NEAR EAST UNIVERSITY

Faculty of Engineering

Department of Electrical and Electronic Engineering

DATA MANAGEMENT FOR MOBILE COMPUTING (MANAGEMENT LOCATION)

Graduation Project
EE – 400

Student: Mahmoud Al-Shayeb(20020890)

Supervisor: MR. Jamal Fathi

Nicosia - 2006

ACKNOWLEDGMENT

First of all I am happy to complete the task which I had given with blessing of God and also I am grateful To whom I can not imagine myself to be without their guidance and continuous support, to my parents who are enduring these all expenses and supporting me in all events. I am nothing without their prayers.

I wish to thank my advisor, Dr. Jamal Fathi, for intellectual support, encouragement, and enthusiasm, which made this project possible, and his patience for correcting both my stylistic and scientific errors. And he delivered me too much information and did his best of efforts to make me able to complete my project.

My sincerest thanks must go to my friends, Ahmad Ibrahim, Mohammed Mowaswes, Khaled Gazali, Ezzat Natsheh , Ziad Abo Helal, Fadi Kuwaik, Rand Marzouq, Rama and Leen Tarabishi who shared their suggestions and evaluations throughout the completion of my project. The comments from these friends enabled me to present this project successfully.

And above, I thank God for giving me stamina and courage to achieve my objective

ABSTRACT

Location management is a key issue in personal communication service networks to guarantee the mobile terminals to continuously receive services when moving from one place to another. The purpose is to provide a thorough and cohesive overview of recent advances in wireless and mobile data management.

Mobile Computing is a new emerging computing paradigm of the future. Data Management in this paradigm poses many challenging problems to the database community.

CONTENTS

INTRODUCTION

The focus of Data Management for Mobile Computing is on the impact of mobile computing on data management beyond the networking level. Data Management for Mobile Computing provides a single source for researchers and practitioners who want to keep current on the latest innovations in the field.

It can also serve as a textbook for an advanced course on mobile computing or as a companion text for a variety of courses including courses on distributed systems, database management, transaction management, operating or file systems, information retrieval or dissemination, and web computing.

The aims of this project are to know what management location is, and to understand how it operate and to know its applications.

The first chapter provides an introduction to mobile wireless network system and its main subsystems.

Chapter two covers the management location in general, with some information about Replication and Placement of Databases.

Chapter three provides the Location Management For 3G Networks, and compares the location in a different 3G states like (DYNAMIC-3G and STATIC-3G).

Chapter four gives an idea about the applications for management location and why people need it and how they can use it.

Chapter five describes the mobility modeling and location tracking, and also gives some information and trajectory prediction in wireless ATM networks.

And finally, chapter six contains an overview of locating in cellular mobile internet, which also contains information about mobile IP protocol and addresses sliding window .

1. INTRODUCTION TO MOBILE WIRELESS NETWORK SYSTEM

1.1 Overview

When developing mobile wireless network systems (i.e., wireless networking algorithms, node architectures, and network infrastructures), the designer is presented with numerous design alternatives. There are numerous factors which can impact the analysis, performance and validation of these design alternatives. These Factors range from having to support different patterns of node mobility to integrating the traffic generators, networking algorithms, and operating system capabilities.

A few operating system kernels and languages have been designed to support wireless and mobile communication, and a number of protocols have been devised to solve the numerous topology setup and maintenance, media access control, and transmission problems in the mobile environment .Commercial radios designed to be hooked up with laptops for wireless multimedia transmissions are available in the market. Thus although solutions to different facets of the wireless mobile information system design are appearing, relatively little effort has been devoted to understanding the performance impact of the interactions among different components of the system.

Traditionally, analysis, simulation and measurement have all been used to evaluate the performance of network protocols and multimedia systems. Measurement-based approaches are useful only after the system has been deployed. Although they offer the most accurate evaluations of performance problems, they are often inadequate because it may be infeasible to modify the deployed system to experiment with a large range of design parameters. Even when such modifications are feasible, the cost of the necessary software and hardware modifications may be exorbitant. Analytical models offer the opportunity to quickly examine a large parameter space to identify efficient configurations; however for complex systems with many interacting components, analytical models may either be inaccurate or computationally intractable. For complex, heterogeneous systems, simulations are often the only realistic alter-native to performance prediction.

The primary drawback with detailed simulation models is that they are frequently slow. Experience with many existing network simulators has shown that a performance study of wireless proto cols for even small networks (tens of nodes) can take many days; running such simulations for networks involving a large number of mobile elements is clearly infeasible. Recent experience with parallel execution of models for personal communication systems has shown that parallelism offers significant potential to improve the execution time for these models; it is likely that these techniques can also be exploited to improve the execution time for simulation models of wireless networks. This paper describes such an environment.[1]

1.2 Mobile Wireless Systems

There are numerous ways to design and examine mobile wireless systems. In order to provide a common reference model to analyzing these systems, we break the system down into three integrated levels: network, node, and algorithm. The network level is used to describe the architecture of the network and details of each node such as its communication capability, location, and impact on the network. The node level is used to describe. the details of the node such as its hardware and software capabilities and interaction such as with the operating system and among algorithms. The algorithm level describes the details of a specific algorithm or layer of the protocol stack.

1.2.1 Mobile Wireless Networks

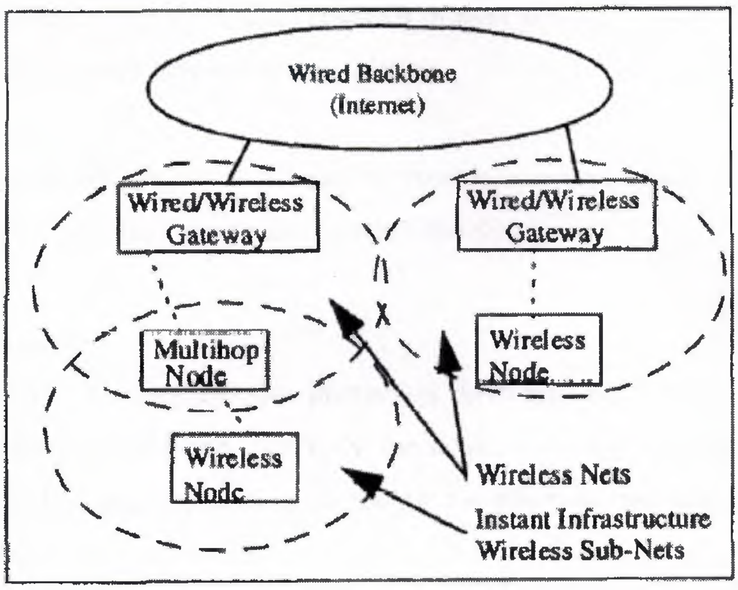


Figure 1.1 Mobile Wireless Networks

In figure.1 we see an example of a mobile wireless network. This network is composed of not only a static wired backbone and a few wireless cells, but also a set of nodes which am able to support instant infrastructure, and multi-hop packet radio networks. We include throughout this paper the study of instant infrastructure net- works, nodes and their algorithms since support for this architecture requires additional flexibility upon the simulation environment and illustrates the complex environment mobile wireless network systems can operate in.

The network nodes shown in figure 1 are comprised of numerous software components which can be used to support self-configuring, multi-hop, multimedia networking architectures and can be added to the capability of each node as shown in figure 2.

1.2.2 Mobile Wireless Nodes

The design of mobile wireless nodes/terminals has been studied by various groups. In this section we describe the components which make up the node architecture and the implementation of the network control functions, multimedia suppon communication substrates, and the interfaces between them. The node functionality, as shown in figure 2, is supported in the nodes being developed in the Wireless Adaptive Mobile Information System (WAMIS) research project at UCLA. These nodes are used as a test-bench For experimentation and validation.

In the following subsections, we will describe various components and algorithms which make up this typical instant infrastructure node's functionality.

1.2.2.1 Applications

Applications are needed for interaction between the system and the user. Multimedia support is necessary not only for acquisition and presentation of video, speech, and data but also for coding/ decoding for efficient transmission through the wireless network.

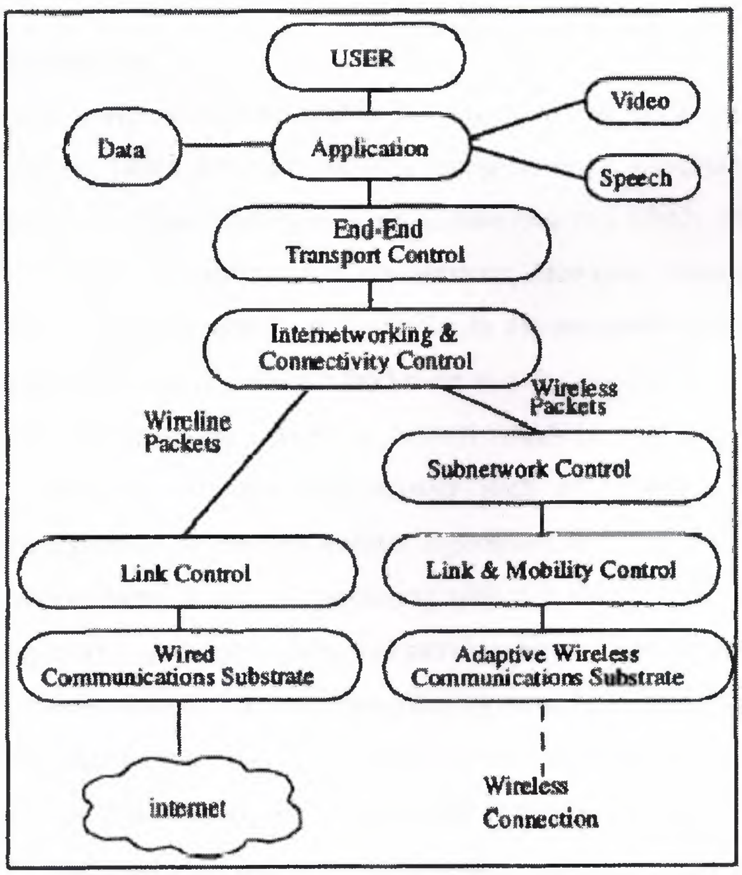


Figure 1.2 Node Functionality

The standard set of TCP/IP protocol suite applications support text based services like remote login or file transfers. New applications am now appearing which support multimedia (Netscape and video conferencing applications). In order to see the effect and demonstrate multimedia over mobile wireless networks, a video conferencing application was developed on the test-bench. This application (Video TALK) brings together video, which uses UDP, and data, which uses TCP, into a single application on the laptop. In order to test the performance of the system, testing tools were developed to measure throughput, delay, packet loss, and track adaptive parameters in the communication device (radio) such as code, power, and spreading factor (i.e. chips/bit). A topology analyzer program (TOPO) was developed which can be used in the simulation environment or in the implemented system to graphically analyze the virtual topology of the wireless multi-hop s&net..

1.2.2.2 Operating System

The operating system is responsible for integrating all these net- work control components together. There are numerous operating systems available today such as Microsoft Windows, PC-Disk Operating System, Mac OS, and UNIX which can have a big impact on the node's capabilities and performance. However, these systems are not designed for ease of programmability or flexibility in the implementation and validation of networking algorithms and thus do not lend them to a flexible mobile wireless network system test-bench. An operating system is desired which is compatible with existing platforms today but still provides functionality such as multi-tasking and packet processing capability useful to network control algorithms and can be easily modeled in the simulation environment. A network operating system is able to function on a layer on top of an existing native operating system and provide the required network functionality and services. A public domain network operating system, NOS (also known as KA9Q developed by Phil Kam), has readily available source code and meets the flexibility requirements. We use this network operating system in our test-bench [Figure 31. It runs on top of DOS and includes its own multitasking scheduler. The benefit of this multitasking operating system is that each algorithm or protocol necessary to support this network can be developed as its own process. The multitasking kernel allows these algorithms and protocols to multitask, sharing the CPU, and yet provide semantics such as wait and signal semaphores for inter-process (inter-algorithm) communication. Time processing routines, such as TDMA, are able to sleep a process for a defined period of time, and can be used to allow other protocols and algorithms to run without halting or consuming unnecessary CPU processing time. Memory buffers (mbufs as found in BSD UNIX system buffers) are used to minimize overhead by allowing memory blocks to be linked together for performing encapsulation, packetization, etc.

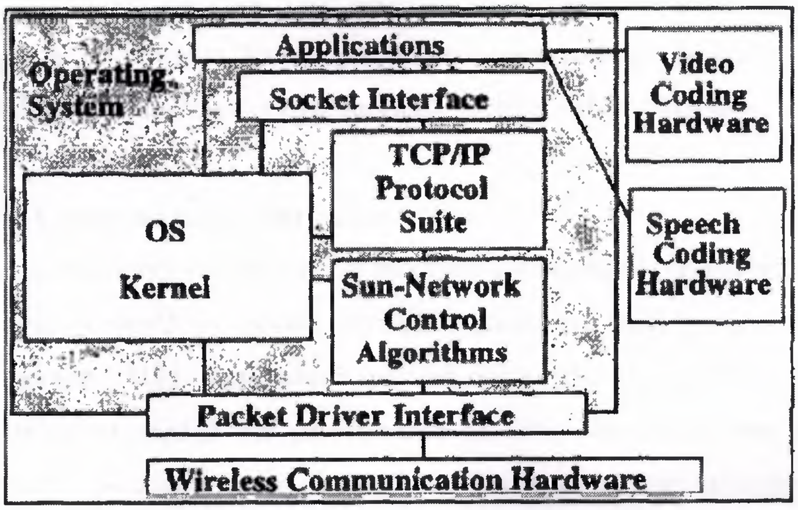


Figure 1.3 Network Operating System Components

Our current test-bench uses a NEC Versa 486 33 MHz laptop and a docking station to support custom interfaces and hardware. The WAMIS Network Operating System is able to run on any laptop as long as that laptop supports DOS and the required interface cards. A Packet Interface (PI) card is used as the network interface card to integrate the wireless communication hardware into the system. In order to provide a standard interface to the network operating sys- tem, a packet driver interface is used. The packet driver interface is based upon FTP's packet driver specification. This interface allows various network interface cards (like the PI card or a PCMCIA card) to be used in place of one another without having to change the details of the network operating system in order to support new or different communication substrate. A packet driver is loaded which corresponds to the correct Network Interface Card (NIC) and its capability. There are also other communication hardware drivers/interfaces such as the NDIS or ODI drivers which can be used to integrate the communication hardware with the operating system.

1.2.3 Multimedia Support

Various multimedia hardware support, such as speech (DSP) and video (Frame Grabber) cards, is now available for laptops. As more multimedia devices are made available for the mobile wireless net- work nodes, the greater impact and demand on the performance, capabilities and functions will have on the design and integration of such systems. The system integration and networking support issues and analysis will become critical since these multimedia devices place greater demands on the system architecture, such as bus bandwidth, and networking services, such as virtual circuits.

1.2.4 Wireless Communication Hardware

There is numerous wireless radio modems commercially available. Many of the algorithms being designed for mobile wireless systems are built to support a particular device/manufacturer. Algorithms which are not designed for a specific radio face the problem of trying to predict the performance of their algorithms over such a wide parameter space of available radio alternatives. The best way to validate over a wide parameter space of various radios is to utilize the models of the various radios in the simulation environment and do experimentation with those which are feasible to experiment with.

The UCLA WAMIS test-bench not only supports commercially available radios, such as the Proxim RangeLan 2, but also uses a specialized direct sequence spread spectrum radio designed and implemented at UCLA. This radio is used to support instant infrastructure networking through adaptive hardware control and feedback with the networking algorithms. This radio is currently able to operate at speeds from 7 to 32 Kbps depending on the spreading factor desired. Although other radios are able to support higher data rates, this radios provides a unique ability to control various hardware parameters such as the spreading (chips/bit), code, power, and even acquisition time. In Table 1 we can see the spreading factor (chips/bit), data rate, and acquisition time trade-off. It should take anywhere from 500 to 1000 data bits to acquire the signal so a preamble is sent before each packet according to the desired acquisition time. Since the radio transmits at a fixed rate of I Mchips/sec. and we are able to vary the number of chips/bit, then we are able to achieve the various data rates as described above. The reason one would not always necessarily want to use the fastest data rate is that the lower the spreading, the less resilience to noise and interference. By using more chips/bit (slower data rate) we are able to have more capacity of the network and less interference. It is up to the network control algorithms, with development and analysis support from the simulation environment, to determine what these parameters should be set at for optimum network efficiency.

Table 1.1 UCLA Radio Parameters

chips per bit	Data Rate (kbps)	Optimistic ACQ Time	Conservative ACQ Time
31	32.258	15.5 ms	31 ms
63	15.873	31.5 ms	63 ms
127	7.824	63.5 ms	127 ms

1.2.5 Mobile Wireless Algorithms

1.2.5.1 Transport and Internetworking Control

Since internetworking requires compatibility with existing net- works and TCP/IP is so widely used through the Internet. The TCPI IP protocol suite has been implemented without need for modifications. Since the Internet Protocol can be used in conjunction with various communication substrates, much of the new mobile wire- less algorithm development takes may place below the network layer. The network layer is responsible for supporting various communication substrates such as internet routing, segmentation, etc. Above the network layer, the transport protocols (TCP and UDP) provide the required support for end-to-end reliability, congestion control, etc. These transport protocols interact with the applications described in the: previous section by using sockets to buffer the bit stream so packetization can take place. Additional services are also being developed to support multimedia over mobile hosts.

Although wireless communication is useful to support mobile communication, wired connections can support much higher band- width and are less prone to errors then wireless radios. Therefore, wired connections should be utilized whenever possible. Wired connections, such as Ethernet, can utilize standard communication hardware, such as a PCMCIA card, for networking. In order to support a combination of wired and wireless communication, provide wireless multi-hop functionality, and support instant infrastructure networking, a node needs to be able to function in three different modes (gateway, multi-hop, or end node) as shown in figure 4. A node functions as a gateway when both wired and wireless connections are available. In the gateway mode, it will forward packets between the wired and wireless domains as necessary. In the multi-hop mode, it will follow the sub-network routing protocol to provide wireless multi-hop communication within the subnet. Other mobile wireless network systems are not be focused on instant infrastructure networks, but upon support mobility throughout the intemet.

1.2.5.2 Mobile IP

The IETF Working Group for Mobile IP has developed an Internet Draft for IP Mobility Support. Much of the focus of this group has been on protocol functionality and standards not on performance analysis. By incorporating the Mobile IP type protocol into this simulation environment, feedback can be provided to vendors interested providing implementations of these protocols on its performance as a function of various mobility environments, network connectivity substrates (wireless & wired), and various traffic loads. The Mobile IP protocol can also be integrated with numerous other system components.

The analysis of the Mobile IP protocols in this simulation environment will be useful to validate and enhance the simulation environment and prototyping implementation path. In addition to protocol designers, the prototype can provide immediate feedback to other groups in industry and academia that are developing protocols in conjunction with Mobile IP to support other network and operating system functionalities.

1.2.5.3 Instant Infrastructure Sub network Control

The functionalities which support instant and reconfigurable networks are new and have been added into the TCP/IP stack (Figure 3) on the UCLA test-bench. Many of the proposed schemes for sup- porting instant and reconfigurable network topologies are based upon TDMA to control channel contention. A clustering algotlthm was implemented which is heavily based on TDMA control and synchronization to test the feasibility and overhead of implementing this functionality in software.

1.2.6 Link Layer Control

Algorithms developed for link layer control fall into a separate category from other networking algorithms. These algorithms are usually not implemented inside the operating system, they usually exist in hardware or programmable processors as part of the NC. For maximum flexibility, simplicity of implementation, and provide a path between simulation and implementation, these algorithms could be implemented as part of the other algorithms in the operating system. To experiment and determine where an algorithm should be implemented, the simulation environment can utilize models or actual code of the link layer control algorithms.

The link layer control components typically include algorithms such as media access control (e.g., CDMA, TDMA, and CSMA/ CA). The link and mobility control layer shown in figure 2 supports new function unique to instant infrastructure mobile wireless net- working.

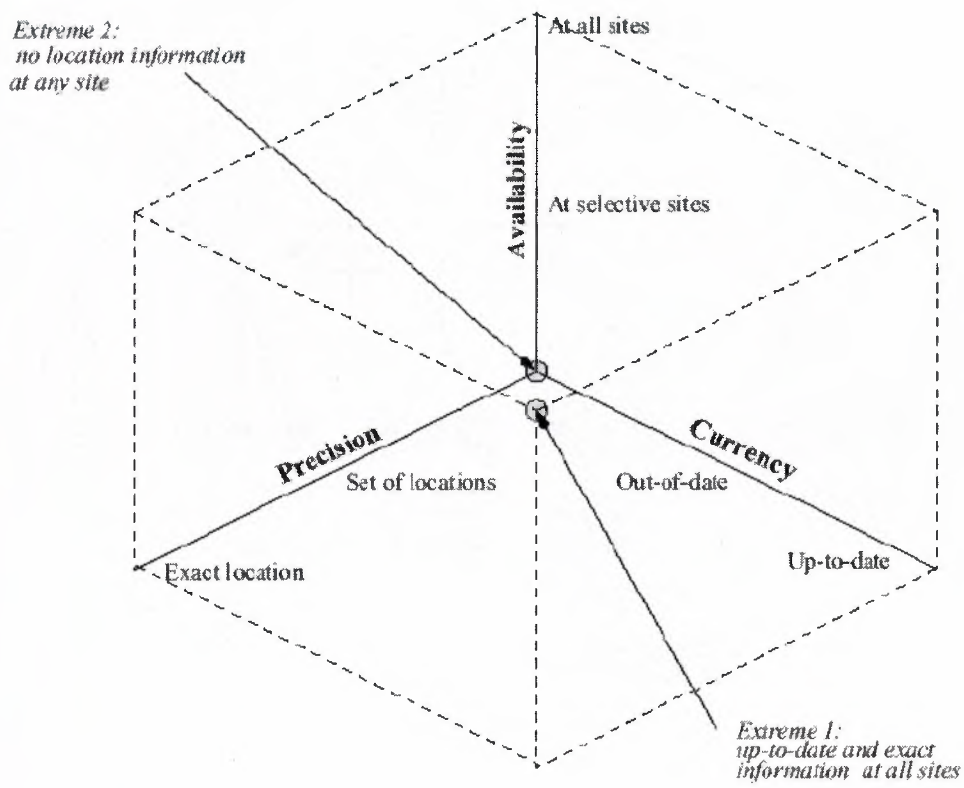
Mobility support is provided by setting appropriate hard- ware parameters such as the CDMA code or transmit power level dynamically. Measurements such as Signal to Interference Ratio SIR) are fed back from the radio into the link control algorithms to do power control and minimize the power consumption of the Link, reduce interference, and provide admission control such as described in. [2].

2. MANAGEMENT LOCATION

2.1 Applications

- Tied to wireless hardware (mobile users)
- Mobile software, i.e., code or data (migration, mobile agents, ubiquitous computing)

2.1.1 Taxonomy



2.1.2 Infrastructure

- Cellular Architecture - WAN – LAN
- GPS

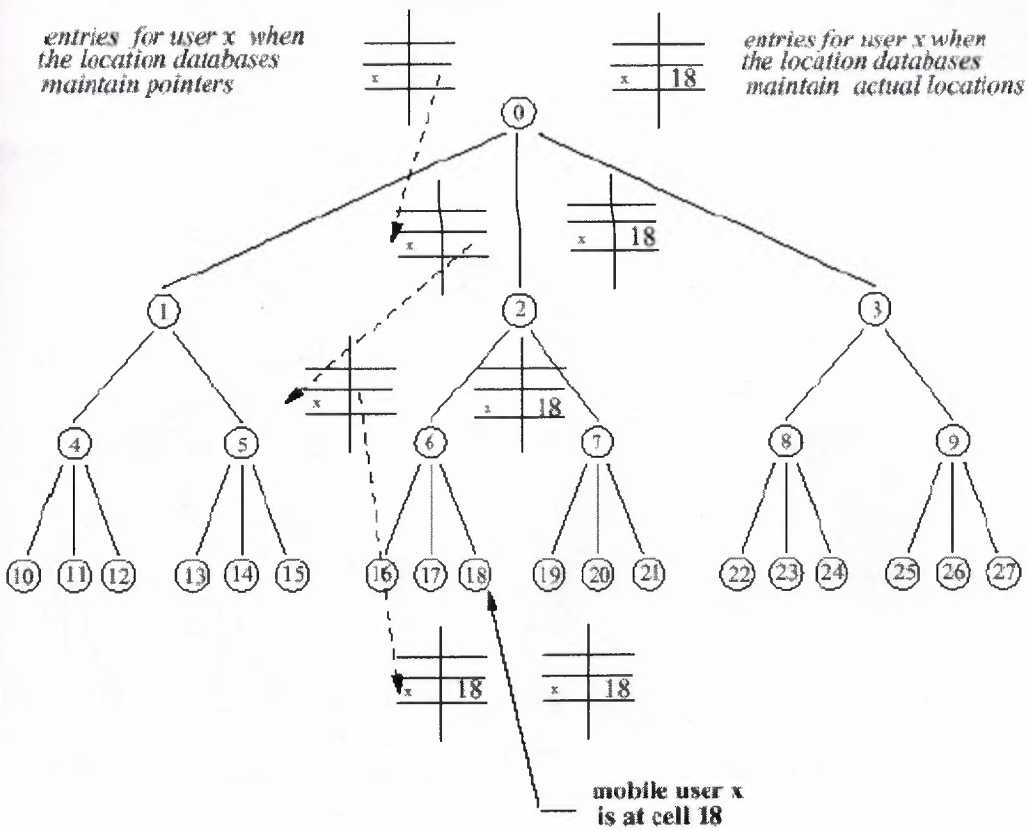
2.2 Architectures

2.2.1 Two-Tier

Home Location Register (HLR)

Visitor Location Registers (VLR)

2.2.2 Hierarchical Schemes



Comparison

(+) No need for life-long numbering (no pre-assigned HLR)

(+) Support for locality

(-) Increased number of operations (database operations and communication messages)

(-) Increased load and storage requirements at higher-levels

2.3 Placement of Databases

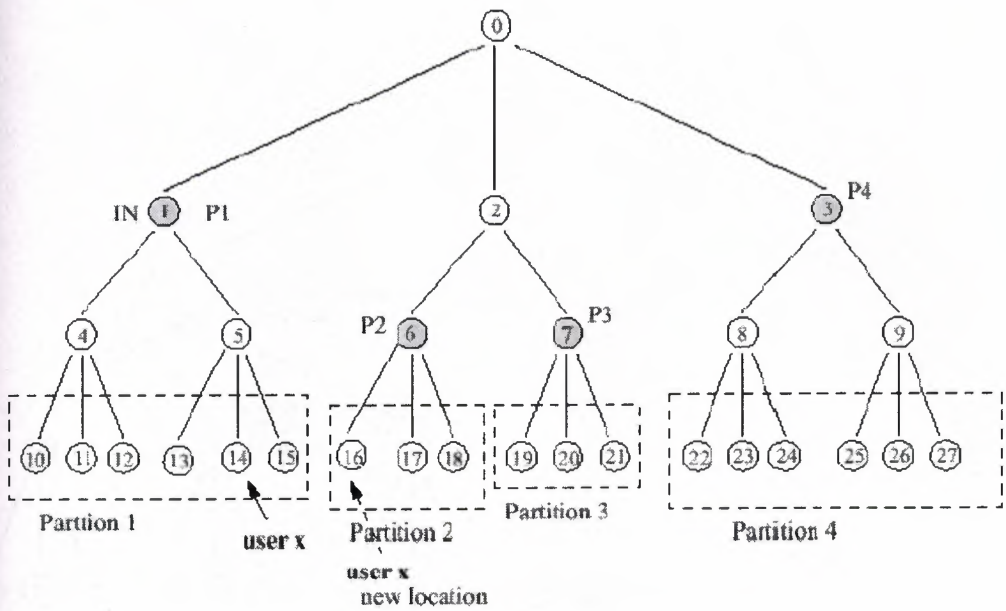
2.3.1 Entries at the Leaves (VLRs)

Flat, expanding, hybrid

Optimization

Objective functions: (a) the number of database updates and accesses, (b) the communication cost, (c) the sum of the traffic on the network link or links. *Constraints*: (a) database capacity (b) link capacity, and (c) storage.

Partitions



2.4 Caching

Two-Tier

After a call, save location at the caller's VLR

Invalidation

Eager caching, Lazy caching

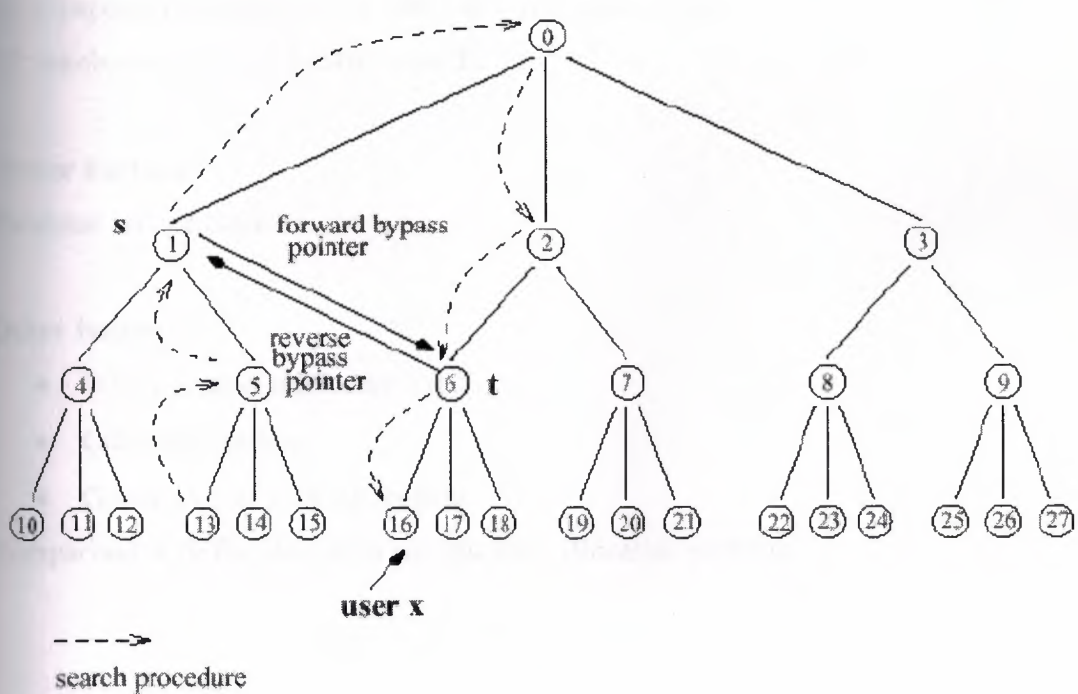
Performance

A hit ratio threshold pT = CH=CB, where CH is the cost of a lookup when there is a hit and CB the cost of the lookup in the non-caching scheme. Among other factors, CH and CB depend on the relative cost of querying HLR's and VLR's. In practice, it is expected that LCMRT > 7

Other

Replacement, Initialization

Hierarchical



Variations

- Simple caching - level caching
- Lazy
- Exact locations - pointers

Performance

Regional Call-to-Mobility Ratio (RCMR) for users with *RCMR* >5, a 30% reduction when considering only the number of database operations. More on granularity: caching and partitions

2.5 Replication

Replicate the location of specific users at selected sites.

Judicious

Replication of *i* at *j*

$\alpha * C_{i,j} >= \beta * U_i$

α: cost savings when a local lookup, as opposed to a remote query, succeeds

β: replica update cost

$C_{i,j}$: expected number of calls from j to i over time T, and

U_i: number of moves made by i over T.

Other Factors:

Database service capacity, storage

Other Issues:

- Where to keep replication set
- Other applications
- Granularity of location replicas

Comparison with file allocation and database allocation problem.

2.5.1 Per User Profile Replication

Problem Formulation

Let M: the number of users and N: number of zones. Find a replication assignment of a user's profile Pi to a set of zones R (Pi) such that the system cost is minimized:

$$\Sigma_{i=1}^{N} \Sigma_{j=1, Z_j \in R(P_i)}^{M} \beta * U_i - \alpha * C_{i,j}$$

Given constraints on the maximum number pj of replicas per zone Zj and on the maximum number of replicas ri per user Pi.

Solution: Construct a flow network F

Vertices: source vertex s, sink vertex t, users Pi and zones Zj

Edges:

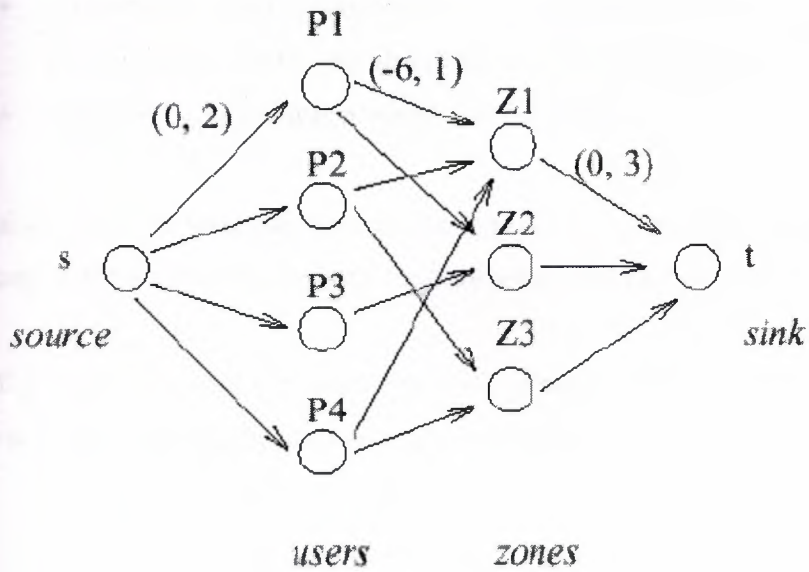
A pair (c, p) of attributes with each edge

S => Pi, with (c, p) = (0, ri)

Zj => t with (0, pj)

Pi Zj with (c, p) = (β * Ui - α * Ci,j , 1) iff it is judicious to replicate Pi at Zj ,

Compute a minimum-cost maximum- flow on F



Adaptation to changing calling and mobility patterns Compute Fτnew from Fτold .

2.5.4 The Adaptive Data Replication (ADR) Algorithm

Presents a solution to the general problem of determining an optimal (in terms of communication cost) set of replication sites for an object in a distributed system, when the objects read write pattern changes dynamically.

Preliminaries

- Tree-structure architectures
- R: the current replication set of object x
- A site i is an R-neighbor, if it belongs to R but has a neighbor site that does not belong to R.
- When site R is not a singleton set, a site i is an R-fringe site, if it is a leaf at a sub graph induced by R.

The Algorithm

- R is updated periodically every T, specifically every T three tests are performed:
- The expansion test performed by each R-neighbor site I Site i invites each of its neighbor j not in R to join R, if the number of reads that i received from j during the last T is greater the number of writes that i received during T from I itself or from a neighbor other than j.

The ADR Algorithm (continue)

- The contraction test executed by each R-fringe site i. Site I requests permission from its neighbor site j in R to exit R, if the number of writes that i received from j during T period is greater than the number of reads that i received during T.
- If site i is both an R-neighbor and an R - fridge, it executes the expansion test first, and if the test fails (i.e., no site joins R), then it executes the contraction test.

- The switch test is executed, when R is a singleton test and the expansion test that the single site i in R has executed fails. Site i asks a neighbor site n to be the new singleton site, if the number of requests received by i from n during T is larger than the number of all other requests received by I during T .

The ADR algorithm is shown to be convergent-optimal: starting at any replication scheme, it converges to the replication scheme that is optimal to the current read-write pattern. The convergence occurs within a number of time periods that is bounded by the diameter of the network.

2.6 Forwarding Pointers

When the number of moves that a user makes is large relative to the number of calls it receives, defer updating database entries holding the user's location.

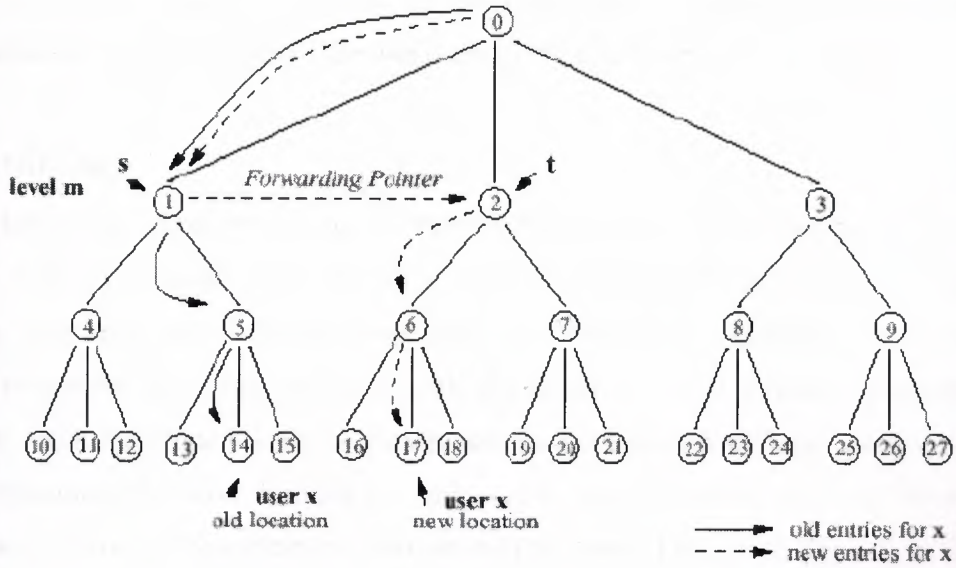
Two-tier Architectures

x's HLR is not updated, each time x moves to a new location. Leave a forwarding pointer at the VLR at x's previous location to point to the VLR at the new location. Calls follow a chain of forwarding pointers. The length of the chain of forwarding pointers grows up to a maximum value of K. Since the approach is applied on a per-user basis, the increase in the cost of call operations affects only the specific user.

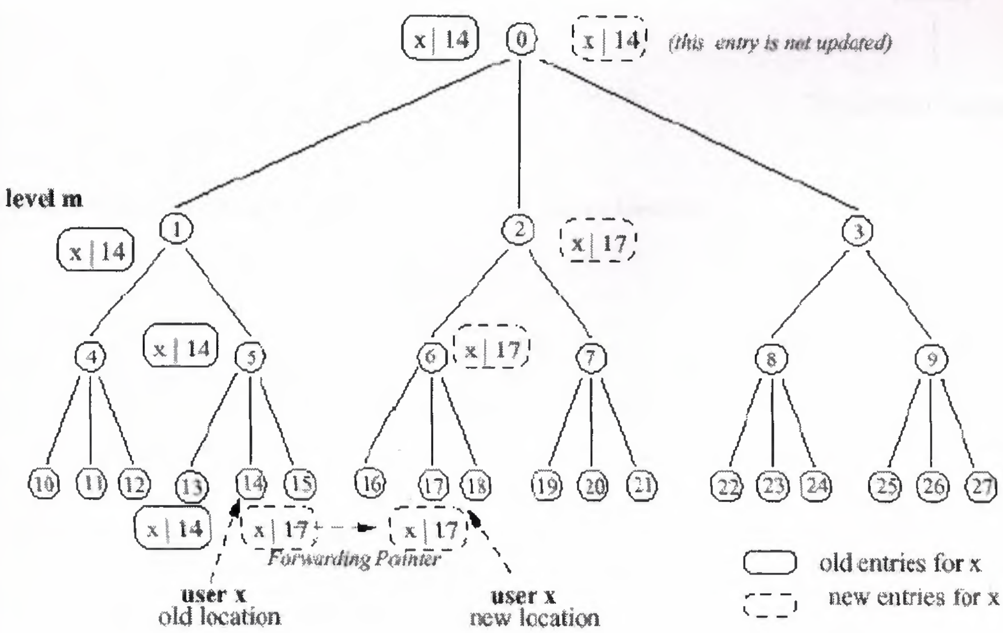
The router optimization extensions to IEFT Mobile IP protocol include pointer forwarding in conjunction with lazy caching. Performance depends on the cost of setting up and traversing pointers relative to the costs of updating the HLR. An analytical estimation. Under certain assumptions and if pointer chains are kept short ($K < 5$), forwarding can reduce the total network cost by 20%-60% for users with CMR < 0.5

Hierarchical Architectures

When x moves from i to j, instead of updating all databases on the path from j through LCA (j, i) to i, only the databases up to a level m are updated. A forwarding pointer is set from node s to node t, where s is the ancestor of i at level m, and t is the ancestor of j at level m.



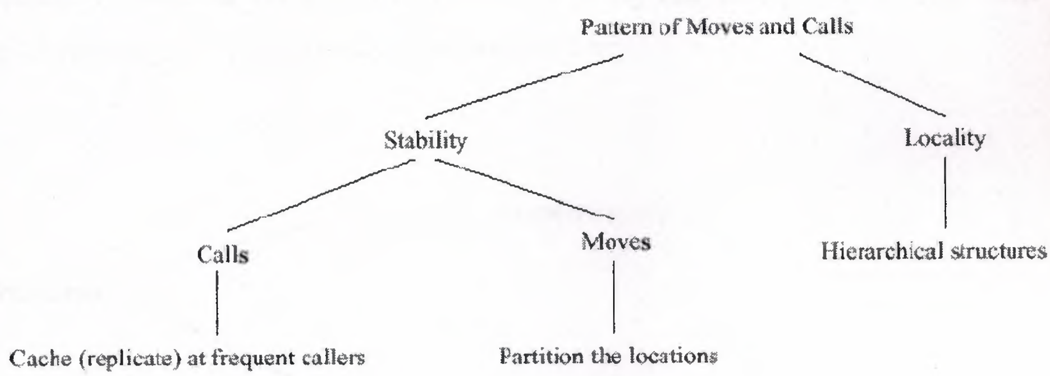
Simple forwarding vs. level forwarding when entries at the internal nodes are actual addresses.



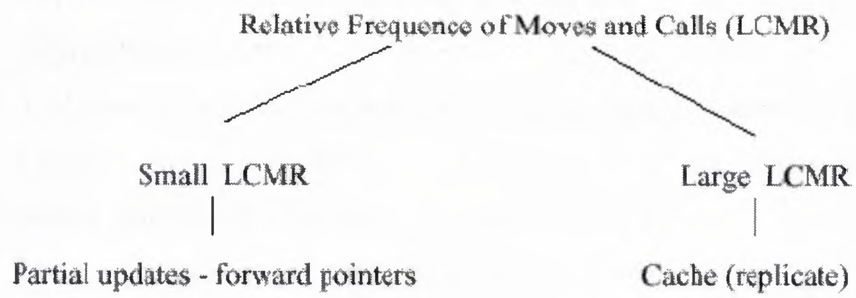
An analysis of a forwarding method when entries are actual addresses [10] along with caching based on the degree of mobility (CMR) host (low or high) and on whether it has a large number of frequent callers. Updating obsolete entries in databases at levels higher than m: e.g., after a successful lookup, or each node sends a location update message to all location servers on the path to the root during off-peak hours.

2.7 Taxonomy

Exploit knowledge about the calling and moving behavior of mobile objects: stability and locality. Stability of calls: most calls for a user originate from the same set of locations. Stability of moves: users tend to move inside specific regions. Locality: the cost of a lookup or update operation increases with the distance. Local operations (moves to neighbor locations or calls from near-by places) are common and should cost less than remote operations. Relative frequency of calls and moves, since often decrease the cost of either the move or call operation in the expense of the other. [3]

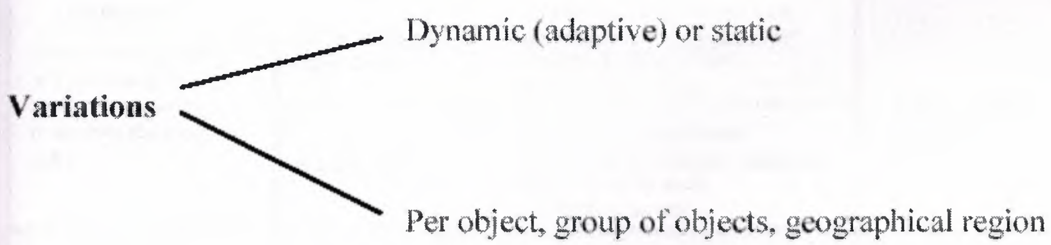


Management Location



More specific types of movement and calling: e.g., follow a certain mobility pattern or there is an epicenter (e.g., home location) of movement. Models of movement can be used in guiding the search for the current location of a mobile object. For instance, search candidate location in descending order of the probability of the user being there.

Dynamic adaptation to the current pattern and ratio.
Employment on a per user basis - overall - per group of users (e.g., based on their geographical location or on their mobility and calling characteristics) all users that receive a large number of calls) or a combination of both.



The topology of network sites, how they are populated and their geographical connectivity. Scales with the number of mobile objects, operations and geographical distribution.

Estimation of the current value of the CMR

- The running average algorithm, Maintain for every user the running counts of the number of incoming calls and the number of times that the user changes location.

- Store information about the CMR, for instance in the HLR, and download it during off-peak hours.
- Analytical estimations For instance, if the coming call stream to a user is consider a Poisson process with arrival rate λ and the time a user resides in a region has a general distribution with mean $1/\mu$, then LCMR = λ/μ.
- Traces of actual moving users (for example, (SUMATRA)[19].

Evaluation based on database operations: Minimizing (a) the total number of database updates and queries, (b) the database load and size, and (c) the latency of each database operation. And communication: Reduce among others (a) the total number of messages, (b) the number of hops, (c) the distance traveled, (d) the number of bytes generated, and (e) the sum of the traffic on each link or over all links.

Two-Tier Schemes

Method	Variations		Applicable when:
Caching When x is called by y, cache x's location at y's zone	*Eager caching:*	Cache update overhead occurs at moves	Large LCMR Call Stability
	Lazy caching:	Cache update overhead occurs at calls	
Replication Selectively replicate x's address at the zones from which it receives the most calls	*Per-user Profile Replication:*	Additional constraints are set on the number of replicas per site and on the number of replicas per user	Large LCMR Call Stability
	Working Set:	Adaptive and distributed: the replication sites are computed dynamically by each mobile host locally	
Forwarding Pointers When x moves, add a forwarding pointer from its old to its new address	Restrict the length of the chain of forwarding pointers		Small LCMR

Hierarchical Schemes

Method	Issues/Variations	Appropriate when:
Caching When x at zone i is called by user y at zone j, cache at a node on the path from j to LCA(i, j) a pointer to a node on the path from i to LCA(i, j) to be used by any subsequent call to x from zone j.	Up to which tree level to maintain cache entries When to update cache entries	Large CMR Call Stability
Replication Selectively replicate x's location at internal and/or leaf databases.		Large CMR Call Stability
Forwarding Pointers When x moves from cell i to cell j, instead of updating all databases on the path from i to LCA(i, j) and from LCA(i, j) to j, update all databases up to some level m and add a forwarding pointer at the level m ancestor of i to point to the level m ancestor of j.	When and how to purge the forwarding pointers Setting the level m	Small LCMR
Partitions Divide the locations into sets (partitions) so that the user moves inside a partition frequently and crosses the boundary of a partition rarely. Keep information about the partition in which the user resides instead of its exact location		Move Stability

2.8 Concurrency Control

Moves and calls are issued asynchronously and concurrently and each results in number of database operations => concurrency control to ensure correctness.

Leave a forwarding pointer to the new location

- When a call reads obsolete data and fails, it is reissued. No upper bound on the number of attempts.
- Traditional database concurrency control techniques such as locking or timestamps.
- Impose a specific order on the execution of the operations

- First, add entries at the path from j to LCA(i; j) in a bottom-up fashion

- Then, delete the entries at the path from the LCA(i; j) to I in a top-down fashion.

- Special care so that during the delete phase, an entry at a level k - 1 is deleted only after servicing all lookups from higher-level databases.

-[2]: application to the regional matching method

When replication => coherency control protocols to maintain the replicas consistent

- An HLR or a master copy that is always consistent

- Use forwarding pointers to handle any incoming calls directed there from obsolete replicas.

2.9 Failure Recovery

2.9.1 VLR Failure Restoration

2.9.1.1 Periodic Check pointing

- If the VLR is check pointed, the backup record is recovered.
- But if the backup is obsolete, then all areas within the VLR must be paged to identify the mobile users currently in the VLR's zone. Thus no improvement.
- GSM exercises periodic location updating: the mobile users periodically establish contact with the network to confirm their location.
- Periodic confirmation does not improve the restoration process, if the confirmation frequency < 0.1 times of the portable moving rate.

2.9.1.2 Location Update on Demand

- Eliminates the need for periodic confirmation messages.
- After a failure, a VLR restoration message is broadcasted to all mobile users in the area associated with the VLR.
- The mobile users then send a confirmation message. To avoid congesting the base station, each such message is sent within a random period from the receipt of the request.

2.9.2 Failure Recovery

2.9.2.1 HLR Failure Restoration

In GSM,

- The HLR database is periodically check pointed. After an HLR failure, reloading the backup restores the database.
- If the backup is obsolete, calls are lost.
- Obsolete data are updated by either a call origination or a location confirmation

In IS-41,

- After an HLR failure, the HLR sends an \Unreliable Roamer Data Directive" to all associated VLRs.
- The VLRs remove all records of associated with that HLR.
- Later, when a portable is registered at a VLR, the VLR sends a registration message to the HLR allowing it to be incrementally reconstructed. Before, calls are lost.

Aggressive Restoration

- HLR restores its data by requesting all the VLRs referenced in its backup copy to provide exact location information
- An algorithm to identify VLRs that are not mentioned in the backup; e.g., VLRs such that there are portables that moved in between the last HLR checkpointing and the failure and not out

2.10 Location Queries

Advanced queries that involve the location of moving objects

Examples: finding the nearest service, or identifying the shortest route with the best traffic conditions.

- May not include location directly, but may require tracking mobile objects indirectly, e.g., queries that involve data produced and located at mobile hosts.

- May be imposed by either static or mobile users and may include databases located at both static and mobile sites.
- Have both a spatial dimension, e.g., involve the position of a user and a temporal dimension, e.g., involve time.
- May include transient data, which is data whose value changes while the queries are being processed, e.g., a moving user asking for nearby hospitals.
- Continuous queries, e.g., a moving car asking for hotels locating within a radius of 5 miles and requesting the answer to the query to be continuously updated. Issues related to continuous queries include when and how often should they be re-evaluated and the possibility of a partial or incremental evaluation.
- Imprecision

2.10.1 Bounded Ignorance

How to derive an optimal execution plan for locations query that will acquire only the missing information necessary to answer it. [4]

2.10.2 Partitions

The system guarantees bounded ignorance: in that the actual and stored location of a user is always in the same partition. To determine the actual location of a user, searching in the partition of its stored location is sufficient. Deriving an optimal execution plan reduces to determining an optimal sequence in which to search inside the partitions of the users involved in the query.

2.10.3 Continuous Queries

The position of a moving object is represented as a function of time.
Thus, position changes *continuously* with time even without an explicit update through a database operation. A new data model, called MOST, is proposed to incorporate such dynamic attributes.

MOST enables queries that refer to future values of dynamic attributes, e.g., retrieve all the airplanes that will come within 30 miles in the next 10 minutes. The answer to future queries is tentative.

2.10.4 Routes

Objects move on predefined routes. The current position of an object is modeled as the distance from its starting point along a given route. Indexing the location of moving objects.

3. DESIGN OF LOCATION MANAGEMENT FOR 3G CELLULAR NETWORKS

3.1 Overview

Mobility management is one of the most important issues in Personal Communications Service (PCS) networks. In the ANSI-41 and Global System for Mobile communication (GSM) Mobile Application Part (MAP), which are the 2G cellular networks, the two-tier mobility databases, Home Location Register (HLR), and Visitor Location Register (VLR), are utilized to support mobility management for Mobile Terminals (MTs). The service area is partitioned into Location Areas (LAs). Within each LA, there are a number of cells. In each cell, there is a Base Station (BS) and many MTs. All the BSs within one LA are connected to a Mobile Switching Center (MSC). All the MSCs are finally connected to the Public Switching Telephone Networks (PSTN). Each LA is associated to a VLR, which is used to store the temporary records of MTs' profiles and location information. An HLR is used to record mobile users' permanent subscription information.

With increasing rates of international travel, the number of roaming users increases. Therefore, the signaling traffic on "short-haul" and "long-haul" international links increases. In order to reduce the international/remote roaming signaling traffic, the Gateway Location Register (GLR) within the Universal Mobile Telecommunication System (UMTS) Core Network is proposed in specification 3GPP 23.119. The GLR is a node between the VLR and/ or SGSN (Serving GPRS Support Node) and the HLR. It handles location management of roaming subscribers in visited network without involving the HLR in every change of LAs. Therefore, the signaling traffic between the visited mobile system and the home mobile system will be reduced and the location updating and the handling of user profile data across network boundaries are optimized. The GLR is located in the visited network. It stores the roamer's information, and handles location management within the network. Note that gateway location registers are optional in the architecture of 3G cellular networks.

There are two basic operations in location management, location update and paging. Location update is a process through which a system keeps track of the location of mobile terminals that are not in conversations. Paging is a search process conducted in a Paging Area (PA). A PA may include one or more cells. When an incoming call arrives, the system searches for the mobile terminal by sending polling signals to cells in the PA. Many location management schemes have been proposed for PCS cellular networks with two-tier mobility databases. Basically, there are two categories of location management: static schemes and dynamic schemes. In a static location update scheme with two-tier mobility databases (STATIC-2G), the HLR location update and VLR location update are performed when an MT enters an LA and the PA is the same as the LA. Therefore, the PA size is fixed. There are basically three kinds of dynamic location update schemes in which the PA size is variable movement-based location update, distance-based location update, and time-based location update. Similar to the static location update scheme, the HLR location update is performed when an MT enters an LA in a dynamic location update scheme. In the distance based location update scheme, the VLR location update is performed when the distance between the current cell and the last cell where the VLR location update is performed reaches a threshold d in terms of number of cells. In the time-based location update scheme, the VLR location update is performed in each d units of time. The movement- based location update scheme is the most practical one among the three kinds of dynamic location update schemes.

In a movement-based scheme (DYNAMIC-2G), a VLR location update is performed either when an MT crosses an LA boundary or when the MT completes d movements between cells, where d is the movement threshold. The PA is the area within both the LA, where the last VLR location update is performed, and a circular area with the diameter d-1 and with the center where the last VLR location update is performed. Therefore, a PA size is a variable. Most existing movement-based schemes only consider that a VLR location update occurs when the MT completes d movements between cells, and fail to consider the case that a VLR location update also occurs when the MT crosses an LA boundary. It is reasonable that a VLR location update also occurs when the MT crosses an LA boundary since an HLR location update occurs anyway [6]. The problem becomes much more complex when both of the above cases are considered. The difficulty exists since it is very hard, if not impossible, to derive the number of cell boundary crossings between two LA boundary crossings when the residence time in an LA follows a general distribution. In this paper, we consider both cases. Moreover, there is not any location management scheme proposed for 3G cellular networks, where the GLRs are deployed. In this paper, we study two location management schemes: a dynamic movement-based location management (DYNAMIC-3G) and a static scheme (STATIC-3G) for 3G cellular networks, and the cost functions of HLR location update, GLR location update, VLR location update, and paging are formulated. Furthermore, when an MS crosses a G-LA, HLR location updates, GLR location updates, and VLR location updates are all performed; and when an MT crosses an LA, GLR location updates and VLR location updates are all performed.

Therefore, the model itself becomes much more complex. Note that the paper presents location management in the presence of (an "extra layer" of) GLRs, while not all aspects of 3G location management are addressed. With the presence of the GLR in 3G networks, formulation of cost functions becomes extremely difficult as shown in later sections. Furthermore, we prove analytically that there is an optimal movement threshold for DYNAMIC-3G that minimizes the total cost of HLR location updates, GLR location updates, VLR location updates, and paging. An effective searching algorithm is also proposed to find the optimal movement threshold. Finally, we compare the following schemes: DYANMIC-3G, DYNAMIC-2G, STATIC-3G, and STATIC-2G under different parameters.

In the distance-based location update scheme, an MT needs to record the distance it has moved since the last location registration and performs an update when the distance exceeds a certain threshold. The device to record the distance is more complex and more difficult to be accurate than that for a movement-based scheme, in which the number of crossing cell boundaries is simply counted.

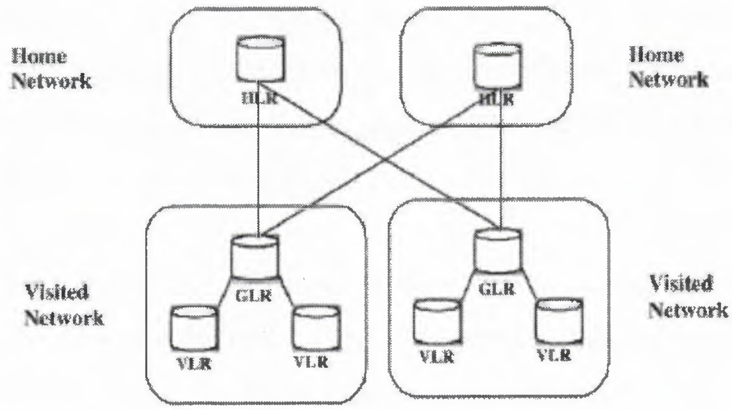


Figure 3.1 Mobility database architecture.

The idea inside the strategy is similar to that in movement based scheme, but the threshold they compared is different. The authors consider the distance-update as an optimization problem. Under a one-dimensional linear model, the optimal distance threshold is determined by dynamic programming. The authors propose an approach to compute the optimal threshold in a two-dimensional hexagonal model based on the assumption of symmetric random walk pattern. In the timer-based location update strategy, an MT updates its location every T time interval. This scheme does not require the MT to record or process location information during the time interval between updates. For implementation, a hardware or software timer can program the timer threshold into the MT. the authors introduce an analytical model for the time-based scheme. In this model, time-varying Gaussian user distribution and a Poisson page-arrival model are used to formulate the paging/registration optimization problem.[5]

3.2 Location Management For 3G Networks

In this section, we describe the DYNAMIC-3G and STATIC- 3G location management schemes for 3G wireless cellular systems (particularly UMTS). Fig. 1 shows simplified network architecture for a UMTS system with the GLR deployment at the edge of the visited networks. A GLR contains roamer's subscriber profile and location information. At the first location update procedure under the GLR, the subscriber profile information is downloaded from the HLR to the GLR. The GLR handles Update Location messages from the VLRs as if it were the HLR of the subscribers at the subsequent location updates. It enables the location update procedure to be handled locally for the movement within the visited network so that the costly Intervisited-network signaling for location management can be minimized. It keeps the profile information until a Cancel Location message is received from the HLR.

The relationship between the GLR and the HLR in 3G wireless systems is the same as that between the VLR and the HLR in the 2G wireless cellular systems (such as in GSM) in terms of the signaling traffic for location management. From the viewpoint of the VLR at the visited network, the GLR can be treated as the roaming user's HLR located at the visited network. From the viewpoint of the HLR at the home network, the GLR can be treated as the VLR. A GLR can interact with multiple VLRs. We define the Ring concept for the movement-based location update schemes as follows: If we treat a cell as the center cell, Ring-0 includes only one cell (the center cell). Ring-1 includes all the cells that surround Ring-0, and Ring-1 has six cells. Similarly, Ring-2 includes all the cells that surround the Ring-1, and so on. We can easily know that Ring-r has 6r cells except that Ring-0 has only one cell, where r = 0; 1; 2....

There are three kinds of location updates in 3G cellular networks: HLR location updates, GLR location updates, and VLR location updates. Location updates and paging procedures will cause a significant amount of cost such as wireless bandwidth and processing power at the mobile terminals, the BSs and mobility databases. In both DYNAMIC-3G and STATIC-3G schemes, unlike the 2G networks, the service area is partitioned into Gateway Location Areas (G-LAs). A G-LA is further partitioned into Location Areas (LAs). An LA consists of a group of cells. An HLR location update is performed when an MT crosses a boundary of a G-LA; a GLR location update is performed when an MT crosses a boundary of an LA.

For the DYNAMIC-3G scheme, a Paging Area (PA) includes a number of cells within an LA (which is also within a G-LA), while the PA's size is variable. A VLR location update is performed when an MT completes d movements between cells, where d is the movement threshold. An HLR location update involves both a GLR location update and a VLR location update, and a GLR location update involves a VLR location update. A PA is the area within the LA, where the last VLR location update is performed, and the circle area with the diameter being d-1 and the center where the last VLR location update happens.

For the STATIC-3G scheme, a VLR location update is performed when an MT crosses a boundary of an LA. A PA is the same as the LA where the last VLR location update is performed and the size of the PA is fixed, i.e., the size of the LA.

3.3 PERFORMANCE EVALUATIONS

In this section, we provide performance evaluation and comparison among different schemes. Similar to, we adopt the call-to-mobility ratio (CMR) as the mobility measure of an MT. The CMR ratio is defined as the ratio of the residence time in an LA and the interarrival time for phone calls on average, that the CMR equals one over the number of crossing of LAs between two phone calls.

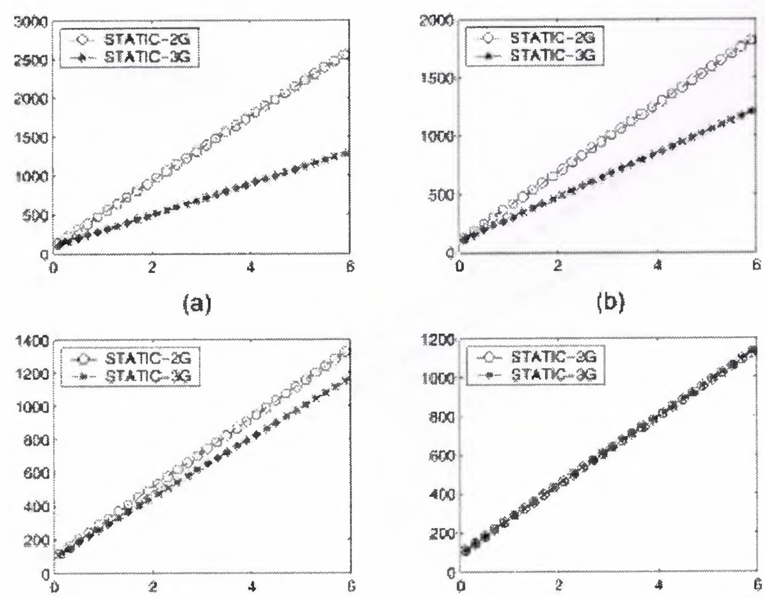


Figure 3.2 Comparing the costs of the static schemes with different RLCRs

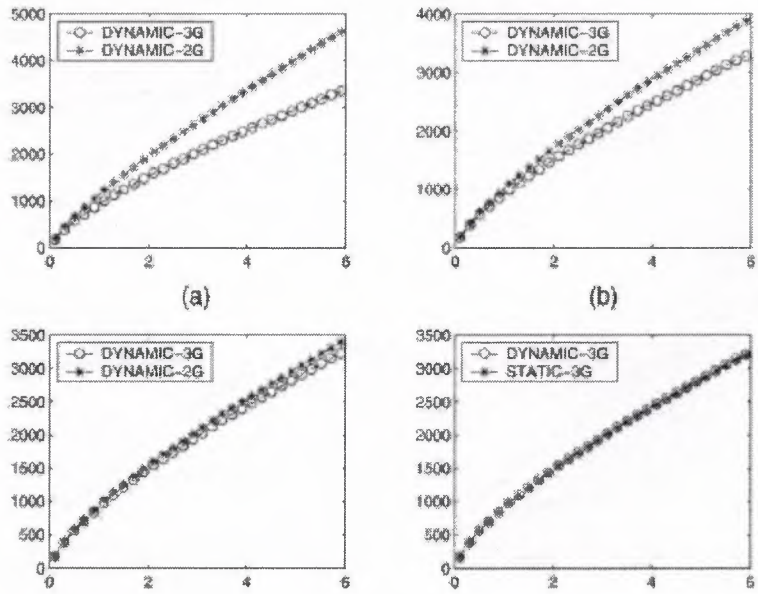


Figure 3.3 Comparing the costs of the dynamic schemes with different RLCRs

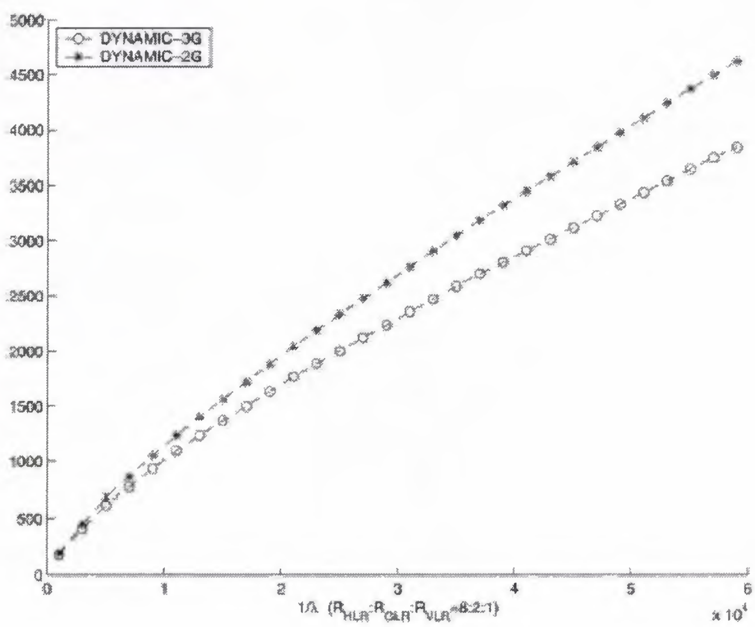


Figure 3.4 Comparing the costs of the dynamic schemes

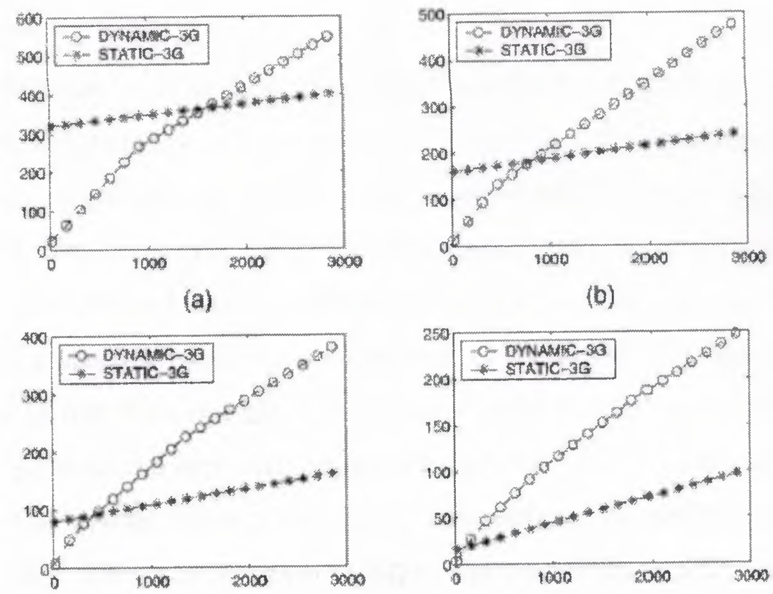


Figure 3.5 Comparing DYNAMIC-3G and STATIC-3G

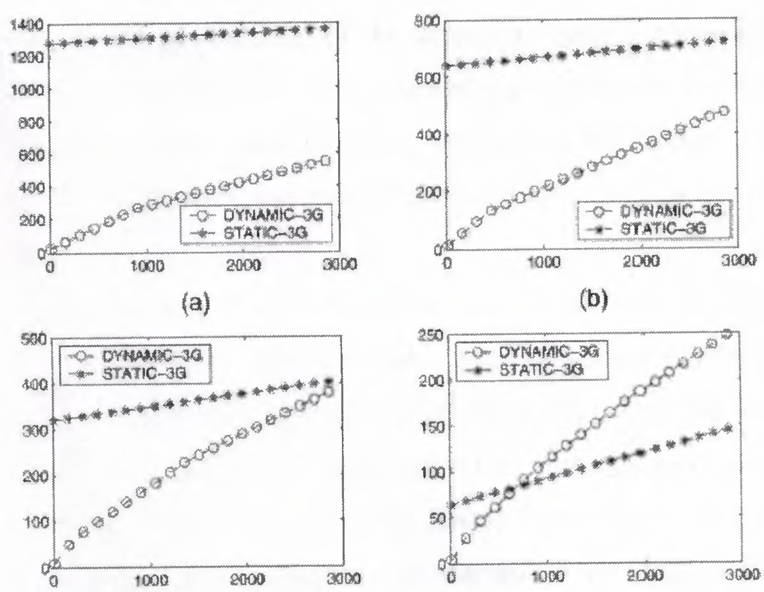


Figure 3.6 Comparing the dynamic scheme and the static scheme

4. MANAGEMENT LOCATION APPLICATIONS

4.1 Overview

In mobile environments, all the elements of the context of use may vary a lot. Users are different and they may use the services for many different tasks, even for tasks that were not anticipated in the design. The variety of mobile devices is growing and the users expect to be able to use the same or the same kind of services on the different devices. The technical and service infrastructure may differ and they may even change in the middle of a usage session, e.g. the network or the positioning system may change when the user moves from one place to another. Similarly, the service infrastructure, i.e. the available services and applications, may change. The physical context may vary a lot in terms of illumination, background noise, temperature and weather. The use of the device may affect the social situation in which the user finds him/her self or the social situation may affect the way the user uses the system.

An efficient way of improving the usability of mobile services and applications is to adapt the contents and presentation of the service to each individual user and his/her current context of use. In this way, the amount of user interaction will be minimized: the user has quick access to the information or service that (s) he needs in his/her current context of use. The information can even be provided to the user automatically.

A system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task. The main problem with context adaptation is that the context cannot be easily identified or measured. The location of the user is an element of the context that currently can be measured more or less accurately depending on the positioning system in use. In this paper, location-aware services are defined as context-aware services that utilize the location of the user to adapt the service accordingly. Location-based services are services that are related as such or by their information contents to certain places or locations. Thus location-aware services are a special case of location-based services.

So far, context-awareness has mainly been studied from the technical point of view and the studies have concentrated on location. Different experimental systems have been set up but only a few user evaluation results from small-scale trials are available. Location aware services are a concrete step towards context awareness. Other aspects of context-awareness will follow as soon as the corresponding elements of the context, such as weather or the social situation, can be measured and the adaptively needs can be identified.

This paper studies location-aware mobile services from the user's point of view. The paper draws conclusions about key issues related to user needs, based on user interviews, laboratory and. Field evaluations with users, and expert evaluations of location-aware services.

4.2 Methods of locating the user

From the point of view of the service, the simplest method of locating the user is to let him/her tell the location. From the point of view of the user, this method requires extra effort because the user needs to define his/ her location and input it to the system as a part of the search.

The user can be located with different positioning systems. If the user device includes a GPS (Global Positioning System) module, the user's location can be defined very accurately (2–20 meters). A GPS cannot be used indoors and it may not work in 'urban canyons' either. The location is calculated in the user device and it has to be sent to the service provider in order to get location-aware services. The range of commercial products currently available include mobile phones with integrated GPS modules, separate GPS modules for PDAs (Personal Digital Assistant), and GPS devices with integrated mobile phone and data features.

A mobile phone can be located by the telecom operator in the network. The positioning is based on identifying the mobile network cell in which the phone is located, or on measuring distances to overlapping cells. In urban areas the accuracy can be down to 50 meters, whereas in rural areas the accuracy may be several kilometers. The advantage of the cell-based positioning method is that the user needs no extra equipment – an ordinary mobile phone will do. If the user wants to use location-aware services from other service providers, the location has to be transferred to the other service provider and the telecom operator must get permission for this from the user. The location data is possessed by the telecom operator, which may not be willing to pass it on free of charge. Possibly because of these data transaction needs, current cell-location-based services are provided mainly by telecom operators.

The user can also be identified at a service point, utilizing e.g. WLAN (Wireless Local Area Network), Bluetooth TM or infrared technologies. These kinds of proximity positioning systems require a dense network of access points. The density of the network depends both on the required location accuracy and on the range of the access points. The accuracy can be down to 2 meters. The user needs special equipment, although WLAN and Bluetooth, for instance, are becoming increasingly common in current mobile devices. Because of the required infrastructure, such systems can only be used in a predefined area, e.g. a shopping centre, an exhibition area or an office building. The location of the user is available only when the user is in the service area.

2.3 Recent researches on context- and location-aware services

Context-awareness can be implemented as an adaptation of the user interface or the contents of the service. Services can also be invoked based on the identified context.

A context-aware user interface can select the appropriate modes for service interaction. A context-aware user interface can also be implemented e.g., as context aware text prediction or a location-aware remote control for the environment. A major challenge for The context-aware user interface is that the context may be continuously changing. This raises the problem of integrating changes into the user interface in such a way that the user remains in control. Moving can also be seen as one mode of interaction with the system. This interaction mode is quite challenging because it is difficult to know the user's intention: is the moving really taking place in order to interact with the system.

In recent research, context-aware contents have been studied in different application areas, e.g. tourist guidance, exhibition guidance, e-mail, shopping, mobile network administration, medical care and office visitor information. In these studies, the location of the user is the main attribute used in the context-adaptation. In well-defined application areas, it is possible to predict the other elements of the context according to the location of the user. Designing for more general user groups and wider contexts of use will be much more challenging.

4.4 The empirical studies

4.4.1 Scenario evaluations

We have carried out several empirical studies to study user attitudes, needs and preferences for location-aware services. We started with scenario evaluations in group interviews. The aim of this evaluation was to study broadly the attitudes of the potential users towards different personal navigation services. Future possibilities of personal navigation products and services were presented to the interviewees as pictured scenarios of everyday life.

In addition to location-based services, the scenarios also introduced route guidance services, services for tracking property and services based on locating other people.

We had 13 evaluation groups, each with 3 to 7 people, totaling 55 persons of different ages, different backgrounds and from different parts of Finland. The groups were selected so that they broadly represented the potential users of personal navigation services. The groups included four families, three hobby groups (football players, boaters and hunters), two youth groups, senior citizens, a group of motor-disabled people, and a group of visually impaired people and students of well-being technology. The groups were somewhat male-oriented, mainly because of the selected hobby groups. We wanted to include boaters and hunters because these groups are already familiar with navigation devices and thus might be early adopters of new personal navigation services.

Each group evaluated between three and we scenarios that were selected so that they presented the different aspects of personal navigation and were targeted according to the group.

The original scenarios were written in 1999 by a multidisciplinary team of experts, as a part of setting up the Personal Navigation Research and Development Programme in Finland. We modified the scenarios so that they reflected the present situation in the research and industrial fields, and so that they covered different aspects of personal navigation. We also wrote some brand new scenarios, targeted specially at different age and/or hobby groups.

The scenarios described location-aware advertising in the form of junk mail, a visit to an exhibition, different holiday and working trips, meeting friends in the evening, going to work and shopping. The scenarios were short stories of everyday life, illustrated with pictures of the context of use and imaginary mobile devices and services.

The scenarios were delivered to most of the groups in advance, so that the participants could read the scenarios before the interview. In the semi-structured group interview, the scenarios were discussed one at a time. The evaluators presented the scenario to the group and started the discussion by asking the interviewees how credible they considered the scenario, and why. Figure 2 illustrates the group interview with the senior citizens.

4.4.2 User evaluations

We have also evaluated with users different commercial location-aware services in Finland. The aim of these evaluations was to identify good solutions in current services as well as user needs for future services. Benefon Esc! When the Benefon Esc! Is used together with a Yellow Pages short message service (SMS), the user can get information on nearby services as well as their location, which the Benefon Esc! Can display on the map screen. The Benefon Esc! Was evaluated outdoors in the city of Tampere with six male users: three students and three middle-aged men (aged 59–62) having fishing and hunting as their hobby. We ended up with a male user group because of the hobby group selected.

In Finland, Sonera Pointer was one of the first location-aware WAP (Wireless Application Protocol) services that utilized cell-based positioning. Pointer Bensa (Gasoline) gave information on the cheapest gasoline stations in the vicinity of the user. Pointer Opas (Guide) offered information about the district around the user: sightseeing and attractions, events, activities, accommodation and tourist information. Pointer Fakta (Facts) gave statistical facts about the city or municipality in which the user was located. The Pointer services were evaluated in our usability laboratory with five users, three women and two men, aged 25–64, and with different backgrounds. Only one of the users had previous WAP experience. One of the users did not own a mobile phone himself.

The evaluations of both Benefon Esc! And Sonera Pointer were carried out in four phases. The evaluation started with filling in a background form to collect user data and information on his/her experience, e.g. with mobile phones and services. Then the user carried out test tasks that were combined with a frame story (''You are driving from Helsinki to Jyva¨ skyla¨ . Near Tampere you notice that you would need to .ll up your car...''). After the test tasks, the user could use the system freely, if there was time left and if the user wanted. Finally, the user was interviewed about his/her general impressions of the evaluated service. One evaluation session took about two hours. The laboratory evaluations were recorded on video, whereas the outdoor evaluation was recorded on audio and as photographs of usage situations.

4.4.3 Expert evaluations

We also reviewed some services in expert evaluations because these services were not available in Finland. CeBIT 2001 Mobile Fair Guide was a tradeshow guide for PDA computers. The software together with the Fair Catalogue could be downloaded from the web.

During the fair, there was also a demonstration on using the guide together with a Bluetooth-based indoor navigation system. Pocket Street map (www.pocketstreetmap.com) is map software for PDAs, which also includes some location-based services. Vindigo is a location-aware service guide for local entertainment, shops and restaurants in major cities in the USA and in London. We reviewed the Cebit 2001 Guide at the trade fair in Hanover, whereas Pocket Street map and Vindigo were reviewed in London.

The group interviews were carried out in spring 2001 and were followed by the user and expert evaluations from summer 2001 to spring 2002. In the following section, the results of all the above-mentioned evaluation activities are discussed as key issues to consider when defining user needs for location-aware services.

4.5 User needs for location-aware services

This section presents our main findings about user needs of location-aware services. The findings are grouped under five main themes: contents, interaction, personalization, service entities and privacy.

4.5.1 User attitudes

The attitudes of the users towards location-aware services were quite positive both in the group interviews and in the user evaluations. Location-aware information was expected to be especially useful in special situations, e.g. in unfamiliar environments, when looking for a specific service or in emergency situations. Location-aware information on parking lots, detours and contemporary events such as concerts and football matches was found useful. The disabled users would appreciate accessibility information, and the visually impaired pointed out the importance of speech-based systems.

However, criticism of new technology was brought up in many group interviews. A predestined and over controlled environment was seen as dubious, and the interviewees

During the fair, there was also a demonstration on using the guide together with a Bluetooth-based indoor navigation system. Pocket Street map (www.pocketstreetmap.com) is map software for PDAs, which also includes some location-based services. Vindigo is a location-aware service guide for local entertainment, shops and restaurants in major cities in the USA and in London. We reviewed the Cebit 2001 Guide at the trade fair in Hanover, whereas Pocket Street map and Vindigo were reviewed in London.

The group interviews were carried out in spring 2001 and were followed by the user and expert evaluations from summer 2001 to spring 2002. In the following section, the results of all the above-mentioned evaluation activities are discussed as key issues to consider when defining user needs for location-aware services.

4.5 User needs for location-aware services

This section presents our main findings about user needs of location-aware services. The findings are grouped under five main themes: contents, interaction, personalization, service entities and privacy.

4.5.1 User attitudes

The attitudes of the users towards location-aware services were quite positive both in the group interviews and in the user evaluations. Location-aware information was expected to be especially useful in special situations, e.g. in unfamiliar environments, when looking for a specific service or in emergency situations. Location-aware information on parking lots, detours and contemporary events such as concerts and football matches was found useful. The disabled users would appreciate accessibility information, and the visually impaired pointed out the importance of speech-based systems.

However, criticism of new technology was brought up in many group interviews. A predestined and over controlled environment was seen as dubious, and the interviewees did not accept the rational and purpose oriented attitude to life that they identified in the scenarios. In addition, some of the scenarios, for instance proactive shopping and exhibition guides, were seen as going too far beyond the real needs of people. The Exhibition scenario presented a system that guided the visitor along a predefined route, pointed out people that the visitor should meet, and offered a possibility to make appointments to popular stands. Almost all interviewees commented that these kinds of systems might be useful for 'some businessmen' but not for them. The interviewees were mainly familiar with quite small exhibitions where the idea is to 'just wander around anyhow'. The Boater group and Family 1 also pointed out that setting up the system before the exhibition visit might become a major task. In some groups, the Exhibition scenario catalyzed discussions where the interviewees identified useful applications for this kind of service. In the Senior citizens group, one lady told about using an automatic guidance system in a submarine museum in the US. The group agreed that the system should provide local information when needed but should not keep to a predefined route. The group also innovated a city guidance system based on the same idea.

Participants in the user groups mainly wanted solutions to ease their life in certain functions, but the aim of ultimate effectiveness itself was seen as bad. The interviewees thought that the location-aware reminder presented in one of the scenarios created a feeling of haste; the servant becomes a master that starts to give commands to the user. Fear of radical changes in human interaction, the usability of systems and the narrow use of new services and products were also raised in the groups.

4.5.2 Contents

4.5.2.1 Topical information

In the interviews and user evaluations, topical information turned out to be important to the users. This is the kind of information that may change while the user is on the move, in which case the information checked previously from other media (e.g. newspaper, TV or Web) may no longer be valid. Examples of such topical information are traffic information, weather forecasts, last-minute theatre ticket deals, or on-line chat. Unfortunately, in the evaluated services the only somehow topical information was included in Pointer Bensa (Gasoline), where gasoline prices at different service stations were updated once a day. In the group interviews, the Senior Citizens group as well as the Boaters group pointed out emergency situations, e.g. roadside help in a situation when the car breaks down. In this kind of situation, the traveler would need information on where (s)he could get help immediately, rather than a general service catalogue. Three Sonera Pointer test users also expressed a need for roadside services.

In addition to topical information, the users will need guidance on how to proceed in the changed situation. For instance, a train schedule as such can be obtained elsewhere but once on the move, the user will need information on delays and estimated arrival times. Furthermore, the service could suggest alternative routes in case of delays. The need for topical information is much greater than the Web is currently providing. This sets high expectations for the systems that maintain information as well as for the mechanisms by which the user will be informed and alerted.

4.5.2.2 Comprehensive contents

Although a mobile service on a small device can give only small portions of information to the user at a time, the user should have the possibility to access further information as exhaustively as (s)he needs. This need has been identified in other studies as well.

In the user evaluation of the Yellow Pages service, a user was looking for the closest restaurant with a dance floor and found that it was over 100 km away. The service did not include any of the dozens of restaurants with this facility in the city where the service was evaluated. The limited content of the services available was one of the main reasons why users considered the services not useful. The users expected to get some idea of the extensiveness of the contents as well as the geographic area that the service covers to be able to assess the usefulness of the service. Starting up a service will not be easy: on the one hand, a critical mass of service providers will be needed to get the users in, and, on the other, a critical mass of users will be needed to get the service providers in. It should be considered carefully how much content a pilot service should be included to be worth releasing.

The Cebit 2001 Guide did not have such problems because the exhibitors at a trade fair are an entity that can quite easily be managed. The earlier user evaluation results of the CeBIT2000 exhibition guide by Bieber and Giersich point out that users wanted to have both more extensive information about the trade fair itself (more information about the exhibitors and more extensive search possibilities), and more information about the immediate surroundings (e.g. a shuttle plan of the fairground and overall travel information). It is not wise to restrict the available information only to the current location and time: the users may also need to plan their next activities or to return to previous activities.

Although the user groups in our evaluations were small, the results point out that information and service needs vary, not only according to the location but also according to the user and the usage situation. In one test task, the users were looking for accommodation information from the Sonera Guide. All the users missed some additional information about the hotels, but the type of information required varied. Some users would have liked to get information on price and room availability; some would have liked to get more information about the location, some about the quality and facilities. Similar variations in user preferences could be identified in making a reservation and in getting route information. The users will need comprehensive information both in breadth (number of services included) and depth enough information on each individual service).

4.5.3 Interaction

Location-aware services differ from many other types of mobile services because they are not just mobile in the sense that they can be carried with the user but they are actually used on the move. The usage situations are demanding because the user can often devote only partial attention to the device, concentrating on his/her primary task of moving. In addition, the physical environment (e.g. background noise, illumination, weather) may disturb the usage situation.

4.5.3.1 Push or pull

In the group interviews, most users thought that they would not mind having the service or information pushed to them as long as the provided service or information was really what they needed in the situation. The attitudes of the young users (Youth groups 1 and 2; Students of well-being technology) towards location aware ads was very positive, although the teens pointed out that they wanted the ability to select what kind of ads, from whom and when they received. The attitudes of the older interviewees were also mainly positive but more prudent. Some of the older interviewees said that they just do not like ads in general. Although the attitudes of the users were positive, location-aware push services should be designed carefully. Attitudes will soon change if the users .nd themselves continuously receiving information that they do not need.

In most cases, location as such is not enough to activate push but location should be complemented with personalization. Excess content must be avoided, but at the same time the user must get an interesting, up-todate and useful service from his/her personal point of view.

4.5.3.2 Detailed search options

When evaluating the Yellow Pages service, one person was looking for a restaurant with a dance floor, another for a vegetarian restaurant, and a third for a moderately priced lunch restaurant. In our trial, the Yellow Pages service made no distinction between different kinds of restaurants. The probability of satisfying the user need in the search was near to chance. When getting information about gasoline stations with Sonera Bensa (Gasoline), the users in the evaluations said that in many situations they would have a greater need for information on the nearest or next gasoline station on the route than on the cheapest one provided by the service.

Vindigo provided shopping guidance by classifying the shops into women, music, sports, electronics etc. On the basis of our user interviews and evaluations, this kind of classification seems to be on too general a level. The users would bene.t more from a classification based on the items that the user is looking for, e.g. tennis shoes, music scores or mobile phones.

Location-aware guidance services should provide accurate searches, both at the level of the search criteria and the classification of information. Similar needs have been identified in mobile services in general.

4.5.3.3 Planning versus spontaneity

An important issue is whether the user needs the information when (s)he is at the given location, before getting to the location, or when planning the visit. Some information or options may be needed only after the visit. Flexibility of use requires that the services should support both pre-trip planning and on-route information on occasionally found points of interest. In general, the user should be able to utilize the services without any premade plans.

Our scenarios of use presented active users who plan their visits carefully and configure their devices accordingly. While traveling around, the users are ready to fine-tune the settings and respond to all kinds of inquiries from other travelers or service providers. The same attitude could be recognized in some of the evaluated services. This is partly because of the slow mobile networks that require the user to pre-load the necessary maps and other information using faster fixed-line connections. In practice, people may not be willing to spend their time on something from which they do not get immediate benefit. In particular, people tend to be busy while on holiday; they do not want to waste a single minute of their holiday time.

In the user evaluations, most users thought that they would use location-aware services occasionally and mainly in unfamiliar environments or in emergency situations. These needs indicate that the services should be easily available when the spontaneous need for them arises. The services should be easy to find, easy to take into use and use thereafter.

The users in our interviews said that location-aware systems should not lead to a predestined and over-controlled environment. The user should feel and be in control. Similar needs have been identified in other studies as well. Cheverst et alstate that the user should be able to override the recommendations of the system. Fano and Espinoza et al.point out that the user should have the alternative of using the system in an explorative way, i.e. having a look around without any predefined goals.

4.5.4 Personalization

4.5.4.1 Personal options and contents

The contents of the evaluated services were quite limited and thus it is understandable that the services did not include any personalization options. However, when the users in the evaluations described the kinds of information that they were missing from the services, it was evident that there are personal variations in preferences for both the contents and the presentation.

Also, location-awareness as such seemed to create an impression of personal entry to the service, e.g., quoting one of the test users of Sonera Pointer: ''Does this phone know that we are in Tampere?''

Personalization in location-aware services is a good way of improving the usability of the services by providing the most essential information and the most probable options easiest available. Earlier studies on personalization indicate that, although the users were interested in personalized services, the motivation to actually do something for this is often quite weak. Hollensberg and Vind Nielsen suggest that with WAP (Wireless Application Protocol), personalization is a highly advanced feature, which WAP users will explore and use later, if ever. On the other hand, Ramsay and Nielsen point out those personal and localised services are where WAP should be able to make an impact. It is a big design challenge to design personalized location- awareness so that it does not require too much effort on the part of the users.

The user may personalize the system according to his/ her preferences but if the preferences are different in different locations, configuring the system for all these locations becomes a major task. Ways of assisting the user in the personalization of the system are worth considering. Cheverst et al. point out the need to make the user profile persistent, enabling the use of the same profile with other services as well, e.g. in the next city to be visited. Collaborative filtering can be used to identify similar users and to adapt the system according to the group profiles. The user should also be able to use the same profile with different devices.

4.5.4.2 User-generated contents

Far too often users are seen as passive information consumers. However, letting the users participate and provide their opinions and recommendations could enhance many services. Such contents may enrich the service, bringing in additional users and encouraging a sense of community among users. The dynamically changing information generated by other users at a particular location may be better suited to the needs of the next visitors than a more general type of information that is provided externally. The users would also benefit from the possibility to store, access and share with others their own location-based information: notes, photos, etc. The system can also identify locations frequently visited by the user and then assist the user in associating information with these places. Shared contexts can encourage a sense of community among visitors.

In the user evaluation of the Benefon Esc!, the users appreciated the possibility to generate and store their own information related to locations that were personally important to them. In the group interviews, information provided by other visitors at the same place was deemed interesting. However, most users did not accept the idea of being contacted by strangers.

4.5.5 Seamless service entities

4.5.5.1 Consistency

With the Sonera Pointer Services, a number of usability problems arose because of inconsistencies in different Pointer services. The logic employed to navigate backwards in the services did not stay the same in different parts of the services. Some pages were organized so that links to the previous and front page were in the middle of the page. Most users did not notice that the page was continuing after those links. In one test task, the users got information on age distribution in the city of Tampere from Pointer Fakta (Facts). When reading the long list on a small screen, the users easily lost the focus and could not say which information was belonging to which age class. In this kind of lists, individual data fields should be marked consistently.

When the choice of services grows, and since the contents may come from several separate service providers, consistency becomes even more important. The user learns to use the services more easily when the information contents and structure remain similar in the different services. When traveling abroad, the user would appreciate familiar-looking services that are consistent with the services that (s)he is using at home. The internationalization and localization of location based services is challenging because the local services will probably have both local users and domestic as well as foreign visiting users.

4.5.5.2 Seamless solutions support the whole user activity

In our evaluation, some of the evaluated products were designed for route guidance and some for location- aware services. On the one hand, when using route guidance services, the users often missed information about places and nearby services. On the other hand, when using location-aware services, the users expressed a need to get route guidance to the place of interest. They also wanted to have easy access to phone numbers, e.g. the possibility to call the given hotel to make a reservation. Ideally the user should see all the necessary information for a given task in a single view. The information and the options should be accessible directly from the point in the service where the need for that piece of information or option arises. The design should aim for a seamless solution whereby the user is supported throughout the whole usage situation.

Cheverst et al. point out the need to be able to use the mobile system both on and o. line. Connectivity affects functionality, e.g. the availability of interactive services. In our evaluations we identified similar needs – the connection may not always be on, but nevertheless the user should be provided with as much functionality as possible. For example, the users of Sonera Pointer wanted to save the search results for later reference. [7]

4.5.6 Privacy

Privacy protection in location-aware services is related to the right to locate a person, use the location, store the location and forward the location. Current legislation is the basis for privacy protection but social regulation can also create rules and norms for different situations in which location-aware services are used. Espinoza et al. emphasize the right of the user to remain anonymous. Marmasse and Schmandt suggest that problems with privacy could be avoided by performing location tracking and analysis solely on the client device. Ljungstrand and Ackerman, Darrel and Weitzner point out the trade-o.Between privacy intrusion and user benefit. They think that, if the benefit is perceived as large enough, some degree of privacy loss will probably be accepted. Continuous requests for permission may overwhelm the users and may disturb the user in his/her activity.

The World Wide Web Consortium (W3C) has de- fined the Platform for Privacy Preferences Specification (P3P). This specification allows web browsers to interpret the privacy practices of the Web sites and compare these with the predefined user preferences for privacy. Then the user does not have to consider privacy issues separately with each individual Web site. This kind of approach could also be utilized with location-aware services.

Privacy can also be affected by the selected input/ output modes of the system. Speech-based interaction may create privacy problems when used in public places. Bisdikian points out those using public terminals as private displays may also create privacy problems.

In our group interviews, people were worried about their privacy and the ''big brother'' phenomenon when considering services enabling people to be located. However, the interviewees were not worried about privacy issues with location-aware services. It did not occur to most of the interviewees that they could be located while using the service. If they did understand this, they seemed to put a great deal of faith in the current telecom operators: ''the telecom operators will guard that kind of information. They already have all kinds of information about me but do not distribute it around''. It was also commented that there would be regulations and legislation protecting those who use location-aware services. The danger of someone somehow abusing their knowledge of the user's location came up only occasionally in the group interviews. In the Sonera Pointer evaluation, two users spontaneously mentioned invasion of privacy as a potential threat in location-aware services.

The location-aware services should inform the users of what kind of data is collected, how it is used and who has access to it. The faith that the users have in the technology, the service providers and the policy-makers should be regarded highly. Any abuse of personal data can betray that trust and it will be hard to win it back again. The user should be able to flexibly control the release of private information such as his/her location at a given time. The user should be allowed to remain anonymous when (s) he wants.

5. MOBILITY MODELING, LOCATION TRACKING, AND TRAJECTORY PREDICTION IN WIRELESS ATM NETWORKS

5.1 Overview

The extension of broadband ATM networks to the wireless domain presents a challenging set of problems for both network designers and managers. One of the key problems within this set is mobility management.

Mobility management entails both connection management and location management. Since ATM is a connection-oriented technology, it contains both a connection establishment phase prior to data exchange and a connection release phase after data exchange. In a wireless network, as terminals move, segments of connections have to be torn down and reestablished with a frequency that corresponds to the speed of the mobile. Meanwhile, data integrity in terms of cell sequence preservation, duplicate cell prevention, and cell loss avoidance has to be provided. Additionally, quality-of-service (QoS) guarantees have to be maintained regardless of the terminal's mobility.

At the most fundamental level, quality of service, a cornerstone of ATM networks, can only be provided if the system is able to maintain connectivity with the mobile terminal, even when the terminal frequently changes its physical location. It is possible to maintain connectivity and guarantee QoS to the mobile if the system knows, prior to the mobile's movement, the exact trajectory it will follow. With this information, the system can determine if there are enough resources available along the mobile's path for the lifetime of the connection. If such is the case, the system can plan in anticipation of the mobile's demands, and take appropriate steps such as setting up end-to-end routes from base stations in the mobile's path, reserving resource along these routes and planning quick handoffs between the involved base stations. With these kinds of preparations, QoS can be guaranteed.

Recently, there has been some work on providing QoS to mobile terminals. Acampora and Naghshineh propose a virtual connection tree (VCT) scheme in which multiple connections are reestablished between a fixed (root) switch and a set of base stations with whom the mobile could potentially connect. While providing a good starting point, the VCT algorithm in its native form suffers from the lack of inclusion of accurate knowledge of the mobile's trajectory. There is a danger of underutilizing resources in base stations to which the mobile never connects, and a potential of overloading base stations when a large number of mobiles connect at the same time.

Overloading can lead to congestion, which can result in ATM cells being either dropped or buffered or both. Buffering can cause temporary violation of delay and cell loss guarantees. As a possible solution to these kinds of problems, Levine et al. have proposed the concept of a shadow cluster. A shadow cluster defines the area of influence of a mobile terminal (i.e., a set of base stations to which the mobile terminal is likely to attach in the near future).

Like a shadow, this set moves along with the mobile, incorporating new base stations while leaving the old ones as they come under and out of the mobile's influence. Each base station in the shadow cluster anticipates the mobile's arrival and reserves resources for it. A close association exists between the mobile's arrival prediction and reservation of resources for it. The accuracy of the mobile's path prediction determines the number of base stations that reserve resources, and consequently determines the overall system efficiency.

Location management or location tracking incorporates the set of mechanisms with which the system can locate a particular mobile at any given time. Two strategies are possible: location updating and location prediction. Location updating is a passive strategy in which the system periodically records the current location of the mobile in some database that it maintains. Tracking efficiency is based on the frequency of these updates which, in most systems, are initiated by the mobile. Location prediction is a dynamic strategy in which the system proactively estimates the mobile's location based on a user movement model. Tracking capability depends on the accuracy of this model and the efficiency of the prediction algorithm. While most recent studies have focused on the update method, relatively little has been done on the prediction side. As a consequence, location management or tracking is generally perceived as a purely database updating and querying procedure .If accurate movement prediction per mobile was possible, the task of locating mobiles given their last location would become substantially efficient in terms of both speed and system resources used.

Thus, to sum it up, the above discussions on connection and location management make a compelling case for placing emphasis on developing algorithms and techniques for mobile trajectory prediction .Being able to determine the mobile's future locations and access points as it moves inside the network while being connected can result in significant improvement in system efficiency and connection quality.

One way for the system to know the future direction of the mobile is to have a formal mechanism in place that allows the mobile to indicate to the system its intended destination and the duration of the connection. The system can combine this information with its knowledge of the geography of the terrain, and the location of the base stations within the terrain, to determine the path of the mobile.

Unfortunately, this not a conclusive solution since there can be multiple paths to the destination, and even the mobile cannot exactly know its precise future movement patterns. It is not unreasonable have the mobile diverge without warning from its stated path in order to adjust to its dynamically changing environment. Without the system dynamically adapting to such unanticipated but reasonable trajectory changes, the amount of resources required to provide improved connectivity would be prohibitive, and consequently unattractive .In this paper, we propose a novel hierarchical location prediction (HLP) algorithm, which substantially increases the system's probability of providing uninterrupted service to the mobile user while consuming minimal resources from the network.

Derived from some classical and well established stochastic signal processing techniques, HLP is a two-tier scheme that combines location updating with location prediction to offer enhanced connection management functions. HLP raises the level of intelligence within the wireless ATM system so that the system aggressively and effectively maintains connectivity (essential for providing QoS features) with the mobile.

Some previous works in the area of mobility prediction includes Tab bane's proposal, which suggests that the mobile's location may be determined based on its quasideterministic mobility behavior represented as a set of movement patterns stored in a user profile. This method was further pursued by Liu and Maquire , in which a user's moving behavior is modeled as repetitions of some elementary movement patterns. Based on these movement patterns, pattern matching/recognition-based mobile motion prediction algorithm (MMP) is proposed which is used to estimate the future location of the mobile. The main drawback of the MMP algorithm is its high sensitivity to so called "random movements." Any movement that cannot be classified by the simple mobility patterns defined is classified as random movement. As reported in, prediction performance of MMP decreases linearly with the increase in the random factor . Other methods for predicting speed and trajectory have also been proposed in the literature, but these have generally been limited in scope as they consider rectilinear ("highway") movement patterns only.

In order to develop a prediction algorithm for mobiles with different mobility characteristics, we require a movement model that allows us to explore regularity and rationality in the seemingly random movement. We avoid using mobility models that incorporate deterministic mobility but instead propose a novel pseudo stochastic movement model which integrates deterministic behavior with randomness in an attempt to mimic actual human behavior.

To achieve this, we model the user's quasi-deterministic intercell movement by editing her deterministic movement patterns with insertion, deletion, and changing operations. Additionally, we model the mobile's micro movement as a non stationary process with dynamic states that are nonlinearly related to a time-correlated Gaussian process whose mean value behaves as a semi-Markov process. Based on this model, we develop a set of low-complexity recursive mobility prediction algorithms that are trivial to implement in real-time systems. HLP is composed of an approximate pattern-matching algorithm that extracts any existing regular movement pattern to estimate the global intercell direction, and uses an extended self-learning Kalman filter that deals with "unclassifiable" random movements by tracking intracell trajectory and predicting the next-cell crossing. We study the performance of our prediction algorithms in the presence of path loss, shadow fading, and random user movements. Simulation results and performance analysis show that our algorithm is robust in the presence of noisy input, being able to predict the speed and direction of travel of the mobile with high degree of accuracy.

With good next-cell prediction, algorithms that improve handoffs, relieve congestion, provide advance resource reservations and advance optimal route establishment, and which improve the overall QoS in wireless ATM networks can be built easily.
The rest of the paper is organized as follows. In Section II, we describe the proposed stochastic movement model. In Section III, we introduce the hierarchical location-prediction HLP) algorithm. In Sections IV and V, we present our simulation results, and provide prediction performance analysis. In Section VI, we present potential ways of incorporating HLP for connection and location management in wireless ATM systems. Finally, we conclude with some discussions in Section VII.

5.2 User Mobility Model

The mobility model we advocate in this paper attempts to mimic human (operator) movement behavior. Our model is built as a two-level hierarchy in which the top level is the global mobility model or GMM whose resolution is in terms of cells crossed by the mobile during the lifetime of the connection, and the bottom level is the local mobility model (LMM), whose resolution is in terms of a 3-tuple sample space (speed, direction, position) that varies with time. Stated another way, GMM is a deterministic model that is used to create intercell movements, while LMM is a stochastic model with dynamically changing state variables to model intracell movement. The LMM model interacts with the GMM model to create a semi random movement trajectory for the mobile. Our complete model (LMM + GMM) is based on the observation that the directional movement of mobiles is generally not ad hoc.

The variations (or perturbation) from a particular path occur when the mobile dynamically adjusts trajectory to compensate for the changing environment (e.g., changing traffic conditions, unexpected roadblocks, weather conditions, etc.). These perturbations are recreated in our model by the intracell stochastic behavior which affects the intercell movement pattern with varying degrees of influence.

5.2.1 Global Mobility Model

The global mobility model, as shown in Fig. 1(a), is motivated by the fact that most mobile users exhibit some regularity in their daily movement, and this regularity can best be characterized by a number of user mobility patterns (UMP's), recorded in a profile for each user and indexed by the occurrence time. The UMP's we proposed are similar to the movement patterns, but are more robust in the sense that we decrease the UMP's sensitivity to small deviations from the user's actual path (UAP). We are able to do this while maintaining their effectiveness for estimating the mobile's intercell directional movement intention by modeling UAP as the edited version of a UMP, and employing an approximate pattern-matching technique to find the UMP that most resembles UAP as explained in Section III. Consequently, the number of UMP's needed to span the network is greatly reduced, which in turn substantially reduces the time needed for pattern classification.

The degree of resemblance of a UAP with a UMP is measured by the edit distance, a well-known metric for finite string comparison. The simplest way to find this distance is by determining the smallest number of insertions, deletions, and changes by which the two finite cell sequences can be made alike. In order to reflect the geometric relationship between UAP and UMP in the physical domain, we assign a nonnegative number to each edit operation as the spatial weight. As a result, the edit distance between a UAP and a UMP becomes the sum of the weights of the editing operations which are chosen not only to make the two cell sequences alike, but also to have the smallest possible total weight. If the edit distance is less than a matching threshold, an approximately matched UMP is found, indicating the general moving intention of the user.

For large systems or systems with complex network topologies, the calculation of the spatial weights can be quite involved. In this work, for the sake of clarity and without losing generality, we limit the temporary deviation of a mobile user within the neighboring cells of the edited UMP. Specifically, we define the weight as follows.

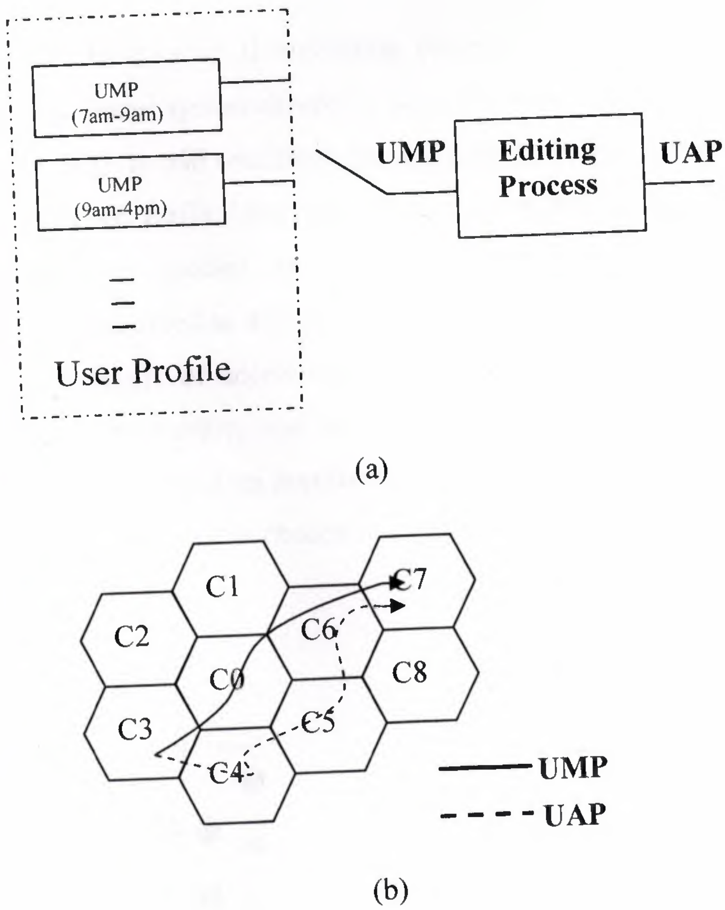
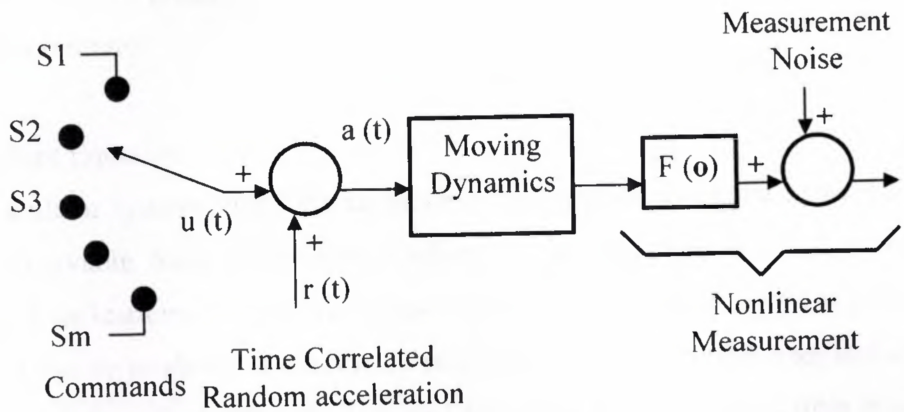


Figure 5.1 Global mobility model.

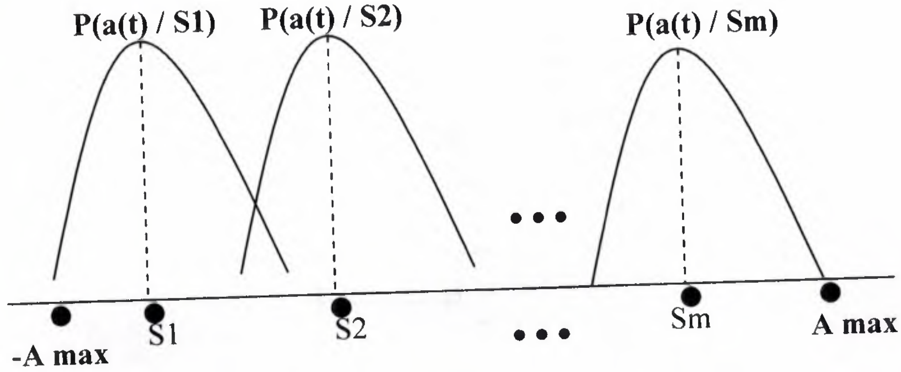
5.2.2 Local Mobility Model

Our motivation behind creating a local mobility model is based on the observation that the seemingly random choice of intercell movement is actually a logical function of the user's position, speed, direction, and cell geometry. The user mobility models found in the literature assume straight line movement or constant speed, which does not reflect reality.

In order to develop time-varying movement patterns, we model a moving user as a dynamic linear system driven by deterministic command and random acceleration shown in Fig. 2(a). In real situations, the acceleration range for a moving user can be fairly wide. Furthermore, traffic lights and road turns can lead to abrupt changes in speed. In order to recreate such sudden and unexpected changes, while covering the wide acceleration range, is modeled as a semi-Markov process with a finite number of "states" as possible discrete levels of acceleration. A semi-Markov process implies the Markovian state transition probability and random duration of time in one state prior to switching to another state. Random acceleration is modeled as a zero-mean Gaussian random variable with a variance that is chosen to cover the "gap" between adjacent acceleration states.



(a)



(b)

Figure 5.2 local mobility models.

The following two subsections present the mathematical description of the local mobility model in terms of dynamical equations and measurement equations.

1) Dynamical Equations for a Moving User:

Based on the model described above, the dynamical equations are derived for continuous-time movement, and are then expressed in discrete time according to the standard discretization procedure, thereby providing accurate statistical representation of the movement behavior.

2) Measurement Equations:

In existing cellular systems, the distance between the mobile and a known base station is practically observable. Such information is inherent in the forward link RSSI (or received signal strength indication) of a reachable base station. Measured in decibels at the mobile station, RSSI can be modeled as the sum of two terms: one due to path loss, and another due to shadow fading. Fast fading is neglected assuming that a low-pass filter is used to attenuate Rayleigh or Rician fade.

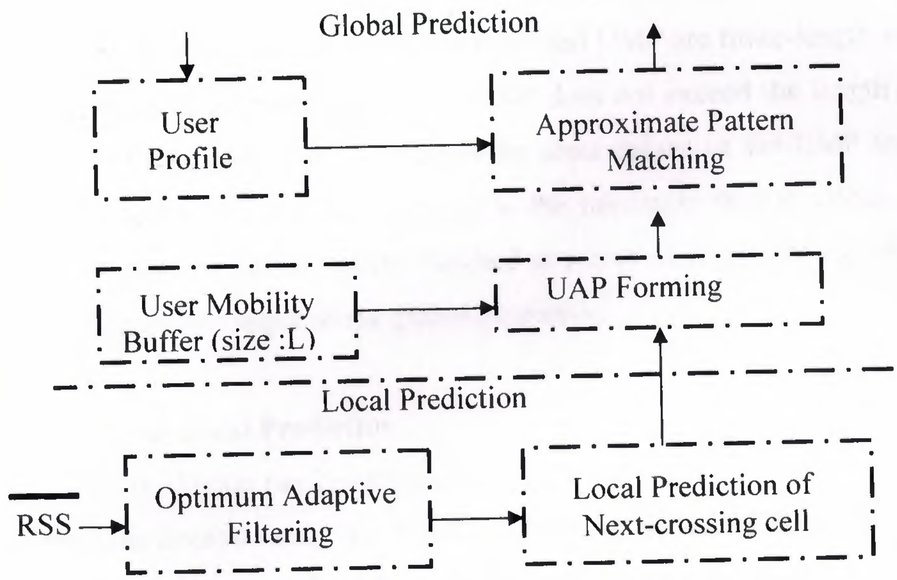


Figure 5.3 Mobility predictions.

5.3 The Hierarchical Location Prediction Algorithm

The algorithm proposed for mobility prediction is illustrated in Fig. 3. As shown, user mobility prediction is carried out at two levels—local prediction (LP) and global prediction (GP). LP provides a best estimate of the next cell to be crossed based on instantaneous trajectory tracking and cell geometry, while GP identifies the overall movement pattern of the mobile. With LP, a high degree of accuracy for next-cell prediction is achieved without any assumption of the user's mobility history. Furthermore, the UMP identification error is greatly reduced since GP uses the prediction data from LP to look ahead before making a decision on the best matched pattern.

5.3.1 High-Level Global Prediction

For global prediction, we solve the following problem: given a number of UMP's for the current time interval, a UAP which reflects the current movement trend of the mobile, find the UMP that best describes the UAP.

A UAP is composed of recently crossed cells, stored in the mobile's user mobility buffer, and an LP for the next cell (if any). Both UAP and UMP are finite-length sequences, and we assume that the memory length of a UAP does not exceed the length of the largest stored UMP. Consequently, it is possible for some subset of the UMP to resemble the UAP in the sense that the edit distance is the minimum and less than the matching threshold. If such an approximately matched sequence is found, the remaining sequence in UMP becomes the output of the global prediction.

5.3.2 Low-Level Local Prediction

Local prediction can be achieved by two steps:

1) Estimate the dynamic state of a moving user using subsequent RSSI measurements,

2) Select the neighboring cell with maximum cell-crossing probability as LP output based on dynamic state estimation and cell geometry.

For the first step, an optimum adaptive filter is needed for real-time dynamic state estimation. As discussed previously, since the instantaneous sampled trajectory of the mobile is a nonstationary process, a modified Kalman filter seems to be the best candidate for achieving low-level local prediction. The conventional Kalman filter has to be modified since it is not designed for deterministic input which we represent as a semi-Markovian process with possible (mobile acceleration) states. We overcome the problem of dealing with the randomness within the states by using a bank of filters with each filter operating on a possible state. Fortunately, as discussed in, when certain practical assumptions are made, the filter bank can be reduced to a single Kalman filter augmented by a recursive technique of estimating. The adaptive state estimator then becomes (see part B of the Appendix for derivation)

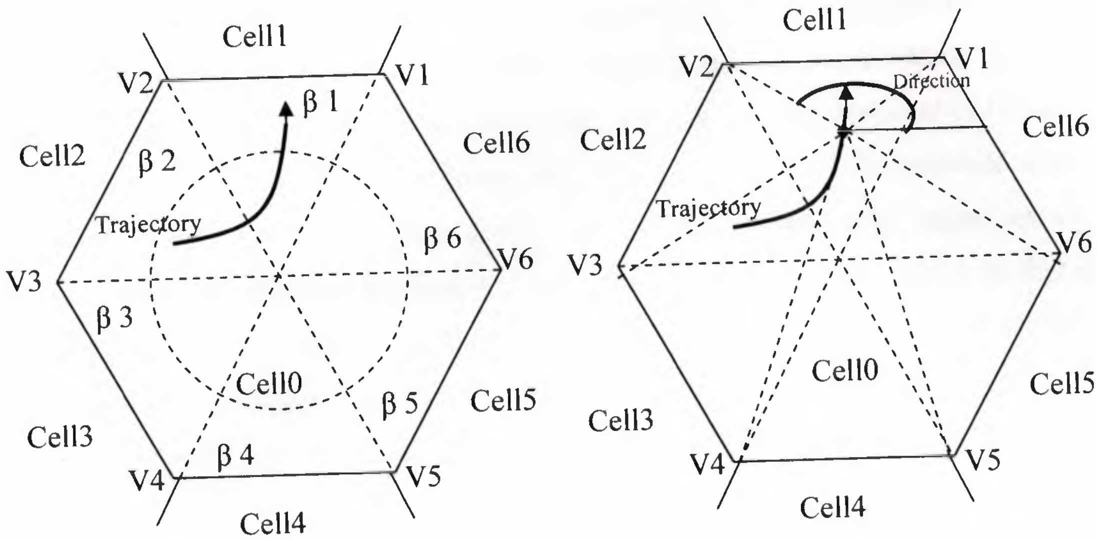


Figure 5.4 Cell geometry.

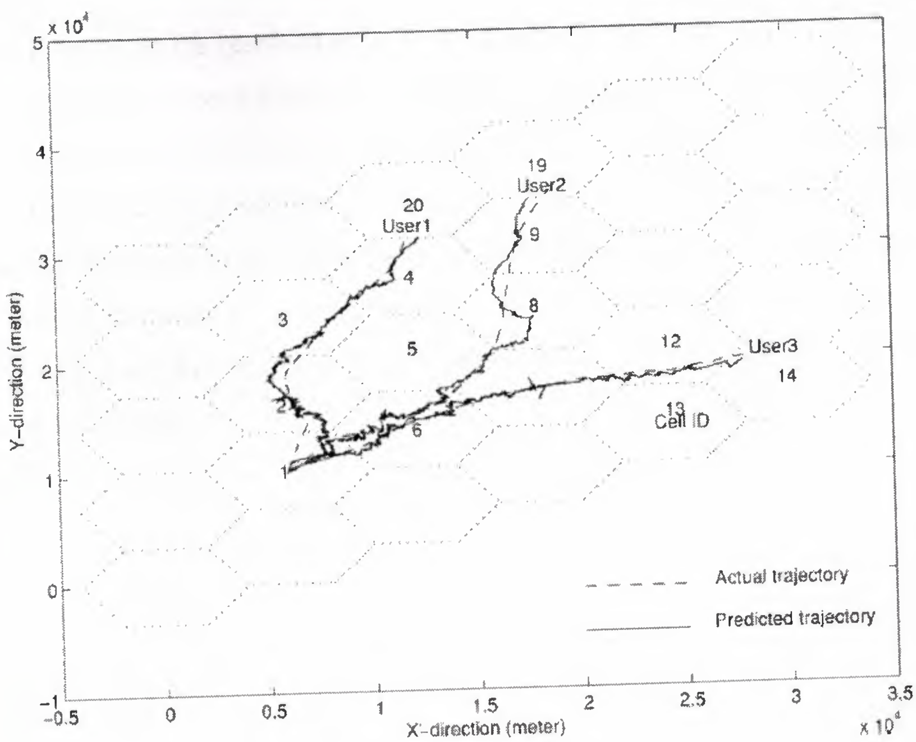


Figure 5.5 Actual and predicted user trajectories.

5.4 Simulation and Results

To examine the performance of HLP, a simple simulation is carried out for the conventional hexagon cell environment. The simulated service area contains 30 base stations with cell radius of 4 km, as shown in Figs. 6 and 8. No constitutional constraints are assumed, and the mobile is permitted to move to any cell in the network along random trajectories with no constant speed. The moving dynamics are based on the movement model discussed in Section II, except that the acceleration is preset so that a known trajectory can be achieved for the purpose of performance tests. On-line mobility-related information includes recently crossed cells stored in a user mobility buffer and subsequent forward link RSSI measurements of the current cell and two neighboring cells.

5.4.1 Results from Local Prediction

Figure 5.5 shows the results from trajectory tracking of three moving users. The dashed curve depicts the actual trajectory UAP), and the solid curve shows the predicted trajectory.

Fig. 6 demonstrates the result of time-varying velocity prediction. The initial value of the dynamic state is estimated from the averaged RSSI value with a position error up to 1000 m and a speed error up to 5 m/s. Because of the strong "pullin" power of the filter, it turns out that the adaptive filter is relatively insensitive to the initial conditions. The result of LP for the next-cell crossing is summarized in Table II with prediction ratios = 75, 80, 100% for users 1, 2, and 3, respectively. Here, the prediction ratio is defined as the ratio of the number of cells correctly predicted to the total number of cells need to be predicted in the path.

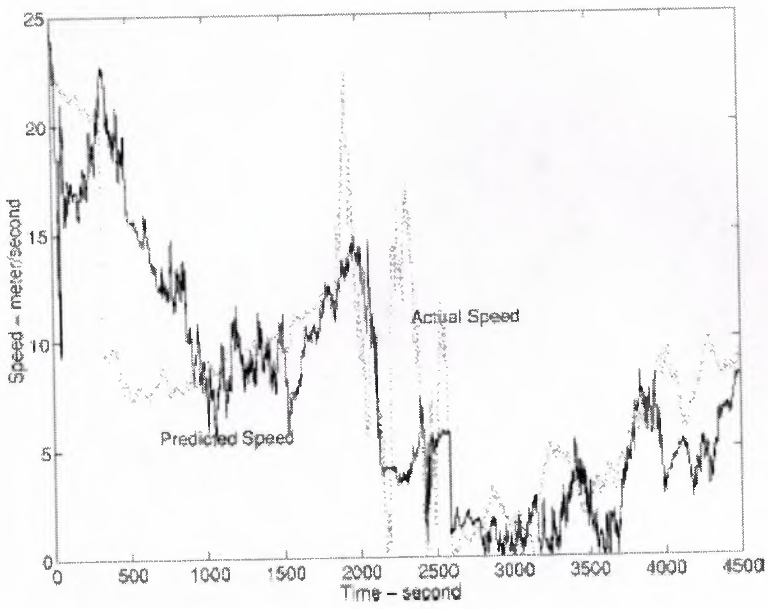


Figure 5.6 Actual and predicted user speed.

5.4.2 Results from Global Prediction

Figure 5.7 shows the results from global prediction, where a moving user is simulated with two possible UMP's, UMP and UMP. In order to identify the UMP from the tracked UAP, a recursive approximate pattern-matching algorithm is implemented with the matching threshold as 3 and the mobile's user mobility buffer size as 4, which means that the GP output is available when the user reaches the fourth cell .

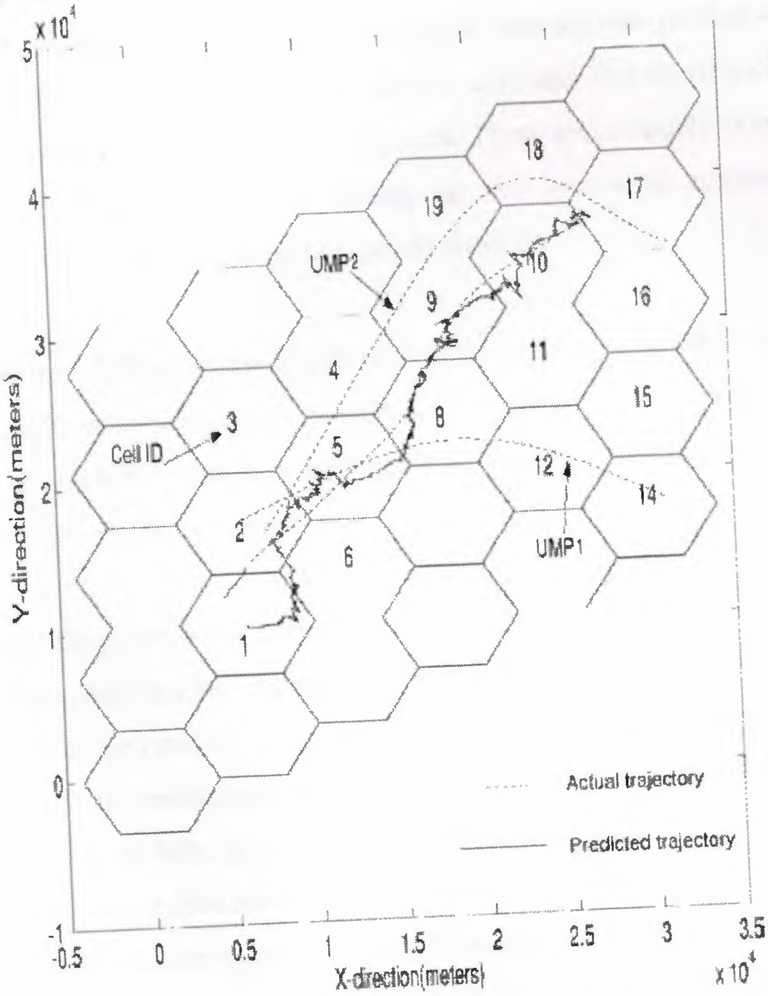


Figure 5.7 Actual and predicted user trajectories with multiple potential UMP's.

5.5 Systems Implementation

We now present two examples to illustrate how HLP can be applied within a wireless ATM system. The examples include a prediction-based dynamic virtual connection tree PVCT) strategy for providing enhanced connection management functionality, and a prediction-based dynamic location update strategy.

For the sake of this explanation, minimal system architecture is assumed. Specifically, we assume that the wireless ATM system consists of: 1) a wired backbone infrastructure that supports ATM packet transport; 2) a collection of radio base station Tran receivers, each of which is connected to the fixed wired infrastructure, and each of which supports on-demand packet access to a shared radio channel the base stations provide connectivity to all mobiles within their area of influence called a cell; and 3) a set of mobile users who are equipped with portable RF wireless terminals. These users (mobiles) are free to roam anywhere in the wireless network, relying on the system to maintain continuous connectivity even as they move from one cell to another.

To implement the HLP algorithm in wireless ATM systems, we notice that both Kalman filtering and the approximate pattern-matching algorithm have a recursive nature and modest use of memory storage, which make it feasible to be implemented as a low-power, embedded software process.

To reduce signaling traffic, distribute computation load, and improve system security while still achieving location and speed prediction, we suggest implementing the HLP algorithm within the mobile hosts rather than in the base stations. By storing the historical long-term movement patterns (UMP's) and the current movement history (UAP) within the mobile, security against unwanted predictors and locators can be provided. Additional information needed by the HLP algorithm to function correctly, such as cell identification numbers and cell topology, can be obtained from the base station broadcasting on a protected control channel at the beacon frequency in each cell. The results from HLP are transmitted to the network infrastructure for predictive mobility management.

5.5.1 Prediction-Based Dynamic Virtual Connection Trees

As explained in Section I, the virtual connection tree (VCT) approach maintains QoS by minimizing latency during handoffs initiation and completion, by minimizing cell loss, and by reducing processing delay during handoff. The strength of the VCT is in the admission control and rerouting algorithms which are fairly simple to implement. The primary drawback of the VCT approach is its inefficient use of network resources, a potential for causing system overload, and a long setup processing time needed for assigning virtual connection (VC) numbers on a typically large area. System inefficiencies occur since the preassigned VC's do not accurately account for the mobile's current and projected movement patterns. Consequently, many preassigned VC's are wasted, and efficient resource reservation cannot be achieved. While keeping the advantages of the VCT approach, we propose to reduce its disadvantages by dynamically allocating VC's based on location prediction. When a matched UMP is found, the mobile informs the current serving base station of the results of the HLP algorithm, which include identifiers for the cells in the path of the mobile's predicted trajectory and the mobile's predicted speed (cell-dwell time) in these cells.

Using this information, the system sets up end-toned connections with appropriate resources from the base stations in the predicted cells. A set of virtual circuit identifier numbers (VCN's) is associated with these connections and passed on to the mobile. Each of these connections is maintained by the system (VCN stays valid) for a time duration determined from the mobile's predicted velocity and predicted cell dwell time. In the case when no UMP is found to match the UAP and only information from LP is available, VCN's of the most likely neighboring cells are assigned to the mobile. By following the fairly accurate hints about the mobile's trajectory, provided by the HLP, unwanted cell overloading is reduced, admission control is faster, and system efficiency is superior since only a subset of the VCT footprint (cluster of base stations) is involved in resource reservation for each ATM connection. Thus, the PVCT approach trades off minimal QoS reliability for vastly improved system efficiency when compared to the native exhaustive VCT approach.

5.5.2 Prediction-Based Dynamic Location Update

HLP is based on the belief that, at a global level, a user's movement pattern is fairly regular, and can be approximated by a representative UMP. (Implicit in this statement is the fact that, with time, the user "learns" and stores the regular patterns of its movement.) Therefore, for the purpose of location tracking, the predefined zones, as proposed in the two-tier architecture, can dynamically be mapped to the matched UMP obtained from the GP.

As long as the mobile follows the assumed UMP, no location update is necessary since the current location server can preinform the rest of the predicted location servers through the wired network. Therefore, the mobile can continue to use the services without conventional registration and location update procedures when it arrives at the zone. In the special case when no UMP is found to match UAP, the conventional location update method can be enabled to collect location information with a lower periodicity. As most mobile users are quite regular in their daily movement, signaling traffic due to location update can be significantly reduced. [9]

1.6 Conclusion

We have explored the fundamental problem of providing lifetime connectivity to ongoing sessions initiated by mobile users in a cell-based wireless ATM network. The motivation behind this research is derived from the recognition that the performance of the mobility management subsystem is key to any QoS-based wireless ATM network. The approach we took was to develop a robust algorithm for predicting the future locations and speeds of the mobiles. In order to develop such an algorithm, we proposed a novel human-centric pseudo stochastic mobility model that rejects the notion that all movement is ad hoc. With this refined observation of human mobility behavior, we developed a two-level hierarchical location prediction (HLP) algorithm for accurate prediction of the cells the mobile will cross during the lifetime of its connection. The HLP algorithm had two main components in it: 1) global prediction and 2) local prediction.

For global prediction, we proposed an approximate pattern matching algorithm good for any finite sequence comparisons. We showed that this algorithm improves the prediction accuracy by abstracting the geometric similarity between two cell sequences which may otherwise seem dissimilar. Our method effectively enlarges the area covered by a UMP, making pattern classification flexible and robust, while reducing memory requirements significantly. Only a few UMP's have to be stored as they can represent a large number of mobility patterns, needed for accurate intercell prediction.

For local prediction, we applied classical stochastic signal processing techniques to extract user mobility information from noisy measurements. Analysis and simulation results proved that a self-adaptive extended Kalman filter provides a high degree of accuracy for next-cell location and instantaneous speed prediction. Local prediction in HLP was kept independent of global prediction (vice versa is not true) so that reasonably accurate short-distance prediction can be obtained even when the system has no knowledge of the user's historical mobility patterns.

Finally, HLP is independent of the architecture of the underlying wireless ATM system. Two strategies, prediction-based dynamic virtual connection trees (PVCT) and prediction-based dynamic location update, were presented as examples of deployable predictive mobility management in wireless ATM networks.

6. LOCATING IN CELLULAR MOBILE INTERNET

6.1 Overview

The global Internet is growing at a tremendous rate. At the same time, portable computing devices such as laptop and palmtop computers, featuring powerful CPUs, large main memories, hundreds of megabytes of disk space, multimedia sound capabilities, and color displays, are becoming widely available at very affordable prices, and many new wireless networking products, protocols, and services are becoming available based on technologies such as spread spectrum radio, infrared, cellular, satellite, CDPC, GPRS and WAP. High-speed local area wireless networks are commonly available with speed up to 2 megabits per second, and wide-area wireless networks are available that provide metropolitan or even nationwide service.

With these dramatic increases in portability and ease of network access, it becomes natural for users to expect to be able to access the Internet at any time and from anywhere, and to transparently remain connected and continue to use the network as they move about. However, these "mobile hosts" in mobile Internet can not currently interoperate easily or conveniently with internetworking protocols such as IP [1] due to the operation of existing internet work addresses and routing algorithms. For example, in IP, the network address of a host is divided into two levels of hierarchy: a network number identifying the network to which the host is connected, and a host number identifying the particular host within that network. IP expects to be able to route a datagram to a host based on the network number contained in the host's IP address. If a host changes its point of connection to the Internet and moves to a new network, IP datagram's destined for it will no linger reach it correctly.

To address this need in the Internet, the Mobile IP Working Group of the Internet Engineering Task Force (IETF) has been working over the past few years to develop standard protocols to support mobile hosts operating in the Internet [2, 3, 4, 5]. Locating in mobile Internet of mobile hosts is a main problem in this protocol. The mobile host will attempt to discover a foreign agent within the network being visited, using an agent discovery protocol. Locating hosts in Cellular Mobile Internet using "Agent Discovery", "Vibrating Registration" problem may arise. Section 2 of this paper describes the infrastructure and agent discovery of the mobile IP protocol. Section 3 presents the basic concept of cellular mobile Internet and the register processes of mobile hosts. Section 4 gives the method named "Addresses Sliding Window" to solve the "Vibrating Registration" problem. Section 5 gives the summary.

6.2 The Mobile IP Protocol

6.2.1. Infrastructure

In the basic IETF Mobile IP protocol, each mobile host is assigned a unique home address in the same way as any other Internet host, within its home network. Hosts communicating with a mobile host are known as correspondent hosts and may, themselves, be either mobile or stationary. In sending an IP packet to a mobile host, a correspondent host always addresses the packet to the mobile host's home address, regardless of the mobile host's current location.

Each mobile host must have a home agent on its home network that maintains a registry of the mobile host's current location. This location is identified as a care-of address, and the association between a mobile host's home address and its current care-of address is called a mobility binding, or simply a binding. Each time the mobile host establishes a new care-of address, it must register the new binding with its home agent so that the home agent always knows the current binding of each mobile host that it serves. A home agent may handle any number of mobile hosts that share a common home network.

A mobile host, when connecting to a network away from its home network, may be assigned a care-of address in one of two ways. Normally, the mobile host will attempt to discover a foreign agent within the network being visited, using an agent discovery protocol. The mobile host then registers with the foreign agent, and the IP address of the foreign agent is used as the mobile host's care-of address. The foreign agent acts as a local forwarder for packets arriving for the mobile host and for all other locally visiting mobile hosts registered with this foreign agent. Alternatively, if the mobile host can obtain a temporary local address within the network being visited (such as through DHCP [4]), the mobile host may use this temporary address as its care-of address.

While a mobile host is away from its home network, a mobile host's home agent acts to forward all packets for the mobile host to its current location for delivery locally to the mobile host. Packets addressed to the mobile host that appear on the mobile host's home network must be intercepted by the mobile host's home agent, for example by using ``proxy" ARP or through cooperation with the local routing protocol in use on the home network.

For each such packet intercepted, the home agent tunnels the packet to the mobile host's current care-of address. If the care-of address is provided by a foreign agent, the foreign agent removes any tunneling headers from the packet and delivers the packet locally to the mobile host by transmitting it over the local network on which the mobile host is registered. If the mobile host is using a locally obtained temporary address as a care-of address, the tunneled packet is delivered directly to the mobile host.

Home agents and foreign agents may be provided by separate nodes on a network, or a single node may implement the functionality of both a home agent (for its own mobile hosts) and a foreign agent (for other visiting mobile hosts). Similarly, either function or both may be provided by any of the existing IP routers on a network, or they may be provided by separate support hosts on that network.

6.2.2 Agent Discovery

The agent discovery protocol operates as a compatible extension of the existing ICMP router discovery protocol. It provides a means for a mobile host to detect when it has moved from one network to another, and for it to detect when it has returned home. When moving into a new foreign network, the agent discovery protocol also provides a means for a mobile host to discover a suitable foreign agent in this new network with which to register.

On some networks, depending on the particular type of network, additional link-layer support may be available to assist in some or all of the purposes of the agent discovery protocol. A standard protocol must be defined for agent discovery, however, at least for use on networks for which no link-layer support is available. By defining a standard protocol, mobile hosts are also provided with a common method for agent discovery that can operate in the same way over all types of networks. If additional link-layer support is available, it can optionally be used by mobile hosts that support it to assist in agent discovery.

Home agents and foreign agents periodically advertise their presence by multicasting an agent advertisement message on each network to which they are connected and for which they are configured to provide service. Mobile hosts listen for agent advertisement messages to determine which home agents or foreign agents are on the network to which they are currently connected. If a mobile host receives an advertisement from its own home agent, it deduces that it has returned home and registers directly with its home agent. Otherwise, the mobile host chooses whether to retain its current registration or to register with a new foreign agent from among those it knows of.

While at home or registered with a foreign agent, a mobile host expects to continue to receive periodic advertisements from its home agent or from its current foreign agent, respectively. If it fails to receive a number of consecutive expected advertisements, the mobile host may deduce either that it has moved or that its home agent or current foreign agent has failed. If the mobile host has recently received other advertisements, it may attempt registration with one of those foreign agents. Otherwise, the mobile host may multicast an agent solicitation message onto its current network, which should be answered by an agent advertisement message from each home agent or foreign agent on this network that receives the solicitation message.

3. Cellular Mobile Internet

3.1 Basic concept

Mobile network is a kind of network in which hosts can move about freely, and his movement is transparent to users. These hosts are called mobile hosts (MH). Mobile nternet is based upon the existing Internet and wireless communication technologies that he hosts can roam within the Internet. Cellular mobile Internet is applying the cellular echnology such as GSM to extend the existing Internet structure. The network topology f cellular mobile Internet is depicted in Figure 1.

Cellular mobile Internet is composed of many basic Cellular Areas (CA), and these areas may overlap each other. The shape and arrangement method of these CAs can be learned from GSM. Each CA is composed Wireless Access Point (WAP) and MHs which are now visiting this CA. WAP may be Home Agent or Foreign Agent. MH can move about within a single CA and can traverse a CA to an adjoining CA.

6.3.2 Registration process

Now we assume that WAP2 and WAP3 are MHi's foreign agents. When MHi moves along the route R1 in Figure 6.1, MHi may receive agent advertisement message from WAP2 and WAP3. Because the R1 is a curve line, MH may first receive agent advertisement message from WAP2 and after a while may receive agent advertisement message from WAP3.

According to the agent discovery in Mobile IP, when establishing with a new foreign, a mobile host must register with that foreign agent, and must also register with its home agent to inform it of its new care-of address. When instead establishing a new temporarily assigned local IP address as a care-of address, a mobile host must likewise register with its home agent to inform it of this new address. Finally, when a mobile host returns to its home network, it must register with its home agent to inform it that it is no longer using a care-of address.

To register with a foreign agent, a mobile host sends a registration request message to the foreign agent. The registration request includes the address of the mobile host and the address of its home agent. The foreign agent forwards the request to the home agent, which returns a registration reply message to the foreign agent. Finally, the foreign agent forwards the registration reply message to the mobile host. When registering directly with its home agent, either when the mobile host has returned home or when using a temporarily assigned local IP address as its care-of address, the mobile host exchanges the registration request and reply messages directly to its home agent.

Each registration with a home agent or foreign agent has associated with it a lifetime period, negotiated during the registration. After this lifetime period expires, the mobile host's registration is deleted. In order to maintain continued service from its home agent or foreign agent, the mobile host must reregister within this period. The lifetime period may be set to infinity, in which case no re-registration is necessary. When registering with its home agent on returning to its home network, a mobile host registers with a zero lifetime and deletes its current binding, since a mobile host needs no services of its home agent while at home.

6.3.3 Problem arising from alternate registration

Following is the registration process of MHi traverse the overlapped area along the route R1. At the time t1, MHi receives the agent advertisement message from WAP3, then sends a registration request message to the foreign agent WAP3. At the time t2 (t2-t1>lifetime), Mhi receives the agent advertisement message from WAP2, then sends a registration request message to the foreign agent WAP2 and deletes its registration of WAP3. At the time t3, MHi receives the agent advertisement message from WAP3 again, then sends a registration request message to the foreign agent WAP3 and deletes its registration of WAP2.

During the travel time along the R1, MHi registers to WAP2 and WAP3 alternately, so adds networks burden and wastes its power and CPU time. Because in the overlapped area of WAP2 and WAP3, MHi can receive the agent advertisement message from both of them. So in this case and similar case, MHi should judge to connect which foreign agent adaptively. To solve this problem named "Vibrating Registration", we present a new method named "Addresses Sliding Window".

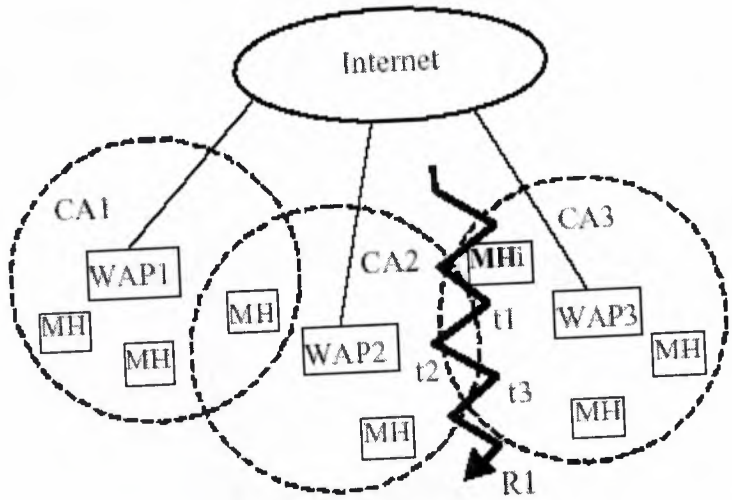


Figure 6.1 The topology of cellular mobile Internet

6.4 Addresses Sliding Window

The main idea of this method is that when A MH receives an agent advertisement message, it does not send a registration request message to the foreign agent immediately. After define time period, it may receive several agent advertisement messages. To realize the method, we need to define an "Addresses Queue" data structure in MH's memory. Addresses Queue stores the addresses of foreign agents. During a time period, such as ti to tj, MH puts several addresses into the Addresses Queue. Then MH judges the queue to determine to whether register or not.

For example, a MH's addresses queue is depicted as Figure2. In this example, we assume the addresses of two foreign are A and B separately. At the time ti, the current address in addresses queue of a MH is B, the MH does not send a registration request message to the foreign agent which address is B. At the time tj, the MH sends a registration request message to the foreign agent which is obtained from the judgement according to the addresses queue. For example, as depicted in Figure 2, the addresses queue is "B->A->B->A->B- >A->B->A", so MH can know that it's location may be in overlapped area of two agents. If a MH is locating in the overlapped area of two agents, it can receive datagram's from each of them, and then the MH may select an agent as its foreign agent to connect according to some strategies.

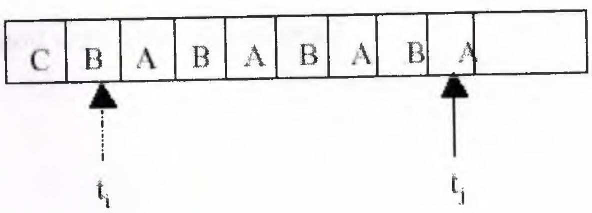


Figure 6.2 Addresses Queue

The size of the "Addresses Sliding Window" is determined by several factors such as the speed of wireless channel, the lifetime of a registration and air protocols.

In our research project named "HBH concise data network", we have implement the part of Mobile IP in Linux kernel. To solve the "Vibrating Registration" problem, we adopt "Addresses Sliding Window" and regard that the method is very effective. [9]

6.5 Summary

Recent increases in the availability of mobile computers and wireless networks provide the opportunity to integrate these technologies seamlessly into the Internet. Mobile users should be able to move about, transparently remaining connecting to the Internet, utilizing the best available network connection at any time, whether wired or wireless. In this paper we have described the infrastructure and the locating method named "Agent Discovery" of the Mobile IP protocol.

Locating hosts in Cellular Mobile Internet using "Agent Discovery", "Vibrating Registration" problem may arise. To solve this problem, we have presented a new method named "Addresses Sliding Window". The main idea of this method and detail example is given. In our research project named "HBH concise data network", we have implemented the method within the Linux kernel.

6. CONCLUSION

Mobile Computing is a new emerging computing paradigm of the future. Data Management in this paradigm poses many challenging problems to the database community.

The focus of Data Management for Mobile Computing is on the impact of mobile computing on data management beyond the networking level. Data Management for Mobile Computing provides a single source for researchers and practitioners who want to keep current on the latest innovations in the field.

Location management is a key issue in personal communication service networks to guarantee the mobile terminals to continuously receive services when moving from one place to another. The purpose is to provide a thorough and cohesive overview of recent advances in wireless and mobile data management.

7. REFERENCES

[1] Cox, D., “Wireless Personal Communications: What Is It? “, IEEE Personal Communications, Vol. 2, No. 2, pp. 20-35, April 1995.

[2] Kleinrock, L., “Nomadic Computing - An Opportunity”, ACM SIGCOMM, Vol. 25, No. 1, pp. 36-40, January, 1995.

[3]http://www.cs.ucy.ac.cy/courses/EPL651/Schedule___Readings/Chapter_4/body_chapter_4.html.

[4] http://www.privateline.com/archive/DataManagement.pdf.

[5] J. Li, Y. Pan, and X. Jia, “Analysis of Dynamic Location Management for PCS Networks,” IEEE Trans. Vehicular Technology, pp. 1109-1119, Sept. 2002.

[6] U. Madhow, M. Honig, and K. Steiglitz, “Optimization of Wireless Resources for Personal Communications Mobility Tracking,” IEEE/ACM Trans. Networking, vol. 3, no. 4, pp. 698-707, Dec. 1995.

[7] http://www.management.com/managementlocation/archive/application.pdf .

[8] http://www.sparknotes.com/wireless_atmnetwork/jsac98.doc.

[9]http://www.cs.ucy.ac.cy/courses/EPL651/Schedule___Readings/Chapter_2/body_chapter_2.html.