that all the conclusions drawn earlier from the case $F(k)= \sqrt{k}$ are valid.

The performance gain is in the amount of communication cost saved per time unit per user, so the saving due to a proper selection of the best service area will have significant impacts since the cumulative effect for all mobile users over a long time period would be significant.

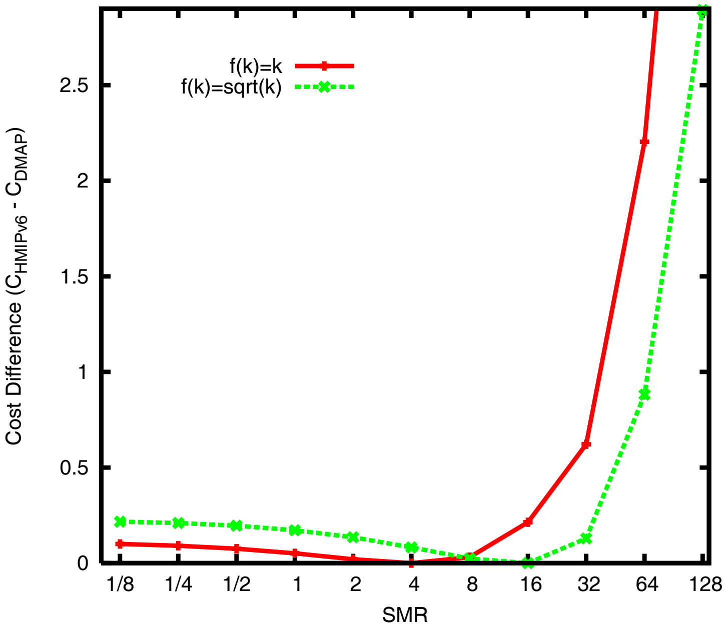

Figure 4.5: Effect of $F(k)$ on $C_{HMIPv6} - C_{DMAP}$.

# Chapter 5

# IMSA: Integrated Mobility and Service Management Architecture

In this chapter, we consider the case that the access routers are quite powerful and flexible in future all-IP based wireless networks. Mobile proxies can be dynamically loaded to run on ARs and can roam among ARs routers to perform network layer and application layer functions on behalf of MNs and applications. The architecture proposed for this case is called IMSA (Integrated Mobility and Service Management Architecture) under which client-side proxies are employed to run on ARs to perform network-layer as well as application-layer functions on behalf of MNs. Due to architectural differences between MIPv4 and MIPv6, we disucss IMSA for MIPv4 and MIPv6 separately.

Under our proposed ISMA, a *client-side* proxy is created when a MN starts to serve as a GFA as in the MIP-RR protocol [18] to maintain the location information of the MN. Moreover when the MN invokes a server application (or a CN), the client-side proxy will communicate with the CN on behalf of the MN as if it were the MN. The proxy is per-user based. Data packets from a server application are sent to the proxy to forward to the MN. From the perspective of a CN, the proxy represents the MN. Conversely, from the perspective of the MN, the proxy represents the server applications (or the CNs) engaged by the MN. All communications between the MN and CNs will go through the proxy.

Mobile server applications would benefit from proxies carrying service context information and performing context-aware functions for services engaged by the MN. For example, the proxy can perform caching for data items frequently accessed by the MN and supply data items to the MN. If the data items are large and impractical to cache, the proxy can store cache invalidation reports and supply cache invalidation information to the MN either when the MN is connected, or after the MN just wakes up from a disconnection to service user queries for data access [9][24]. The proxy can buffer video frames for multimedia applications and keeps service context information regarding the video title, minutes played, current frames being buffered at the server and played at MN. The proxy can also apply optimization techniques such as execution of Web processing units [56], format transformation, data compression/decompression and differencing to reduce the amount of traffic in the network. Our architecture allows context sensitive information to be stored by the service proxy based on the application requirement. The optimal service area, among other factors, depends on the amount of service context information carried by the service proxy. This is so because the context transfer cost for moving the proxy from one location to another as the proxy moves increases as the amount of context information increases. Our architecture allows the amount of context information to be determined by the application and supplied as part of service characteristics for the determination of the optimal "service area" for a MN.

## 5.1 IMSA-MIPv4: IMSA for IPv4

When the MN crosses a FA boundary in MIPv4 environments, thus incurring a location handoff, the proxy acting as a GFA will be informed of the new FA address but may not move with the MN. The proxy will move only when the MN crosses a service area. The size of the service area in terms of the number of FA areas covered depends on the mobility and service characteristics of the MN such that the network cost associated with mobility and service handoffs will be minimized. Once a proxy moves into a new service area, it again acts as a GFA for mobility management and a service proxy for all server applications that the MN currently engages.

Figure 5.1 illustrates the system model where a FA area corresponds to an IP subnet area.

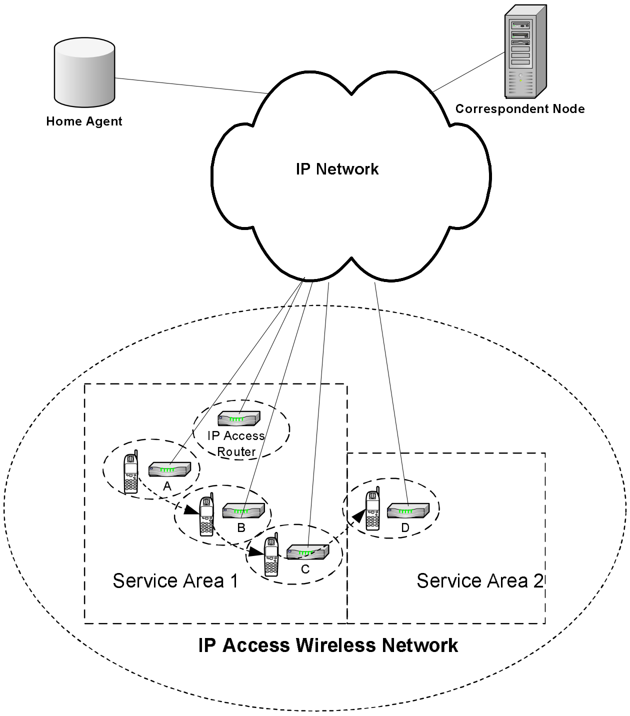

Figure 5.1: IMSA on Mobile IPv4.

The client-side proxy initially runs on the FA node of subnet A. When the MN moves within service area 1 from subnet A to subnet C through B (with 2 location handoffs), the proxy, acting as a GFA for mobility management, remains at the same location and is informed of the address change by the FA's. The correspondent node and the HA are not informed of these address changes due to location handoffs in this case. When the MN crosses a service area boundary into service area 2, a service handoff occurs by which the proxy moves into subnet D and runs on the FA node of subnet D. The proxy after the move will stay closer to the MN, so the communication cost for data delivery along the path of CN-proxy-FA-MN is lower. Note that the first FA (i.e., node D) that the proxy moves into is actually at the center of the new service area since starting from that FA, the proxy will move again only when the MN moves across a number of subnets starting from the first FA in the new service area. A proxy move involves the cost of informing the HA and all the correspondent nodes of the proxy address change, as well as the context transfer cost. Therefore, there is a trade-off between these two cost factors. The optimal service area size is dictated by the

MN's mobility and service characteristics. One should note that the service area size of the proxy is not necessarily uniform. Figure 5.1 shows that service area 1 is larger in size than service area 2 since the mobility and service characteristics of the MN may be drastically different in different service locations. Figure 5.1 also shows that a service handoff coincides with a location handoff as the MN moves from subnet C into subnet D.

From the perspective of service management, we like to keep the proxy close to the MN since this will reduce the communication overhead for data delivery. This factor favors a small service area. From the perspective of mobility management, we like to keep a large service area to reduce the cost of mobility and service handoffs.

The proxy acting on behalf of the MN to communicate with application servers keeps service context information for services engaged by the MN. Since the proxy is created by the MN, the MN could supply the proxy its mobility and service characteristics the moment it crosses a service area. When the MN moves across a FA boundary, the proxy will check if the service area is crossed. If yes, after the proxy moves into the new service area, a new optimal service area size can be determined by executing a computational procedure developed in this chapter based on the MN's mobility and service characteristics in the new service area.

Figure 5.2 shows the process by which a MN submits a service request. The MN will submit the service request to the proxy through its current FA. The proxy forwards the service request to the correspondent node on behalf of the MN. The responses from the CN are returned to the proxy first and then forwarded to the MN through the current FA. The proxy acting as a GFA knows the location of the current FA and the MN all the time, so data delivery is very efficient in our scheme without incurring the overhead of triangular routing through the HA.

Figure 5.3 shows the process by which a MN performs a location handoff within a service area during a service session with a CN. In this case, the proxy is informed of the address change of the FA which the MN moves into but itself does not migrate. The HA and the CN are not involved in the location handoff; they still know the MN by the proxy's current

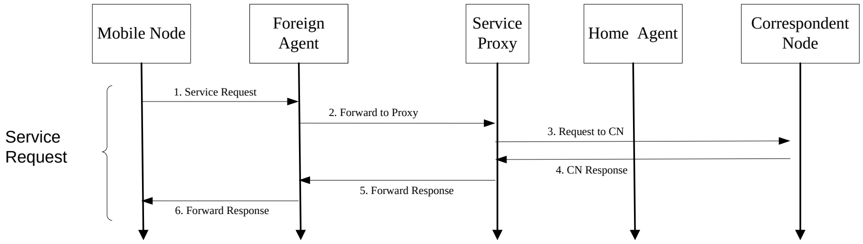

Figure 5.2: Service Request Process in IMSA on Mobile IPv4.

address which is not changed.

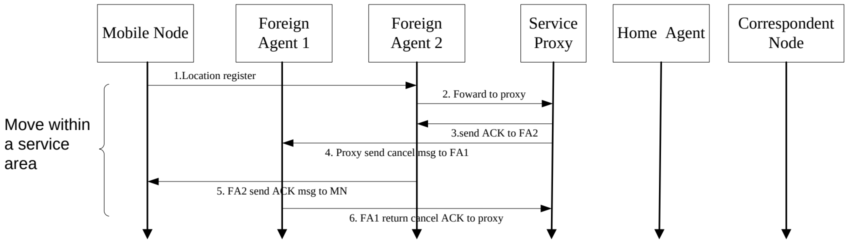

Figure 5.3: Location Handoff Process within a Service Area by a MN in IMSA on Mobile IPv4.

Figure 5.4 shows the process by which a MN moves across a service area thus incurring a service handoff during an ongoing service session with a CN. The MN first updates its location information with the proxy through the FA it moves into (FA2 shown in the figure). When the proxy realizes that a service area has been crossed, it initiates a migration process to move to FA2. Once the proxy moves into the new service area and runs on FA2 with a new IP address, it informs the HA and the CN of its new address to complete the service handoff.

The implementation of the client-side proxy for a MN can be realized by a middleware running on the MN which initially creates a proxy acting as a GFA to interact with Mobile IP when the MN starts up. The client-side proxy has two components. One component is at network layer dealing with mobility management in Mobile IP systems by interacting with Mobile-IP software running on the network router. The other component is at the application layer dealing with service management with respect to services currently engaged

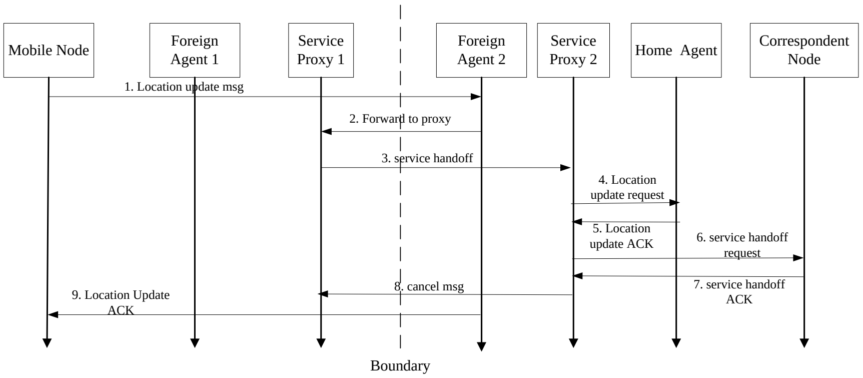

Figure 5.4: Service Handoff Process when Crossing a Service Area by a MN in IMSA on Mobile IPv4.

by the MN. These two components cooperate with each other as necessary for a cross-layer implementation of the proxy.

The service and mobility characteristics of a MN are summarized by two parameters. The first parameter is the residence time that the MN stays in a subnet. This parameter can be collected by each MN based on statistical analysis [10]. We expect that future MNs are reasonably powerful for collecting data and doing statistical analysis. The residence time would be characterized by a general distribution in general. We use the MN's mobility rate ($\sigma$) to represent this parameter. The second parameter is the service traffic between the MN and server applications. The MN can also collect data statistically to parameterize this. We use the data packet rate ($\lambda$) between the MN and CNs to represent this parameter. Both of these parameters will be periodically determined by the MN. For efficiency, the MN could also build a table to lookup its mobility and service rates as a function of its location, time of the day, and day of the week, based on statistical analysis.

When a MN moves across a service boundary, the service proxy moves, thus incurring a service handoff. This overhead involved includes the reconnection cost and the service context transfer cost. The reconnection cost refers to the communication cost for the proxy to inform the HA and the server applications of the new network address. The service context transfer cost is the communication cost to move the service context to the new proxy. We denote the number of packets carrying the context information for context transfer by $n_{CT}$,

| Symbol | Meaning |
|---|---|
| $\lambda$ | data packet rate, i.e. the data packet rate for all services currently engaged by the MN. |
| $\sigma$ | mobility rate at which the MN moves across FA boundaries. |
| SMR | service rate to mobility rate ratio, i.e., $\lambda/\sigma$. |
| $n_{CT}$ | number of packets required for content transfer. |
| $N$ | number of server applications currently engaged by the MN. |
| $F(k)$ | a general function relating the number of subnets $k$ to the number of hops. |
| $K$ | number of subnets (or FAs) in one service area. |
| $\tau$ | 1-hop communication delay per packet in wired networks. |
| $\alpha$ | average distance (in hops) between the HA and the proxy |
| $\beta$ | average distance (in hops) between a CN and the proxy |
| $\gamma$ | ratio between the communication cost in a wireless network to the communication cost in a wired network. |

Table 5.1: Parameters in IMSA on Mobile IPv4.

the magnitude of which is application-dependent. The system parameters that characterize the mobility and service characteristics of a MN in a Mobile IP system are summarized in Table 1 for easy reference.

### 5.1.1 Performance Model

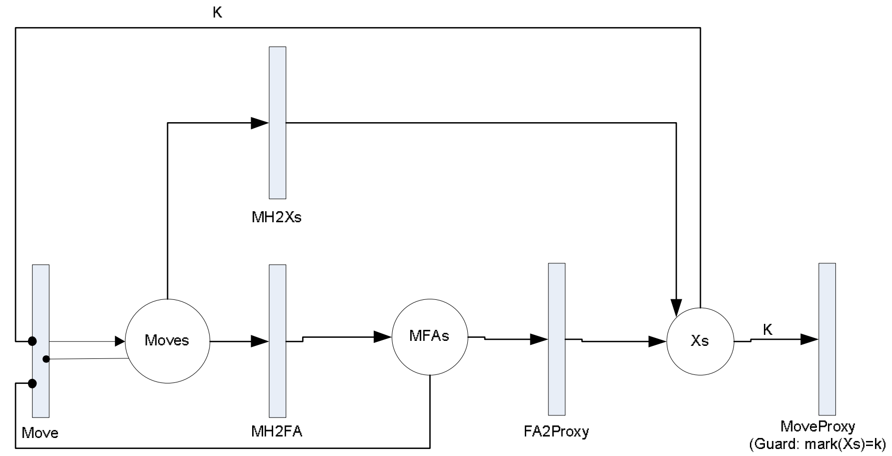

Figure 5.5: SPN Model for Proxy-based Integrated Location and Service Management in IMSA on Mobile IPv4.

A stochastic Petri Net (SPN) model as shown in Figure 5.5 is developed to analyze the

| Symbol | Meanings |
|---|---|
| `Moves` | Mark(`Moves`)=1 means that the MN has just moved across a FA area. |
| `Move` | A timed transition for the MN to move across FA areas. |
| `MFAs` | Mark(`MFAs`)=1 means that the MN has just changed its FA. |
| `MN2FA` | A timed transition for the MN to register with a new FA. |
| `FA2Proxy` | A timed transition for the FA to communicate with the proxy. |
| `Xs` | Mark(`Xs`) indicates the number of FA's crossed in a service area. |
| `MoveProxy` | A timed transition for the proxy to move into a new server area. |
| `K` | number of FA's crossed after which a service handoff will occur. |
| Guard:Mark(`Xs`)=K | enabled if the number of tokens in place `Xs` is K. |

Table 5.2: Meanings of Places and Transitions in the SPN Model for IMSA on Mobile IPv4.

behavior of a MN in a Mobile IP system under our proxy-based regional registration scheme. For ease of referencing, Table 5.2 gives the meaning of places and transitions defined in the SPN Model. In particular, $K$ represents the number of FAs under a regional registration area for which we want to obtain its optimal value $K_{opt}$ through SPN modeling. The justification of having a different $K_{opt}$ value for each MN in regional registration is that each MN has its specific mobility and service characteristics, thus requiring the use of a different $K_{opt}$ value (through a proxy) to define its optimal regional registration area to minimize the overall cost due to mobility signaling and packet delivery.

A token in the SPN model represents a FA area crossing (or a location handoff) event by the MN. The function Mark(`P`) returns the number of tokens in place `P`. The number of tokens accumulated in place `Xs`, that is, Mark(`Xs`), represents the number of FA's crossed (or the number of location handoffs) in a service area. SPN modeling is state-based, with a state being defined by the token distribution into places in the net. Thus, in a particular state we know exactly how many FAs a MN has already crossed by looking at the number of tokens in place `Xs`. Below we explain how the SPN model is constructed.

- When a MN moves across a FA area, thus incurring a location handoff, a token is put in place `Moves`. The rate at which location handoffs occur is $\sigma$ which is the transition rate assigned to `Move`.

- The MN will register with the new FA. This is modeled by enabling and firing transition `MN2FA` while disabling transition `Move`, after which a token flows from place `Moves` to place `MFAs`, meaning that the MN has just registered with the new FA.

- The MN's new FA will communicate with the proxy which acts as a GFA. This is modeled by enabling and firing transition `FA2Proxy` while disabling transition `Move`. After `FA2Proxy` is fired, a token in place `MFAs` flows to place `Xs`, representing that a location handoff has been completed and the proxy has been informed of the FA address change.

- If the number of tokens in place `Xs` has accumulated to K, a threshold set by the system to represent the size of a service area, then it means that the MN has just moved into a new service area and a service handoff ensues. This is modeled by assigning an enabling function that will enable transition `MoveProxy` when there K tokens in place `Xs`. After transition `MoveProxy` is fired, all K tokens are consumed and place `Xs` contains no token, representing that the proxy has just moved into a new service area.

### 5.1.2 Parameterization

Here we parameterize (give values to) transition rates of transitions `MN2FA`, `FA2Proxy`, and `MoveProxy` based on the set of base parameters defined in Table 1:

- The firing time of transition `MN2FA` stands for the communication time of the MN registering with the current FA through the wireless network. Thus, the transition rate of transition `MN2FA` is calculated as:

  $$\frac{1}{\gamma\tau}$$

  where $\tau$ stands for the one-hop communication delay per packet in the wired network and $\gamma$ is a proportionality constant representing the ratio of the communication delay in the wireless network to the communication delay in the wired network.

- When transition `FA2Proxy` fires, the FA under which the MN resides will inform the proxy of the FA address change. The transition rate of transition `FA2Proxy` depends on how far the MN's current FA is away from the proxy in terms of the number of hops. The number of hops the current FA is away from the proxy depends on the number of FA's crossed by the MN since the last time the proxy moved into the service area. Therefore we calculate the transition rate of transition `FA2Proxy` as:

  $$\frac{1}{F(Mark(\mathrm{Xs})+1)\times\tau}$$

  where $F(Mark(\texttt{Xs})+1)$ returns the number of hops between the current FA and the proxy separated by $Mark(\texttt{Xs})+1$ FA's (or subnets). The argument of the $Mark()$ function is added by 1 to satisfy the initial condition when $Mark(\texttt{Xs}) = 0$ in which the proxy has just moved into a new service area, so at the first FA crossing event, the distance between the proxy and the FA is one FA apart. Note that this transition rate is state-dependent because the number of tokens in place `Xs` changes dynamically over time.

- When transition `MoveProxy` fires, the proxy will move into a service area involving a context transfer cost. The proxy also needs to inform the HA and CNs of the address change. The transition rate of transition `MoveProxy` thus can be calculated as:

  $$\frac{1}{n_{CT}F(K)\tau+(\alpha+N\beta)\tau}$$

  where $F(K)$ returns the number of hops for two FA's separated by K subnets, $n_{CT}$ is the number of packets required to carry the service context information during a proxy transfer, $\alpha$ is the average distance between the proxy and the HA, $N$ is the number of server applications (or CNs) which the MN engages concurrently, and $\beta$ is the average distance between the proxy and a CN. The proxy could determine the values of $\alpha$ and $\beta$ based on statistical data collected on the fly to provide the best estimates of these two parameters.

### 5.1.3 Cost Model and Measurement

The cost metric considered in this chapter is the communication cost due to mobility management (i.e., the use of the proxy as a GFA in Mobile IP) and service management (the use of the same proxy for data communication activities with the application servers).

The stochastic model underlying the SPN model is a continuous-time semi-Markov chain[1] with the state representation of (a,b,c) where a is the number of tokens in place `Moves`, b is the number of tokens in place `MFAs`, and c is the number of tokens in place `Xs`. The stochastic process will reach equilibrium eventually such that there is a non-zero probability that the system will be found in one of the states in a finite set. Let $P_i$ be the steady state probability that the system is found to contain $i$ tokens in place `Xs` such that Mark(`Xs`) $= i$.

Let $C_{i,service}$ be the communication cost for the network to service a data packet when the MN is in the $i^{th}$ subnet in the service area. Let $C_{service}$ be the average communication cost to service a data packet weighted by the respective $P_i$ probabilities. This service management cost includes the communication cost between the CN and proxy, the cost between the proxy and the current FA (which depends on the number of hops separating the proxy and the current FA), and the cost for wireless communication between the current FA and the MN. Thus $C_{service}$ is calculated as follows:

$$\begin{aligned} C_{service} &= \sum_{i=0}^{K}(P_i \times C_{i,service}) \\ &= \sum_{i=0}^{K} P_i(\beta\tau + F(i)\tau + \gamma\tau) \\ &= \beta\tau + \gamma\tau + \sum_{i=0}^{K}(P_i \times F(i)\tau) \end{aligned} \qquad (5.1)$$

Let $C_{i,location}$ be the communication cost to service a location handoff operation given that the MN is in the $i^{th}$ subnet in the service area. If $i < K$, then the location handoff would only involve the communication cost for the MN to register with the new FA and for

[1] If all times are exponentially distributed, then the underlying model is a Markov chain. However, since our SPN model allows times to be generally distributed, the underlying model is semi-Markov.

the FA to inform the proxy of the address change. On the other hand, if $i = K$, then the location handoff also triggers a service handoff, which in addition to the FA registration cost mentioned above, will also incur a context transfer cost to move the proxy to the new service area and the communication cost for the proxy to inform the HA and the CNs (or application servers) of the address change of the proxy. Let $C_{location}$ be the average communication cost to service a move operation by the MN weighted by the respective $P_i$ probabilities. Then, $C_{location}$ is calculated as follows:

$$\begin{aligned} C_{location} &= \sum_{i=0}^{K}(P_i \times C_{i,location}) \\ &= \sum_{i=0}^{K-1}\{P_i \times (\gamma\tau + F(i)\tau)\} \\ &\quad +P_k \times (\gamma\tau + \alpha\tau + N\beta\tau + F(K)\tau n_{CT}) \end{aligned} \tag{5.2}$$

The total cost *per time unit* for the Mobile IP network to service operations associated mobility and service management of the MN is calculated as

$$C_{total} = C_{service} \times \lambda + C_{location} \times \sigma \tag{5.3}$$

where $\lambda$ is the data packet rate and $\sigma$ is the mobility rate.

To calculate the total communication cost $C_{total}$ based on Equations 5.1, 5.2 and (5.3), we need to obtain the steady state probability that $i$ tokens are found in the place `Xs`. We utilize SPNP [52] to define and evaluate the SPN, and to obtain $P_i$'s, when given a set of parameters characterizing the MN's mobility and service conditions. Specifically, we use the following reward assignment to calculate $P_i$:

$$r_i = \begin{cases} 1 & \text{if Mark(Xs) = i} \\ 0 & \text{otherwise} \end{cases}$$

### 5.1.4 Results and Analysis

Here we present numerical data obtained based on the SPN model developed and Equations 5.1, 5.2 and (5.3) with physical interpretations given. The numerical analysis is used to

(a) show that there exists an optimal service area for Mobile IP systems operating under the proposed proxy-based regional registration scheme for network cost minimization; (b) compare our proxy-based scheme with basic Mobile IP and MIP-RR and (c) study the effect of model parameters, including the SMR and $n_{CT}$, on the optimal service area size.

First we observe from numerical data that there exists an optimal proxy service area size $K_{opt}$ to minimize the overall communication cost when given a set of parameter values characterizing the mobility and service behaviors of the MN and the network conditions of the Mobile IP network. Figure 5.6 shows that there exists an optimal size $K_{opt} = 17$ at SMR $= 2$, $N$=1, $\alpha = \beta = 30$, $\gamma = 10$ and $n_{CT} = 4$, and an optimal size $K_{opt} = 15$ at SMR $= 2$, $N$=1, $\alpha = \beta = 30$, $\gamma = 10$ and $n_{CT} = 1$ under which the overall cost for the network to service the associated mobility and service management operations with the MN is minimized.

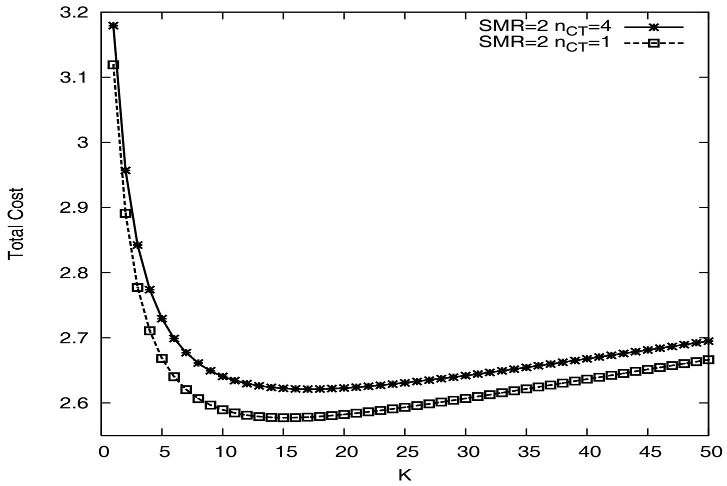

Figure 5.6: Optimal Service Area Size $K_{opt}$ with varying SMR and $n_{CT}$ in IMSA on Mobile IPv4.

To provide a better sense of the performance improvement of our proposed proxy-based scheme for integrated mobility and service management, we compare our proxy-based scheme with basic Mobile IP without route optimization and MIP-RR with route optimization. Below we first compare our scheme with basic Mobile IP. For the basic Mobile IP scheme, there is no proxy. Thus, for a client-server application, the CN (server application) itself keeps the service context information without the overhead of context transfer. The communication cost $C^{MIP}_{service}$ for servicing a packet delivery in basic Mobile IP includes a communication delay from the CN to the HA, a delay from the HA to the current FA, and a delay in the wireless link from the FA to the MN as follows:

$$C^{MIP}_{service} = \beta\tau + \alpha\tau + \gamma\tau \tag{5.4}$$

where the average distance between the CN and the HA is assumed to be about the same as that between the CN and the proxy in our proxy-based regional registration scheme.

The cost $C^{MIP}_{location}$ for servicing a location handoff under basic Mobile IP consists of a delay in the wireless link from the MN to the FA that it enters into, and a delay from the current FA to the HA as follows:

$$C^{MIP}_{location} = \gamma\tau + \alpha\tau \tag{5.5}$$

The total cost assuming $N$=1 is

$$C^{MIP}_{total} = C^{MIP}_{service} \times \lambda + C^{MIP}_{location} \times \sigma \tag{5.6}$$

Figure 5.7 compares the cost of our proxy-based scheme with basic Mobile IP under various SMR. All other parameters are fixed ($N$=1, $\alpha = \beta = 30$, $\gamma = 10$ and $n_{CT} = 2$) to isolate the effect of SMR. Both the residence time and packet interarrival time are assumed to be exponentially distributed with varying rates of $\sigma$ and $\lambda$, respectively. We observe that our proxy-based regional registration scheme incurs much less communication overhead, the effect of which is especially pronounced when SMR is high. Another observation is that the total cost increases with the increase of SMR for both schemes. We also observe that basic Mobile IP performs comparably well under very low SMR values. The reason is that when SMR is low, the data packet rate is low compared with the user mobility rate. Thus, the overhead of basic Mobile IP due to triangular routing (CN-HA-FA) for data packet delivery is negligible. Under a low SMR, our proxy-based scheme would favor a large service area (to be

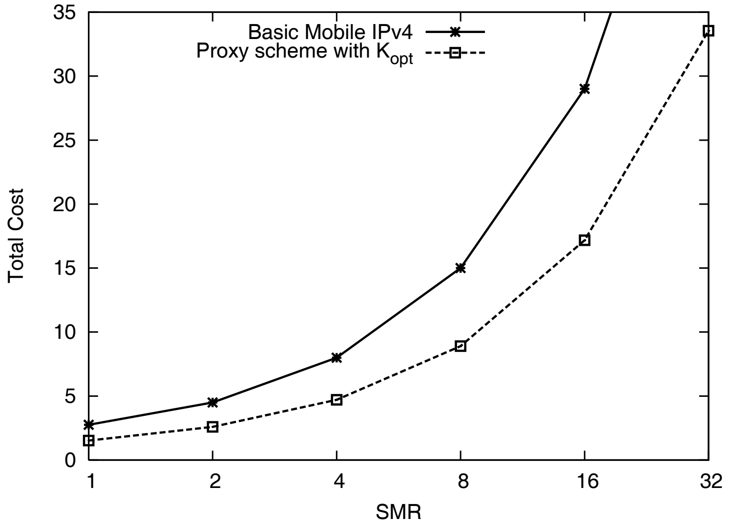

Figure 5.7: Comparison of the Proxy-based Scheme with Basic Mobile IP.

shown later) under which a service handoff incurs a high cost. These two factors thus make our proxy-based regional registration scheme performs comparable to basic Mobile IP under low SMR values. When SMR increases, the cost due to service management also increases. In basic Mobile IP, the service management cost quickly dominates the mobility management cost because of the high overhead associated with triangular routing for data packet delivery. In our proxy-based regional registration scheme, the balance between mobility and service management is obtained by moving the proxy close to the MN more frequently (thus favoring a smaller service area) to avoid the costly triangular routing (which in our case is CN-proxy-FA) for data packet delivery. Consequently, when SMR is reasonably high, our scheme grossly outperforms basic Mobile IP.

Our proxy-based regional registration scheme can be viewed as an extension of MIP-RR with route optimization, except that for each individual MN, we determine the optimal placement of its *personal* GFA through proxy migration to minimize the network cost. For MIP-RR, the placement of GFAs is pre-determined at fixed locations being applied to all MNs, and each GFA covers a fixed number of subnets, say, $K_H$. Thus, when a MN crosses a subnet within a GFA area $K_H$ subnets, the MN only informs its CoA change to the GFA. When the MN crosses a GFA domain of $K_H$ subnets, the address binding of the new GFA's

CoA is also sent to the HA and CNs through route optimization in MIP-RR to minimize the signaling cost for data delivery.

Figure 5.8 compares the total network signaling cost of our proxy-based scheme with MIP-RR with $K_H = 4$ under different combinations of $\lambda$ and $\sigma$ in the same setting. The cost is normalized with respect the per-hop communication cost $\tau = 1$. We first observe that the total network signaling cost increases with the increase of either the mobility rate or the data packet arrival rate. We next observe that our proxy scheme always incurs less communication overhead than MIP-RR, the effect of which is especially pronounced when the MN's mobility rate is high and/or when the data packet arrival rate is high. It is noteworthy that a small cost difference is considered significant because the metric is expressed in terms of the network signaling cost per time unit (normalized to the per-hop cost $\tau = 1$) so the cumulative effect would be significant over the lifetime of a single MN. It is also noteworthy that the cumulative effect of cost saving for a large number of MN would be even more significant.

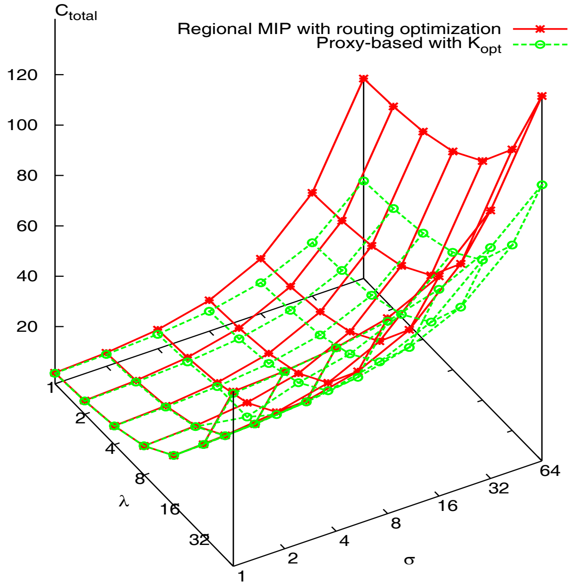

**Figure 5.8:** Comparison of the Proxy-based Scheme with MIP-RR with Route Optimization.

Below we analyze the effect of several model parameters on the optimal service area size $K_{opt}$. Figure 5.9 shows the effect of $n_{CT}$ on the optimal proxy area size $K_{opt}$ with SMR fixed at 2 for the case $N = 1$. Other cases exhibit similar trends are not shown here. Figure 5.9 shows that the optimal size $K_{opt}$ initially increases as $n_{CT}$ increases. The reason is that as $n_{CT}$ increases, the context transfer cost becomes high upon a service handoff. Thus, the proxy likes to stay in a large service area to avoid the high cost associated with a service handoff. However as $n_{CT}$ continues to increase past a threshold value, $K_{opt}$ decreases and eventually $K_{opt} = 1$. This is because the context transfer cost is proportional to $n_{CT}F(K)$, so a large $n_{CT}$ makes the context transfer cost very large by a factor of $F(K)$. Thus if we allow $K > 1$ during a proxy move operation, the cost of context transfer would dominate the cost of informing the CN and the HA of the address change and the effect of cost saving for informing the CN and the HA of the address change only after a few FA crossing events have occurred would not be significant. Consequently, allowing $K > 1$ does not gain much cost saving as far as the cost of proxy move is concerned when $n_{CT}$ is very large. This factor, coupled with the factor that when $K > 1$ the cost of informing the proxy of the current FA and the cost of delivering packets both would increase by a factor of $F(K)$, favors a small $K_{opt}$. As shown in Figure 5.9 when $n_{CT}$ is sufficiently large the system is better off with $K_{opt} = 1$.

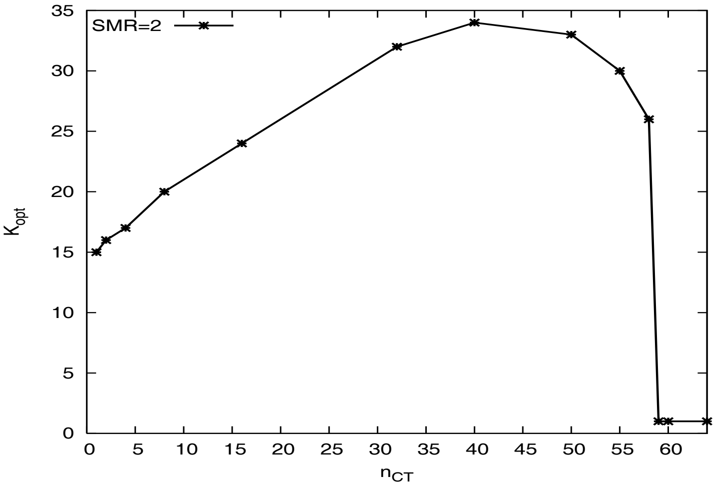

Figure 5.9: Optimal Service Area Size $K_{opt}$ as a Function of $n_{CT}$ in IMSA on Mobile IPv4.

Figure 5.10 shows the effect of SMR on the optimal service area size $K_{opt}$ with $n_{CT}$ fixed at 2 and $N = 1$. Figure 5.10 shows that the optimal service area size $K_{opt}$ decreases as SMR increases. The reason is that when SMR is small, the mobility rate is high compared to the data packet rate; thus, the mobility management cost is much larger than the service management cost. The proxy likes to stay at a large service area to reduce the location handoff cost such that a location handoff will most likely only involve informing the proxy of the location change without incurring a service handoff to migrate the proxy. However, as the data packet rate increases compared with the mobility rate (i.e., as SMR increases), a small service area will reduce the packet delivery cost because the proxy will tend to stay closer to the MN, thus reducing the overhead of triangular routing (CN-proxy-FA) for data packet delivery. As a result, when SMR increases, $K_{opt}$ decreases.

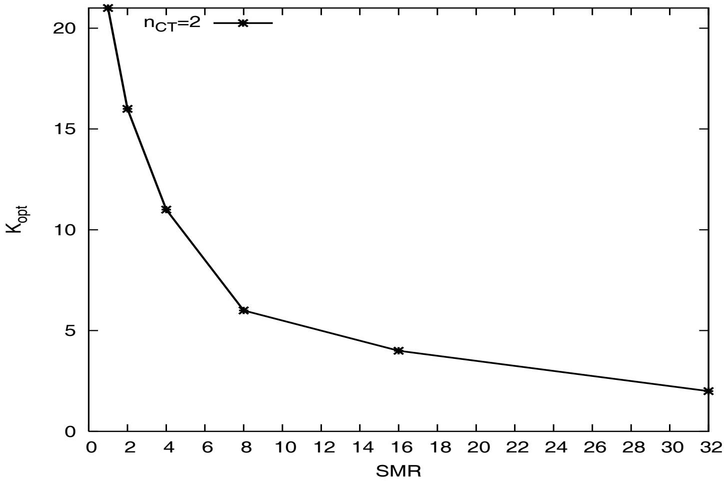

Figure 5.10: Optimal Service Area Size $K_{opt}$ as a Function of SMR in IMSA on Mobile IPv4.

All the above results are obtained based on the assumption that the average number of hops between two FA's separated by $k$ subnets is given by $F(k)= \sqrt{k}$ adopted from the fluid flow model [59]. Below we test the sensitivity of the results with respect to the form of $F(k)$. Figure 5.11 shows the total cost obtained under the optimal service area size $K_{opt}$ for $F(k)= \sqrt{k}$, $F(k) = k$ and $F(k) = k^2$. The trends exhibited by these three forms are very similar and are not sensitive to the form of $F(k)$. Thus, all the conclusions drawn earlier from the case $F(k)= \sqrt{k}$ are valid. Here we observe that, however, as the number of hops increases from $\sqrt{k}$, $k$ to $k^2$ for two FA's separated by $k$ subnets, the total cost also slightly increases. Thus this function $F(k)$ does affect the performance of our proposed proxy-based scheme for integrated mobility and service management. Since the performance metric is a rate parameter (amount of cost incurred per time unit), a small difference is also not negligible.

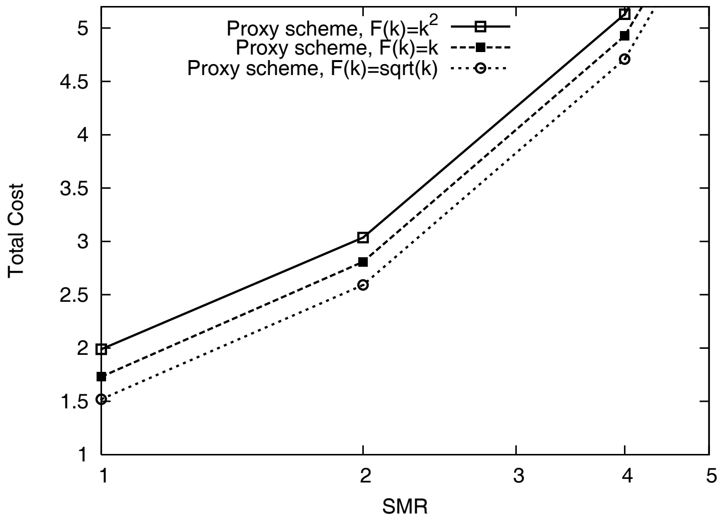

Figure 5.11: Effect of $F(k)$ on Communication Cost in IMSA on Mobile IPv4: A Comparison.

Figure 5.12 shows the sensitivity of $F(k)$ on the optimal service area size $K_{opt}$ The data show that as the number of hops separating two FA's involved in a service handoff increases, e.g., $F(k) = k^2$, the system would favor a smaller service area because a smaller service area tends to mitigate the negative effect of $k^2$ so as reduce the cost of context transfer in a service handoff.

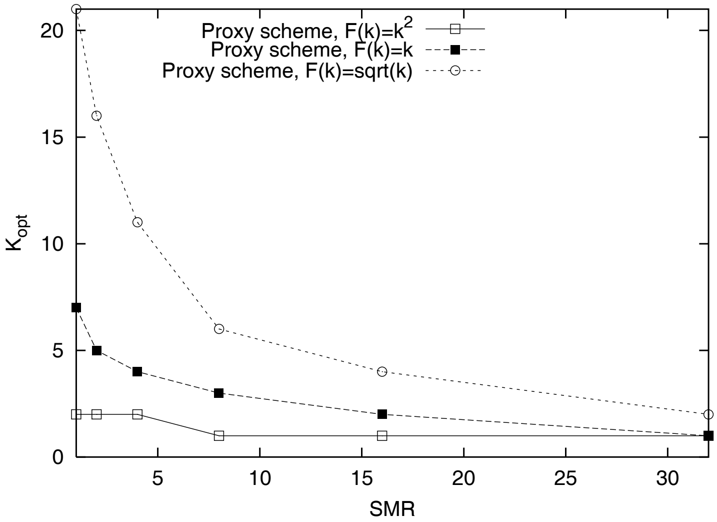

Figure 5.12: Effect of $F(k)$ to $K_{opt}$ in IMSA on Mobile IPv4.

## 5.2 IMSA-MIPv6: IMSA for IPv6

The IMSA-MIPv6 design is also based on the notion of proxies. Specifically, a client-side proxy is created on a per-user basis to serve as a gateway between the MN and all application services engaged by the MN.

IMSA-MIPv6 and DMAP introduced in Chapter 4 have a similar architecture except that IMSA-MIPv6 uses a proxy to maintain the mapping of a MN's RCoA to its CoA. Specifically, in the DMAP design, since ARs are not able to run proxies, the mapping of the MN's RCoA to its CoA is done by having the current MAP maintain an internal table, so the MAP can intercept a packet destined for a MN's RCoA and forward it to the MN's CoA. In IMSA-MIPv6, the mapping of RCoA to CoA is done by having a proxy run on the MAP directly receive a packet destined for the MN's RCoA, so the proxy can in turn forward the packet to the MN's CoA. In addition, in IMSA-MIPv6 the user proxy carries context information to facilitate service management. For both designs, the concept of dynamic service area applies. That is, the MN will dynamically determine the size of the regional registration service area depending on the MN's current mobility and service characteristics.

In our IMSA design, a central issue is to dynamically determine an optimal service area when given runtime mobility and service characteristics of a MN to minimize the network signaling cost. A MN's service area can be modeled as consisting of $K$ IP subnets. As in IMSA-MIPv4, the computational procedure developed allows each MN to determine the optimal service area size in terms of $K$. If $K = 1$, the proxy will run on the MN, in which case our scheme degenerates to the basic MIPv6 scheme. If $K > 1$, the proxy will run on the first AR after $K$ subnets are crossed by the MN. The optimal $K$ value is determined at runtime dynamically so as to minimize the network signaling cost. For the special case in which $K$ is constant for all MNs, our scheme degenerates to HMIPv6 with a two-level MAP-AR structure.

### 5.2.1 Modeling and Analysis

We use a performance model as shown in Figure 5.13 to analyze IMSA-MIPv6. As expected, the model is remarkably similar to that for DMAP in Figure 4.2. However because in IMSA-MIPv6, a proxy is used to serve as the regional register and to carry service context information, the performance model is parameterized very differently. Here we describe the part that is different from DMAP.

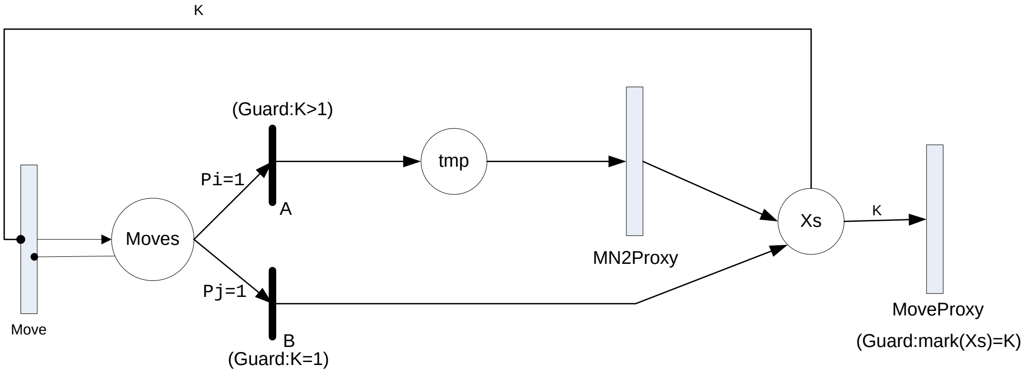

Figure 5.13: Stochastic Petri Net Model for IMSA-MIPv6.

Specifically, when transition `MoveProxy` fires, the user proxy will move into a service area after invoking a service handoff. The cost involved includes informing the HA and CNs of the CoA address change, and transferring service context information if the proxy carries context information. The transition rate of transition `MoveProxy` therefore is calculated as:

$$\begin{cases} \dfrac{1}{n_{CT}F(K)\tau + (\alpha + N\beta)\tau} & K > 1 \\ \dfrac{1}{(\alpha + N\beta)\tau} & K = 1 \end{cases}$$

where $F(K)$ returns the number of hops for two subnets separated by $K$ subnets, $n_{CT}$ is the number of packets required to carry the service context information during a proxy transfer, $\alpha$ is the average distance between the proxy and the HA, $N$ is the number of server applications (or CNs) which the MN engages concurrently, and $\beta$ is the average distance between the proxy and a CN. The proxy could determine the values of $\alpha$ and $\beta$ based on statistical data collected on the fly.

### 5.2.2 Numerical Results

Figure 5.14 shows that under IMSA-MIPv6, there exists an optimal proxy service area size $K_{opt}$ to minimize the overall communication cost when given a set of parameter values characterizing the mobility and service behaviors of the MN and the network conditions in the MIPv6 network. The example is based on $F(K) = \sqrt{K}$ [59], $\alpha = \beta = 30$, and $\gamma = 10$. We observe that $K_{opt}$ exists for all $n_{CT}$ values, including the special case in which $n_{CT} = 0$ when the proxy does not carry service context information. We see that $K_{opt}$ increases as $n_{CT}$ increases to amortize the cost of service context transfer due to proxy migration. However when $n_{CT}$ is sufficiently large, the service context transfer cost is too high, and $K_{opt}$ converges to 1 (i.e., the proxy resides on the MN, or the MAP is always the current AR) to eliminate the service transfer cost.

### 5.2.3 Discussion

For HMIPv6, the placement of MAPs is pre-determined. Thus, when a MN crosses a subnet within a MAP domain of $K_H$ subnets, it only informs the MAP of its CoA. On the other hand, when the MN crosses a MAP domain of $K_H$ subnets, it changes the MAP, obtains a

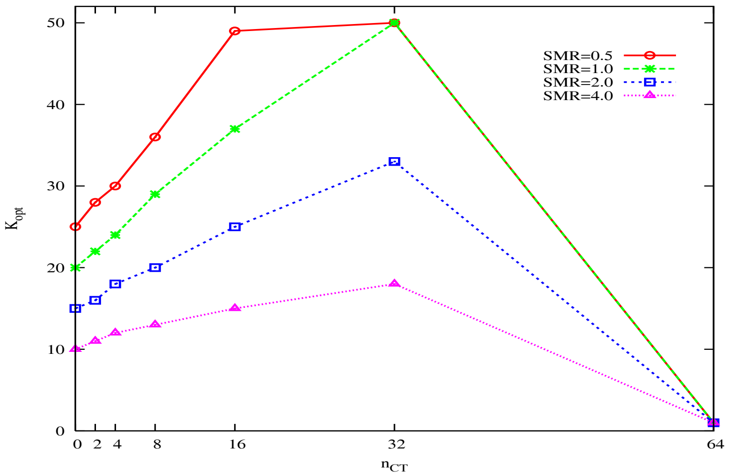

Figure 5.14: Optimal Service Area Size $K_{opt}$ as a Function of $n_{CT}$ and SMR in IMSA-MIPv6.

new RCoA and informs the HA and CNs of the new RCoA. The performance comparison of IMSA-MIPv6 with HMIPv6 and basic MIPv6 remains to be done. Also IMSA-MIPv6 now is based on a two-level architecture. We plan to investigate Hierarchical IMSA-MIPv6 with not only an optimal regional area size at each level but also an optimal number of levels and compare its performance with HMIPv6.

# Chapter 6

# PICMM: Proxy-Based Integrated Cache Consistency and Mobility Management Scheme in Mobile IP Systems

In this chapter, we demonstrate the applicability of our IMSA-MIPv6 design with a class of mobile data query applications. For efficiency purposes, a MN engaged in such applications can cache a set of frequently accessed data objects such that if a data object queried by the MN is up-to-date, the MN can answer the query instantly using the cached copy. A cache consistency management scheme being considered in the dissertation is based on a stateful strategy by which cache invalidation messages are asynchronously sent by the server to a mobile node (MN) whenever data objects cached at the MN have been updated. Further, we use a per-user proxy to buffer invalidation messages to allow the MN to disconnect arbitrarily and to reduce the number of uplink requests when the MN is reconnected.

For integrated data consistency and mobility management, we let the user proxy also take the responsibility of mobility management to further reduce the network traffic. Specifically, in our design the MN's proxy serves as a gateway foreign agent (GFA) as in the MIP Regional Registration protocol to keep track of the address of the MN in a region. The proxy migrates with the MN when the MN crosses a regional area. We identify the *optimal* regional area size under which the overall network traffic cost, due to cache consistency management, mobility management, and query requests/replies, is minimized. We demonstrate that the integrated cache consistency and mobility management scheme outperforms both no-proxy and/or no-cache schemes in Mobile IPv6 environments.

## 6.1 Background

The design considered here follows IMSA-MIPv6. We call it the Proxy-based Integrated Cache and Mobility Management (PICMM) scheme to make clear that the service context is for supporting mobile client-server database applications in which the MN queries the server for dynamic data. For example, a MN may query dynamic data such as stock prices, weather reports, or traffic information. To avoid sending a query to the server and receiving a reply through the expensive and often unreliable wireless communication network, a MN can cache data objects on the local storage and then answer queries for data that are up-to-date. Caching reduces the server access cost and improves user-perceived response time [26].

To process user queries correctly based on caching, a MN must ensure that its cached data are up-to-date. In the literature, cache invalidation strategies can be classified into two categories [51] [54]. The first category is based on the *stateful* strategy by which the server knows which data objects have been cached by the MNs. When there is an update to a data object, the server will send an invalidation message to those MNs that keep a cache copy. The second category is based on the *stateless* strategy by which the server has no knowledge of cache contents of MNs. The server will broadcast information on data objects that have been updated either periodically or asynchronously. A MN would listen to the broadcast to determine whether its cached objects are up-to-date. Both strategies have the problem that if a MN misses invalidation reports while it is disconnected, it will have to discard the cache content after it reconnects.

The PICMM scheme proposed in this chapter is based on the stateful strategy by which a cache invalidation message is asynchronously sent by the server to the MN, whenever there is an update to a data object cached at the MN. The MN uses invalidation reports received to determine the validity of its cache content before answering a query. If a query asks for a data object that has been invalidated, then a request is sent uplink to the sever to ask for a fresh copy of the data object accessed by the query before the query can be answered. Moreover, to support service continuity in cases the MN is disconnected and then reconnected again, we use a per-user proxy to buffer invalidation messages to allow the MN to disconnect arbitrarily and to reduce the number of uplink requests to check cache status when the MN is reconnected.

The generated network traffic cost associated with cache consistency management thus includes the cost of receiving and buffering invalidation messages at the proxy and the cost of forwarding them from the proxy to the MN, as well as the cost of sending requests to the server and receiving responses in case cached data objects are not up-to-date to answer queries. This cost is considered as part of the "service" management cost which we like to minimize. Another major network traffic cost in MIPv6 systems is due to mobility management to maintain the location of the MN.

To minimize both the "service" and "mobility" network traffic costs, we let the user proxy mentioned above also take the responsibility of mobility management to further reduce the network traffic. The PICMM design uses a MN's proxy as a GFA as in the MIP-RR protocol [18, 35] to keep track of the address of the MN in a region. The proxy migrates with the MN when the MN crosses a regional area. We aim to identify the *optimal* regional area size under which the overall network traffic generated due to cache consistency management, mobility management, and query requests/replies, is minimized.

The basic idea is that we use a client-side proxy to support caching and mobility management in Mobile IPv6. The proxy has three functions: (1) working as a GFA as in regional registration to keep tracking MN's location; (2) acting as a service proxy for services engaged by the MN; (3) allocating a buffer space to store service context information for each MN. The proxy will receive invalidation reports from the server on behalf of the MN. If the MN is connected, the proxy will forward the invalidation report to the MN. If the MN is disconnected, the proxy will store invalidation reports in the proxy's buffer. Once the MN is reconnected, the MN will get the latest invalidation reports from the proxy.

We consider mobile client-server applications in Mobile IPv6 environments in which mobile clients query database servers for data objects. To speed up query processing, a client stores a copy of frequently used data objects in its local cache. When a query is issued by the client, the validity of cached data objects accessed by the query is checked first. If these data objects are valid, then the query is immediately processed with very little access time. Otherwise, a fresh copy of invalid data objects will be retrieved to the local store before the query is answered.

## 6.2 Algorithm Description

To support cache management, the client-sider proxy maintains a buffer to store invalidation reports received from the server. The proxy communicates with the server on behalf of the MN. Thus, invalidation reports as well as data packets sent to the MN from the server will be routed directly to the proxy, even when the MN is disconnected.

To support mobility management, the client-sider proxy also serves as a GFA as in the MIP-RR protocol to maintain the location information of the MN. When the MN moves across a subnet boundary within the GFA area, it obtains a new CoA. However, the MN does not need to inform the HA and CNs of the CoA change; instead only the proxy is informed of the CoA change. On the other hand, when the MN moves across a GFA area[1], the proxy also moves to run on the AR of the first subnet in the new GFA area. This incurs the cost of transferring the proxy from one GFA to another GFA area, and also the cost of informing the HA and CNs of the address change of the proxy. The size of the service area is defined as the number of subnet it covers. The optimal size of the service area depends on a MN's runtime mobility and service characteristics to minimize the mobility and service management cost.

To improve query performance, the MN may store frequently used data objects in its

[1] A GFA area is termed a "service area" in this chapter because we consider integrated mobility and service management

cache (called *MNCache*). The MN will use invalidation reports received from the CN (through the proxy) to validate the content of its cache. We consider that the system operates based on the asynchronous *stateless* strategy, that is, when a data object is changed, the CN immediately sends out an invalidation report to those MNs that keep a cached copy. When a service proxy is created to run on an AR, the proxy is allocated with a buffer space (called *ProxyCache*) to cache invalidation reports and possibly application context sensitive information. The proxy acting on behalf of the MN receives invalidation reports from the CN. If the MN is connected, the proxy forwards them to the MN. If the MN is disconnected, the proxy stores them in *ProxyCache*. Once the MN wakes up, it gets invalidation reports from *ProxyCache* to check if its cache content is current. When the proxy moves because the MN moves across a service area, *ProxyCache* moves with the proxy, thus incurring a *context transfer* cost.

Figure 6.1 illustrates our integrated cache and mobility management scheme in Mobile IPv6 environments. The per-MN proxy works as a GFA for mobility management. It migrates only when it crosses a service area. Initially the proxy runs on the first AR of service area 1. When the MN moves within service area 1, the proxy stays put and the MN will inform the proxy of its CoA change. When the MN moves to service area 2, the proxy also migrates to service area 2 and then runs on the first AR of service area 2. The proxy's *ProxyCache* also moves. In addition, the HA and CNs are informed of the CoA change of the proxy.

### 6.2.1 Cache Invalidation

The cache invalidation process runs as follows:

```
The client-side proxy receives an invalidation report from the CN
whenever there is an update to a cached data object;
if the MN is connected to the wireless network
then
  the proxy forwards the invalidation report to the MN;
```

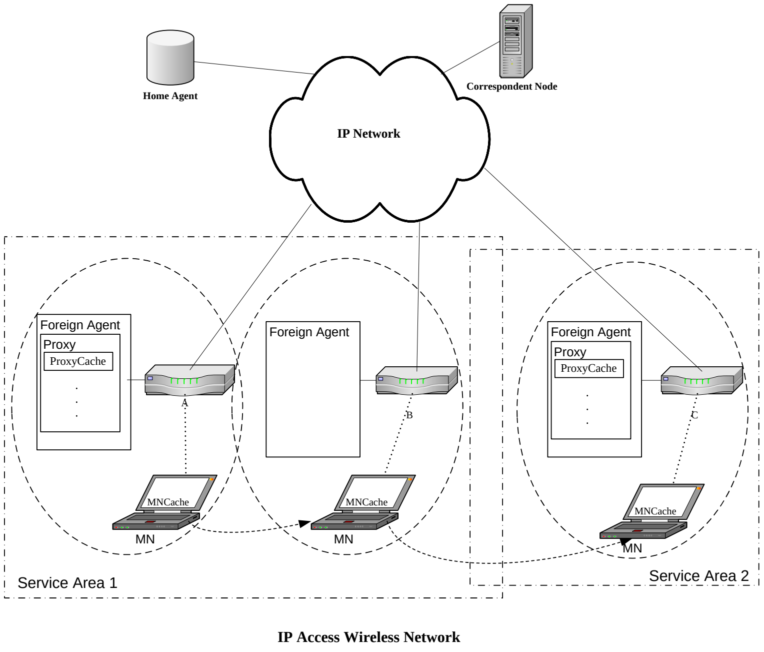

Figure 6.1: Integrated Cache and Mobility Management System Infrastructure.

```
else
  the proxy stores the invalidation report in ProxyCache;
  when the MN wakes up, the proxy forwards the invalidation report.
```

### 6.2.2 Query Processing

The MN querying process runs as follows:

```
The user issues a query request to the MN for a data object;
if the data object requested is in the MN's cache and valid
   there is a cache hit;
   the MN returns the query result;
else
   there is a cache miss;
   user request is sent to the proxy;
   proxy forwards the request to the CN;
   CN returns a copy of the requested object to the proxy;
   proxy forwards the data object to MN;
   MN stores the data object in MNCache and returns query result.
```

### 6.2.3 Disconnection Support

When a MN reconnects after a sleep, it performs the following process:

```
MN just reconnects to wireless network;
if the MN is in the same subnet as the proxy
    the MN gets invalidation reports from ProxyCache;
else
    proxy moves to MN's current subnet;
    ProxyCache is moved with the proxy;
    the proxy informs the HA and CNs of its address change;
    the MN gets invalidation reports from ProxyCache.
```

A trade-off exists to balance the mobility management cost and the cache consistency management cost. A large service area size means that the proxy will not move often, and the query cost due to cache misses could be high because of the triangular routing path CN-Proxy-MN. The cache invalidation cost is also high because of the larger distance between the proxy and the MN when the service area is large. A small service area means that the proxy moves often. This increases the cost of moving ProxyCache with the proxy and to inform the HA and CNs of address change. Thus, an optimal service area size exists. The proxy determines the *dynamic* optimal service area size at runtime. When the MN moves across a subnet boundary, the proxy will check if the service area is crossed. If yes, a new optimal service area size is determined by executing a computational procedure developed in this chapter based on the MN's mobility and service characteristics in the new service area.

## 6.3 Modeling and Analysis

### 6.3.1 Parameters

We summarize *dynamic* connection/disconnection, service and mobility characteristics of a MN by several parameters. The first parameter describes the on/off (or wake/sleep) behavior of the MN. We assume that while the MN is in a wake state, it will go to sleep with rate $\omega_w$, and, conversely, while the MN is in a sleep state, it will wake up with rate $\omega_s$.

The second parameter is the residence time that the MN stays in a subnet while it is in a wake state. This parameter can be collected by each MN based on statistical analysis techniques [11]. We assume that future MNs are adequately powerful for collecting data and performing simple statistical analysis. The residence time in general would be characterized by a general distribution. We use the MN's mobility rate ($\sigma$) to represent this parameter.

The third parameter is the service traffic between the MN and server applications while the MN is in a wake state. The service rate depends on the query rate for data objects needed by queries, and, if data objects are cached at the MN, the miss ratio. While a query may access several data objects, we assume it is possible to break up a query into sub-queries each accessing a single data object. Thus, it is possible to know a priori the average query rate for a data object $i$. We characterize the frequency at which the MN accesses data object $i$ by a query rate $\lambda_{q,i}$. Suppose that the MN prefetches $N_{data}$ into its cache. Then the cumulative query arrival rate to access $N_{data}$ data objects, denoted by $\lambda_Q$, is given by:

$$\lambda_Q = \sum_{i=1}^{N_{data}} \lambda_{q,i}. \qquad (6.1)$$

The ratio of $\lambda_Q$ over $\sigma$ is called the service to mobility ratio (SMR). In general, the SMR of a MN is dynamically changed. The integrated cache and mobility management scheme proposed in this chapter allows the optimal regional area to be determined dynamically based on a MN's runtime mobility and service traffic characteristics to minimize the generated network traffic. For efficiency purposes, the MN could build a table to lookup its mobility and service rates as a function of its location and time. We assume that the MN will process a query only when it is in a wake state because otherwise the MN would not be able to ascertain if data objects requested are up-to-date.

A query accessing data $i$ stored in a MN's local cache will result in a miss if the data object has been invalidated. Let $P_{miss,i}$ represents the miss ratio for data object $i$. Then, since an uplink request will be generated by the MN to send to the server only when there is a miss, the overall query rate under cache management to the server is given by:

$$\lambda_Q^{cache} = \sum_{i}^{N_{data}} \lambda_{q,i} P_{miss,i} \tag{6.2}$$

Lastly, the fourth parameter summaries how often a data object cached at the MN is modified by the server in the mobile application. We use the per data object update rate $\mu_i$ to denote this. Assume there are $N_{data}$ data objects cached by the MN in a time window. Assume for simplicity, the MN always queries these $N_{data}$ data objects in a time window. When a query accessing a data object results in a miss, we assume a packet is sent from the MN to the server to retrieve a copy of the requested data object. After the data object is received, the query can be answered by the MN. We assume that a reply will take $n_D$ packets to hold a copy of the requested data object.

When a MN moves across a service boundary, the service proxy moves. We call this a *service handoff*. A service handoff incurs a mobility signaling cost and a service cost. The mobility signaling cost refers to the network traffic cost generated for the proxy to inform the HA and the CNs of the new CoA address. The service cost refers to the network traffic cost generated for moving the service context and invalidation reports stored to the new proxy. We let $n_{CT}$ represent the number of packets required to move the service context.

The client side proxy is created when the MN starts in MIPv6. We assume that a cross-layer design is being employed for the implementation of the client-side proxy. The cross-layer design has two components. The first component is at the network layer dealing with mobility management and the second component is at the application layer dealing with service management such as storing invalidation reports or service context information.

The system parameters that characterize the mobility and service characteristics of a MN

| Symbol | Meaning |
|---|---|
| $\lambda_{q,i}$ | query arrival rate for data object $i$ |
| $\lambda_Q$ | cumulative query arrival rate to a MN |
| $\sigma$ | mobility rate, i.e., how fast a MN moves across subnet boundaries |
| $\mu_i$ | data update rate to data object $i$ |
| SMR | service rate to mobility rate ratio, i.e., $\lambda_Q/\sigma$ |
| $\omega_w$ | disconnection rate to go from awake to asleep |
| $\omega_s$ | reconnection rate to go from asleep to awake |
| $n_{CT}$ | number of packets required for content transfer of *ProxyCache* |
| $n_D$ | number of packets to hold a data object |
| $N$ | number of server applications currently engaged by the MN |
| $N_{data}$ | number of data objects cached at the MN |
| $F(K)$ | a general function relating the number of subnets $K$ to the number of hops |
| $K$ | number of subnets (or ARs) in a service area |
| $\tau$ | 1-hop **round trip** communication delay per packet in wired networks |
| $\alpha$ | average distance (in hops) between the HA and the proxy |
| $\beta$ | average distance (in hops) between a CN and the proxy |
| $\gamma$ | ratio between communication time in a wireless network<br>to communication time in a wired network |
| $P_{miss,i}$ | probability of cache miss for data object $i$ causing an uplink request |
| $P_{wake}$ | probability of MN in the wake state; it is equal to $\omega_s/(\omega_w + \omega_s)$ |

Table 6.1: Integrated Cache and Mobility Management Parameters.

in a Mobile IPv6 system are summarized in Table 6.1 for easy reference.

### 6.3.2 Performance Model

We develop a performance model based on Stochastic Petri nets to analyze the cache management cost and mobility management cost incurred due to the employment of the integrated mobility and cache management scheme. The objective is to derive an equation from the analytical model to allow the overall network traffic cost incurred to be calculated as a function of the number of subnets covered ($K$) in a service area, when given a set of parameters as input to characterize a MN's mobility and service behaviors, including the mobility rate ($\sigma$), query arrival rate ($\lambda_{q,i}$), disconnection and reconnection rates ($\omega_s$ and $\omega_w$), and data object update rate $\mu_i$. This computational procedure allows us to determine the optimal service area size ($K_{opt}$) to be deployed at runtime to minimize the overall network traffic cost due to cache and mobility management.

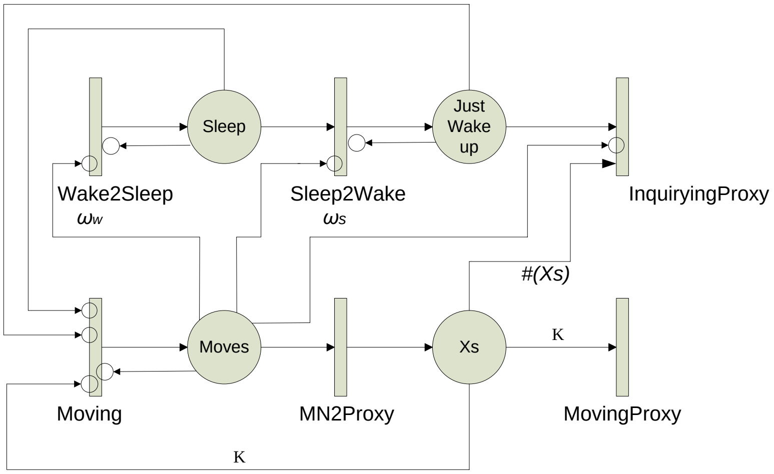

Figure 6.2: Integrated Cache and Mobility Management Petri Net Model.

Figure 6.2 shows the performance model based on Stochastic Petri nets. Table 2 gives the meaning of places and transitions defined in the model. The function Mark(`P`) returns the number of tokens in place `P`. The number of tokens accumulated in place `Xs`, that is, Mark(`Xs`), represents the number of subnets crossed by the MN since the MN enters a new service area. We allow it to accumulate to $K$, at which point we perform a service handoff. We construct the SPN model to describe the behavior of a MN operating under our proposed integrated cache and mobility management scheme. By varying $K$, the SPN model allows us to compute the cache and mobility management cost as a function of $K$, thereby allowing the optimal $K$ to be determined. Below we describe how we construct the SPN model:

- When a MN moves across a subnet area, a token is put in place `Moves`. The mobility rate at which location handoffs occur is $\sigma$ which is the transition rate assigned to `Moving`.

- After moving into a subnet, the MN obtains a new CoA and informs the proxy (that acts as a GFA) of the CoA change. This is modeled by enabling and firing transition `MN2Proxy` while disabling transition `Moving`. After `MN2Proxy` is fired, a token in place

| Symbol | Meanings |
|---|---|
| `Sleep` | If Mark(`Sleep`)=1, the MH has just entered sleep state. |
| `Wake2Sleep` | a timed transition for the MH to change from wake to sleep. |
| `Just Wake up` | If Mark(`Just Wake up`)=1, the MH has just entered wake state. |
| `Sleep2wake` | a timed transition for the MH to change from sleep to wake. |
| `InquiryingProxy` | a timed transition for the MH to query the proxy. |
| `Moves` | If Mark(`Moves`)=1, the MH has just moved across a subnet area. |
| `Moving` | a timed transition for the MH to move across subnet areas. |
| `MH2Proxy` | a timed transition for the MH to communicate with the proxy. |
| `Xs` | Mark(`Xs`) indicates the number of subnets crossed in a service area. |
| `MovingProxy` | a timed transition for the proxy to move into a new server area. |
| `K` | number of subnets crossed after which a service handoff occur. |

Table 6.2: Meanings of Places and Transitions in the SPN Model for PICMM.

`Moves` flows to place `Xs`, representing that a location handoff has been completed and the proxy has been informed of the CoA address change of the MN. The rate at which `MN2Proxy` is fired depends on the number of subnets separating the MN and the proxy.

- If the number of tokens in place `Xs` has accumulated to $K$, a threshold representing the size of a service area, then it means that the MN has just moved into a new service area and a service handoff ensues. This is modeled by assigning an enabling function that will enable transition `MovingProxy` after $K$ tokens have been accumulated in place `Xs`. After transition `MovingProxy` is fired, all $K$ tokens are consumed and place `Xs` contains no token, representing the action that the proxy has just moved into a new service area. The rate at which transition `MovingProxy` fires depends on the cost of informing the HA and CNs of the proxy CoA change and the cost of transferring $ProxyCache$ to the new proxy location.

- The MN alternates between "sleep" and "wake" states due to connection and disconnection. Initially the MN is in the wake state. After a time is elapsed representing the online connection time, the MN goes to the sleep state. This is modeled by having a token flow into place `sleep`. The transition rate to transition `wake2Sleep` is $\omega_s$. Note that if the MN is already in place `sleep`, transition `wake2Sleep` cannot fire.

- While the MN is in sleep mode, after a time is elapsed representing the disconnection time, the MN goes to the wake state. This is modeled by having a token flow from place `wake2Sleep` to place `just-wakeup`. The transition rate to transition `sleep2Wake` is $\omega_w$. If the MN is already in the wake state, transition `sleep2Wake` cannot fire.

- After the MN wakes up, it will check with the proxy of the status of its cached data objects. If the proxy is not in the same subnet, then the proxy moves to the subnet that the MN currently resides. The cost involved includes the cost of service context transfer of *ProxyCache* and the signaling cost of informing the HA and CNs of the address change. This is modeled by firing the `inquiryProxy` transition with a transition rate reflecting the cost (see how we do parameterization below). Firing this transition will flush all the tokens in place `Xs` as if a service handoff had happened. This is modeled by a variable input arc from place `Xs` to transition `inquiryProxy`.

- While the MN is in state `sleep`, it won't move, so we use an inhibitor arc from place `sleep` to transition `Moving` to prohibit transition `Moving` from firing. Similarly we use an inhibitor arc from place `just-wakeup` to transition `Moves`, and an inhibitor arc from place `Moves` to all other transitions.

Below we describe how we parameterize the SPN model. Transitions `Moving`, `wake2Sleep` and `sleep2Wake` have transition rates of $\sigma$, $\omega_w$ and $\omega_s$, respectively. Transition rates of the three remaining transitions need to be parameterized (i.e., given values to) given basic parameter values. The firing time of transition `MN2Proxy` stands for the communication time of the MN registering with the proxy through the wireless network. This time depends on the number of hops separating the MN and its proxy. Thus, the transition rate of transition `MN2Proxy` is calculated as:

$$\frac{1}{\gamma\tau + F(Mark(\mathrm{Xs}) + 1) \times \tau}$$

where $\tau$ stands for the one-hop communication delay per packet in the wired network and $\gamma$ is a proportionality constant representing the ratio of the communication delay in the wireless network to the communication delay in the wired network. $F(Mark(\texttt{Xs})+1)$ returns the number of hops between the current subnet and the proxy separated by $Mark(\texttt{Xs}) + 1$ subnets. The argument of the $F(x)$ function is added by 1 to satisfy the initial condition that $Mark(\texttt{Xs}) = 0$ in which the proxy has just moved into a new service area, so at the first subnet crossing event, the distance between the proxy and the subnet is one subnet apart. Note that this transition rate is state-dependent because the number of tokens in place `Xs` changes dynamically over time.

When transition `MovingProxy` fires, the proxy will move into a service area after invoking a service handoff. The cost involved includes informing the HA and CNs of the CoA address change, and transferring service context information stored in *ProxyCache*. The transition rate of transition `MovingProxy` thus is calculated as:

$$\frac{1}{n_{CT}F(K)\tau + (\alpha + N\beta)\tau}$$

where $F(K)$ returns the number of hops for two subnets separated by $K$ subnets, $n_{CT}$ is the number of packets required to carry the service context information during a proxy transfer, $\alpha$ is the average distance between the proxy and the HA, $N$ is the number of server applications (or CNs) which the MN engages concurrently, and $\beta$ is the average distance between the proxy and a CN. The proxy could determine the values of $\alpha$ and $\beta$ based on statistical data collected on the fly.

When transition `InquiryProxy` fires, the MN will contact the proxy. If the proxy is in the same subnet as the current MN resides, the cost is the transmission delay from the AR to the MN, that is, $\gamma\tau$. If the proxy is not in the same subnet, the cost includes the delay for contacting the proxy, moving the proxy to the current AR, and informing the HA and CNs of the proxy's CoA address change. Thus, the transition rate of transition *InquiryProxy* is calculated as:

$$\begin{cases} \dfrac{1}{\gamma\tau} & \text{if Mark(Xs)} = 0; \\ \dfrac{1}{F(\text{Mark(Xs)})\tau + n_{CT}F(\text{Mark(Xs)})\tau + (\alpha + N\beta + \gamma)\tau} & \text{if Mark(Xs)} > 0; \end{cases}$$

### 6.3.3 Cost Function Derivation

A MN and its proxy determine the service area dynamically to minimize the overall cache and mobility management network traffic cost incurred by the MN. There are three costs to be considered: a cost for forwarding a user query uplink to the server to obtain a copy of the data object because of cache miss, a cost for forwarding invalidation reports from the proxy to the MN, and a cost for mobility management, including handling location and service handoffs. The overall network traffic cost that we aim to minimize is the sum of these three costs *per time unit*.

Let $C_{query}$ be the average communication cost to service a query. Let $C_{mobility}$ be the average communication cost to service a location handoff, including one that can trigger a service handoff. Let $C_{invalidation}$ be the average communication cost to forward an invalidation report. Finally let $C_{total}$ be the overall cost incurred *per time unit*. Then, $C_{total}$ is the sum of the product of the respective communication cost multiplied with the rate at which the respective event occurs, that is,

$$C_{total} = \lambda_e \times C_{query} + \sigma_e \times C_{mobility} + \mu_e \times C_{invalidation} \tag{6.3}$$

Here $\lambda_e$ is the effective data query rate, $\sigma_e$ is the effective mobility rate, and $\mu_e$ is the effective data update rate.

The effective query arrival rate $\lambda_e$ is the query arrival rate multiplied with the probability of the MN is being awake because queries are issued only when the MN is awake. Thus,

$$\lambda_e = \lambda_Q \times P_{wake} \tag{6.4}$$

where $\lambda_Q$ is the aggregate query arrival rate as given in Equation 6.1 and $P_{wake}$ is the probability of the MN being in the wake state, given by:

$$P_{wake} = \frac{\omega_s}{\omega_s + \omega_w} \tag{6.5}$$

The effective mobility rate $\sigma_e$ is the mobility rate multiplied with the probability of the MN being awake again because the MN does not trigger mobility handoff events while it is in sleep model. Thus,

$$\sigma_e = \sigma \times P_{wake} \tag{6.6}$$

On the other hand, the effective data update rate $\mu_e$ is simply the aggregate data update rate, calculated as:

$$\mu_e = \sum_{i=1}^{N_{data}} \mu_i \tag{6.7}$$

Here the expression for $\mu_e$ is not multiplied with the probability of the MN being awake because updates to data occur regardless of if MN is in sleep or wakeup mode.

Below we derive $C_{query}$, $C_{mobility}$ and $C_{invalidation}$. The stochastic model underlying the SPN model is a continuous-time semi-Markov chain with the state representation of $(a, b, c, d)$ where $a$ is the number of tokens in place `Moves`, $b$ is the number of tokens in place `Xs`, $c$ is the number of tokens in place `Sleep`, and $d$ is the number of tokens in place `Just Wake Up`. The distribution of tokens in these four places makes up the states of the system. In general let state $i$ represent a particular state represented by $(a, b, c, d)$. Let $P_i$ be the steady state probability that the system is in state $i$. $P_i$'s can be obtained by applying numerical analysis methods such as SOR or Gauss Seidel to solve the underlying model.

Let $C_{i,query}$ be the communication cost for answering a query given that the MN is in state $i$. Then, $C_{query}$ can be calculated as a weighted average of $C_{i,query}$'s as follows:

$$C_{query} = \sum_i (P_i \times C_{i,query}) \tag{6.8}$$

Deriving $C_{i,query}$ requires knowledge of $P_{miss,j}$, i.e., the probability of cache miss of data object $j$, because this cost depends on whether the requested cached data object is up-to-date and can be used to answer a query. A cache miss happens if an update has occurred to object $j$ prior to the query requesting object $j$ arriving at the MN. Inevitably this depends on the relative magnitudes of $\lambda_{q,j}$, $\mu_j$ and $\omega_w$ and $\omega_s$. The calculation of $P_{miss,j}$ essentially hinges on the competition between the *effective query arrival rate* and the update rate (with respect to time), since the cached data object will be invalidated when an update occurs before a query arrives. Since the MN will not issue queries while it is in sleep mode, the effective query arrival rate for object $j$ is equal to the query arrival rate for data object $j$ multiplied with the probability of being awake, i.e., $P_{wake}\lambda_{q,j}$, or $[\omega_s/(\omega_s+\omega_w)]\lambda_{q,j}$. Thus, $P_{miss,j}$ is calculated as:

$$P_{miss,j} = \frac{\mu_j}{(\mu_j + [\omega_s/(\omega_s+\omega_w)]\lambda_{q,j})} \tag{6.9}$$

Suppose that a query or a subquery is asking for data object $j$. If data object $j$ being queried is in the MN's cache and is valid, then there is a cache hit and the query cost is zero. Otherwise, there is a cache miss, and the query cost will include a communication cost between the MN to the proxy, and a cost from the proxy to the CN. Thus, the average query cost for a query asking for data object $j$ when the MN is connected is given by $(\gamma\tau + \beta\tau + F(\text{Mark(Xs)})\tau) \times n_D \times P_{miss,j}$.

On the other hand, when the MN just wakes up, i.e., in state “just wake up”, the MN will first check with the proxy regarding the cache status when answering a query, and, if the cached data object is not valid, will get a copy from the server. Thus, the cost is $\gamma\tau + (\gamma\tau + \beta\tau + F(\text{Mark(Xs)})\tau) \times n_D \times P_{miss,j}$, where the first term is the cost for the MN to get invalidation reports from the proxy (which has moved to local) to check the cache status and the second term is for the cost to get a copy of object $j$ from the CN if there is a miss. Lastly, when the MN is in sleep mode, there is no query cost because no query is issued while the MN is in sleep.

Thus,

$$C_{i,query} = \begin{cases} 0 & \text{if Mark(Sleep)} \\ \gamma\tau + (\gamma\tau + \beta\tau + F(\text{Mark(Xs)})\tau) \times n_D \times P_{miss,j} & \text{elseif Mark(Just Wake Up)} > 0 \\ (\gamma\tau + \beta\tau + F(\text{Mark(Xs)})\tau) \times n_D \times P_{miss,j} & \text{otherwise.} \end{cases}$$

Let $C_{i,mobility}$ be the communication cost to service a location handoff given that the MN is in state $i$. $C_{mobility}$ is calculated as a weighted average as follows:

$$C_{mobility} = \sum_i (P_i \times C_{i,mobility}) \tag{6.10}$$

If in state $i$, $Mark(Xs) < K$, then the MN will only inform the proxy of the CoA address change. On the other hand, if $Mark(Xs) = K$, then the location handoff also triggers a service handoff. A service handoff will incur a context transfer cost of $ProxyCache$ while moving the proxy to the new service area and a communication cost to inform the HA and $N$ CNs (or application servers) of the CoA address change of the proxy. When the MN is in the sleep state, there is no location handoff cost since the MN is disconnected from the system. When the MN just wakes up, it will look for its proxy. If the proxy is in the same subnet as MN currently resides because the MN does not move during sleep, the cost involved is only for contacting the current AR for the proxy location. Otherwise, the proxy is moved to the current subnet, and the cost is that of a service handoff, i.e., a context transfer cost and a cost to inform the HA and $N$ CNs of the CoA address change of the proxy.

Therefore,

$$C_{i,mobility} = \begin{cases} 0 & \text{if Mark(Sleep)} > 0 \\ \gamma\tau & \text{else if Mark(Just Wake Up)} > 0 \text{ and Mark(Xs)} = 0 \\ \gamma\tau + \alpha\tau + N\beta\tau + F(\text{Mark(Xs)})n_{CT}\tau & \text{else if Mark(Just Wake Up)} > 0 \text{ and Mark(Xs)} > 0 \\ \gamma\tau + F(\text{Mark(Xs)})\tau & \text{else if } Mark(Xs) < K \\ \gamma\tau + \alpha\tau + N\beta\tau + F(K)n_{CT}\tau & \text{else if } Mark(Xs) = K. \end{cases}$$

Lastly, let $C_{i,invalidation}$ be the cost to forward an invalidation report from the proxy to the MN when the MN is in state $i$. Similarly, the weighted cost per invalidation report is calculated as:

$$C_{invalidation} = \sum_i (P_i \times C_{i,invalidation}) \tag{6.11}$$

When an update occurs, the CN sends an invalidation report to the MN. If the MN is in sleep or just wake-up mode, then the invalidation report is buffered in the proxy, so the cost is from the CN to the proxy. Otherwise the cost is from the CN through the proxy to the MN. Thus,

$$C_{i,invalidation} = \begin{cases} \beta\tau & \text{if Mark(Sleep)} > 0 \text{ or Mark(Just Wake Up)} > 0 \\ \beta\tau + F(\text{Mark(Xs)})\tau + \gamma\tau & otherwise \end{cases}$$

Summarizing above, Equation 6.3 along with other formulas derived in this section allows one to calculate the network traffic incurred per time unit for a MN characterized by a set of parameter values.

## 6.4 Numerical Results

In our proposed scheme, a MN and its proxy would apply Equations 6.3, 6.8, 6.10, and 6.11 to calculate $C_{total}$ as a function of $K$ and determine the optimal $K$ representing the optimal *service area* size that will minimize the network signaling cost. Below we present numerical results.

To provide a better sense of the performance improvement of our proposed proxy-based scheme for integrated cache and mobility management, we compare our scheme with three schemes: a *no-proxy no-caching* (NPNC) scheme, a *proxy no-caching* (PNC) scheme, and a *no-proxy caching* (NPC) management scheme. The NPNC scheme essentially is the basic MIPv6 scheme without using a proxy for either mobility or cache management. The PNC scheme is the proxy-based regional registration scheme using a proxy for mobility management. The proxy serves as a gateway between the mobile node (MN) and all services engaged by the MN [8]. In these two schemes, the MN does not cache data objects. The NPC scheme uses basic MIPv6 for mobility management and cached data objects maintained by the MN for cache management, but there is no proxy being used.

The default parameter values used in our performance study are given in Table 6.3. These are typical values for practical mobile applications. We analyze other parameters by varying their values and observing their effects on $C_{total}$ and $K_{opt}$. Without loss of generality, we consider the case that $\lambda_{q,i}$ is the same for all $N_{data}$ objects, and that $\mu_i$ is the same for all

| | |
|---|---|
| $\alpha$ | 30 |
| $\beta$ | 30 |
| $\gamma$ | 5 |
| $\tau$ | 0.025 |
| $N$ | 1 |
| $N_{data}$ | 10 |
| $n_{CT}$ | 2 |
| $n_D$ | 1 |
| $\lambda_{q,i}$ | in the range 1/100, 1/50, 1/10, 1/5, 1 |
| $\sigma$ | in the range 0.005, 0.01, 0.05, 0.1, 0.15 |
| $\mu_i$ | in the range 1/500, 1/100, 1/50, 1/10 |
| $\omega_w/\omega_s$ | in the range 1/16, 1/8, 1/4, 1/2, 1.0, 2.0, 4.0, 8.0 |

Table 6.3: Parameters Values in the Case Study in Integrated Cache and Mobility Management.

cached data objects. The numerical data reported are obtained based on Equations 6.3, 6.8, 6.10, and 6.11.

For ease of disposition, we normalize $C_{total}$ with respect to $\tau = 1$ (that is, the per-hop cost is set to 1), so the cost corresponds to the number of hops that packets need to travel per sec. Note that the cost is on a per-sec and per-MN basis so the cumulative cost saved for all MNs over time is significant. The goals of the numerical analysis are to (a) show that there exists an optimal service area under our cache scheme for network traffic cost minimization in Mobile IPv6 systems; (b) compare our integrated cache and mobility management scheme with existing schemes to demonstrate its benefit and identify conditions under which our scheme performs the best; and (c) study the effect of model parameters, including the query arrival rate $\lambda_{q,i}$, mobility rate $\sigma$, sleep pattern $\omega_w/\omega_s$ and data update rate $\mu_i$, on the optimal service area size.

### 6.4.1 Optimal Service Area

Figure 6.3 shows that under our integrated scheme there exists an optimal proxy service area size $K_{opt}$ to minimize the overall network traffic cost, when given a set of parameter values characterizing the mobility and service behaviors of the MN in Mobile IP networks.

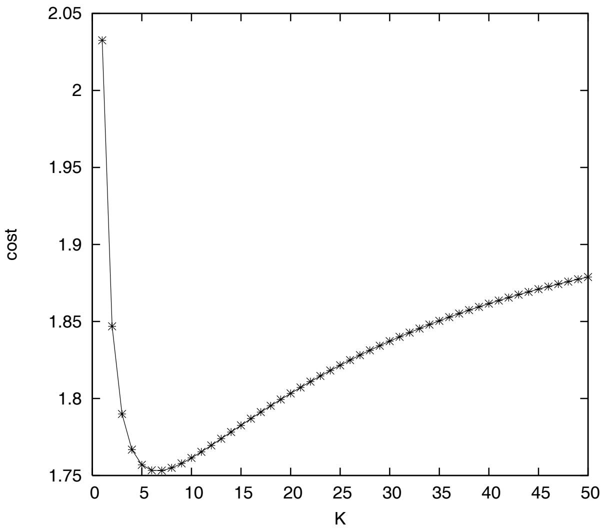

Figure 6.3: Cost vs. $K$ in Integrated Cache and Mobility Management.

We observe from Figure 6.4 that $K_{opt}$ exists for all $\lambda_{q,i}$ values. We see that $K_{opt}$ decreases slowly as $\lambda_{q,i}$ increases. The reason is that as $\lambda_{q,i}$ increases, the query cost increases, and subsequently the MN prefers a small service area size to reduce the query cost.

We observe from Figure 6.5 that $K_{opt}$ increases as $\sigma$ increases. The reason is that when the mobility rate is high, the mobility management cost is also high. Therefore, the proxy likes to stay at a large service area to reduce the location handoff cost such that a location handoff will most likely only involve informing the proxy of the location change without incurring a service handoff to migrate the proxy. As a result, when $\sigma$ increases, $K_{opt}$ increases.

We observe from Figure 6.6 that $K_{opt}$ decreases as $\mu_i$ increases. The reasons is that as $\mu_i$ increases, the data in the local cache are more likely to be out-of-date. The MN in this case will more likely to send queries to the server to obtain the latest version of data through the proxy. Also more invalidation reports will be sent from the CN to the MN through the proxy. Therefore, the MN will stay close to the proxy to reduce the triangular CN-proxy-MN communication cost.

Figure 6.7 shows the relationship between $K_{opt}$ vs. the sleep ratio. We observe that $K_{opt}$ decreases as the MN sleeps longer. The reason is that the data in MN's local cache are more

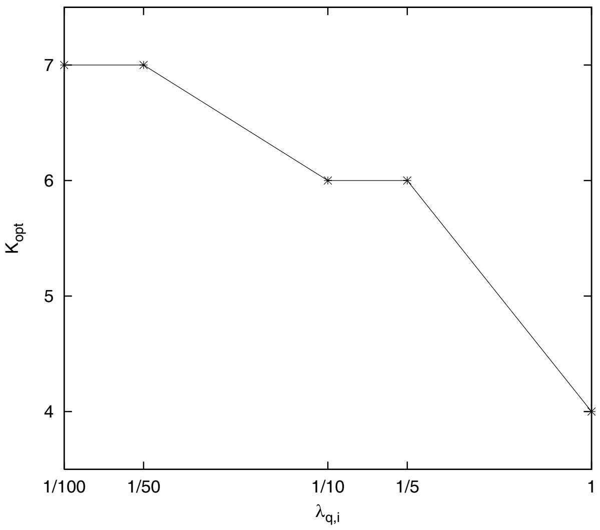

Figure 6.4: $K_{opt}$ vs. $\lambda_{q,i}$ in Integrated Cache and Mobility Management

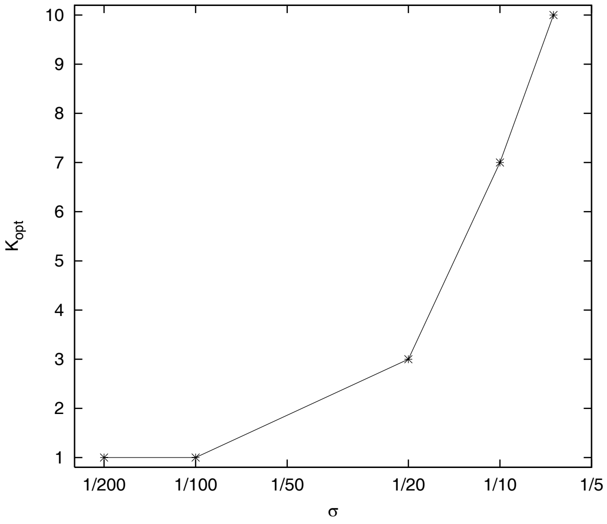

Figure 6.5: $K_{opt}$ vs. $\sigma$ in Integrated Cache and Mobility Management.

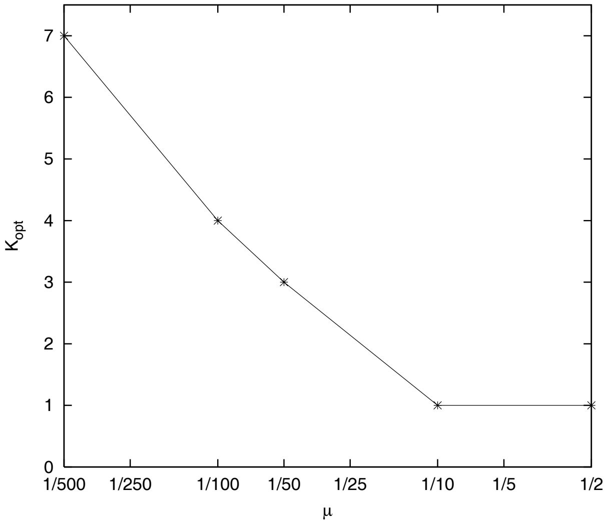

**Figure 6.6:** $K_{opt}$ vs. $\mu_i$ in Integrated Cache and Mobility Management.

likely to be out-of-date when the MN sleeps longer. Consequently, the MN will stay close to the proxy to reduce the triangular CN-proxy-MN communication cost for sending the query to and receiving replies from the server.

Figure 6.8 shows the relationship between $K_{opt}$ vs. cache size. We observe that $K_{opt}$ decreases as the number of cached data objects increases. The reason is that as the number of cached data objects increases, more invalidation reports will be sent from the CN to the MN, given the same update rate for all data objects. In order to reduce the triangular CN-proxy-MN cost for routing invalidation reports, the MN tends to stay closer to the proxy.

### 6.4.2 Performance Comparison

We compare our proxy-based integrated cache and mobility management (PICMM) scheme with the no-proxy caching (NPC), proxy no-caching (PNC) and no-proxy no-caching (NPNC) management schemes. We first derive the total cost incurred *per time unit* for these schemes.

Let $C_{total\_NPC}$ be the total cost incurred *per time unit* under the no-proxy caching scheme

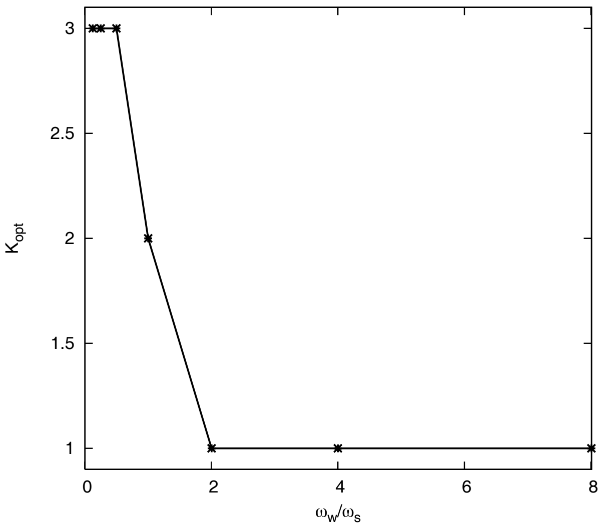

Figure 6.7: $K_{opt}$ vs. $\omega_w/\omega_s$ in Integrated Cache and Mobility Management.

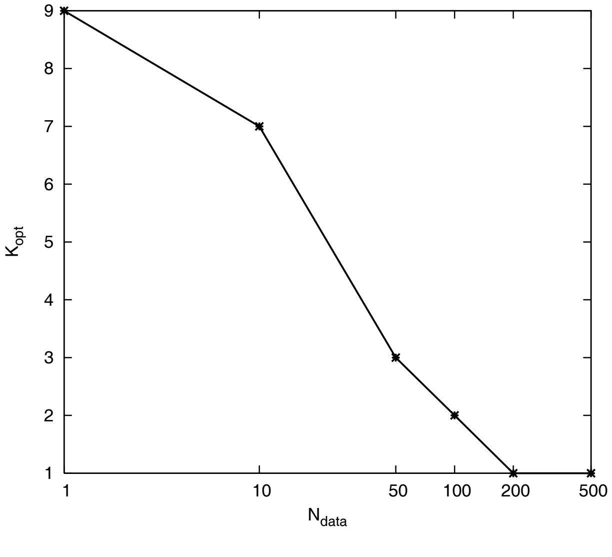

Figure 6.8: $K_{opt}$ vs. Cache Size $N_{data}$ in Integrated Cache and Mobility Management.

in Mobile IP networks. Similar to our integrated scheme, $C_{total\_NPC}$ can be computed by:

$$C_{total\_NPC} = \lambda_e \times C_{query\_NPC} + \sigma_e \times C_{mobility\_NPC} + \mu_e \times C_{invalidation\_NPC} \quad (6.12)$$

where $\lambda_e$, $\sigma_e$ and $\mu_e$ are defined as before by Equations 6.4, 6.6 and 6.7, respectively. For the query cost, the MN will check with the CN when there is a cache miss. Thus,

$$C_{query\_NPC} = (\beta\tau + \gamma\tau) \times P_{miss,j}$$

For the mobility cost, when the MN triggers a location handoff, it will inform the HA and CNs. Thus,

$$C_{mobility\_NPC} = \alpha\tau + N\beta\tau + \gamma\tau$$

For the invalidation cost, the CN will always inform the MN whenever there is an update. Thus,

$$C_{invalidation\_NPC} = \beta\tau + \gamma\tau$$

For the proxy no-caching (PNC) scheme, let $C_{total\_PNC}$ be the total cost incurred *per time unit*. Then, $C_{total\_PNC}$ can be computed by:

$$C_{total\_PNC} = \lambda_e \times C_{query\_PNC} + \sigma_e \times C_{mobility} \quad (6.13)$$

where $C_{mobility}$ is computed based on Equation 6.10 as before, and $C_{query\_PNC}$ is computed much the same way as $C_{query}$ is computed in Equation 6.8 except that if the MN is awake, the MN always sends the query to the CN through its proxy since the MN does not cache data objects. Thus,

$$C_{i,query\_PNC} = \begin{cases} 0 & \text{if Mark(Sleep) or Mark(Just Wake Up)} > 0 \\ \gamma\tau + \beta\tau + F(\text{Mark(Xs)})\tau & \text{otherwise.} \end{cases}$$

Note that there is no cache invalidation cost under PNC since there is no cache used.

Finally for the no-proxy no-caching scheme (i.e., basic MIPv6), let $C_{total\_NPNC}$ be the total cost incurred *per time unit*. Then $C_{total\_NPNC}$ can be computed as:

$$C_{total\_NPNC} = \lambda_e \times C_{query\_NPNC} + \sigma_e \times C_{mobility\_NPNC} \quad (6.14)$$

where $C_{query\_NPNC}$ is the cost from the MN to the CN, i.e., $\beta\tau + \gamma\tau$ on average, and $C_{mobility\_NPNC}$ is the cost to inform the HA and CNs, computed as $\alpha\tau + N\beta\tau + \gamma\tau$ on average.

Figure 6.9 compares our proposed PICMM scheme vs. NPNC, PNC and NPC management schemes in the network traffic cost generated, as a function of $\lambda_{q,i}$ with all other parameters fixed. For the PNC scheme, the cost shown is the minimized network traffic generated at the optimal service area. From Figure 6.9 we see that caching based schemes (NPC and PICMM) achieve much better performance for non-caching based schemes (NPNC and PNC), especially when $\lambda_{q,i}$ is large. The cost of either NPC or PICMM increases slowly as $\lambda_{q,i}$ increases, while the cost for either PNC or NPNC increases drastically as $\lambda_{q,i}$ increases. The reason is that at high $\lambda_{q,i}$ values, the dominating network traffic is due to sending queries to the server. Caching-based schemes can greatly reduce the magnitude of server-querying network traffic because caching allows some of the queries to be processed by the MN based on up-to-date cached data objects. This trend is true when the update rate to cached data objects ($\mu_i$) is low to modest so that the cost saving due to query processing for sending queries to the server outweighs the extra cost due to triangular routing for receiving invalidation reports and query replies. On the other hand, between the two caching-based schemes, our PICMM scheme performs consistently better than NPC due to the use of a proxy for integrated cache and mobility management for extra cost saving.

Figure 6.10 compares our proposed PICMM scheme vs. NPNC, PNC and NPC management schemes in the generated network traffic, as a function of $\sigma$ with all other parameters fixed. Figure 6.10 shows that our proposed PICMM scheme outperforms all other schemes. Compared with the NPNC scheme, the performance of our scheme is especially pronounced when $\sigma$ is high. The reason is that PICMM uses a proxy to serve as a GFA to reduce the cost of location handoffs and the benefit is especially pronounced when the mobility rate of the MN is high. Compared with the NPC scheme, PICMM considers integrated cache management and mobility management. Consequently, the best service area determined can minimize the overall network traffic cost better than that by NPC which runs mobility

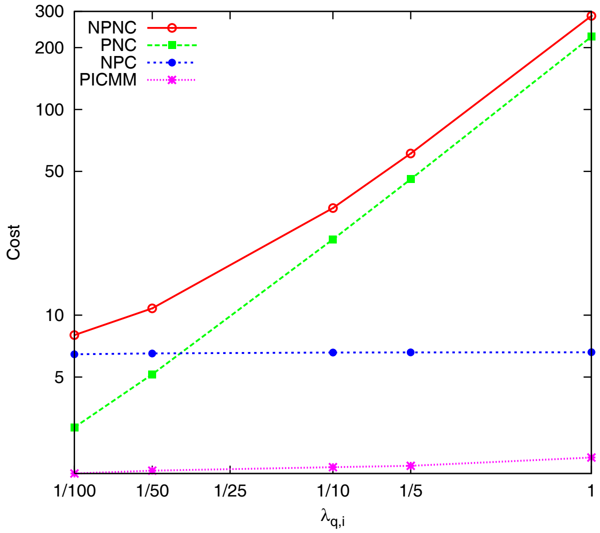

**Figure 6.9:** Generated Network Traffic as a Function of $\lambda_{q,i}$ in Integrated Cache and Mobility Management.

management independently of cache management.

In Figure 6.11 we compare PICMM vs. other schemes in the generated network traffic, as a function of the cache size. The total cost incurred under PICMM increases as the number of cached data items increases. The main reason is that the invalidation cost increases as the number of cached data items increases. Compared with non-caching based schemes (NPNC and PNC), PICMM achieves much better performance especially when $N_{data}$ is large because caching saves much of the uplink cost for query processing.

Figures 6.9, 6.10, 6.11 have shown that our proposed PICMM scheme performs better than NPNC, PNC and NPC, particularly as the query rate ($\lambda_{q,i}$), MN mobility rate ($\sigma$), or the cache size ($N_{data}$) increases. Below we reveal conditions under which PICMM could perform worse than non-caching based schemes (NPNC and PNC).

In Figure 6.12 we compare PICMM vs. NPNC, PNC and NPC management schemes in the network traffic generated, as a function of $\mu_i$ with all other parameters fixed. Under a non-caching based scheme (NPNC or PNC), the MN will always send queries to the server for processing, since the MN does not cache data objects. As a result, the generated network traffic is insensitive to $\mu_i$. On the other hand, when caching is used as in PICMM or NPC,

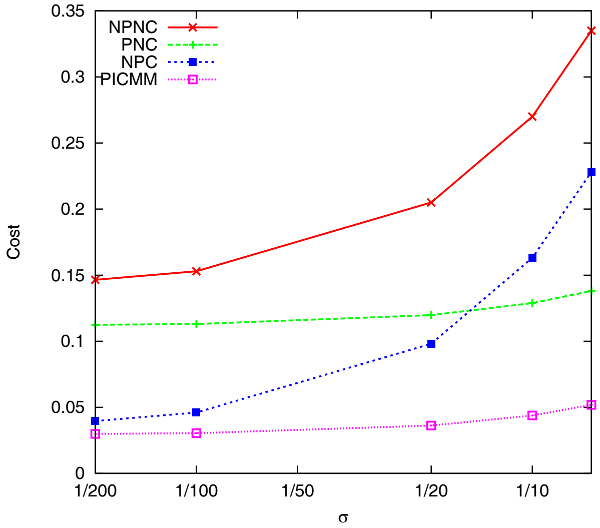

**Figure 6.10:** Generated Network Traffic as a Function of $\sigma$ in Integrated Cache and Mobility Management.

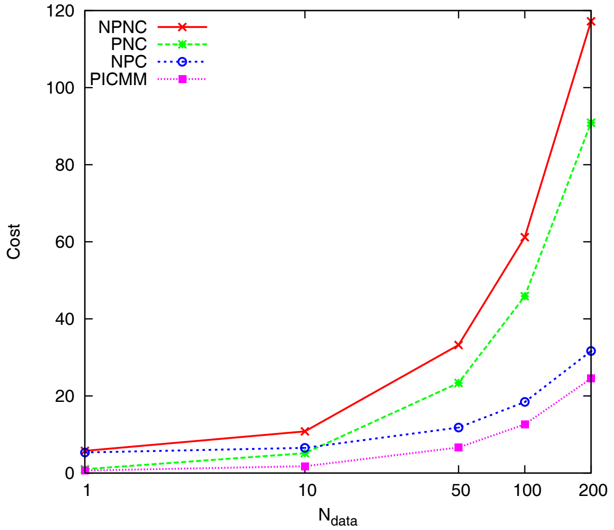

Figure 6.11: Generated Network Traffic as a Function of Cache Size in Integrated Cache and Mobility Management.

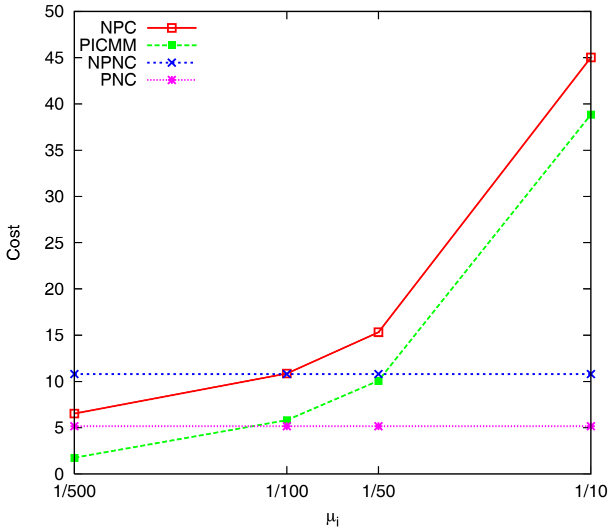

**Figure 6.12:** Generated Network Traffic as a Function of $\mu_i$ in Integrated Cache and Mobility Management.

the generated network traffic becomes sensitive to $\mu_i$ because the query traffic to the server depends on if cached data objects are valid. We see that PICMM consistently performs better than NPC. However, when the data update rate is very high, most cached data objects are invalid, so queries will need to be routed to the CN. In this case, there exists a cross-over point in the update rate beyond which PICMM would perform worse than a non-caching scheme because of the CN-proxy-MN triangular cost for routing query inquiries/replies and invalidation reports. In general in cases data are being updated frequently, a caching scheme would not perform better than a non-caching scheme, even with the use of a smart proxy as in PICMM for optimized, integrated cache and mobility management. In this extreme case, the system is better off with the proxy no-caching scheme (PNC).

In Figure 6.13 we compare PICMM vs. NPC, PNC and NPNC in the generated network traffic, as a function of the sleep ratio $\omega_w/\omega_s$. When the sleep ratio $\omega_w/\omega_s$ is extremely large, MN is mostly sleeping all the time. The cache is mostly invalid due to long sleep. In this extreme condition, PICMM incurs a higher query cost than non-caching schemes although the cost is small since $P_{wake}$ is small. Again we see that there is a cross-over point in the sleep ratio beyond which PICMM would perform worse than a non-caching scheme such as

PNC because of the CN-proxy-MN triangular cost for routing query inquiries/replies and invalidation reports under PICMM. We see that, however, under a reasonable range of sleep ratio ($< 2$), PICMM outperforms all other schemes.

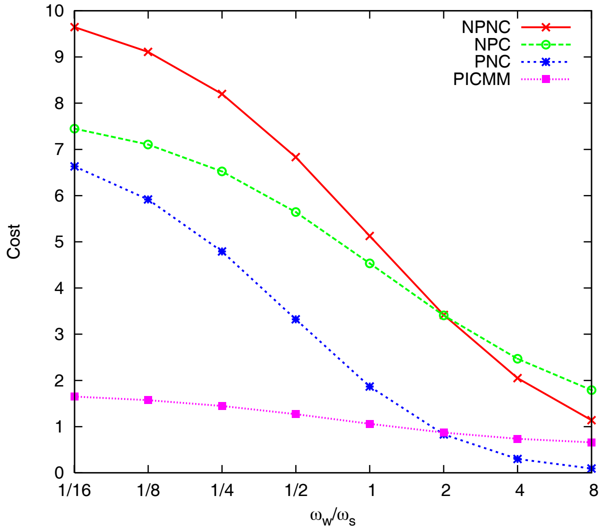

Figure 6.13: Generated Network Traffic as a Function of $\omega_w/\omega_s$ in Integrated Cache and Mobility Management.

In summary, we observe that our proposed PICMM scheme outperforms NPC, PNC and NPNC in terms of the network traffic cost generated over a wide range of parameter values, the effect of which is especially pronounced when the data update rate is low, the mobility rate is high, the query rate is high, the cache size is large, and/or the sleep ratio is low. PICMM greatly reduces the network traffic of mobile query-based applications, except under extreme conditions in which data update rates are exceptionally high and the MN disconnects more than 2/3 of the time during the execution of mobile applications. These extreme conditions, however, rarely happen in practical mobile applications.

PICMM represents a specific implementation of IMSA-MIPv6 to support mobile data query applications. Other than performing simulation validation of analytical results reported in this chapter, we plan to investigate conditions under which it will be more beneficial for the CN to send copies of objects instead of invalidation reports to the MN for cache management. Finally, we also plan to investigate other mobile applications to which IMSA is applicable and can outperform existing schemes.

# Chapter 7

# Completed, To Be Completed, and Future Work

## 7.1 Completed Work

So far we have completed design and analysis of several algorithms for integrated mobility and service management for future all-IP based wireless networks. These algorithms include DMAP based on routing-only ARs, IMSA based on powerful ARs that can support mobile agents, and, finally PICMM based on IMSA for supporting mobile data query application with cache management capability. The completed work has resulted in the following publications:

### Conferences

- I.R. Chen, *W. He*, and B. Gu, DMAP: A Scalable and Efficient Integrated Mobility and Service Management Scheme for Mobile IPv6 Systems, *2nd IEEE International Workshop on Performance Management of Wireless and Mobile Networks,* Tampa, FL, November 2006 [8].

### Submitted papers

- I.R. Chen, *W. He*, and B. Gu, Proxy-based Regional Registration for Integrated Mobility and Service Management in Mobile IP Systems, submitted to *The Computer Journal*, 2006.

- I.R. Chen, *W. He*, and B. Gu, DMAP: Integrated Mobility and Service Management in Mobile IPv6 Systems, submitted to the *International Journal in Wireless Personal Communications*, 2006.

- *W. He*, I.R. Chen, and B. Gu. A Proxy-Based Integrated Cache Consistency and Mobility Management Scheme for Mobile IP Systems, submitted to the *IEEE 21st International Conference on Advanced Information Networking and Applications*, May 2007.

## 7.2 To be Completed and Future Work

There are several work items remained to be done in the dissertation research.

- We plan to evaluate our DMAP, IMSA and PICMM designs by simulation with ns-2 [13]. We choose ns-2 because it is a discrete-event simulator designed to target networking research. It provides substantial supports for simulation of routing protocols over wired and wireless networks. The simulation study will first setup a number of ARs, each situated in an IP subnet in a topology randomly generated. The locations of the HA and CNs for a MN will be randomly determined with the MN and CNs exchanging data packets. The mobility path of a MN will be randomly generated through these ARs. Upon a handoff the MN will determine the optimal regional registration area based on the computational procedure developed. It would determine if a local regional registration is needed, or if a change of the regional register is required. In the latter case, the AR of the subnet that the MN roams into will be designated as the regional register. The overall network-signaling overhead and packet delivery cost by the MN would be collected as a metric to measure the system performance. This will be compared against analytical data for validation. The HMIPv6 design as well as the basic MIPv6 will also be simulated and their results will be compared against those obtained from DMAP and IMSA-MIPv6 for simulation validation.

- We plan to extend DMAP and IMSA based on two-level regional registration to more than two levels as in HMIPv6 and identify the optimal level to use to maximize the system performance in terms of cost incurred per time unit due to mobility handoff and packet delivery. The simulation will be extended to cover the new design.

- We plan to investigate context-aware database applications that can benefit from knowledge of the MN's location and service context. For each application, we plan to investigate what context information a MN's proxy can carry and/or the MN can collect during runtime to improve the system performance. We plan to obtain quantitative data from modeling and analysis, as well as from simulation to demonstrate that such applications can benefit from our DMAP or IMSA designs.

- We plan to add fault tolerance and recovery into DMAP, IMSA and PICMM designs such that they are tolerant of faults of ARs and MNs. In the case of IMSA and PICMM, they are also tolerant of failures of user proxies. To tolerate failures of ARs, for example, once the proxy detects a failure in an AR, another AR can be dynamically selected to backup the faulty AR. The proxy then can initiate a handoff to move the MNs being served by the faulty AR to the service areas of the backup AR. The data querying capability for the MNs can be continuously supported by backup failure-free ARs.

- Lastly, for the experiment with mobile database query application and the design of PICMM, we plan to investigate conditions under which it is more beneficial for the server application to forward a copy of the data object instead of an invalidation report to the MN whenever a data object cached by the MN is updated. The condition may involve the data object write and read frequencies, the MN's connection/disconnection pattern, distance separating the MN from the server application, and/or the sizes of data objects and invalidation reports. Our objective is to support the hypothesis that our research is particularly beneficial to mobile database services in the next-generation mobile Internet.

## 7.3 Schedule and Milestones

Table 7.1 presents the schedule for completing my PhD dissertation. I will perform proposed research activities using a cyclic process which alternates between critical reflection, algorithm design, modeling and analysis, and experiment. In the later cycles, I will be continuously refining the proposed algorithms based on lessons learned in the earlier cycles. I am planning to submit several papers based on our research to well-known international conferences and journals. The anticipated dates for my Research and Final defenses are Dec 2007 and May 2008, respectively. My PhD work will be carried out in the Systems and Software Engineering Lab at the Department of Computer Science, Virginia Tech, Northern Virginia Center.

| ***Deadline*** | ***Research Activity*** |
|---|---|
| January 2005 | Surveyed and analyzed existing mobility/service management in wireless networks. |
| May 2005 | Studied characteristics of future all-IP based wireless networks. |
| September 2005 | Proposed and analyzed IMSA-MIPv4: proxy-based integrated mobility and service management scheme in Mobile IPv4 systems |
| January 2006 | Proposed and analyzed IMSA-MIPv6: proxy-based integrated mobility and service management scheme in Mobile IPv6 systems |
| April 2006 | Proposed and analyzed DMAP for the case in which ARs can perform only network-layer functions: DMAP |
| November 2006 | Proposed and analyzed PICMM for the case in which ARs can perform application-layer functions allowing proxies to carry service context information regarding cached data objects for mobile database applications. |
| December 2006 | Preliminary Defense |
| March 2007 | Completion of design and analysis of hierarchical DMAP and IMSA for MIPv6 |
| June 2007 | Completion of design and analysis of fault tolerance and recovery of DMAP and IMSA for MIPv6 |
| September 2007 | Completion of extension to PICMM to deal with the case that data objects are forwarded to the MN instead of invalidation reports for mobile database query applications |
| Nov. 2007 | Completion of the applicability study by identifying context-aware database applications that can benefit from DMAP and IMSA designs; completion of algorithm design and analysis for such applications identified. |
| December 2007 | Research Defense |
| March 2008 | Completion of simulation studies based on ns-2 to validate analytical results as well as to compare DMAP, IMSA and PICMM and new algorithms extended against basic MIPv6, MIP-RR, HMIPv6 and other existing algorithms in MIPv6 for mobility and service management. |
| May 2008 | Final Defense. |

Table 7.1: PhD Work Schedule and Milestone.

# Bibliography

[1] F. Adelstein, S. K. Gupta, G. R. III, and L. Schwiebert, *Fundamentals of Mobile and Pervasive Computing.* McGraw Hill, 2005.

[2] I. Akyildiz, J. McNair, J. Ho, H. Uzunalioglu, and W. Wang, "Mobility management in next-generation wireless systems," *Proceedings of the IEEE*, vol. 87, no. 8, pp. 1347–1384, 1999.

[3] S. Ardon, P. Gunningberg, B. LandFeldt, Y. Ismailov, M. Portmann, and A. Seneviratne, "March: a distributed content adaptation architecture," *International Journal of Communication Systems*, vol. 16, no. 1, pp. 97–115, 2003.

[4] Y. Bejerano and I. Cidon, "An efficient mobility management strategy for personal communication systems," in *The Fourth Annual ACM/IEEE International Conference on Mobile Computing and Networking*, October 1998, pp. 215 – 222.

[5] C. Castelluccia, "Extending mobile ip with adaptive individual paging: a performance analysis," in *IEEE International Symposium on Computers and Communications*, 2000, pp. 113–118.

[6] I. R. Chen, T. Chen, and C. Lee, "Agent-based forwarding strategies for reducing location management cost in mobile networks," *Mobile Networks and Applications*, vol. 6, no. 2, pp. 105–115, 2001.

[7] I. R. Chen, B. Gu, and S. Cheng, “On integrated location and service handoff schemes for reducing network cost in personal communication systems,” *IEEE Transactions on Mobile Computing*, vol. 5, no. 2, pp. 179–192, 2006.

[8] I. R. Chen, W. He, and B. Gu, “Dmap: A scalable and efficient integrated mobility and service management scheme for mobile ipv6 systems,” in *2nd IEEE International Workshop on Performance Management of Wireless and Mobile Networks*, 2006.

[9] I. R. Chen, N. Phan, and I. Yen, “Algorithms for supporting disconnected write operations for wireless web access in mobile client-server environments,” *IEEE Transactions on Mobile Computing*, vol. 1, no. 1, pp. 46–58, 2002.

[10] I. R. Chen and N. Verma, “Simulation study of a class of autonomous host-centric mobility prediction algorithms for cellular and ad hoc networks,” in *36th Annual Simulation Symposium*, Orlando, USA, 2003, pp. 65– 72.

[11] I. R. Chen, O. Yilmaz, and I. Yen, “Admission control algorithms for revenue optimization with qos guarantees in mobile wireless networks,” *Wireless Personal Communications*, vol. 38, no. 3, pp. 357–376, 2006.

[12] F. Chiussi, D. Khotimsky, and S. Krishnan, “Mobility management in third-generation all-ip networks,” *IEEE Communications Magazine*, vol. 40, no. 9, pp. 124– 135, 2002.

[13] A. collaboratoin between researchers at UC Berkeley, LBL, USC/ISI, and X. PARC., *The Network Simulator - ns-2*, http://www.isi.edu/nsnam/ns/.

[14] L. Dimopoulou, G. Leoleis, and I. Venieris, “Introducing a hybrid fast and hierarchical MIPv6 scheme in a UMTS-IP converged architecture,” *29th Annual IEEE International Conference on Local Computer Networks*, pp. 42–49, 2004.

[15] M. Endler, D. Silva, and K. Okuda, “RDP: A result delivery protocol for mobile computing,” in *ICDCS Workshop on Wireless Networks and Mobile Computing*, 2000, pp. D36–D43.

[16] J. Gomez, C.-Y. Wan, S. Kim, Z. Turanyi, and A. Valko, *Cellular IP*, http://www.ietf.org/proceedings/00jul/I-D/mobileip-cellularip-00.txt.

[17] 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)*, pp. 453–463, 2004.

[18] E. Gustafsson, A. Jonsson, and C. Perkins, *Mobile IPv4 Regional Registration*, http://www.ietf.org/internet-drafts/draft-ietf-mip4-reg-tun nel-02.txt, IETF, Work in Progress, 2006.

[19] M. Handley, H. Schulzrinne, E. Schooler, and J. Rosenberg, *SIP: Session Initiation Protocol*, http://www.ietf.org/rfc/rfc2543.txt.

[20] W. He and A. Bouguettaya, "Using hashing and caching for location management in wireless mobile systems," in *MDM '03: Proceedings of the 4th International Conference on Mobile Data Management*, 2003, pp. 335–339.

[21] J. Ho and I. Akyildiz, "Dynamic hierarchical database architecture for location management in pcs networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 5, no. 5, pp. 646 – 660, 1997.

[22] Y. Huh and C. Kim, "New caching-based location management scheme in personal communication systems," in *Information Networking*, 2001, p. 649.

[23] IEEE, *IEEE Trial-Use Recommended Practice for Multi-Vendor Access Point Interoperability via an Inter-Access Point Protocol Across Distribution Systems Supporting IEEE 802.11 Operation*, http://standards.ieee.org/getieee802/download/802.11F-2003.pdf, 2003.

[24] R. Jain and N. Krishnakumar, "Network support for personal information services to pcs users," in *Proceedings of IEEE Conf. Networks for Personal Comm. (NPC)*, 1994, pp. 1–7.

[25] D. Johnson, C. Perkins, and J. Arkko, *Mobility Support in IPv6*, http://www.ietf.org/rfc/rfc3775.txt, IETF, Work in Progress, 2004.

[26] A. Joshi, "On proxy agents, mobility, and web access," *Mobile Networks and Applications*, vol. 5, no. 4, pp. 233–241, 2000.

[27] C. Lee, C. Ke, and C. Chen, "Improving location management for mobile users with frequently visited locations," Performance Evaluation 43, pp. 15–38, January 2001.

[28] D. A. Maltz and P. Bhagwat, "MSOCKS: An architecture for transport layer mobility," in *INFOCOM (3)*, 1998, pp. 1037–1045.

[29] P. Maniatis, M. Roussopoulos, E. Swierk, K. Lai, G. Appenzeller, X. Zhao, and M. Baker, "The mobile people architecture," *ACM Mobile Computing and Communications Review*, vol. 3, no. 3, pp. 36–42, 1999.

[30] J. Manner and M. Kojo, *Mobility Related Terminology*, http://www.faqs.org/rfcs/rfc3753.html.

[31] H. D. Meer, A. Corte, A. Puliafito, and O. Tomarchio, "Programmable agents for flexible qos management in ip networks," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 2, pp. 256–267, 2000.

[32] H. Minh and H. As, "User profile replication with caching for distributed location management in mobile communication networks," in *Proceedings of the 16th ACM SAC2001 Symposium on Applied Computing*, March 2001, pp. 381 – 386.

[33] A. Misra, S. Das, A. Dutta, A. McAuley, and S. K. Das, "Idmp-based fast handoffs and paging in IP-based cellular networks," *International Conference on 3G Wireless and Beyond*, pp. 427–432, 2001.

[34] A. Misra, S. Das, A. McAuley, A. Dutta, and S. K. Das, *IDMP: An Intra-Domain Mobility Management Protocol using Mobility Agents*, draft-mobileipmisra-idmp-00.txt, IETF, Work in Progress, 2000.

[35] S. Mtika and F. Takawira, “Mobile IPv6 regional mobility management,” in *ACM 4th international symposium on Information and communication technologies*, Cape Town, South Africa, 2005, pp. 93–98.

[36] J. Mysore and V. Bharghavan, “A new multicasting-based architecture for internet host mobility,” in *MobiCom ’97: Proceedings of the 3rd annual ACM/IEEE international conference on Mobile computing and networking.* New York, NY, USA: ACM Press, 1997, pp. 161–172.

[37] H. Omar, T. Saadawi, and M. Lee, “Supporting reduced location management overhead and fault tolerance in mobile-ip systems,” in *IEEE International Symposium on Computers and Communications*, 1999, pp. 347–353.

[38] S. Pack, Y. Choi, and M. Nam, “Design and analysis of optimal multi-level hierarchical Mobile IPv6 networks,” *Wireless Personal Communications*, vol. 36, no. 2, pp. 95–112, 2006.

[39] B. Patil, Y. Saifullah, L. Aravamudhan, S. Kularatna, S. Faccin, R. Mononen, S. Sreemanthula, S. Sharma, and L. Lyengar, *IP in Wireless Networks.* Prentice Hall Ptr (ISBN: 0130666483 ), January 2003.

[40] C. Perkins, *IP Mobility Support for IPv4, revised*, http://www.ietf.org/internet-drafts/draft-ietf-mip4-rfc3344bis-01.txt.

[41] Z. Petros, Z. Gary, C. Jerry, L. Haiyun, L. Songwu, and L. Jefferey Jia-Ru, “Dirac: a software-based wireless router system,” pp. 230 – 244, 2003, 939009230-244.

[42] E. Pitoura and G. Samaras, *Data Management for Mobile Computing.* Kluwer Academic Publishers, 1998, vol. 10.

[43] R. Prakash and M. Singhal, “Dynamic hashing + quorum = efficient location management for mobile computing systems,” in *Proceedings of the Sixteenth Annual ACM Symposium on Principles of Distributed Computing*, August 1997, p. 291.

[44] R. Ramjee, K. Varadhan, L. Salgarelli, S. R. Thuel, S. Wang, and T. Porta, "Hawaii a domain-based approach for supporting mobility in wide-area wireless networks," *IEEE/ACM Transactions on Networking*, vol. 10, no. 3, pp. 396–410, 2002.

[45] H. Schulzrinne and J. Rosenberg, "The session initiation protocol: Internet-centric signaling," *IEEE Communications Magazine*, vol. 38, no. 10, pp. 134 – 141, 2000.

[46] J. Scourias and T. Kunz, "An activity-based mobility model and location management simulation framework," in *Proceedings of the 2nd ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, August 1999, pp. 61 – 68.

[47] A. C. Snoeren and H. Balakrishnan, "An end-to-end approach to host mobility," in *MobiCom '00: Proceedings of the 6th annual international conference on Mobile computing and networking*. New York, NY, USA: ACM Press, 2000, pp. 155–166.

[48] H. Soliman, *Mobile IP the Internet Unplugged*. Prentice Hall PTR. (ISBN: 0138562466), January 1998.

[49] ——, *Mobile IPv6: Mobility in a Wireless Internet*. Addison-Wesley Professional. (ISBN: 0201788977), April 2004.

[50] H. Soliman, C. Castelluccia, K. El-Malki, and L. Bellier, "Hierarchical Mobile IPv6 mobility management," *http://www3.ietf.org/proceedings/03mar/I-D/draft-ietf-mobileip-hmipv6-07.txt, IETF, Work in Progress*, 2002.

[51] K. Tan, J. Cai, and B. C. Ooi, "An evaluation of cache invalidation strategies in wireless environments," *IEEE Transactions on Parallel and Distributed Systems*, vol. 12, no. 8, pp. 789–807, 2001.

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

[53] K. Wang and J. Huey, “A cost effective distributed location management strategy for wireless networks,” *Wireless Networks*, vol. 5, no. 4, pp. 287 – 297, 1999.

[54] Z. Wang, M. Kumar, S. Das, and H. Shen, “Dynamic cache consistency schemes for wireless cellular networks,” *IEEE Transactions on Wireless Communications*, vol. 5, pp. 366– 376, February 2006.

[55] E. Wedlund and H. Schulzrinne, “Mobility support using SIP,” in *WOWMOM*, 1999, pp. 76–82.

[56] D. Wisely, P. Eardley, and L. Burness, *IP for 3G: Networking Technologies for Mobile Communications.* John Wiley & Sons Ltd. (ISBN: B00008ZQ5F), June 2002.

[57] V. Wong and V. Leung, “Location management for next generation personal communication networks,” *IEEE Network*, vol. 14, no. 5, pp. 18–24, Sept./Oct. 2000.

[58] J. Xie and I. F. Akyildiz, “A novel distributed dynamic location management scheme for minimizing signaling costs in mobile ip,” *IEEE Transactions on Mobile Computing*, pp. 163–175, 2002.

[59] X. Zhang, J. Castellanos, and A. Campbell, “P–MIP: Paging extensions for Mobile IP,” *Mobile Networks and Applications*, vol. 7, no. 2, pp. 127–141, Mar. 2002.