LBS Interoperability For Commercial Aspects

A Thesis Report For Submission To The

Central Department Of Computer Science and Information Technology (CDCSIT)

Tribhuvan University, Kathmandu, Nepal

Studied As Partial Fulfillment Of The Requirements For The Course CSC689
In
Master's Degree of Computer Science and Information Technology

By: Sudeep Manandhar (Roll No: 10)
[CDCSIT-TU, Nepal]
April, 2007

Tribhuvan University

Central Department of Computer Science

And Information Technology

(CDCSIT)

Date: _______________

LETTER OF RECOMMENDATION

I certify that I've read the dissertation entitled "**LBS Interoperability For Commercial Aspects**" , carried out by Mr. Sudeep Manandhar, and that in my opinion it is fully adequate, in scope and quality, as a dissertation for the degree of Masters of Computer Science and Information Technology. This dissertation is carried out direct under my supervision and guidance. I, therefore recommend for further evaluation.

____

Mr. Lochan Lal Amatya

Deputy Manager

Central Office Bhadrakali Plaza

Nepal Telecom (NTC), Nepal

(Supervisor)

Tribhuvan University

Central Department of Computer Science

And Information Technology

(CDCSIT)

Date: ________________

LETTER OF APPROVAL

We certify that we have read this project document and in our opinion it is satisfactory in the scope and quality as a project work in the partial fulfillment for the requirement of Master's Degree In Computer Science And Information Technology.

Evaluation committee:

____

Acting Head, Mr. Min Bahadur Khati
CDCSIT-TU, Nepal

____

Mr. Lochan Lal Amatya
Central Office Bhadrakali Plaza
Nepal Telecom (NTC), Nepal

____

(Internal Examiner)

____

(External Examiner)

Abstract

Geographical Information System (GIS) and Global Positioning System (GPS) technologies are expanding their traditional applications to embrace a stream of consumer-focused, location-based applications. Through an integration with handheld devices capable of wireless communication and mobile computing, a wide range "Location-Based Services" (LBS) may be offered to mobile users.

A location-based service is able to provide targeted spatial information to mobile workers and consumers depending upon their location. These include utility location information, personal or asset tracking, concierge and route-guidance information, is possible because of LBS. The technologies and applications of LBS will play an ever increasingly important role in the modern, mobile, always-connected society.

LBS aims to provide specific, targeted information to users based on each specific user's location at any time. The security and safety are also important considerations for a "mobile society". The second type of application is concierge services where location-specific information on something nearby is sought. The requested information may be related to points-of-interest such as hospitals, restaurants, cinemas, car parks, ATMs, and so on. Such a service may provide information about the point-of-interest, or route-guidance to find it.

There are also many examples of applications in typical work practices which might not be viewed as a form of commercial aspects, but which nevertheless benefit from using procedures that have a "spatial component", useful for M-business purpose. A mobile device can be used to query a GIS database that would permit the isolation of the customer requirements at the appropriate location. In essence, LBS can only be provided through the integration of wireless communications and computing technologies, with 'spatial elements' such as positioning technologies and spatial data sets. These components form a network using wireless communication standards to transfer service requests and information between a mobile user and a service (or server) facility. The location-based service facility is able to perform spatial functions based on the user's location, generally with the aid of a Geographical Information System.

Acknowledgements

I would like to express my sincere gratitude to the Central Department of Computer Science and Information Technology (CDCSIT) for insisting and encouraging me to prepare this dissertation report in the field of Location Based Services and Mobile GIS.

I am also greatly indebted to my thesis advisor Mr. Lochan Lal Amatya, Chief Executive Officer, of Nepal Telecom (NTC), Jawalakhel, Lalitpur, Nepal, for his invaluable assistance, suggestion, guidance, and full support. He was always there to listen and to give advice. He is responsible for leading me up to this position to create this dissertation report.. He taught me how to deal with the topic I've chosen and to express the ideas. He showed me different ways to approach a research problem and the need to be persistent to accomplish any goal.

Besides my advisor, I would like to thank from the bottom of my heart to the rest of my thesis evaluation committee: Prof. Dr. Devi Dutta Paudyal and Prof. Dr. Subarna Shakya, who asked me good questions and rescued me from various red tape crisis, encouragement and hard questions, and to all my professors, lecturers and teachers namely Prof. Dr. Laxmi Pd. Gawali, Prof. Dr. Shree Nath Shrivasa, Mr. Sudarshan Karanjit , Mr. Arun Timilsina, Laxmi etc. who gave insightful comments and reviewed my work.

I express deep gratitude to my wife, who have always encouraged and supported for this achievement. This small piece of work is affectionately dedicated to her. I cannot leave without thanking my deep appreciation especially to my friends Mr. Dipesh Man Joshi, Mr. Rajendra Maharjan and to all my well wisher friends.

Finally, I anticipate this dissertation will serve to explore the ideas on implementing Location Based Service using Mobile GIS.

Sudeep Manandhar

Roll No. 10

CDCSIT-TU, NEPAL

List of Figures

List of Tables

List of Abbreviations

CDPD	:	Cellular Digital Packet Data
GSM	:	Global System for Mobile Communication
GPRS	:	General Packet Radio Service
OO	:	Object Oriented
LBS	:	Location Based Services
WWW	:	World Wide Web
FCC	:	Federal Communications Commission
GIS	:	Geographic Information System
SA	:	Selective Availability
GPS	:	Global Positioning System
DGPS	:	Differential GPS
AGPS	:	Assisted GPS
E-OTD	:	Enhanced Offset Time Division
TOA	:	Time Of Arrival
AOA	:	Angle Of Arrival
IN	:	Intelligent Network
SMS	:	Short Message Services
CDMA	:	Code-Division Multiple Access
TDMA	:	Time Division Multiple Access
PDC	:	Personal/Pacific Digital Cellular
CGI	:	Cell Global Identity
W-CDMA	:	Wideband-CDMA
UMTS	:	Universal Mobile Telecommunications System
3G	:	Third Generation
PDE	:	Position Determination Equipment

GIS	:	Geographic Information System
ANI	:	Automatic Number Identification
ALI	:	Automatic Location Information
E-OTD	:	Enhanced Observed Time Difference
OTDOA	:	Observed Timed Difference of Arrival
MSC	:	Mobile Switching Center
HLR	:	Home Location Register
VLR	:	Visitor Location Register
STP	:	Signal Transfer Points
MA	:	Mobile Agent
MIDP	:	Mobile Information Device Profile
J2ME	:	Java 2 Micro Edition
J2EE	:	Java 2 Enterprise Edition

Table of Contents

1. Introduction

Location Based Services (LBS) provide personalized services to the subscriber based on their current position. In this age of significant telecommunication competition Location based services open new horizons for cellular operator for provisioning of innovative lucrative value added services. Location services provide information specific to a location. Location based services employ accurate, real-time positioning to connect users to nearby points of interest, advise them of current conditions such as traffic and weather, or provide routing and tracking information--all via wireless devices. The Location of mobile radios first appeared in military systems developed during the Second World War. The idea was simple: to find people in distress, or to detect and eliminate people causing distress. US Department of Defense launched a series of Global Positioning Satellite (GPS) to support Military operations. Recently, a number of commercial GPS system have been developed to provide location information to mobile user, with application for navigation, military targeting, and emergency assistance.

1.1 Background

A category of applications, which is known variously as Location-Based Services (LBS), Location Commerce (or L-commerce), mobile commerce, mobile location services, wireless location, and similar terms, is now emerging rapidly in the Geospatial Information marketplace. By any name, the purpose and character of LBS remains the same: employing accurate real-time position information of users to connect them to nearby points of interest (such as retail businesses, public facilities, or travel destinations), to advise them of current conditions (such as traffic and weather), or to provide routing and tracking services. At the intersection of Web, wireless communication and Geographic Information System (GIS) technologies, Location Based Services are aimed at giving everyone the ability to exploit location information anywhere, anytime, and on any device. LBS are expected to create a new global market – in both business-to business and business-to-consumer services – with annual revenues well into double the initial capital within a few years. The market for Location Based Services is rich with commercial services for global markets. The applications for LBS are numerous, such as E911, logistics, vehicle automation, real estate, field service, travel service, real-time navigation, and so on . LBS technology is creating an emerging market with huge revenue potential. According to the research firm Analysys Inc., revenues from the provision of Location Based Services will be worth $18 billion worldwide by 2006. The report of Allied Business Intelligence Inc (ABI) indicates that global LBS revenues will grow from approximately $1 billion in 2000 to over $40 billion in 2006. This growth will represent a compound annual average growth rate of 81%

The explosion of LBS should be attributed to the revolutionary advancements in Global Positioning Satellite System (GPS), distributed GIS, handheld client device, database, wireless network, communication protocol and the Internet in recent years. With the integration of these technologies, Location Based Services open the door to opportunities in virtually every discipline of every industry. Among all the foresaid technologies, wireless communication is regarded as key for LBS, since the essential of LBS is using location to deliver targeted applications to users, most of which are mobile, at their moment of need. The explosion of LBS results in fast increasing requirements for software. The diversification of the market significantly increases requirements for software. Moreover, wireless communication technologies evolve so fast that the corresponding software has to be updated frequently to catch up with the advancements. Furthermore, the fact that it is lack of semi-custom solutions for wireless communication in the market forces the application developers to develop their programs from scratch. As a result of continuously increasing software requirements, the growth of LBS applications will result in a software crisis if no action is taken. The outcomes of a software crisis, such as lack of Highly Qualified Personnel (HQP), increased development costs and time, and degradation of software quality, will make the LBS application developers incompetent to respond to the market requirements. The best solution for the software crisis up to now is to increase software reusability, which has been demonstrated successful by practice. The LBS wireless communication software has lots of potential for software reuse. First, although wireless communication technologies are quite diverse, those that dominate the markets are relatively monotonous. For example, the commercial cellular telephone system dominates the wireless communication market, and fortunately, it can provide a relatively cheap service for both voice and data. It is not difficult to combine only a few popular wireless communication technologies to serve almost every type of LBS applications.

Second, the different types of wireless communication available on the market are highly complementary to each other, and this stimulates software developers of LBS applications to support various communication methods in their programs. For example, Cellular Digital Packet Data (CDPD) has a much greater effective transmission range via the widespread commercial telephone system network than by wireless radio modem, but one need not to pay for running the wireless radio modem except the capital investment on the modems, while CDPD will charge a monthly fee. Moreover, in program developers view, most of the wireless communication methods can be abstracted into similar user interfaces. For example, CDPD and Global System for Mobile (GSM) currently are the two most important methods for wireless Internet. They are quite different from each other technically, but after installation both can provide the same interface to program developers. Program designed for one can be used for the other without any modification. Since LBS Wireless Communication software shares a lot of common features and supports similar user interfaces, it makes itself a perfect target for software reuse. Software reuse is the process of creating software systems from existing software rather than building them from scratch. Software reuse is still an emerging discipline. It appears in many different forms from ad-hoc reuse to systematic reuse, and from white-box reuse to black-box reuse. Traditional software reuse paradigm supports code reuse only, which is also called white-box reuse . In order to achieve code reuse, the programmers have to study the source codes of previous software and grasp the details. The code reuse process takes time and is far from easy since the reused codes are possible incompatible with other codes in the new software. Traditional software reuse paradigm does not support other forms of software reuse, such as components, design document, and patterns. Recent advances in Object Oriented (OO) technology and Application Framework make it possible to take full advantages of multiple forms of software reuse, and at the same time save the work to study source codes.

The main goal of this research is to develop a framework to promote software reuse of LBS applications by using Object-Oriented Application Framework technology, the product of recent advances in Object Oriented (OO) technology and Application Framework. Object Oriented (OO) technology is a unique way of thinking about problems and their solutions. OO attempts to break a problem into its component parts instead of tackling the problem in a top-down and linear fashion as in traditional approaches and can significantly improve the efficiency of software development as well as the maintenance, reusability and modifiability of the developed software [Goraj, 1999]. OO is suitable for LBS wireless communication software development. Different types of wireless communication such as Radio pair, CDPD, GSM, Internet, Compression, Encryption are treated as objects; their attributes, like advantages and disadvantages, are treated as the constant value of these objects; their potentials, like protocols support, are treated as variables of the objects; their performances, like sending or receiving, can be treated as methods. Object-Oriented application framework, or framework for short, is a newly booming and very important branch of Object-Oriented technology. According to Johnson and Foote (1988), a framework is a reusable, semi-complete application that can be specialized to produce custom applications. Frameworks are targeted for particular business units (such as data processing or cellular communications) and application domains (such as user interfaces or real-time avionics) [Johnson and Foote, 1988]. In contrast to earlier OO (Object-Oriented) technology based on class libraries, framework describes not only the component objects but also how these objects interact by describing the interface of each object and the flow of control among them.

This special character makes framework an ideal candidate for the development of wireless communication software. Object-Oriented Application Framework is on its way to become the industry standard for LBS wireless communication software development.

1.2 Objectives

The objective of this research is to investigate and develop an Object-Oriented Application Framework to improve the software reusability of the wireless communication framework for Location Based Services applications for commercial aspects. A wireless communication framework for LBS applications, thereafter called wireless framework for short, is developed to provide LBS application developers with an efficient, simple, and reliable way to take advantage of the benefits of wireless communication technologies. Listed below are the specific objectives for this research:

- ❖ Investigate the current wireless technology available to determine the best suited candidates for wireless objects.
- ❖ Investigate the Location Based Services applications to determine the class structure of wireless objects and their interfaces.
- ❖ Develop independent wireless objects that can run in different Operating Systems (OS) with Java language.
- ❖ Develop a wireless framework based on wireless objects.
- ❖ Apply the developed wireless framework to a Mobile Equipment Management System.
- ❖ Apply the developed wireless framework to a wireless Internet-Based Real- Time Kinematic GPS Positioning System.

1.3 Thesis Outline

The thesis consists of nine chapters. Brief introductions of the remaining chapters are as follows. In Chapter 2, fundamental aspects of Location Based

Services (LBS) are briefly introduced that is Location Based Services theory and its concept.doc. Then, the system architecture of LBS is described. LBS are composed of three most important parts: Wireless Communication, Client, and Server. Wireless Communication, Client, and Server are then compared according to their role definitions, functions, and possible choices in the market. As a result, this chapter concludes that Wireless Communication is the most suitable candidate of these three to improve software reusability. Chapter 3 concentrates on investigating and analyzing the current advance of Wireless Communication technology. A discussion of how well these technologies can be served and introduced. In Chapter 4, the methodology is introduced along with the prototype development are first introduced. It includes designing and implementation of GIS is discussed along with how spatial data can be designed, and designing of database schema is introduced.

Chapter 5 focuses on the GIS Data Distribution through the World Wide Web. The principles of distribution of GIS data is expressed in WWW (World Wide Web) is introduced. In Chapter 6, Location management for location traffics, wireless routing techniques for the proper mobilization of data and along with location management policies is also described in Chapter 7. In Chapter 8 the future trends of LBS is discussed in detail and finally in Chapter 9, conclusions and recommendations for further research are finally presented

2. Location Based Services-Theory and Concept

Location Based Services (LBS) use location to deliver targeted applications to users at their moment of need. The applications for LBS are numerous. Progressive industry leaders are building solid foundations today to support well conceived solutions for new location applications and value-added services. The foundation of Location Based Services was laid by the FCC (Federal Communications Commission) in the US. FCC required wireless network operators to supply public emergency services with the callers' location and callback phone number. This generated the emergence of a new and dynamic field called LBS, where the service was based on the geographical location of the calling device. Further, advances in the field of Positioning Systems, Communications and GIS fueled the imagination of the industry people with regards to LBS. This ability to provide the user with a customized service depending upon his or her geographical location could be used in services such as advertising, directory services, tracking, emergency services, billing, and social/entertainment. The leading driver for LBS comes from wireless carriers and associated hardware and software developers. These companies hope to build value-added, revenue-generating services out of a Federal Communications Commission (FCC) mandate to provide the location of wireless emergency callers automatically to public safety agencies. In their wake come positioning technology providers (both GPS and non-GPS network-based solutions), base map and geocoding product/service providers, portable device manufacturers, LBS application service providers, LBS application software developers for both server and client devices, and a multitude of on-line information services.

Location Based Services have been seen as a key for differentiating between the mobile and fixed Internet worlds since LBS capitalize on the nature of mobility by bringing together the user and his or her immediate environment. A survey conducted by Mobile Internet in April, 2000 revealed that 50% of operators thought LBS would be the killer app for mobile Internet services, significantly ahead of all other categories [4]. Location Based Services will serve both consumers and network operators. For consumers, they meet the demand for greater personal safety, more personalized features, and increased communication convenience; for network operators, Location Based Services help differentiate service portfolios, improve network efficiency and create greater pricing flexibility to address discrete market segments. Although the market potential is enormous, Location Based Services cannot begin with the most complex, technically demanding and feature-rich offerings. Instead, network operators must use today's technology to gain market leadership and hone critical technical skills. With a head start, they will be ready to create new services quickly when more accurate location and wireless personal digital assistants arrive [6]. The implementation of Location Based Services depends on two cutting-edge technologies, Wireless Location and Mobile Internet.

There are a number of technologies currently available for locating mobile devices, which can be classified into handset centric and network centric solutions. The former builds significant intelligence into the handset to achieve location while the latter builds more intelligence into the mobile network infrastructure [6]. Positioning accuracy of COO(CELL OF ORIGIN) generally depends upon the size of the cell. It is possible to achieve accuracy within 150 meters in urban areas with the deployment of pico-cell sites. As more network-based location finding schemes are deployed and Global Positioning System (GPS) capability is integrated into wireless devices, the improved accuracy of location fixing will not only improve current services, but will also allow for the introduction of new services. GPS is the most commonly discussed option in recent years. GPS is a RF satellite-based navigation system that was developed by the United States Department of Defense. After Selective Availability (SA) was switched off on May 1, 2000, the accuracy of stand-alone GPS positioning is about a few tens of meters for civilian users, even when the solar activity is high [4]. The positioning accuracy can be further improved to centimeter level with Differential GPS (DGPS) technology. Assisted GPS (AGPS) uses fixed GPS receivers that are placed at regular intervals on a network to reduce the time needed for users' GPS receivers to calculate the location. For locating mobile devices, the common alternatives available are Enhanced Offset Time Division (E-OTD), Time Of Arrival (TOA), Angle Of Arrival (AOA) and Intelligent Network (IN) solutions. These different types of technologies are summarized and compared in Table 2.1 [Nguyen, 2001]. The world of Mobile Internet is not simply an advanced stage of Internet evolution, but rather an entirely new world shaped by mobility. Fixed Internet and mobile telephony have been deemed as two of the most influential technological developments in the past five years. The convergence of fixed Internet and mobile telephony ultimately results in the birth of Mobile Internet. In fact, simple Mobile Internet services have existed in European markets for some time in the form of Short Message Services (SMS) [Nguyen, 2001].

Technology Scheme	Technology Dependence	Advantages	Disadvantages
COO	Network	No modifications needed to networks or handsets	Relatively low accuracy
E-OTD	Handset		Software modified handsets needed
TOA	Network	Uses existing CDMA network features	Relatively low accuracy
GPS	Handset	GPS is free to use	New handset needed
A-GPS	Handset & Network	GPS is free to use: TTFF time is reduced	New handset needed: Network assistance needed
AOA	Network		Complicated antennae required
IN	Network	Location Finding System independent	

Table 2.1 Characteristics of Different Wireless Location Technologies

Mobile Internet, or Wireless Communication in the broad sense, is now pushed forward by both market demand and technological advancement. From the market side [Lu, 2000]:

- ❖ Users are more and more dependent on the diversified information service provided by the Internet.
- ❖ There is an economic development trend driven by "Mobility + Information".
- ❖ People are more and more mobile than ever before. From the technological side:
- ❖ Network Technology keeps evolving. Compared to the wireless network, the current wireless network is still far from being perfect in terms of bandwidth, delay, error rate and connection stability. However, the growth and application of 1G and 2G technologies like CDPD and GSM, 3G technologies, and protocols like WAP, SOAP and GPRS, have laid a good foundation for wireless Internet applications.
- ❖ Terminal equipment tends to be more diversified. Limitations in CPU computation speed, storage capacity, display size, keyboard size and battery life are being eased. A lot of new handheld equipment like PDA, Palm, and smart phone is adequate for Mobile Internet services. Some of the most powerful and influential companies in the world – Microsoft, Sun, Motorola, 3Com, Hewlett-Packard, Ericsson, Oracle – are developing hardware, software and networking equipment for the new category of smart devices to support Location Based Services. Innovative smaller companies also are creating new platforms and applications for Mobile Internet and Location Based Services. Although still a nascent industry, Locations Based Services are expected to have a major impact on the market.

2.2 LBS System Architecture

All of the LBS applications are similar in nature. They usually have a client/server structure and can be further abstracted into three parts: Client, Server, and Wireless Communication to connect Client and Server (Figure 2.1). These three parts are highly dynamic and interactive, since they are changed by the fast-advancing technologies almost daily, and the advancement in one part will dramatically affect the development of others.

Generally speaking, Client is responsible for sending the user's request and the geographical location of the mobile device to Server, and Server is responsible for providing services based on the geographical location of the mobile device. The role definitions of Client and Server, however, are not always reasonable considering the fact that Client is not only an information consumer but can also be an information provider. Client can make contributions to information acquisition by collecting data in the field or on the spot. Server will put the information collected from the field into the database and will then provide services for all clients based on the database.

In fact, the role definitions of Server and Client are becoming more and more vague. In the future, Location Based Services will benefit from real-time information acquisition at the Client side. Client will be equipped with sensors to collect information automatically and send it back to Server. Server can analyze this vital information and put it into the database for service. The possible applications for information collecting at the Client side include Equipment Management, Asset Track, Intelligent Distribution, Dynamic Working Plan, Traffic Control, On-line Survey and so on. Although it is a trend for Location Based Services to collect information at the Client side, there are still some problems caused by wireless communication. Information acquisition at Client side is likely to be more popular in the near future when 3G is fully implemented[3].

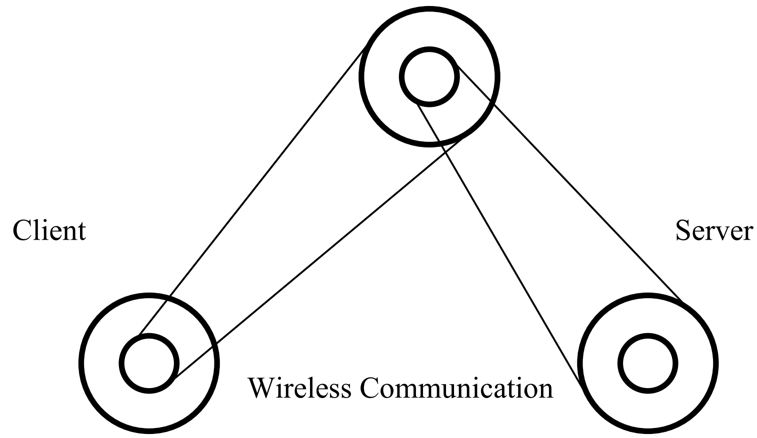


Figure 2.1 Location Based Services Components

Client, Server, and Wireless Communication of Location Based Services can be further divided into an aggregation of functions. While some functions can be intrinsic and indispensable for Location Based Services, the other functions might not. Although the functions of each part are application-dependent, i.e. the functions of a part are fully determined by the specific applications and the functions for one application might be different from those for others, the collective functions of a part can still be generalized and abstracted into a function set, or in other words a function pool. The functions for a certain application will fall into a subset of the function pool. The architecture of Location Based Services is shown in Figure 2.2, and the functions of each component are described in the following.

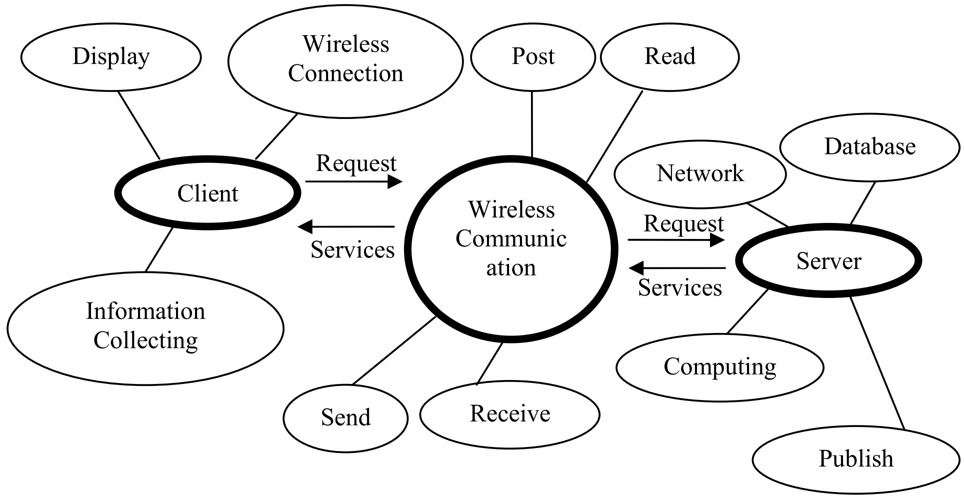


Figure 2.2 Location Based Services Architecture

2.2.1 Client

The function pool of Client is as follows:

- ❖ Display Function: A display device, usually a screen, is used to display the text or multimedia information to users.
- ❖ Information Collecting Function: The ability to collect information from equipment like a GPS receiver or information input manually. In the second case, the handheld device should provide a user-friendly interface.
- ❖ Peripheral Control Function: The ability to control peripheral equipment connected to the handheld devices. The control information can be generated by the local handheld device or received from remote control center.
- ❖ Computing Function: The ability to perform tasks such as mathematical computation, multimedia compression, and information encryption etc.
- ❖ Wireless Connection Function: The ability to connect the server with wireless communication. Save Function: The ability to save information for future use.

- ❖ Multimedia Function: The ability to display multimedia information like voice and pictures.

2.2.2 Server

The function pool of Server is as follows:

- ❖ Network Function: The ability to transfer over multiple protocols, multiple operating systems and web browsers on Internet and Intranet.
- ❖ Database Function: Server should have the ability to manage and utilize the database to save the information and provide service for Client.
- ❖ Computing Function: The ability to perform tasks such as mathematical computation, multimedia compression, and information encryption, etc.
- ❖ Multimedia Function: The ability to display multimedia information like voice and pictures. Business Logic Function: The ability to provide business logic in a distributed network for applications.
- ❖ Wireless Connection Function: The ability to support wireless communication. It is useful when Server is moving and has no fixed Internet access, or Client has no Internet access and has to communicate with Server directly.

2.2.3 Wireless Communication

The function pool of Wireless Communication is as follows:

- ❖ Receive Function: Wireless Communication should have the ability to transfer services information from Server to Client.
- ❖ Send Function: Wireless Communication should have the ability to transfer the request and location information from Client to Server. The send function is not always essential. For example, it is possible for the service provider to detect the

appearance of the mobile device via wireless network, and send the information to the client even without request.

- ❖ Real-time Function: The ability to support real-time Location Based Services. Not all Location Based Services need real-time communication, and not all wireless communication technologies support real-time communication.
- ❖ Post Function: The ability to post data to the web.
- ❖ Read Function: The ability to read data on the web.
- ❖ Compression Function: The ability to compress information before sending and to restore information after receiving. This function needs cooperation from Client and Server.
- ❖ Encryption Function: The ability to encipher messages before sending and to decipher messages after receiving. This function needs cooperation from Client and Server.
- ❖ Information Security Function: The ability to ensure that the only authorized users receive the information. This function classification is the first step for Client, Server, and Wireless Communication to pursue reusability. However, the methods and the procedures used to realize reusability for each of them are different in each case. At the Client side, hardware compatibility is the core problem for application developers to realize reusability. There are so many products available now for Client, such as laptops, handheld PCs, PDAs, pocket PCs, smart phones, GPS receivers, etc. Considering power consumption, computation ability, size, hardware interface, and screen issues, there is not a universal solution to meet the requirements of all users. At the Server side, the thorniest problem lies in network compatibility. The program running on the Server side should support multiple operating systems, web browsers, and protocols that are proliferating rapidly on the Internet and Intranet. Compared to those for Client, the available choices in the market for Wireless Communication

are much less, especially in the market for wide area mobile wireless communication. The most common and dominant method of wireless communication available today is the commercial cellular telephone system. Compared to Server, the protocols for Wireless Communication are much less, although they are still various. Moreover, different types of wireless communication are highly complementary and easily merged. As discussed in Chapter 1, Wireless Communication has many potentials for software reuse. Comprehensively speaking, it is easier to build a framework for Wireless Communication to support Location Based Services than it is to build a framework for either Client or Server, this is the most important reason for us to choose Wireless Communication as our first step toward the aim of software reuse for LBS.

2.3 Theory and Concept

2.3.1 Basic Technologies and Concepts

Some concepts that may help the reader to read the report are described. All definitions are also described later in the report in its right context. The definition of location-based services, which is the heart of this report, is described further[6].

Location-based services	This is an application that will allow mobile users to receive personalized and lifestyle-oriented services relative to their geographic location. These services use the positions (coordinates) provided from a Mobile Location System.
Mobile Location System	A system that has support for location of GSM subscribers based on one or many Geographic Positioning Technologies. It also handles roaming, charging/billing and subscriber privacy management. Many different techniques can be combined in a Mobile Location System.
Geographic Positioning Technology	The concept is very similar to Mobile Location System but here is only the actually location technique considered. They are often divided in network based and terminal based.
Terminal based	This technique requires a new terminal, or a new SIM card, or even both. Such limitation

	makes it harder to make a success on the market. On the other hand large investments in the GSM network are avoided.
Network based	This technology does not require new mobile equipments so it will be available to all members in all GSM networks from dav one that the technique are installed.
Mobile Phone	A mobile phone allows the user to make wireless phone calls. The mobile phones have many features and a great variety in design today. In other literatures one can find synonyms like cellular phone and wireless phone. Sometimes only phone is used.
Mobile Equipment	This is a wider definition than mobile phone. It also includes other different existing and futures devices such as mobile PDA's and Smart Phones. A mobile phone is also an example of mobile equipment. Other synonyms are handheld device, mobile device, mobile station, handset, wireless appliance and mobile terminal.
PDA	Personal Digital Assistant, PDA is a combination of a digital calendar, address books and services such as email, SMS and Internet. Handheld is a synonym.
Smart Phone	A Smart Phone is a combination of a mobile phone and a PDA.
Mode	Active= The mobile phone is on and a call is in progress. Idle= The mobile phone is on but a call is not in progress. Detached= The mobile phone is off. It can t be positioned.
Finder Client	It is an enhancement of Finder Application. The Finder Client is designed and implemented in this thesis. The interface is developed for a PDA.
User	In this report, a user is using the location-based services. For example he/she uses Finder Client.
Cell point	CellPoint is providing GSM operators with end-to-end mobile location services built on CellPoint's GSM positioning technology platform. It is worth noticing that the company has two different platforms. The Mobile Location System, MLS is truly network based. The technology used in this thesis is SIM Card based, which is a hybrid solution based on both network and terminal

Figure 2.3 Basic Technologies and Concepts

2.3.2 Definition of Location-based services

“Location-based services will allow mobile users to receive personalized and lifestyle-oriented services relative to their geographic location.” . Location

Inter-Operability Forum(LIF) also predicts that Mobile Location Services are going to become one of the most compelling value-added services in the future, allowing wireless appliance users to combine mobility with the Internet. The services use the mobile network to do the positioning. One typical application is the yellow pages services where a user can find the nearest companies and organizations such as pharmacies, banks, gas stations, car repairs, flowers, tires, hairdressers, hotels, locksmiths and restaurants. Another location-based service that many companies develop is friend finder applications. The service is reachable via WAP, SMS or World Wide Web. The friends must first accept being positioned by his friend, often for legal reasons, and then he will have them on his personal friend list ready to be located or they can locate him. These services are mainly developed for mobile phones. The answer of the request is often only received as text. This shows a solution to present friends current location on a map [6].

2.4 GSM

To understand the different positioning techniques it is good to know something about the Global System for Mobile Communications, GSM. It was first introduced 1991 and today it is one of the leading digital cellular systems. The other techniques is Code-Division Multiple Access(CDMA), Time Division Multiple Access(TDMA), and Personal/Pacific Digital Cellular (PDC). GSM is most widespread in Europe and Asia[17] .

2.4.1 Geographic Structure of a GSM Network

A GSM network has a specified hierarchical structure. This structure is very important because it makes it possible to route incoming calls to the right subscriber. The structure consists of these parts: cell, location area, MSC Service area, PLMN service area and GSM Service Area. The smallest unit is the cell. It is the area of radio coverage from one transmitter. The cell's size depends on the transmitters level of power, its angle and the amount of obstacles. The size can be between 100 m and 35 km. A cell can be circular or triangular. The cell is identified with its Cell Global Identity, CGI. That identity can be used for positioning purpose. The CGI consists of four parts:

MCC, MNC, LAC and CI. MCC = Mobile Country Code. Identifies the country in which the user is registered. MNC = Mobile Network Code. Distinguishes between each network in a country. LAC = Location Area Code Specifies the particular area concerned. CI = Cell Identity

MCC	MNC	LAC	Cell Identity	Description	Latitude	Longitude
240	5	61443	55707	Location1	59,40xxxx	17,94xxxx
240	7	61462	37130	Location2	63,28xxxx	18,69xxxx
240	7	61462	38203	Location3	65,66xxxx	22,05xxxx

Figure 2.4 Cell ID database used "outside" the GSM networks to convert from CGI to coordinates

The CGI is stored in a hexadecimal system in the GSM network. When the CGI is changed to decimal system by a location system it can be translated into coordinates with the help of a cell id database that are provided from the operators. GSM Service area is the entire geographic area of the GSM network. The area increases as more countries get GSM networks.

2.4.2 Conceptual Components of a GSM Network

The purpose of this chapter is to show the basic components of a GSM network. Many of the nodes are affected when add-on software for network based positioning are installed. As an example the HLR and VLR described below stores the cell-id of the cell the telephone where in when it was last active. The Mobile Location System calculates an approximate position based on the knowledge about the structure of the cells and the base stations. A Mobile Location System must often also be capable to force a mobile phone up in active mode. Because when the mobile phone goes active its location information in the mobile network is updated. The GSM network is divided into two systems. The Switching system, SS is responsible for performing call processing and subscriber related functions. The other part is the Base Station System, BSS that performs all the radio related functions. The two systems are built up by different components.

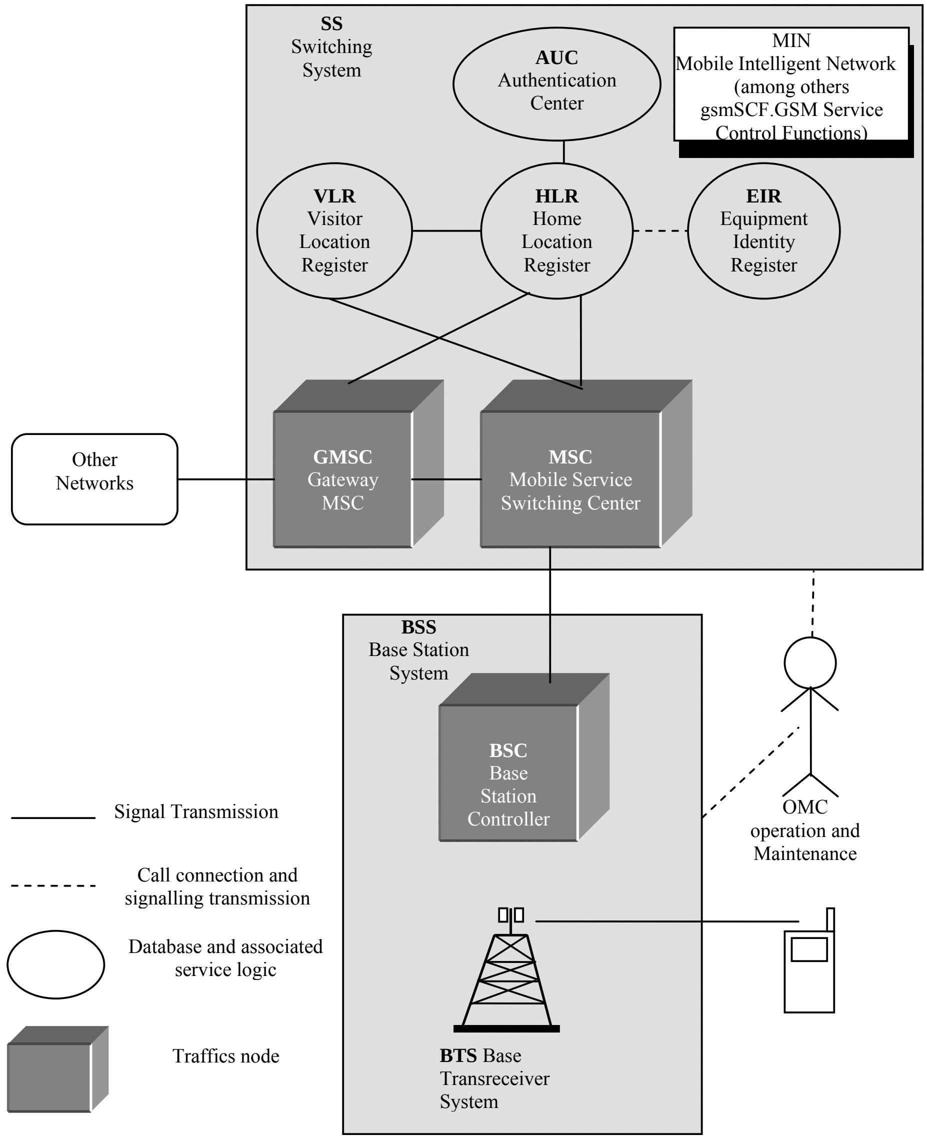


Figure 2.5 Conceptual Components of a GSM Network

The figure shows only the basic structure of a GSM network. Mobile services Switching Center, MSC performs the telephony switching. Calls from other telephony and data systems are also controlled here. It is connected to public data networks, private networks and other mobile networks. Gateway MSC, GMSC is the interface for GSM against other networks. It makes it possible for the MSC to "read" from different networks HLR and therefore makes it possible to route a call to a mobile telephone. Other networks always pass the gateway when they connect to the GSM network. Home Location Register, HLR is a database that stores and manages the subscribers for one operator. An operator's subscribers are stored here permanent until the subscription is cancelled. The HLR also holds associated service logic. The subscriber information stored includes identity, services and last serving cell location (cell id). The Visitor Location Register, VLR, The VLR can be regarded as a distributed HLR. It holds a copy of the subscriber information (cell id etc.) of the HLR. It can also execute commands. But the VLR also stores information of subscribers that usually belongs to other networks. It is necessary because of roaming of subscribers. Roaming is when a mobile phone moves around in a network in idle mode. The VLR is always collocated with a MSC and not as a stand-alone unit. Two components that not are affected by a Mobile Location System are Authentication Center, AUC and Equipment Registry Center, EIR. Base Station Controller, BSC manages all the radio related functions of a GSM network. One BSC can control many Base Transceiver Station, BTS. This is where a mobile phone interfaces to the GSM network via radio (Ericsson Radio Systems, 1998). The Mobile Intelligent Network, MIN provides architecture for the introduction of new services throughout the network. This service can be a positioning system. The MIN concept has many advantages. MIN makes it possible to quicker make more attractive services and it separates the services from the other nodes in the GSM network. GSM Service Control Function, is a part of the Mobile Intelligent Network[18].

2.4.3 Example of Geographic Positioning in a GSM Network

The question is how does the positioning system interact with the GSM nodes. Let us use CellPoint's Mobile Location System, MLS as an example. The aim for MLS is to read the cell id from the database in the VLR. As mentioned earlier this cell id can later be converted to a geographic position. The basic idea is that, from a GSM perspective, MLS acts logically as another GSM node in the network. For instance when MLS sends requests to MSC/VLR, the MSC/VLR accept the commands because it accepts MLS as a GSM node. The MSC believes, in this example, that it is a call from the HLR. When a subscriber roams into a new MSC service area the VLR request the subscribers home HLR for information. The information is stored in the new service area VLR and now the new network has all the information required for call set-up. It means that a location system must read from the VLR to find out in what cell the mobile phone is located. It is also important to know that different manufactures implement their GSM systems in different ways. The figure gives a correct answer if the mobile phone to be located is active and is in the same MSC service area as the user.

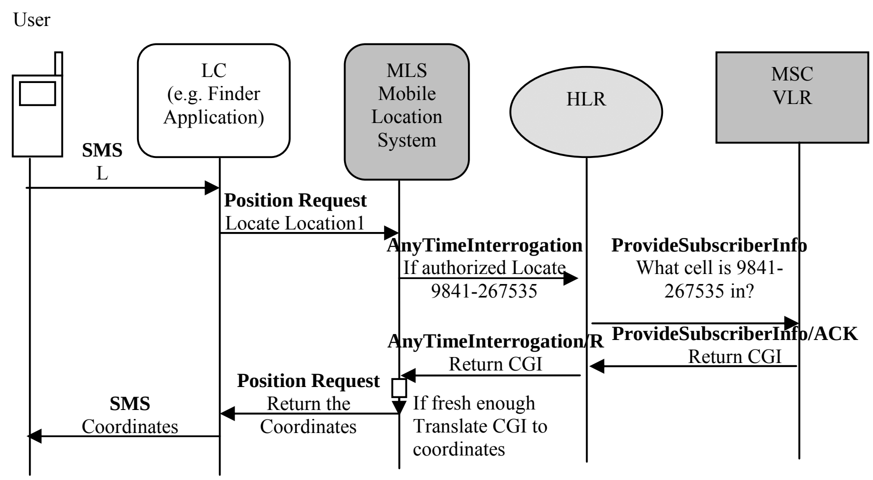


Figure 2.6 the basic idea to read the cell id from the VLR database with the commands and anyTimeInterrogation, provideSubscriberInformation

The user sends for example an SMS to locate another person[14]. MLS do not receive the request directly from the user, instead MLS receives a Position Request via a Location Client, LC. In this case the LC is a middleware service that is the link between the mobile equipments and the MLS. Applications like Finder and Yellow pages are examples of such a Location Client. If the user is authorized to the service MLS uses anyTimeInterrogation, ATI. MLS send the ATI operation towards the HLR. The HLR receives the telephone number of the phone that is going to be located. The database HLR uses the operation provideSubscriberInformation, PSI to find out what Cell Global Identity that is stored in the VLR. The cell id is returned via ATI and PSI result message. Based on the timestamp in the result message, MLS determines that the received position measurement information is fresh enough and therefore computes the coordinates for the location. If the timestamp is too old, further operations must be performed. How long that timestamp is allowed to be must be decided by the application using MLS. MLS uses the operator’s geographic data (cell id database) about the base stations to convert the cell id to longitude, latitude in the reference system WGS84 and an error radius. The location estimate is returned to the Location Client. Finally the user receives the location (CellPoint, 2001). This MLS is now available for the market. Its strength is that it requires no large modifications of mobile devices nor the GSM networks. But the accuracy is likely very low in areas with large cells.

2.4.4 Current and Future representation of Mobile Technology

GPRS has always been regarded as a key stepping stone on the path to 3GSM – not least because it initiates that vital first step with the introduction of packet switching. Today, given slower than forecast next generation deployment, GPRS has become even more important in its own right. While 2.5G implies an interim solution, in reality the introduction of GPRS represents a significant transformation of mobile technology, delivering the benefits of IP connectivity and “always on” service access for the user. The adoption of GPRS is a fast and cost-effective strategy that not only supports the real first wave of mobile internet services, but also represents a big step towards 3GSM networks and services. Customers are already beginning to enjoy advanced, feature rich data services, such as photo messaging on MMS, email on the move and other leading edge wireless applications

3GSM is the latest addition to the GSM family. 3GSM is about having third generation mobile multimedia services available globally. 3GSM focuses on visionary communications, in more ways than one. It's about the new visual ways in which people will communicate and the unique vision of the GSM community, which has always focused on the future needs of our customers.

The technology on which 3GSM services will be delivered is built around a core GSM network with a Wideband-CDMA (W-CDMA) air interface, which has been developed as an open standard by operators in conjunction with the 3GPP standards development organization. Already over 85% of the world's network operators have chosen 3GSM's underlying technology platform to deliver their third generation services. 3GSM is a key element of GSM-The Wireless Evolution.

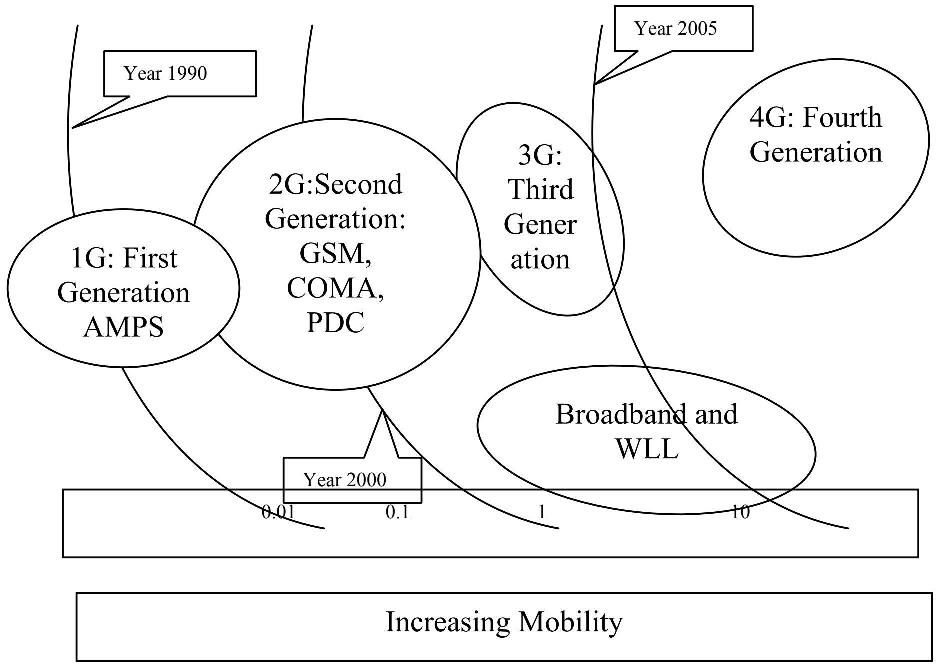


Figure 2.7 Generation of wireless networks

2.4.5 Future Mobile System

The techniques for the new mobile networks are being developed right now. Spokesmen say that it will revolutionary our lives and keep people connected to Internet at all times and in all places. For the first time we will have fully enabled mobile Internet functionality[1]. The new systems are going to support all the existing Internet services like web browsing, File Transport Protocol, FTP, email etc. Others say that the large investments do not correspond to the benefit of being able to work, play games and find information from everywhere. The first step is the General Packet Radio Service, GPRS. It is a new nonvoice value added service. Today's available GSM networks are upgraded with software and routers to create the GPRS. The new services gives opportunities for better applications and fast instant connection to Internet. Information can be sent and received in a GPRS mobile network at a theoretical speed of up to 171.2 kbps, which is about ten times faster than today's communication method in mobile networks. It is not only the speed that makes the difference between GPRS and today's GSM networks. Instant connection is the new concept. If someone surfs the web wireless today a connection is established from the start of the session to the end. Today's mobile phones send signals even when the user does not actually send or receive something on the Internet. It is therefore expensive to stay connected a long time, because users pay for the time connected not the amount of data transmission. The batteries in the mobile phone will also be drained as long as it is connected to Internet a long time from the start of the session to the end. The GPRS is packet based, which means that the connection is established once. That allows the user to be instant connected to Internet. One possible business model is that the users only pay for the amount of delivered data. Access to private network can also be continuously. GPRS is one step towards the next level called Universal Mobile Telecommunications System, UMTS or third generation's mobile system, 3G. 3G can be thought of as GPRS plus entertainment and new mobile equipments. It is packet based such as GPRS but the speed is much higher. The estimated data rates are depending on the mobility. If someone travelling fast (over 120 km/h) in rural areas the speed will approximately be 144 kbps. But when the user is stationary inside a building the speed will increase to at least 2 Mbps. It can become faster and hopefully cheaper to send a video clip to someone than it is to send a SMS today. Example of new entertainment services is games, video and mobile multimedia. The future mobile equipments will probably have larger color screens and a larger storage to save data on. Many analysts say that different devices will collide. Consumer electronics such as video cameras, MP3 players and GPS receivers could be built into the mobile phones or the other way around. It is an exciting challenge for the manufactures to find out what the combination of devices that the consumers really need or demands. One of the aims with 3G is to unify the different mobile standards that today's second generations wireless networks use. The system is based on a different technology platform that is unlike the one that is used in the 2G world. The different organizations want their system to be the raw model. The solution is that 3G are going to support three different optional air interface modes. That's a good thing for the positioning industry that builds solutions based on the cell id. Smaller cells will give higher accuracy for such positioning. The 30 May 2001, the first 3G call was made in Japan. The company NTT DoCoMo will be the first to launch 3G commercially, the 1 October 2001. Throughout 2002 new network operators will launch their systems and more mobile equipments will arrive. But it is first in 2004 that 3G will reach a larger quantity of consumers. All these changes may have following impacts on companies and private persons:

- Mobile equipment manufactures in Japan gets a lead against other countries. They have so far had a small market share outside their home market.
- It is not necessarily the services with the most advance technique that are going to be the winner. Time is no longer money! It is the amount of bytes downloaded that costs. However the best solution for all parts would maybe be a flat monthly rate. That would increase the usage and amount of services.
- Challenge for companies to understand how to make money in the new area.
- The systems will become increasingly vulnerable to attack by malicious crackers when technology becomes more sophisticated and bandwidth increases.
- Spam, unwanted messages, could increase.
- 3G let people to be connected at all times and in all places.
- Experiences from Japan show that entertainment is the entrance gate for many people to the wireless market.
- Many people do not have a fixed telephone at home anymore. The mobile phone is enough. Users will maybe spend more time looking at the mobile equipment (nonvoice services) than talking to it.

2.5 Mobile Phones, PDA's and Operative Systems

PDAs, Personal Digital Assistants is becoming more and more popular. The term handheld is a synonym for PDA. Its origin is from the programmable calculators. In the beginning, a PDA was like a simple filofax. The most used functions were a calendar and an address book. Today other areas like email, SMS and Internet makes handheld computers more interesting. The email box on the normal PC can be synchronized with the one on the PDA. Internet access is normally made via an internal modem or a mobile phone's modem. Many handhelds do not have a keyboard. Instead, input is made by drawing/writing/clicking with a stylus on the screen, just like when using paper and pencil.

2.5.1 PDA Operative Systems

There are three major operative systems for PDAs on the market today. It is EPOC, Windows CE and Palm OS. In the near future Linux will also be available. Palm OS from 3Com, has dominated the market from the beginning, e.g. 1995. In Palm OS only one application can be active at any given time. 3Com competitors' states that the Palm OS is out-of-date, on the other hand, Palm has done the things that the mass market wants to a reasonable price. The Palm PDAs made by 3Com has the largest market share together with the new company Handspring. Handspring is a Palm OS licensee. Development for Palm OS is made in C, C++, assembler or scripting. One of the Palm PDAs primary strength is that it has a large amount of third party applications. A company of Microsoft's size will sure do its best to compete with Palm OS. They have large resources and many widespread Internet and e-mail products such as Explorer and Outlook.

But on the other hand it can be hard for them to focus on the narrow PDA market. Microsoft has a varied portfolio of products while Palm and Handspring only focuses on handheld computing. The operative system from Microsoft is Windows CE. It is a 32-bit thread operating system based on Windows. Even if Windows CE PDA's has the fastest processors, it is not sure it is the best system for a handheld usage because CE has parts of Windows for PC built in. That results in that Windows CE has functionalities that not primarily has been developed for a PDA. Many of these memory-demanding functions are not needed on a PDA. The new Windows CE operating system is called PocketPC. PocketPC is actually version three of CE. There is a Visual Basic toolkit available for development of applications. Examples of PDA's running Pocket PC/Windows CE are Casio Cassiopeia and Compaq iPAQ. EPOC's is designed for small, portable computer-telephones. The company Psion has developed it, mainly in C++. It is a 32-bit multitasking operative system whose Graphical User Interface support writing/drawing/clicking on the screen. A program is normally started by clicking on an icon with a pen (stylus). Today companies like Psion, Ericsson, Nokia and Motorola supports EPOC through the common company Symbian. EPOC is compact operative system, allowing small chips for the read only memory. It is easy to develop programs on EPOC by using C++, Java or OPL. There are emulators available as well for development on an ordinary PC. Famous brands are the PDA's Psion Series 5mx, Ericsson MC218, Ericsson's Smart Phone R-380 and the Nokia 9210 Communicator. During the work three different PDA's have been investigated; Psion5mx, Compaq iPAQ and PalmV. Results from the investigations are listed in table. Some of the viewpoints are from PDAStreet (PDAStreet, 2001). The actual development has been made only on the Psion 5mx. The other two models have been used as calendar and address list. They all have their benefit and drawbacks but the large threshold to overcome is to always bring it with us. Another aspect is that it is uncomfortable to bring two expensive devices, like a PDA and a mobile phone, between home and work.

If the consumers are going to be connected all the time to Internet it is important to have good batteries. Devices like GPS receivers also consume much energy. All the three investigated PDAs can be connected to mains power, but the strength of a device is that it should be wireless. Let us examine if it is possible to make Finder Clients with mapping capabilities on Palm and Compaq. In both cases it is possible to develop products for them and Route Planner is available for both systems. The two brands also have an Infrared link. The Palm processor might be a little bit too slow. Compaq has a faster processor and a colour screen and would therefore be worth a try to develop a Finder Client on. Psion 5mx Compaq iPAQ Palm V Price ~$500 or more

~$500 ~$300 Size (cm) 170x89x24 mm. 350 g 130x84x16 mm. 250g 114 x 88 x 10. 122g CPU speed 36 MHz ARM710T 206 MHZ StrongARM 16 MHZ Dragonball EZ Memory 16 MB ram, 10 MB rom 32 MB ram, 16 MB rom 2 Mb ram, 2 MB rom Operative EPOC32 Window CE/Pocket PC Palm OS 3.1 Input Keyboard. Pen writing. Easier than the Palm. A keyboard on the screen is also available. Pen writing or keyboard on the screen. Screen 640x240 dpi. Grayscale. Sometimes difficult to read from. 240x320 dpi. Colour. Easy to read when the backlight is on. 160x160.Monochrome but easy to read from. Advantage Suitable for development and map software. Color screen. MP3 and video. Small, fast and handy. Disadvantage Uncomfortable to bring as a calendar. To big while not having a keyboard. Not suitable for memory demanding operations. Battery AA batteries ~1 month Chargeable battery ~12 h Chargeable battery ~30 h 46

2.5.2 Smart Phones

Smart Phones is a concept for the future. In the near future it is likely that it will be difficult to note any differences between a mobile phone and a PDA. The third generation mobile systems and GPRS sets up for the new generation of color mobile phones with capability for advanced mobile Internet. An important part of Smart Phones is personal information management functionality and email synchronization. The goal is that a user can get information fast wherever he is located. The market for Smart Phones has been limited but many analyzers thinks that it will increase rapidly with GPRS. Microsoft has produced a Smart Phone platform named Stinger built on Windows CE. Partnership with companies like Mitsubishi, Sendo and HTC has been announced [24]. Ericsson R-380 Smart Phone opponent is the Nokia Communicator series. Both products have typical PDA functions like address book, calendar, email and web. They have also of course a phonebook, voice controlled answering, voice controlled dialing and Short Message Service. It is not possible to develop own products for the two phones yet. However Nokia's EPOC based telephone Nokia Communicator 9210 has a software development kit for both Java and C++. Unfortunately the Nokia Communicator 9210 does not seem to support the language OPL. Nokia

Communicator 9210 should be available during the summer 2001. WAP can be used to show maps on the Smart Phones today. The qualities of the maps are low and also monochrome (black and white).

2.6 LBS communication model

Application Layer	
Middleware Layer	GIS
Positioning Layer	

Figure 2.8 General LBS communication Model

In order to make location applications work, the industry had to overcome several challenges of both a technological and economic nature over the past years[13]. Technologically, the realization of LBS can be described by a three-tier communication model (Figure 2.8), including a positioning layer, a middleware layer, and an application layer.

The positioning layer is responsible for calculating the position of a mobile device or user. It does so with the help of Position Determination Equipment (PDE) and geospatial data held in a Geographic Information System (GIS). While the PDE calculates where a device is in network terms, the GIS allow it to translate this raw network information into geographic information (longitudes and latitudes). The end result of this calculation is then passed on via a location gateway either directly to an application or to a middleware platform.

Originally, the positioning layer would manage and send location information directly to an application that requests it for service delivery. The application layer (which in the LBS industry is often and confusingly referred to as a ‘‘client’’) comprises all of those services that request location data to integrate it into their offering however, as increasingly more LBS applications are being launched, many network operators have put a middleware layer between the positioning and application layer. Primarily, this is because PDE sits very deep in the network of a mobile operator, leading to complex and lengthy hookup of each individual new data service. Also, a middleware layer can significantly reduce the complexity of service integration because it is connected to the network and an operator's service environment once and then mitigates and controls all location services added in the future. As a result, it saves operators and third-party application providers time and cost for application integration. Figure 2.9 illustrates this concept.

Making application integration easy is vital for mobile operators in order to move to a so-called wholesale model for location data. The wholesale approach means that operators offer a kind of bulk access to the location of devices. An advertisement company, for example, can buy access to thousands of mobiles entering a certain location and then contact the devices with a push message. A roadside assistance company can offer its customers an automatic mobile positioning service for emergency purposes, but would have to buy the right to access this data from an operator.

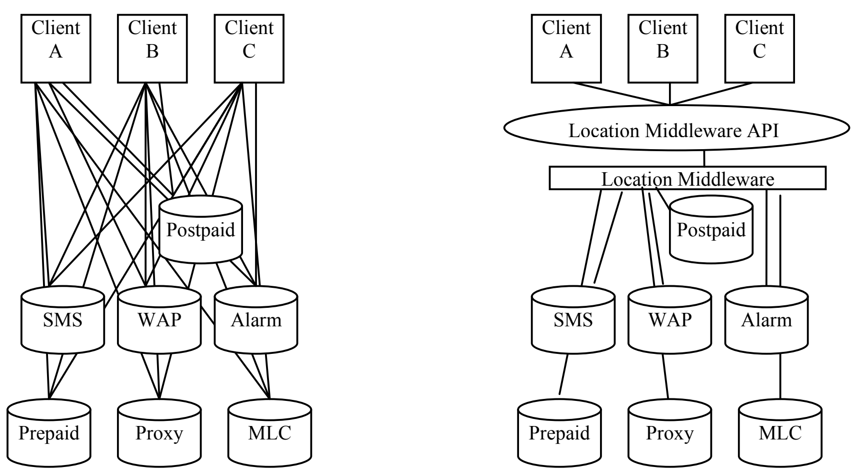


Figure 2.9 Application integration with or without middleware.

Finally, many companies may want to take advantage of fleet management services. If a third-party company rather than an operator offers fleet management, then this company would have to purchase location data in bulk in order to realize the service. The examples show that the wholesaling of location data is an important business area for operators.

For quite a while, operators hesitated to embrace wholesaling, arguing that major privacy concerns would doom this model to failure. Here, location middleware can fulfill another role. On the downstream, it allows users to manage location access rights of third-party applications, while on the upstream it systematically anonymizes location information revealed. Thus, the location middleware takes over a similar role as an anonymizing proxy does on the Internet. In this way, many privacy concerns are addressed by an operator. Also, users get direct access to turn privacy on or off. Finally, location middleware can be used to manage interoperability between networks for location data.

3. Introduction to Location Based Services Technology

The Location Based Services (LBS) is used to identify the location of caller by using various positioning determination technologies. Emergency number uses Location Based Services for identifying the location of caller. In order to understand the concept of LBS let's take example of processing of emergency number – when the emergency numbers such as 911 in U.S. are dialed from any telephone route the call to the jurisdiction where the call is originated, and display the callers telephone number and address in the communication center. First, telephone companies already maintain a subscriber database listing every assigned telephone number, the subscriber's name, address and billing information. Second, the telephone system already identifies the telephone number for every call placed, in order to properly bill the subscriber each month. This is known as Automatic Number Identification (ANI). The third component of LBS system is a Master Street Address Guide (MSAG). This database cross references every assigned telephone number, subscriber's address and the block number ranges for every street, in every jurisdiction served by the telephone company. Figure 3.1 shows the simplified diagram of the system.

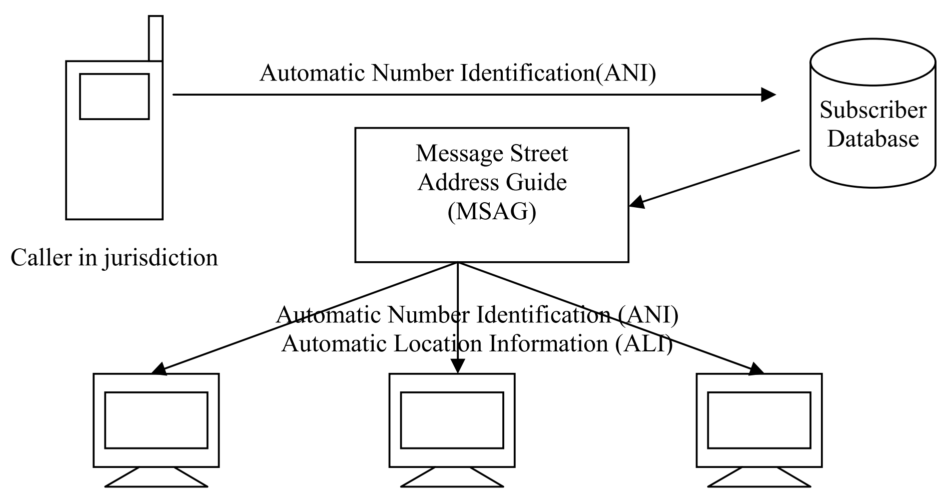


Figure 3.1 LBS System Diagram

Fourthly, most telephone company have built special switches and networks to carry LBS traffic, so that other telephone company traffic will not interfere with LBS operations, and LBS traffic is protected from power failures and other system problems. Whenever telephone company central office switch receives a call regarding position determination such as emergency call then it is routed to the LBS networks. The ANI (telephone number) information is decoded through a subscriber database to obtain the caller's address and other information. Next, the call is processed---sometimes simultaneously---through the MSAG to obtain the ID code of the agency that should handle the call. The LBS network then routes the voice and ANI/ALI (automatic number information and automatic location Information) information to the correct agency. The ANI/ALI information is displayed when the call taker answers and, at some agencies, the call information is printed out when the call is completed.

3.1 Application of Location Based Services

The major LBS applications that are:

- Destination guides with maps, directions
- Location-Based Traffic and weather Alerts
- Wireless advertising and electronic coupons
- Movie, theatre and restaurant location and booking
- Store (offering cheapest prices for brand-name items) locating applications
- Emerging buddy, child, or car (e.g. LoJack) finders.
- Telematics-based Roadside Assistance - OnStar from GM and telematics implementations from other auto vendors offer emergency services that utilize location information (currently provided by the user but in future, provided by GPS or similar technologies)
- Personal Messaging (Live Chat with Friends)

- ❖ Mobile Yellow Pages
- ❖ Information Services (News, Stocks, Sports)
- ❖ Personalized Content: Most of the people know that if they are attacked by intruder, trapped in fire or injured in car accident, they need to dial a few digits for immediate assistance.

3.2 Location Based Services – Techniques

Location of the caller is generally determined by various position determination techniques[6]. These include Cell-ID, Enhanced Observed Time Difference (E-OTD), Observed Timed Difference of Arrival (OTDOA), Wireless Assisted GPS (A-GPS) and hybrid technologies (combining A-GPS with other standard technologies). A location is a place where an object is physically situated in the real world. The location can be expressed in different ways using different reference frames such as absolute spatial location, descriptive location, or relative location. The different ways of expressing location will pinpoint the location of the object to a certain point, area or region somewhere on or close to the earth. Another factor that affect the accuracy of the location is use of position determination technology.

3.3 Background -- Geodetic datum

In spatial reference system, the Geodetic Datum defines the size and shape of the earth, origination and orientation of coordinate systems. The shape of the Earth is irregular different datum attempts to model it. Since the datum describes the earth differently, used datum will affect e.g. how accurately one can express the position of the object Referencing coordinates to wrong datum may result in position errors of hundreds of meters. Widely used geodetic datum is World Geodetic Reference System of 1984 (WSG-84) specified by the United State Mapping Agency. It is also used for Satellite Navigation System Global Positioning Services.

3.3.1 Coordinate Systems

The absolute spatial location can be expressed using many coordinate systems for example the Latitude-Longitude-Altitude system, the Earth Centred Earth Fixed (ECEF) coordinate system and Universal Mercator coordinate system. The most commonly used coordinate system is Latitude-Longitude-Altitude system.

3.3.1.1 Latitude--Longitude--Altitude system

In Latitude-Longitude-Altitude coordinate system location is expressed in term of Latitude Longitude and Altitude (see figure 3.2). The Latitude and Longitude are expressed in degrees. The Latitude is expressed in range of 0-90 degrees were 0 degree is equator and 90 degree is North Pole or South Pole. To differentiate between northern hemisphere and southern hemisphere "+" or "N" is used to indicate Northern hemisphere and "-" or "S" is used to indicate Southern hemisphere. The Longitude is expressed in the range of 0- 180 degree to the west or east from the prime meridian. To express to the west "+" or "W" is used and to express "-" or "E" is used. The altitude (or height) at a point is the distance from the reference ellipsoid to the point in a direction normal to the reference ellipsoid[4].

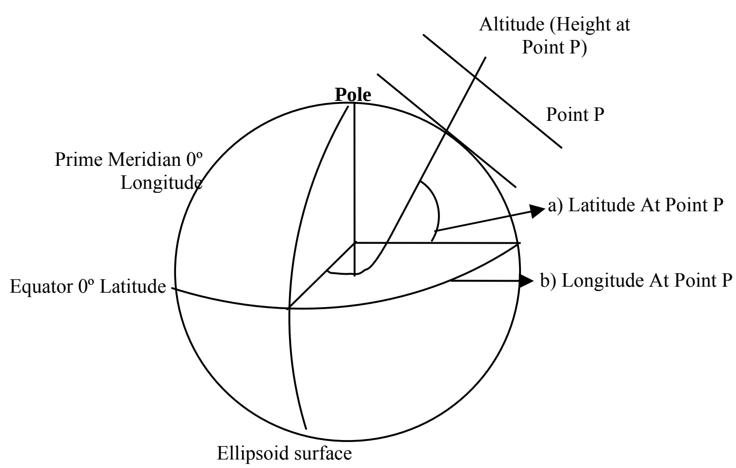


Figure 3.2 Longitude-Latitute-Altitude

3.3.1.2 Earth Centered--Earth Fixed ((ECEF)) coordinate system

Earth Centered, Earth Fixed coordinates (x, y, z) define a three dimensional position with respect to the centre of the mass of the reference ellipsoid (see Figure 3.3). The x-axis is defined by the intersection of the plane defined by the prime meridian and the equatorial plane. The y-axis is in the intersection of a plane 90-degree east of the x-axis and the equator. The z-axis points towards the North Pole.

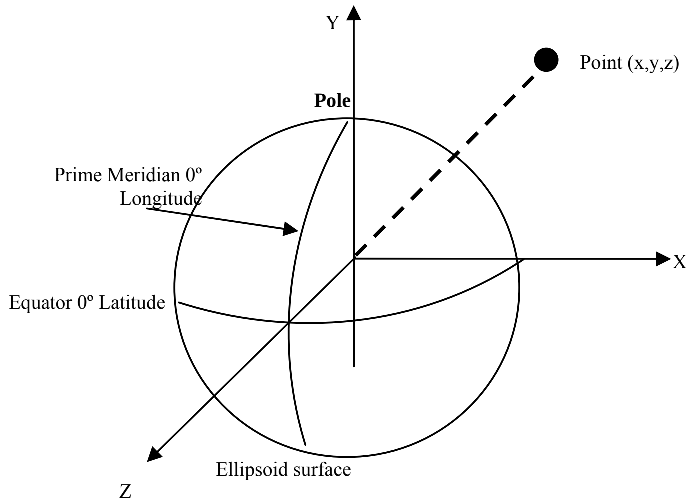


Figure 3.3 Earth Centered-Earth Fixed (ECEF) Coordinate system

3.3.1.3 Universal Transverse Mercator (UTM)) coordinate system

In the Universal Transverse Mercator (UTM) coordinate system the earth is divided into zones indicated by a number and a character. Universal Transverse Mercator (UTM) zone numbers designate 6 degree longitudinal strips extending from 80 degrees south latitude to 84 degrees north latitude. UTM zone characters designate 8 degrees zones extending north and south from the equator. There are special UTM zones between 0 degrees and 36 degrees longitude above 72 degrees latitude and a special zone 32 between 56 and 64 degrees north latitude. UTM coordinates (zone, easting, and northing) define two dimensional horizontal positions. Each zone has a central meridian. Eastings are measured from the central meridian with a 500-km false easting to ensure positive coordinates. Northing is measured from the equator with a 10 000 km false northing for positions south of the equator.

3.4 Positioning Determination Value Chain

A high level framework for location-based services has emerged that contains four critical components. Each component builds on the functionality offered by the previous.

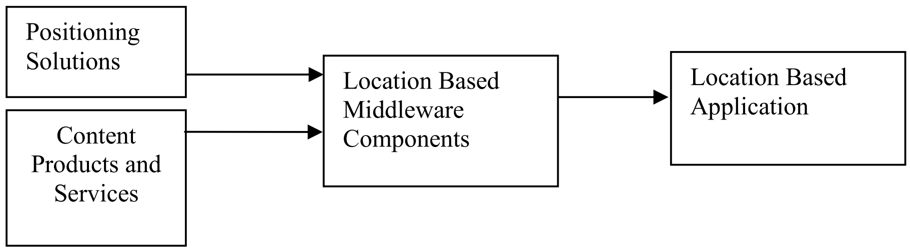


Figure 3.4 Position Determination Value Chain

3.4.1 Positioning solutions

It refers to tools that can locate a wireless device. Their job is to capture the location and covert it into meaningful positional information. Positioning solution uses various positioning services technologies as described in detailed in next section

3.4.2 Content products & services

It refers to the volume of reference information required to support a location sensitive service, including mapping, address information, route models, points of interest and real-time information. A significant quantity of geographic information must be acquired, aggregated, formatted and quality controlled. This process isn't just a once-off effort; it is a constant, cyclical process the objective of which is to improve completeness, conformity, consistency and accuracy. While content is a significant part of the location solution, it is often pushed down the priority chain in preference to technology related decisions.

3.4.3 Location based middleware platform

It is often described as the "Geo- Toolbox" middleware. It interfaces with the XY position content and billing systems to provide a rich API that enables application developers build location sensitivity into services at speed and with ease. This component is the workhorse of the overall solution, in that it undertakes a broad variety of complex tasks on behalf of many applications, including[3]:

1. Geo-coding: Ability to accept partial address information and return the positional information, taking into accounts the nuances of local address conventions and common human error.

2. Reverse Geo-coding: Ability to accept positional information (perhaps from the network) and return an intuitive description of that location in terms of a spread of nearby and relevant landmarks.

3. Location Refinement: Ability to accept positional information (perhaps from the network) and return an intuitive description of that location with an additional refined list of location points, so that the user can select their own location precisely.

4. Spatial Searches: A spatial engine able to support queries for "points of interest" based on geographic proximity, type, name etc. taking into account the local route network.

5. Route Directions: Ability to generate step-by-step navigation directions and route mapping, taking into account mode of transport, one-way streets, traffic lighting sequences and other traffic impedances

6. Map Rendering: Ability to rapidly compile renders and returns geographic information in the form of an optimum map image for a given location, scale, overlay information, number of colors and format.

3.5 Location Determination Technologies

Location Technologies mostly used by wireless carriers are handset-centric and network centric. These involve different levels of positional accuracy, hardware and software investment levels, and implications for the mobile operators[5].

3.5.1 Handset--Centric Location Technologies

3.5.1.1 Cell-ID

Cell-ID operates in GSM, GPRS and WCDMA networks. It requires the network to identify the BTS to which the cell phone is communicating and the location of that BTS. The Cell-ID Location service identifies the MS or UE location as the location of the Base Station and passes this information on to the location services application. Cell-ID was used earlier when high levels of location accuracy were neither mandatory nor necessary. If a handset is being used to make a call, then the information about the cell site that it is in will be updated to the network in real-time. However, if the handset is idle (i.e. switched on but not transmitting), then the last known transmission location will be stored by the network in the Home Location Register (HLR). In order to update the network’s information on the location of a handset, the network will page the device, prompting it to monitor the signal strength of the surrounding BTS, thereby informing the network of it’s Cell ID. The accuracy of this method depends on the cell size, and can be very poor in many cases, since typical GSM Cell is anywhere between 2km to 20KM in diameter. With Pico cells, accuracy of 150 meters can be achieved. Using either one or both of the following techniques – Timing Advance (TA) and Signal Strength (RX Measurement/NMR), can increase the level of accuracy.

3.5.1.2 Cell-ID + Timing Advance (TA)

The time at which a terminal sends its transmission burst is critical to the efficient functioning of a GSM/GPRS network. Every mobile station within a given cell will be at a varying distance from the serving base station, yet the burst from each device must reach the base station at the exact moment that their receptive timeslots become available. Consequently, it is necessary for the mobile station to co-ordinate with the base station at the right time. Even tough the burst arrive either before or after the availability of the allocated timeslot, the mobile station is instructed to “advance” the transmission of its burst accordingly. As the duration of the timing advance for each mobile station is dependent upon its distance from the base station, it is possible to use this information to determine how far away the caller is. TA information is only of any use in increasing the level of positioning accuracy within cells with a radius greater than 550 meters. This is because the adjustments made to the timing of the mobile station’s transmissions are calculated depending on how many multiples of 500-550 meters the mobile station is distant from the base station.

3.5.1.3 CELL-ID + Signal Strength (RX Measurement / NMR)

The Mobile Station continuously measures the signal strength from each of the base station report this information back to the serving base station. This is so that the Mobile Station is able to transmit to – and receive from – the base station that has optimum signal strength, thereby improving the quality of call for the end user and making most efficient usage of network infrastructure. With this signal strength information, it is theoretically possible to calculate the position of the caller, by taking into consideration the rate at which the strength of an RX signal degrades as the distance between the transmitter and receiver increases. There are however number of factors that limit the effectiveness of this method, distance is not the only factor to affect RF waveform propagation. The characteristic of terrain between the transmitter and receiver, as well as the issue of indoor attenuation both has significant impact upon these measurements. The denser the material that a building is made from, as well as the higher the floor that a person is calling from, both have an increasingly negative affect on the strength of the signal received. Signal Strength/RX measurements are sometimes referred to as Network Measurement Results (NMR).

3.5.2 Network--Centric Location Technologies

3.5.2.1 Network Based Triangulation Technologies

A number of different network-based measurement technologies can be used to locate a caller. Some of the major ones are listed as below.

3.5.2.1.1 Enhanced Observed Time Difference

E-OTD operates only on GSM and GPRS networks. In GSM, the MS monitors transmission burst from multiple neighboring BTSs and measures the time shifts between the arrivals of the GSM frames from the BTSs to which it is communicating. These observed time differences are the underlying measurement of the E-OTD radiolocation method and are used to trilaterate the position of the mobile devices. The accuracy of the E-OTD method is a function of the resolution of the time difference measurements, the geometry of Neighboring Base station and the signal environment. The Mobile handset must measure time difference from at least three base stations to support two-dimension position determination (no altitude measurement is provided). E-OTD requires precise time information.

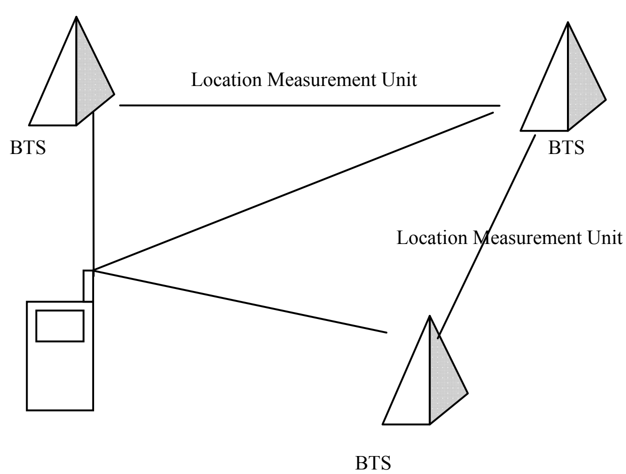


Figure 3.5 E-OTD Operations

Location Measurement Units (LMUs) is required in the GSM and GPRS network for precise time information. Most important requirement for this technology is that BTS in the network is observed by at least one LMU. Further, special software is required in MS to support E-OTD. The need for LMUs introduces significant infrastructure changes, as it requires the installation of thousands of LMUs in GSM/GPRS networks. This needs significant network planning, an assessment of the RF impact to the network, adherence to local ordinances where new sites are involved, and the expense to plan, install, test and maintain the network of LMUs. This level of intricacy complicates the operator's ability to provide roaming support for an E-OTD based location service and extends the time required to deploy network-wide location services. E-OTD offers improved performance relative to Cell-ID, but requires the use of LMUs. This increases the cost and complexity of implementation, as described above. E-OTD also requires that a large number of data messages be exchanged to provide location information. And this information is updated constantly. This message traffic is much greater than used for A-GPS or Cell-ID, and E-OTD uses more network bandwidth than these technologies. The accuracy is affected by multipath and signal reflections as it utilizes at least three base stations. The system is quite inaccurate in rural areas as there is lesser number of BTS.

3.5.2.1.2 Observed Time Difference Of Arrival

OTDOA operates only on WCDMA networks. OTDOA is generally considered a WCDMA version of E-OTD. The OTDOA Location Services Technique estimates the position of a handset by referencing the timing of signals as they are received at the UE from a minimum of three Node B station (BTS). The handset's position is at the intersection of at least two hyperbolas defined by the observed time difference of arrival of the WCDMA frames from multiple Node BTS. The weakness is similar (poor yield in areas without at least three Node Bs, poor accuracy along linear networks, multipath degradation, compatibility with only one network, etc.). As WCDMA network is based on CDMA, it is optimized for low power and the efficient use of communication bandwidth[5]. The handset ability to see and use multiple

Node B station is severely limited. This affects accuracy and overall OTDOA performance in many cases is worse than EOTD. In order to synchronize a network to the degree of precision required to support OTDOA location require using more expensive time units, such as LMU.

3.5.2.1.3 Angle of Arrival (AOA)

AOA utilizes multi-array antennas and tries to estimate the direction of arrival of the signal of interest. Thus a single AOA measurement restricts the source location along a line in estimated AOA. If at least two such AOA estimates are available from two antennas at two different locations, the position of the signal source can be located at the intersection of line bearings from the two antennas. Usually multiple AOA estimates are used to improve the estimation accuracy by using the redundant information.

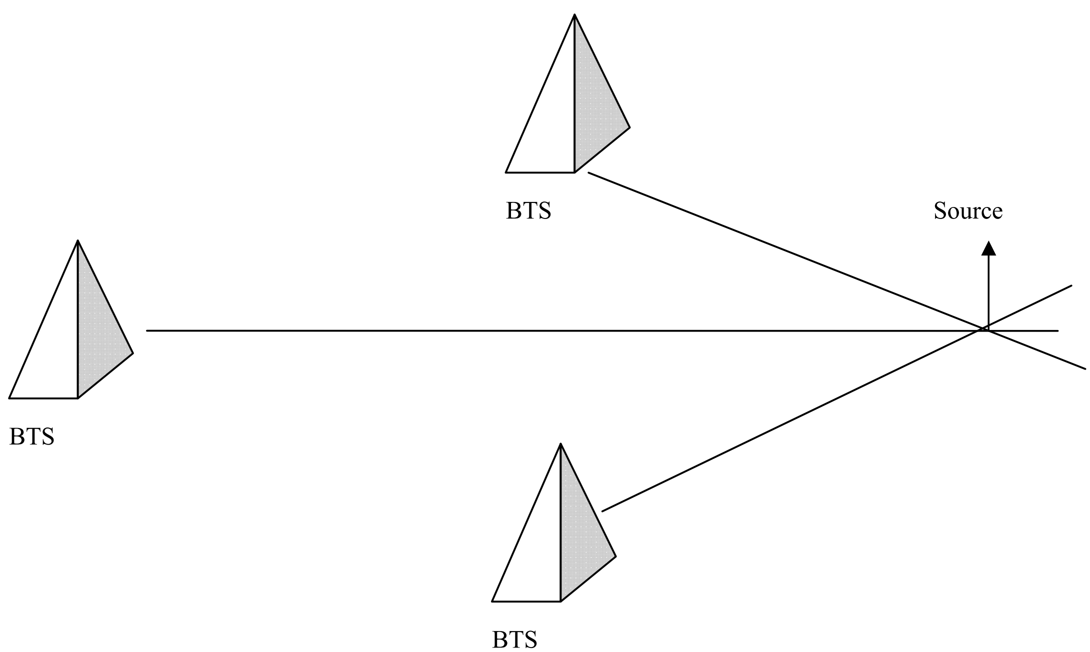


Figure 3.6 Direction finding position location solution

Figure 3.6 shows the method where the source location is found by the intersection AOA of the signal for three Antenna Arrays. AOA estimation is often used with short baseline to reduce or eliminate ambiguities and other time it is used with long baseline to improve resolution. To estimate the AOA, algorithms are used that exploit the phase difference or other signal characteristic between closely spaced antenna element of antenna array and employ phase alignment method for beam/null steering. One of the crucial requirement for accurate estimate of position is that signal coming from source to the Antenna array must be coming from the line of sight (LOS) of direction. However it is not often the case in cellular system, which may operate in heavily shadowed channel. Another drawback is considerable cost of installing antenna arrays. Location service using AOA for Position fix may need regular calibration since a minute change in the physical arrangement of the array because of wind and storm, may result in considerable change in Position location error. Another problem with AOA method is the complexity of AOA algorithms.

3.5.3 Global Positioning Services (GPS)

GPS is used in GSM, GPRS and WCDMA networks. GPS is a Satellite Navigation System funded by and controlled by the US Department of Defense (DoD). Despite the large base of millions of civil users of the system worldwide, the system was designed for and is operated by the US military personnel. GPS provides specially coded satellite signals that may be only processed by a GPS receiver, enabling the receiver to compute position, velocity and time. The basic measurement performed by a GPS receiver is the time required for a signal to propagate from one point in space to another. Because in the general case, the speed that RF signals travel is known with relative accuracy this time measurement can easily be converted to distance - range from the RF source. If the distance from the receiver to four satellites is calculated, the receiver can accurately determine its position anywhere on earth. Four GPS satellite signals are thus used to compute positions in three dimensions and the unknown time offset in the receiver clock. The system allows the military users to make use of an enriched signal set, achieving a much better guaranteed accuracy than civilian receivers may achieve. The system's operation relies primarily on the GPS satellites. A number of 28 LEO-SV (Low Earth Orbit - Satellite Vehicles) are positioned in such orbits as to cover almost all the earth surface, At any time 4 to 6 satellites are on stand-by in orbit to replace malfunctioning. The satellite orbit at an altitude of eleven thousand nautical miles and follow six different orbital paths, each satellite orbiting the earth twice every 24 hours. The satellite transmits on two frequencies, the frequency known as “L1” – which resides at 1575.42 MHz being used by civilians. The GPS satellite also transmits the second ranging signal known as L2 at 1227.6 MHz used for military purposes. The users of the system take advantage of special purpose GPS receivers to convert the signals into position, velocity estimates, while the receiver may be also used as a highly accurate timing source. GPS receivers are used for navigation, positioning, time dissemination, and other research[13].

3.5.4 Wireless Assisted – GPS

As discussed above the GPS receiver measures the distance by measuring the time required for the signal to travel from the satellite to the receiver. For accurate time information the received satellite signal should be relatively strong. To overcome this limitation Assisted GPS (A-GPS) receiver utilizes aiding data from an A-GPS LS that provides the receiver information, which increases the start-up sensitivity by as much as 25 dB (relative to conventional GPS) and reduces the start time by 5 seconds. This approach eliminates the long start times typical of conventional GPS (one to two minutes) and allows the GPS receiver to operate in difficult GPS signal environments. AGPS yield will drops in environment where the satellite signals are relatively blocked. Figure 3.8 shows A-GPS operation. A-GPS operates in two primary mode MS/UE-based and MS/UE-assisted. In MS/UE-assisted mode, the A-GPS receiver calculates only Pseudorange from satellite signals and sends this information back to A-GPS Location Server (LS), which calculate the position. In MS/UE-based, the position calculation is made in the receiver, which requires an extended set of assisted data.

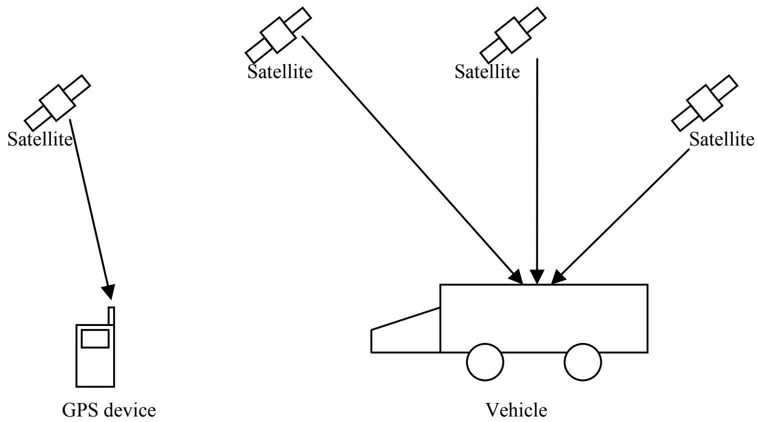


Figure 3.7 GPS operation

A-GPS provides better accuracy than CELL-ID, E-OTD or OTDOA, and expensive element like LMU is not required. An A-GPS has almost negligible affect on the infrastructure and can easily support roaming, but requires GPS circuitry inside the phone.

3.5.5 Hybrid technology

A-GPS hybrid operates on GSM, GPRS and WCDMA networks. Hybrid location technology combines A-GPS with other location positioning technique in a way that allows the strengths of one to compensate for the weakness of the other so as to provide a more reliable and robust location solution. Because A-GPS is air-interface independent, it can be combined with any of the major technologies as discussed in this paper. Hybrid solution provides accurate and reliable positioning even where the independent network solution and unassisted GPS solutions fail. Figure 3.8 shows the operation of hybrid technology. Most common implementation of Hybrid technology for GSM, GPRS and WCDMA is to combine A-GPS with Cell-ID. This improves yield in areas where A-GPS cannot produce position information and provides the accuracy of A-GPS in all other cases. A-GPS accuracy is typically useful and degrades only deep inside buildings or in the dense urban areas where Cell-ID may still be able to produce a position. The combination of A-GPS and Cell-ID also incorporates the roaming advantage defined for both Cell-ID and A-GPS.

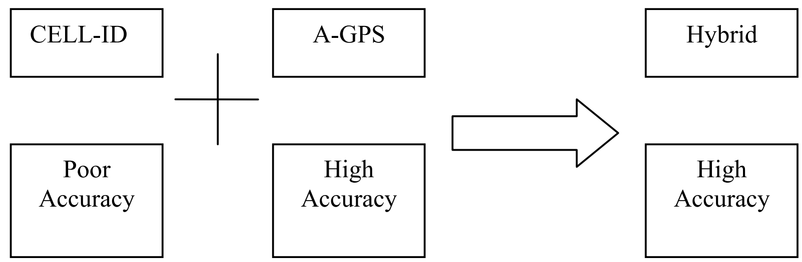


Figure 3.8 Hybrid Solutions

A-GPS can also be combined with E-OTD or OTDOA. This approach requires spot deployment of E-OTD and OTDOA, allowing A-GPS to be used in the majority of the network to provide basis for most location information. Hybrid approach generally improves yield and performance of location technology.

3.6 Performance Characteristics

This section identifies the performance characteristics of the technologies as discussed in the previous section. The most common measures of performance are location accuracy, since accuracy is easy to measure and considered as indicative of the quality of solution. The location technology must produce the location information reliably, quickly and with consistent performance across a variety of networks and diverse geographies. The most important for performance to be adequate for Location services is all these goals should be achieved simultaneously. Figure 3.9 shows performance characteristics of these Position determination technologies.

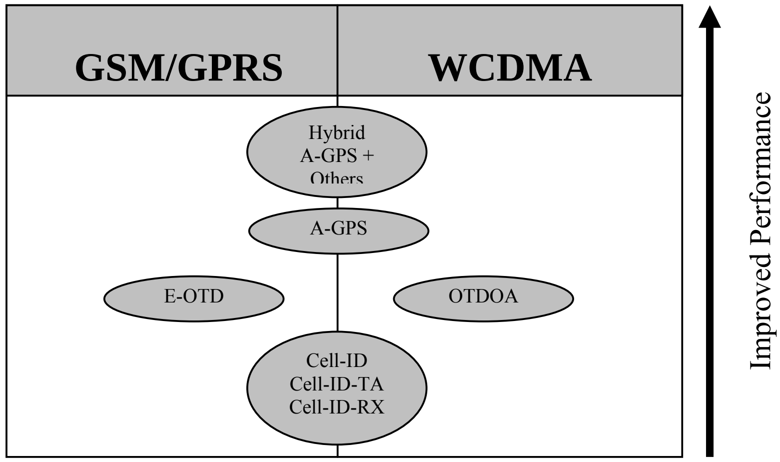


Figure 3.9 shows performance characteristics of these Position determination technologies.

3.7 Basic Location System for GSM/GPRS Network

To prepare a GSM network for the introduction of location based capability, certain modifications must be made to the existing network infrastructure, along with the installation of new network components. The most essential requirement is to upgrade the software component in BTS, as well as the software within the mobile switching centers (MSCs). The basic network element commonly used in providing location information to location services application is listed in figure 3.10

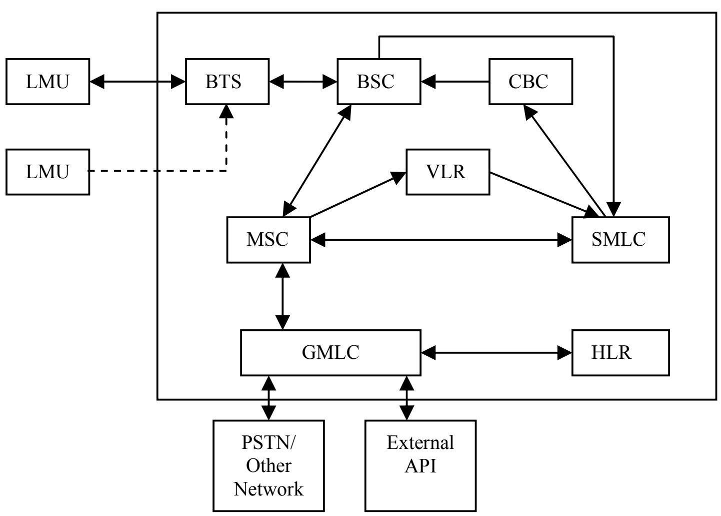


Figure 3.10 Basic Location System elements on GSM/GPRS network

The additional network component required for position calculation is referred to as the Mobile Location Center (MLC). The MLC contains detailed database of network information, including the precise geographic location of all of the networks base stations, plus all relevant cell sector size and coverage data. It is this data that is used to determine the location of a mobile station. There are two variations of the MLC, the Serving Mobile Location Center (SMLC) and the Gateway Mobile Location Center (GMLC). Depending on the amount of location traffic that needs to be processed, an operator may need to deploy regional MLCs (i.e. SMLCs), to help manage the data load more efficiently. GSM specification states that each BSC should be accompanied by an SMLC, however an operator may decide to deploy one SMLC, which, via the MSCs, is capable of serving all the network's BSCs. The functions of these two-network components are explained below:

SMLC – The SMLC co-ordinates the network resources necessary to perform positioning calculations. A request is made to the SMLC for the location data of a particular end user, which is calculated by the SMLC through it's interaction with the BSC, using position determining techniques in next section. This information is used by the SMLC to generate location information data, which is then routed through to the GMLC.

GMLC – The GMLC acts as a gateway connecting the location data of the wireless network to the Public Switched Telephone Network (PSTN), other wireless networks, as well as to providers of remotely hosted content. The GMLC receives location information data from the SMLC and routes it to the operator's suite of mobile applications. The GMLC performs function such as authorization and authentication procedure for positioning requests, safeguards the privacy of the end user's data, generates the billing information for location-based services and returns the requested location enabled information to the end user. MLCs may also incorporate additional PCF (Position Computation Function) units. These are proprietary servers provided by individual position determination entity (PDE) vendors, designed specifically for the determination of longitude & latitude co-ordinates using the specific type of data provided by their particular solution. PCFs may be integrated into the MLC, or simply reside along side it. A location-enabled network also makes use of the following existing GSM network components:

VLR – The Visitor Location Register is connected to the Mobile Switching Centers (MSC) and interfaces with the SMLC. The purpose of the VLR is to store the following subscriber location information (for both home and roaming users), for the purpose of quickly setting up the next call: Mobile Network Code (MNC), Mobile Country Code (MCC), Location Area Code (LAC) and the Cell ID (CI). AGSM network contains multiple regional VLRs.

HLR – The Home Location Register (HLR) is traditionally coupled to the MSC. However in a location enabled network the GMLC resides between the two. The HLR stores the last known location information for each mobile subscriber using the network. Specifically, the HLR stores the Location Area Code (LAC) and the Cell ID, as well as any user profiles/service subscription information about each end user. This allows the HLR to provide routing information for calls and data to and from each end user[15]. The process of making a location request and returning the appropriate information to the end user operates as follows:

1. A request for location-enabled information is sent from the user's handset (MS).

2. The request is received by the base station, along with the appropriate location information (provided either by the handset or by the network of LMUs, depending on the type of position determining technology in use). In this example, the LMU may represent either an E-OTD receiver monitoring

synchronization bursts, a GPS receiver monitoring GPS satellite data, or a TDOA/AOA antenna receiving time of arrival/direction of arrival data.

3. The location information is routed from all BTS' receiving from the MS to the BSC.

4. The BSC then routes any traffic data (e.g. Cell ID information) to the Visitor Location Register (VLR), via the Mobile Switching Center (MSC). The BSC also routes any data from dedicated position determining equipment directly to the Serving Mobile Location Center (SMLC).

5. The traffic data from all the network's VLR is provided to the SMLC for conversion into ECEF coordinate system.

6. The data received by the SMLC enables it to instruct the Cell Broadcast Center (if the network has one) when – and what - information should be transmitted.

7. The SMLC, utilizing either traffic data, or information from dedicated PDE, calculates the Longitude-Latitude of the mobile station using its network information database.

8. This location information data is routed to the GMLC. GMLC interfaces with the operator's suite of location based application and any remotely hosted applications/ content, to return the location-enabled information that the end user requested.

9. The appropriate location-based information is then routed back to the end user by the GMLC

3.8 Comparative Analysis of Location Technologies

Each technology should be evaluated in light of its advantages and disadvantages for the application, considering the simultaneous performance, implementation and cost requirements.

3.8.1 Costs

The cost of implementing location services depends on a large number of factors, such as handset modification, infrastructure modification, maintenance activity, network expansion plans, etc. Table 3.1 summarizes cost factor location based technologies.

Cost Area	Cost Factor for Cell-ID	Cost Factor for E-OTD/OTDOA	Cost Factor for A-GPS
Handset Cost	Low-No modification required	Low- Modifications to existing handsets are required for E-OTD. Special software is also required in OTDOA handsets	Medium-A-GPS circuitry must be added to the handset. This circuitry can be deeply integrated into the phone's component so that its cost is optimal
Infrastructure cost	Low-Requires addition LS software only no other modification required	High –cost depends on the size of deployment at base stations throughout the infrastructure, where location coverage is desired	Low-Requires additional A-GPS LS software only no other modification required.
Expansion cost	Low-This cost is low as long as network expansion is done into a network that	High-Cost is high since the network being expanded must have LMUs at the majority of	Low-A handheld enabled with A-GPS requires

	supports the technology	the base station	no change to move to another network. A small infrastructure change will accommodate expansion
Maintenance cost	Low No special maintenance required	High After being deployed LMU must maintained according to a specified time schedule	Low-Negligible maintenance cost in the infrastructure since locations-server locations are limited and typically centrally located
Overall cost factor	Low-Overall technologies in this category are relatively low cost	High-The cost factor is high to deploy initial system and remains high throughout system deployment for ongoing maintenance	Low to medium-High performance and very low infrastructure costs make this an attractive technology

Table 3.1 Summarizes cost factor location based technologies

3.8.2 Performance, Implementation and Cost Trends

The majority of location service applications require high performance at reasonable cost to optimize return on investment[11]. Table 3.2 provides summary of performance and implementation for each technology.

Trend	CELL-ID	E-OTD,OTDOA	A-GPS
Performance	Cell-ID accuracy varies dramatically	Provides accuracy compared to cell-ID	Provide

	and is often very poor Provide good Coverage	Has Coverage problems where there are limited base stations	optimum accuracy compared to the other Location technologies Has coverage problem deep inside the buildings
Implementation Trends	Easy to implement	Difficult to implement E-OTD requires change in handset Requires LMUs Does not easily support roaming in wide areas and other network	Easy to implement in the infrastructure Requires handset changes Requires no major infrastructure change Easv To roam wide areas and into other networks.
Overall cost evaluation	No change in handset required	Initial cost is high as required to provide 1 LMUs per 1.5 BTS Cost to maintain is high	Initial cost is driven by cost of handset Cost to maintain is negligible
Standard supported	GSM, GPRS and WCDMA	E-OTD-GSM only OTDOA-WCDMA only	GSM. GPRS and WCDMA

Table 3.2 Summary of Performance, Implementation and Cost for Location Technologies

4. Methodology

The research on this thesis focuses on Location Technology, Mobile technology and GIS to some extent. Location technology is used to implement LBS service in mobile with aid of GIS system. Our raw data is content in GIS system from which we extract data to be used in mobile. Coming up chapter discuss in detail about how to design and implement these technology.

4.1 Design & implementation of GIS

4.1.1 GIS

Like any information system, a GIS is an organized accumulation of data and procedures that help people make decisions about what to do with things. In a GIS these things have one characteristic that makes them at least a little special—their location is an important part of what they are. People have been constructing GISs to manage and analyze types of things for which location matters for almost the last 40 years[21].

GIS is composed of

a. People—the users of the system

b. Applications—the processes and programs they use to do their work

c. Data—the information needed to support those applications

d Software—the core GIS software

e. Hardware—the physical components on which the system runs Versions of these components of a GIS

They certainly predate GIS specifically and are applicable to any information system. Figure 4.1 should be read as a sentence, and it proceeds from the most important elements to the least important elements.

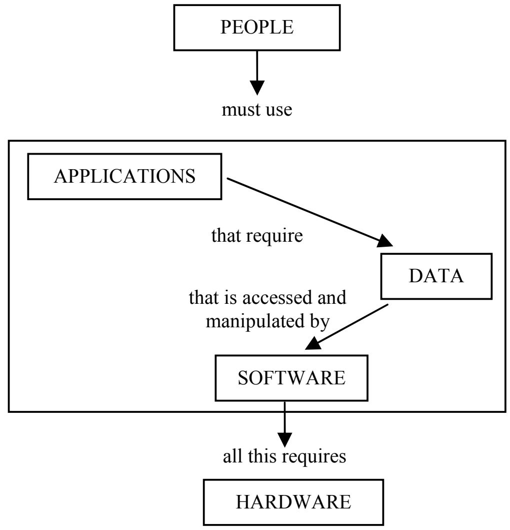


Figure 4.1 Components of an enterprise GIS

The people are the most important component, although some would argue for the data. Information systems, geographic or not, spring from the needs of people in organizations to do work, answer questions, and generally interact with the world and the people and organizations in it. An information system is supposed to support the work, to make it quicker to do with more consistent results, and to provide high levels of confidence in the output. The process of design and implementation of a GIS begins with people and their needs and ends up with applications in the hands of people who do the work. The entire system exists to support them and their tasks. The applications come next in the hierarchy because they define the work that needs to be done. In organizations people need to create all kinds of reports, make all sorts of decisions, and generally apply their skills so the work gets done. The processes they develop to do these things are the applications. Some applications are routine and get done multiple times a day, whereas others are less routine but get done with some regularity, and then there are specific analytical applications that might have to be accomplished only rarely or even just once. The applications arise out of the mission and goals of the organization. In any information system one need to know what applications the system will be expected to support. Applications require data to work. One can't generate a map of sales potential or customer locations without the appropriate data tables necessary to create that type of output. These tables will reside in a database (possibly more than one), and the system will require software to access, manage, and manipulate the data so that the application can generate a useful product. The data support the application, and if there were no software for data storage and retrieval, the application would have to get done somehow.

4.1.2 Corporate or Enterprise Geographic Information Systems

An enterprise GIS is one that is designed to meet the needs of multiple users across multiple units in an organization. Although many organizations have a GIS in one or more units, they have been built to support the needs of those units only and may be of little utility to other departments in the organization. It is common to find an organization where a each department its own GIS license to support to do work. An enterprise GIS is built around an integrated database that supports the functions of all units that need spatial processing or even mapping. That database, whether centralized for real-time access by all users or replicated across many computers, is the engine of the enterprise GIS. In a well-designed system, users in the departments where GIS already existed will interact with the GIS in ways that are not much different from what they had been doing. New users will interact with the system with custom-designed applications that use the centralized data. The system will no longer be a particular department's GIS but will be the organization's GIS[19]. This kind of corporate, or enterprise, GIS is different from a single-unit or project-oriented GIS in several ways:

a. Data are standardized and redundancy is reduced.

In municipal government, a type of organization that is an excellent candidate for an enterprise GIS, an assessment of information needs will almost always reveal that many different units of the government maintain information on addresses. But there will be no standardization of address composition; that is, some units may store addresses in a single field of a data table, others break it into street numbers and streets, and a still others might use the address parsing scheme that came with a packaged information system. As a result, there is a lot of duplication and confusion about what an address is and how to store it. An enterprise GIS for a local government may have standards and a non redundant master data table of addresses.

b. Database integrity is maximized

When people start using information and modifying data in databases, what was once clean and accurate data has a way of getting dirty and inaccurate. Names are misspelled, addresses incorrectly recorded, records deleted that should not have been; the list of things that can happen to corrupt data is very long. An enterprise GIS will have safeguards and procedures to minimize that kind of data loss. All organizations run on a supply of accurate and timely information to make decisions that will move the organization forward, and it should be as good as possible. The progression of data in information systems, GIS included, has been from databases on mainframe computers maintained by centralized staff to distributed databases on individual desktop computers, and now, full circle, back to central data servers. The data are too important and need careful watching and maintenance.

c. Units come together through the database

In a complex organization there are many different departments or units with unique missions and goals. When they come together in the creation of an enterprise GIS, they become aware of other units' needs for information and begin to see their own needs in a different light. People in almost all of the organizations that undertake and successfully complete the implementation of an enterprise GIS will speak positively about the benefits to the organizations that go beyond the cost and times savings the GIS might provide. An enterprise GIS database is a gathering point for different units, and how they come together in its creation and maintenance usually benefits the organization in some unexpected ways.

d. There is a consistent look and feel to output

In the process of design and implementation of this kind of database it is necessary to set standards and requirements on what the output from the database will look like. This results in a consistent-appearing output that is usually important to top management. Typically, management does not want any individual units of the organization looking too much out of step with the others, and an enterprise database with standardized outputs makes uniformity much more likely.

e. Geographic information costs are centralized

A common problem that managers encounter when they are building a case for an enterprise GIS within the organization is that the current costs of obtaining and using geographic information are diffused and hidden within the budgets and operations of many different units. Anyone who tries to conduct a cost benefit analysis around a GIS discovers this. It is comparatively easy to calculate the costs of implementation but much more difficult to compare those costs to ones presently being incurred.

4.1.3 The GIS Strategic Plan

Strategic planning is an activity with which most complex, multiunit organizations are familiar. The organization may have prepared many strategic plans, overall plans around the goals, and specific plans for projects and other important changes the organization has gone through. A strategic plan for GIS is therefore just another working out of a familiar process and should not cause a lot of anxiety, although midlevel management and front-line staff often feel a little put out about participating in yet another planning exercise. A strategic plan for GIS is a document of interest principally to top-level management and outside organizations[20]. Top-level managers needs a document they can refer to when difficult decisions need to be made; is this something that fits our strategic plan or not?

The strategic plan for an enterprise GIS should contain at least the following:

- A concrete discussion of how an enterprise GIS fits within the existing mission statement of the organization.
- A tentative and light discussion of how GIS is going to fit inside the organization with recognition that the design and implementation process may require modification of the plan.
- A timetable with checkpoints.
- A plan without a timetable is a dangerous thing because it gives management no way to monitor
- If the implementation is to be phased (i.e., brought into certain units first and later diffused to the entire organization), those units need to be identified in the strategic plan.
- Initial decisions on what existing staff will do in the design and implementation process and what functions the organization will hire consultants to do the remaining
- Some statement of the resources the organization is willing to contribute to the process.

Steps	Central Questions	Primary Outcomes/Products	Who is involved? (Staffs or consultants)	Secondary goals
Needs assessment requirements/analysis	What are the current and future needs for geographic information? How do people currently use geographic information and how would they like to?	Needs assessment report. This is a central document to guide the implementation plan phase.	Potential users and managers of using departments and departments that may later become users.	To built support at front-line and mid management levels.
Strategic Plan	How will GIS further the mission of the organization? How will the enterprise GIS fit within the organization?	Strategic plan with rough time-table, agreed upon by all stake-holders not just top management.	Top and mid-level managements plus core GIS committee.	To built support in top management.
Implementation Plan	What are the phase steps we need to take to implement an enterprise GIS that will meet our needs and further our goals?	Implementation plan with detailed timetable, including a decision on high-level design questions.	Full GIS implementation committee with technical support.	
Design Phase	What are tables, fields and initial application are	Data schema, data dictionary, applications, flow charts.	Technical staff.	

	required to meet and user needs and how should they be arranged?			
Implementation Phase	What process will we use to populate the schema and implement the initial application?	GIS database and selected applications.	Technical staff and front -line users.	
Pilot Project	Where and when can we test this system?	Output information for pilot project.	Selected front-line staff and management.	To quickly show everyone, particularly management, that it will work.
Application development	How can we add to the existing set of application and improve existing ones?	Updated procedure manuals.	Technical staff working with front -line users.	Convincing reluctant users of the system's value.
Maintenance and Upgrade Plan	How will we keep the data, software and hardware current?		Technical staff with front -line managements.	Ensuring ongoing support.
Training	Who needs what kind of training in what applications?		Technical staff and skilled users	Convincing reluctant users of the system's values
Evaluation	After the reasonable training and working period, is it working as advertised?		All users	Convincing top management of the wisdom of their decision.

Table 4.1 Steps in Design and Implementation Process

4.2 Designing Spatial Data

4.2.1 The Two Principal Data Models

Practitioners of GIS have for years grouped geographic data into two classes of models: raster and vector data models. The choice of raster or vector data models for a GIS has usually been presented as an either-or question and sometimes as a debate over the merits of the two data models[15].This view has presumed that there is one data model that is better than others and that the data models are like prizefighters: each exploits the weaknesses of the other and uses its strengths to represent the world in a better way. More recently, however, practitioners have realized that their needs may incorporate multiple models for spatial data, and it has become unnecessary to view the process as an either or choice. Many, but not all, GIS software systems can move data back and forth between these two data models, and one may discover that certain tasks are easier to perform in one or the other of the data models. But many users find that one or the other of the data models meets their needs so that one is best for their applications. Organizations that require high-quality mapped output from their GIS will find that only the vector data model allows that. Users who will be doing lots of overlaying of different geographic features will find out that the raster data model is more efficient for that type of task. When one is trying to develop a model of a real-world object, it is vital to keep several questions in mind. By knowing the answers to these questions one can ensure that our model of the real-world objects will relate successfully to the object[15]t.

4.2.2 What Is the Purpose of This Model?

If the purpose of the model is solely to be a digital representation of the feature for mapping, the design issues are rather simple. It needs to decide what kind of digital feature that will use to represent the real-world feature and how to symbolize it on the map. In the case of a land parcel, if the sole purpose is mapping of the land parcel for visual representation, the parcel will consist of a set of lines with symbolization and a point inside these lines where some unique identification number is attached. There may be one symbol representing the coincident features of a road right of way and a parcel edge and a second line symbol for the other edges of a property. The point symbol may not be visible at all because its only purpose is to serve as an anchor for some text. The digital line, point, and text features may exist separately in the GIS because the map reader does the linking between them visually. The presence of a number inside a set of bounding lines informs the reader that that number identifies the polygon represented by the bounding lines. This is a common way to visually represent land parcels using CAD tools. If the model is to serve multiple purposes, design concerns become more complex. Using the land parcel example, let us say that in addition to simple mapping we want to be able to link the assessor's information so that this information can be included in a report that provides a quick printout of the important information along with a map of the parcel and adjoining parcels. In this case, the land parcel can no long exist only as a set of lines but must be modeled as a polygonal area in the GIS. In addition, the unique identifier (parcel identification number, or PIN) must be in the GIS database and directly linked to the parcel so that it can be connected with the assessor's information. The system, not having the ability to see the PIN inside the lines bounding the parcel, must know the parcel's PIN and relate it to that parcel and only that parcel. The representation or model of a real-world feature may be different depending on the purpose of the model. Although a land parcel in the real world is a polygon, there may be some value in representing that polygon as a point in some Designing Spatial Data. The fire department may need to have the buildings on land parcel represented as a digital image so that they can see the buildings and where they are relative to each other when planning or evaluating a fire response. But when they want to know what hazardous materials are stored in that building and where in the building they are stored, the building may need only to be a record in a database linked to an address or a point. A business responsible for distributing goods over a wide area may be able to find good locations for a warehouse by modeling their regional demand and supply facilities as points along a network of major highways, but when they have to actually move goods from a particular warehouse to a set of locations, they will need a larger-scale model of the transportation network. So the same real-world feature may exist in our GIS database several times for the different purposes. The need to display or analyze geographic data at very different spatial scales also results in this apparent duplication. A state agency trying to represent all the land parcels in the state may be forced to model each parcel as a point to even begin to display and locate the hundreds of thousands or millions of parcels that exist. At the state scale that may be the right choice, but at the local level modeling the parcels, as point will not suffice. The parcels will have to be modeled as polygons. At one scale we might want to represent only major arterial roads in our database, but at a larger scale we might want to include all roads.

4.2.3 Which Data Model, Raster or Vector is better Suits?

Traditionally the highest-level model choice in GIS has been presented as the choice between the raster (grid) and the vector model. A diagram such as Figure4.2 is standard for any introductory text on GIS. The raster model represents the real world as a tessellation, with grid cells having a certain length and width and covering the entire rectangular area of interest. The value in the cell represents the real-world feature at that location within the cell. The real world is not modeled as set of points, lines, or polygons but as a complete covering of regular (usually square) cells. A point can be located only within a cell, a line is a collection of connected cells underneath the feature, and a polygon or area feature is a set of connected cells touching the polygon. The vector data model represents real-world features as strings of x/y pairs representing spatial information about the features. A point is a single pair of values with an x value representing the longitude or x dimension of the point in some coordinate space and a y value representing the latitude or y dimension of the point. Linear features may consist of a set of connected points or some mathematical function describing the beginning point of the feature and the formula that constructs it. All GISs that support the vector data model can model circles and other curved lines as sets of short, connected straight lines. Some are able to store information in such a way as to represent the feature as a curve. If that is important in our model of the world, we need to ensure that the software we select can treat features that way. Polygon or area vectors model the real-world feature as a connected set of lines that closes on itself. These features are discussed in more detail later in this chapter. For these entire three basic feature types in the vector data model one can imagine that a real-world feature might consist of multiple instances of points, lines, or polygons. In a pavement management system a section of pavement may consist of several segments of street that together make up the management section. If the purpose were only mapping, one could represent each segment between intersections as single lines, but for managing pavement would need to represent those multiple lines as a single thing. In a similar fashion, governments frequently have noncontiguous pieces of territory such as islands off coastline, and we can combine these into a single multipolygon region. Most GISs are able to handle both raster/image data and vector information; so the correct question is what is the proper data model for particular features rather than what is the proper data model for the entire GIS. This discussion ignores the quad tree data model, which is the third principal model for geographic data; some consider this a sub model of the raster data model. This data model is still atopic of research and there are a few commercial software systems that use it for geographic data; advances in computer processors and storage have removed a lot of its early appeal. There are five possibilities for combining the [18]

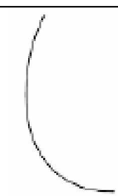
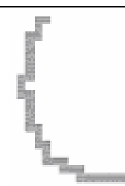
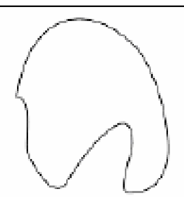
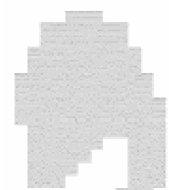
	Vector	Raster	Considerations
Point Feature			Indeterminate location within pixel (raster)
Line Feature			Aliasing or stair steps, length calculation (raster); true arc or approximations with many straight lines (vector)
Area Feature			Raster cell size (resolution)

Figure 4.2 Raster and Vector comparisons[20]

raster and vector data models that we need to consider when implementing a GIS:

a. Principally vector with raster underlay. This model combination uses the vector model as the principal data model, but for context there is a raster back drop. This backdrop is often a digital version of aerial photography, digital orthophoto, but sometimes is processed satellite information. This option is becoming more common as compression techniques for the Point Feature, Line Feature, Area Feature, VECTOR RASTER CONSIDERATIONS Indeterminate location within pixel (raster) Aliasing or stair steps, length calculation (raster);true arc or approximations with many straight lines(vector)Raster cell size(resolution). The Two Principal Data Models background digital images improve, allowing rapid redraw and panning. This option assumes that we will be acquiring the raster backdrops in some standard format and do not plan to do any processing or changing of that backdrop.

b. Principally raster with vector overlay. This model combination is useful when remotely sensed imagery is the primary source of data for our GIS or we need to directly process any raster information. Software systems marketed principally as remote sensing applications all have the capability of draping vector information on top of the raster, and the distinction between GIS and remote sensing software systems of this sort is becoming more blurred. The vectors provide a map-like context for the raster data. Because raster data sets are rectangular, having vector polygons to draw over the raster data allows us to put roads, governmental boundaries, water features, and other base map information to provide location context for viewers. If this is our combination choice, it implies that we have no need of vector modeling capabilities (i.e., topological data sets). It assumes that all the processing and modeling of reality is taking place in the raster data model, and vector information is only symbolized and drawn over the raster pixels in the correct location. Using vector polygons to clip out segments of the raster, for example, would require topology in the vector data set and would mean we would need software to fit the fifth combination of data models.

c. Solely raster. This option, although logically a possible one, really does not exist because all the raster GISs and remote sensing software systems are capable of draping vector data over the grid cells.

d Solely topologic vector. In this data model combination all the data are stored as vectors, and topological relationships of connectivity and adjacency are known or can be computed. This is the most map-like of the combinations in its output but, like the solely raster option, really does not exist anymore because the capability to display raster images behind vector data is so widespread and useful.

e. Full vector and raster. If the organization will be processing its own raster data, whether for backdrop or modeling purposes, and also needs topologic vector information, this combination of data models is necessary. It can deal with both data models, and also the data to be moved between the data models. In some cases it is easier to analyze the real world in the raster data model and transform the results of the analysis to the vector data model for further analysis or display. More than 10 years ago a software vendor claimed that in the near future GISs would have these capabilities, and the system would decide which was the appropriate data model to use, transform the data into that model, and do the analysis or query. This has not yet come true, and it is still up to the users of systems with these dual capabilities to decide what is appropriate in what circumstances. So the raster/vector decision is not a debate with a winner. Nor is it appropriate to state, as has been common in the literature, that the raster data model is a better fit for natural resource applications and the vector data model for artificial information. Rather, certain features of reality are easier and better to handle in one or the other models. Generally, if we have any needs for aerial photography backgrounds or satellite image data in GIS, one will need some raster display capabilities within the vector GIS. If needs for raster information are greater than this (e.g., we need to acquire and process remotely sensed satellite data), we will need software and hardware specific to those tasks[16].

4.2.4 Layers and Objects

Early GISs all used the layer as the principle data structure, and this came from using layers to produce maps. Manual cartographers tend to work in layers even when the final map is going to be entirely in black and white. There may be one layer for each width of line work or a layer with text information and a layer with fill patterns. Drafted on clear acetate, each layer has common registration and the layers are held together with registration pins to combine the layers. The final layer combination is photographed for map reproduction. The principal reason for doing this, even in black and white mapping, is safety. Layering is also part of color map production. The digital GISlayer is an adaptation of the analogue notion of layering in multicolor map production. To produce multiple colors on paper, the information needs to be prepared in layers so that the printing plates for the different colors can be made. To produce the 7.5-minute topographic map, the USGS cartographers need to produce the following layers:

Black

Most artificial features except large roads. These include smallerroads, buildings, political boundaries, place and feature names, and benchmarks.

Red

Built-up urban areas (screened to pink) and larger roads.

Green

Continuous tree cover.

Brown

Topography

Blue

Hydrography (water features and symbols).

Purple

Revisions from aerial photographs not yet field checked. Drafting each of these feature sets in black ink on a different layer of acetate, using these acetate layers to produce printing plates and then applying different colors of ink to the plates produces a multicolor map containing a high density of information. This separation of the world into layers was also adopted by the early versions of CAD, where features would be placed in layers not polygon the basis of their color but also their symbolization on the final drawing. CAD layering is extensive; a single drawing may consist of dozens of layers, and individual features may be made up of elements from multiple layers. In a GIS layering system the entire feature is usually stored on a layer, and the result is many fewer layers to represent the real world. The notion of structuring the data in layers has a lot of advantages and a long tradition in GIS. In fact, it was the ease with which digital layers of suitability criteria could be combined that served as the motivation for the early development of raster GIS to replace the cumbersome methodology of analogue overlay popularized by Ian McHarg (McHarg 1969). On a topographic map, it is difficult to examine the drainage pattern revealed by the blue lines representing streams because the other information obscures what we want to see. If that layer is represented on a computer screen in a GIS, though, we can simply turn off the other layers, and the stream layer is more visible. We can also control the order in which layers display on the screen by dragging them to the bottom or top of the list of available layers. The drawing order on a paper map is determined during the printing process, and we cannot vary what layer overlays what other layers. Whether we are drawing geographic information on a computer screen or printing it on a map, we are drawing on a blank surface, and we will be adding information inlayers that draw in a specified order. At the level of display, all data are presented as layers in a GIS whether or not it is stored as layers in the database. But the question here is one of representation in the database, not display on the screen.

The idea of object-oriented databases developed in the 1990s out of object-oriented programming languages. This style of programming has mostly replaced the traditional style of sequential programming where programming code was long and involved, containing many calls to subroutines, program libraries, and so on. In object-oriented programming systems (OOPS) the software engineers write code in small snippets. Code can be associated with an object, a window, a dialog box, or some other type of object defined in the system. This object then possesses some "intelligence" and can communicate and direct activities of other objects if this is allowed within the programming system. The idea of extending object orientation to databases was a logical outgrowth of this programming style. We can conceptualize features in a database as objects, give them properties, and allow them to communicate with other objects in the database. This functionality elevates the feature from being something passive in the database that we can retrieve, manipulate, and replace in the database to being an object with properties of its own and a set of functions it can perform on itself or other objects. Object-oriented geographic databases are becoming more common in GIS, and each software company has its own terminology for them. Right now many organizations are taking their legacy-layered databases and converting them to object-oriented databases. To be consistent, we are going to refer to this type of database as a spatially enabled relational database system (SERD).

4.2.5 Representing Geographic Features

There are three classes of geographic vector features: points, line, and polygons. In advanced classes they learn that there are many more types of geographic features and that not all geographic information systems support all types. Certain types of geographic features are extremely useful in some applications and completely useless for others. A full topologic linear transportation network is essential for a GIS to support physical distribution but absolutely useless for one that is set up to facilitate the legal transfer of land. Both systems are GISs, but their model of the world is quite different. So it is important to understand the geographic nature of the features we need to represent in the GIS so that we can acquire the right kind of software to model these features. In the 1980s, as different software manufacturers started programming for new types of features, the U.S. federal government became concerned because different organizations within the government were purchasing different GISs, and they were having trouble exchanging data. So in 1980, the USGS was designated the lead agency for dealing with this problem with respect to spatial data, and they initiated process which resulted 12 years later in the publication of the standard for the transfer and exchange of spatial data, the Spatial Data Transfer Standard, or SDTS . The most recent version was published in 1998. Topologic Relationships, and there will probably be additional changes as users come up with different representations of real-world features that need to be exchanged between GISs. It was the need to exchange and transfer real-world features that forced the creators of the standard to think very clearly about exactly how to represent them from the top of the process (the data model) to the bottom (the physical file structure on the computer).Although the many vendors of GIS software often do not use the same terminology as SDTS, each of their supported data types—whatever they call them—must conform to one of the SDTS data types if they wish to sell systems to the federal government, a large market to be sure. The acceptance of these standards has not lived up to expectations, and creative people in the GIS world are regularly developing new ways to represent geographic features. We use SDTS as the foundation for our discussion principally because different GISs implement different features and have their own internal names for these features. What SDTS calls an arc is not what a particular GIS might call an arc. SDTS at least provides a common language for discussing types of geographic features. There are two profiles in the standard, one for raster or grid cell data and one for topologic vector data. The discussion that follows is for the topologic vector profile.

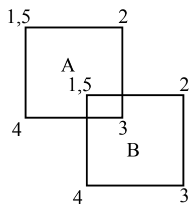
4.2.6 Topologic Relationships

Topology is one of the most confusing and difficult to understand concepts in the design of spatial data sets partly because people often confuse it with topography. They are completely different things. If a data set has topology, it means that relationships of adjacency (for two-dimensional features) or connectivity (connections between linear or one-dimensional features) are explicitly stored in the database. Geographers often make a distinction between absolute and relative location. The absolute location of a feature places it in a two- or three-dimensional coordinate space (e.g., latitude, longitude, and altitude).The relative locations that feature's location relative to other features (e.g., which other features are adjacent to it). It is possible to represent the absolute location of a feature in a digital file with an ordered listing of the number pairs (x/longitude and y/latitude)that describe its location in a coordinate space. This is sometimes referred to as spaghetti vector data because, like a plate of cooked spaghetti, the lines may overlay each other, but the system has no way of knowing that they do because each line is stored independently of the other lines; whether a line crosses another is not stored and therefore not known to the system. He/She, with our birds eye view, can see that the two lines intersect, but the software does not have that view of the data; it sees the line only as a series of points in a file. For topologic vector data, on the other hand, one must explicitly store the relationships of adjacency or connectivity or be able to construct them as needed. Without the knowledge of adjacency and connectivity, no one cannot execute spatial queries. It is possible to query features based on their attributes or properties (i.e., display all the sales territories that had sales increases over 20 percent) but not possible to locate all the sales territories adjacent to one particular territory to aggregate their information. Spatial querying is at the heart of what most users do with their GIS, so one must have a data model that allows it to occur. In the early days of GIS when computer central processing units (CPUs) were small and slow, topology had to be stored explicitly in tables in the database, and spatial queries were executed by making many hits, or reads, into those tables. As the processors became more powerful and faster, it was no longer necessary to store the topology in tables. Now we can execute the spatial queries through sophisticated indexing and searching procedures. The systems moved from complex data structures and simple queries to simple data structures and complex queries, and the user sees the same result. Polygon topology is easier to visualize than to describe. Figure 4.3 shows two polygon data sets, one with topology and one without.

The non topologic data set stores only the coordinates of the points where the line that outlines the polygon changes direction. The first pair of points in each polygon is duplicated at the end to close the polygon. In this data structure the polygon that appears to be where the two overlap is not in the database. The bottom part of Figure 4.3 shows the same polygons in a topologic structure, and it demonstrates the additional database complexity that topology requires. A single file or table can hold the non topologic information, but storing topology requires a linked set of several data tables. The key table in the set of tables shown here is the Chain_node table (often referred to as the Arc_Node table, but we are using the terminology of the SDTS); it is here that the topology exists. Each chain has a specified node it comes from and one it goes to. This produces an implied direction, and that makes it possible to identify the polygon to the right and left of the chain. The G-polygon table contains the identifiers of the chains that make it up. These chains are listed in clockwise order, and if a chain needs to be reversed to keep the clockwise order, a negative sign is added. So, for example, G-polygon B is made up of chains 3 and 4, but to get the chains to go in clockwise order, they need to be reversed. So it does not matter what the implied direction of the chain is, but it does have to have one. This is determined by specifying which is the from and which the to node. The node and chain tables do not contain topology, only the coordinates that describe the location of the nodes and chains. For topology to exist in a G or GT-polygon data layer two conditions need to be met:

There must be a universe or background polygon. Because every chain making up the G-polygons must have a right and left polygon defined, their needs to be a polygon to the right or left of polygons that do not touch any other polygon. The universe polygon provides for that. No feature in the database may extend beyond this universe polygon, so correctly choosing the extent of that polygon is an important early step .Fortunately, if our area of interest extends beyond the current universe polygon, one can create a new one and rebuild the

Non topologic (Spaghetti) Structure



Polygon A

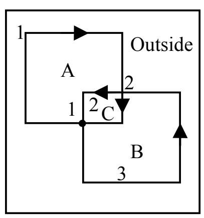
ID	X	Y
1	0.00	2.00
2	1.25	2.00
3	1.25	1.00
4	0.00	1.00
5	0.00	1.50

Polygon B

ID	X	Y
1	1.00	1.50
2	2.00	2.50
3	2.00	0.00
4	1.00	0.00
5	1.00	1.50

Key
XY
1 ID value and coordinate pair mark change of direction or closure of polygon. Notice that the polygon ends with a duplicate of the first point; this closes the polygon.
A Polygon identifier

Topologic Structure



Key
A Polygon Identifier
1 Line segments defining polygons
2 Nodes marking the beginning and ending of line segments
► Direction of line segment.
— Negative sign in chain list means direction of chain must be reversed to create

Node Table

ID	X	Y
1	1.00	1.00
2	1.25	1.50

Chain Table

ID	X	Y
	1.00	1.00
1	0.00	1.00
	0.00	2.00
	1.25	2.00
	1.25	1.50
2	1.00	1.50
	1.00	1.00
	1.10	1.10
3	1.00	0.00
	2.00	1.50
	1.25	1.50
	1.25	1.50
4	1.00	1.50
	1.00	1.00

Chain Node Table

ID	From_node	To_node	Left_poly	Right_poly
1	1	2	Outside	A
2	2	1	C	A
3	1	2	B	Outside
4	2	1	B	C

G Polygon Table

ID	Chain_list
A	1,2
B	-3,-4
C	4,-2

Figure 4.3 Topologic and Non topologic structure polygons

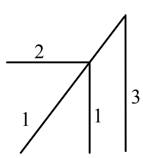
topology. Where ever a chain touches another chain. There must be a node, and that node will be the end point of one chain and the beginning point of the second (or third, fourth, etc.). We can notice that this is not the case in the spaghetti vector example at the top of Figure 4.4, and this is one of the most important requirements to construct topology. A GIS cannot create topology if there are any places in the database where a chain crosses another and there is no node. It is possible to have the system automatically find these locations and place nodes there, however. So what is the advantage of this complex structure? The principal advantage is that many geographic queries and processes can be conducted using only the Chain_node table. The actual coordinates defining the features are only needed for drawing the results on the screen or a map. As an example, say we wanted to remove G-polygon C from the layer shown in the bottom of Figure 4.4.The system would do this by searching the Chain_node table for all chains whose left or right G-polygon is C (chains 2 and 4), removing them from the appropriate tables, and then reconstructing the Chain_node table for the new topology, which would now have only one G-polygon made up of chains 1 and 3. The following common geographic functions on polygons all require knowledge of the topologic relationships between polygons[9]:

Clipping out a subset of polygons with another polygon. Say we acquired a data layer of all the land parcels in a large region, but we are only interested in the parcels within a particular area .We can create a polygon layer of this area of interest and use it like a cookie cutter to clip out the larger parcel layer, keeping only those polygons inside the area of interest.

Overlay of polygon layers. Perhaps we have one layer of wetlands and another of watershed boundaries and we want to know how many acres of wetland are in each watershed. By overlaying the wetland layer with the watershed layer, we can attach the watershed identification to the wetlands inside it. And if a wetland has some peculiar topography so that it drains in two directions into different watersheds, the overlay process will create two wetland/watershed polygons correctly reflecting the situation.

Dissolving lines between adjacent polygons that are similar on some characteristic. If we had a parcel layer of polygons and information on the principal land use for each parcel, we could use a dissolving function, which would remove the lines between adjacent parcels that had the same principal land use. The result would be a much-reduced set of polygons that showed principal land use. Vector GISs that do not have the ability to store or create polygon topology are useful only for map display and have relatively few analytical capabilities; therefore, vector topology is almost standard in GISs

Non topologic (Spaghetti) Structure



Chain 1

X	Y
0	0
1	1
1	0

Chain 2

X	Y
0	1
1	1

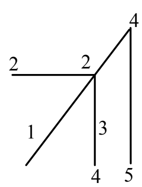
Chain 3

X	Y
1	1
2	2
2	0

Key

1 X Y
ID value for chain; x y pairs for beginning and end of chain and at each change of direction.

Topologic Structure



Notice: Nodes exist at beginning and ends of chains. Where chains touch, there must be a node. A chain is related to only two nodes, a node it originates from and a node it goes to.

Chain_Node Table

Line_Id	From_Node	To_Node
1	1	2
2	3	2
3	2	4
4	2	5

Node_Table

Node_Id	X	Y
1	0	0
2	1	1
3	0	1
4	1	0
5	2	0

Line_Table

Line_Id	N_XY Pairs	XY_Pairs
1	2	0,0,1,1
2	2	1,0,1,1
3	2	1,1,1,0
4	3	1,1,2,2,2,0

Figure 4.4 Topologic and non topologic structure lines

today. It is also possible and useful to create topology for linear features. Although polygon topology requires knowledge of adjacency (which polygons are to the right and left of each chain defining them), line topology requires knowledge of connectivity (which lines touch which other lines). Non topologic, or spaghetti vector, line segments do not carry this information, only the coordinate pairs or mathematical descriptions of the arcs. Topologic linear data layers must have anode where lines touch each other, and the Chain_node table relates the line segments or arcs to their beginning and ending nodes. The value of this kind of topology is principally in network analysis.

4.2.7 Types of Spatial Objects

The creators of SDTS chose to use a classification system based on the number of spatial dimensions the object requires in the database. There is a hierarchical component to this structure as well; we cannot describe a line without knowing the points that make up at least the beginning point of the line and some way to describe where the line goes from there. One-dimensional lines, which have points as their beginnings and ends, bound two-dimensional features, and everywhere they change direction. A GIS is an abstraction of reality, and one of the most important abstractions we have to make is do decide what kind of digital feature, or SDTS data type, we are going to associate with the real-world features we must deal with:

Zero dimensional features (see Table 4.2).These are point features with no length or breadth and therefore appear to take up no space. They may, however, represent a real-world feature that has dimensionality. For example, a point representing a well is an NE point according to this classification scheme, but a well is really a cylinder or possibly a more complex three-dimensional object that changes direction and width at different depths. The well is represented as a point feature principally because that it how it is represented on a map, but if the GIS needed to model in three dimensions, we would need to model it differently. Some of the point entity types defined in

SDTS is for cartographic convenience, such as label and area points and do not represent any real world feature.

One-dimensional features (see Table 4.3). Linear features have length but no breadth. They may represent features with breadth that can be stored as a property of the feature or object. The transfer standard explicitly includes linear features in networks and distinguishes between networks that are planar (fit entirely in two dimensions) and non planar (require more than two dimensions to work). In a planar network if two lines cross, a node must exist. But this is not a requirement of non planar networks. Most transportation systems are non planar networks with underpasses, overpasses, and tunnels that require three dimensions, but there are ways to model them as planar networks, which are simpler computationally.

Two-dimensional features. These are areal or polygon features that have breadth as well as height (see Table 4.4).These features are more complex, and their verbal and graphic descriptions demonstrate this. The profile for vector data in the SDTS is for topologic vector data, the most useful data model for vector data, but data without topology can also be described within this standard. What happens with real-world features that seem to have elements of more than one SDTS data type? Usually the response is to model it in different ways for different purposes. This is necessary even if the database is object-oriented, but theo bjects can be related to each other so that when a centerline is added or removed, the necessary changes to the curb and pavement area objects must be made as well. An object-oriented database would also allow for relationships between the street objects and the infrastructure on, below, and above the street. The standard also allows for composite features, which are groupings of related features of the same type. In transportation planning, for example, a route is defined as a set of linked network chains (LW).A well field for a water utility maybe a composite feature made up of several entity points (NE). Not all GISs can handle composite features, however, so this becomes an important design issue.

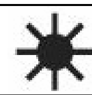
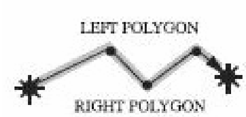
Point Features (Zero Dimensional)			
SDTS Description	SDTS Abbreviation	Figure	Explanation
A zero-dimensional object that specifies geometric location.	NP		The generic point Line segments conned two points. No attribute information is attached to this type of feature; they simply represent geometry.
A point used to identify the location of point features (or areal features collapsed to a point), such as towers, buoys, buildings, places, etc.	NE		Real-world features best represented by points at a specified scale. A point at one scale but a polygon at another scale may best represent a feature (eg. place).
A reference point used for displaying map and chart text (e.g, feature names) to assist in feature identification.	NL		Not a real-world feature; to place labels on features or text on maps.
A representative point within an area usually carrying attribute information about that area.	NA		Usually the centroid (approximate center) of a polygon. Often is the default label point
Topological junction of two or more links or chains.	NO(Planar graph)		An intersection joining two segments (links) of street.

Table 4.2 Spatial Data Transfer Standard Point Features

Line Features (One Dimensional)			
SDTS Description	SDTS Abbreviation	Figure	Explanation
A direct line between two points (NP).			The simplest linear feature, a straight line connecting two points.
Connected non branching sequence of line segments specified as the ordered sequence of points between those line segments. Note: a string may intersect itself or another string.	LS		This is the so-called spaghetti vector data structure. Line segments have no official beginning or end, and if they intersect, a point (node) is not defined. Mostly used for mapping only.
A locus of points that forms a curve that is defined by a mathematical expression.	AC (circular arc), AE (elliptical arc), AU(uniform b-spline), AB (piecewise bezier)		To represent complex linear features, often artificial Some GISs cannot model this type of spatial feature and

			will convert them to strings. A circle as an AC-type feature is described by a point and a radius; a circle as a string is a set of a large number (often 256) strings.
Topological connection between two nodes. A link may be directed by ordering its nodes.	LQ		For modeling networks. A network chain is a more complex representation of a link in a network.
A directed non branching sequence of non intersecting line segments and (or) arcs bounded by nodes, not necessarily distinct, at each end.			Most common feature in topologic vector data layers.
A chain that explicitly references left and right polygons and start and end nodes. It is a component of a two-dimensional manifold.	LE		A part of arc-node topology.
A chain that explicitly references left and right polygons and not start and end nodes. It is a component of a two-dimensional manifold.	LL		A part of arc-node topology.
A chain that explicitly references start and end nodes and not left and right polygons. It is a component of a network.	LW (planar network), LY (non-planar network)		
A sequence of non intersecting chains or strings and (or) arcs, with closure. A ring represents a closed boundary but not the interior areas inside the closed boundary.			

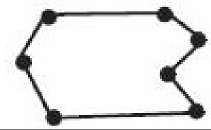
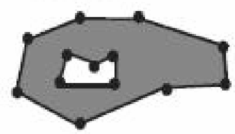
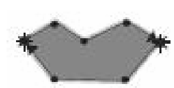
A ring created from strings and (or) arcs.	RS (ring of strings), RA (ring of arcs, RM (ring of mixed composition)		
A ring created from complete and (or) area chains.	RU		

Table 4.3 Spatial Data Transfer Linear Features

Polygon Features (Two Dimensional)			
SDTS Description	SDTS Abbreviation	Figure	Explanation
An area not including its boundary.			
An area consisting of an interior area, one outer G-ring and zero or more non intersecting non-nested inner G-rings. No ring, inner or outer, must be collinear with or intersect any other ring of the same G-polygon.	PG		Allows for do nut polygons with holes inside; the rings may not touch each other, and there may not be an area inside the hole, an island (hole) in a lake (GT-polygon) with a pond on the island
An area that is an atomic two-dimensional component of one and only one two-dimensional manifold. The boundary of a GT-polygon may be defined by GT-rings created from its bounding chains. A GT-polygon may also be associated with its chains (either the bounding set or the complete set) by direct reference to these chains, the complete set of chains associated with a GT-polygon may also be found by examining the polygon references on the chains.	PR (made of rings) PC (made of chains)		The polygon defined by arc-node topology. Does not allow for interior polygons (see PG-polygon).

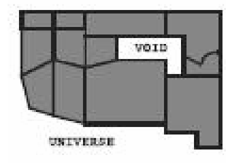
Defines the part of the universe that is outside the perimeter of the area covered by other GT-polygons (covered area) and completes the two-dimensional manifold This polygon completes the adjacency relationships of the perimeter links. One or more inner rings and no outer ring represent the boundary of the universe polygon. Attribution of the universe polygon may not exist or may be substantially different from the attribution of the covered area	PU (made of rings), PW (made of chains)		This is sometimes known as the 'outside' polygon and must exist for adjacency relationships to be defined for the rest of the polygons that make up the covered area It usually have no attribute information but may. Some GISs recognize it but do not display it or list it in tables.
Defines part of the two-dimensional manifold that is bounded by other GT-polygons but otherwise has the same characteristics as the universe polygon. The geometry and topology of a void polygon are those of a GT-polygon. Attribution of a void polygon may not exist or may be substantially different from attribution of the covered area.	PV (made of rings), PX(made of chains)		Allows for intrusions of the universe polygon inside the area covered by GT-polygons.

Table 4.4 Spatial Data Transfer Standard Polygon Features

4.2.8 Accuracy, Precision, and Completeness

4.2.8.1 Accuracy

Ensuring that the coordinates attached to the features in our GIS database are close to where they are in the real world is always a critical design concern. There are a number of factors that are important with respect to accuracy. The first is the positional accuracy of the data in the system. Positional accuracy is defined as the overall reliability of how close a feature is represented relative to its actual position on the surface of the earth. Positional accuracy is usually documented with a statement that a certain percentage (say 95 percent) of the visible features are within a certain distance (say, 2.5 ft) of the real-world location as determined by a more accurate method. Depending on the application of the GIS, the actual physical location of a feature is important to know when we need to make decisions relating a feature on one layer and a feature on another layer in the system. When designing, implementing, and using data in a system, we need to take great care to ensure that measurements and decisions made with the system are based on the layer with the lowest level of positional accuracy. A common error in the use of a system is to combine two layers with differing levels of positional accuracy and make measurements on the combined layer as though it were as positionally accurate as the most accurate, not the least accurate, source layer. Often the result can result in improper decisions. More important to most GIS applications is the relative accuracy of the features in the system. The relative accuracy is defined as the overall reliability of how close a feature is to another feature on the same layer. It is possible to have a layer with a high level of relative accuracy but a low level of positional accuracy. For examples, features can be located using either conventional surveying techniques or GPS and have a very high level of relative accuracy (i.e., the measurements of distances between features could be very close to what they are in the real world).At the same time, unless the data are properly tied into a standardized coordinate system, the level of positional accuracy could be quite low. Again, we need to take care that the data development strategy includes processes to correct the positional accuracy of features and documents the methods used. Finally, the last type of accuracy important in spatial data design is absolute accuracy, which is a combination of positional and relative accuracy. It is a measurement that defines not only the reliability of a feature's position with respect to other features on the same layer, but also with respect to where it is on the face of the earth. There are fundamentally two different levels of accuracy: map- and surveying level accuracy. In most cases in GIS design we are dealing with map-level accuracy, which is always lower than surveying-level accuracy. Getting a GIS database to surveying-level accuracy means that all the features in the database, and there could be hundreds of thousands or more, must be located using precise surveying techniques, and the cost would be prohibitive. This is why accuracy is of such concern; to attain high levels of accuracy is exceedingly expensive, but inaccurate data may reduce the utility of the database. As with many things there is a tradeoff between accuracy and cost. The relationship between accuracy and cost is nonlinear, so if we can calculate the cost of developing data with an accuracy of ±1 m and then decide we would like to improve that accuracy to 1/2 m, the total cost will be more than twice the original. Nobody knows the exact parameters of this relationship; we just know it is not a linear progression and that it goes up at least with the square of the required additional accuracy. For example, if original design standards accepted a positional accuracy of ±2.0 ft and managers decided they wanted to double that accuracy to ±1 ft, we can be reasonably certain that the cost of data development will increase at least 4 times, not twice, because the data are in two dimensions. In the United States, the USGS published map accuracy standards in 1941, revised them in 1947 and they have not changed since. The standards for horizontal accuracy are a function of the map scale (see Table 4.6).The 1947 map accuracy standards were adequate when most geographic information existed in paper form but not for digital spatial data. In the late 1990s the U.S. federal government developed a new National Standard for Spatial Data Accuracy. For horizontal data the accuracy is based on root mean square error(RMSE).To determine the RMSE for a set of coordinate values we calculate the square of the difference between the coordinate values in the data set and the coordinate values as determined by an independent source of higher accuracy. We sum these squared differences across the data set and take the square root of that sum, which is the RMSE. The accuracy is reported in ground distance (in the dataset's distance units) at the 95 percent confidence level. This means that 95 percent of the locations tested to determine accuracy are within that distance of their locations as determined by the higher-accuracy system. In practice, the higher accuracy system is either GPS or surveying. There is also a reporting procedure for elevation data. Because so much of the source material for GIS databases comes from already published paper maps, often the accuracy we can attain in our GIS is determined by the scale of those maps and whether or not they were prepared using the map accuracy standards for that scale. There should be a statement to that effect on the map; if it is not there, we can make no assumptions about absolute location accuracy of that data source. But as is clear from the table, if our data source is existing maps, we need very large-scale mapping to attain high accuracy levels. If we extend that table to a ridiculous extent and state that we want centimeter accuracy, we would have to have a map scale of 1:1 (i.e., the map would be the real world). Generally, any desired accuracy below 1 m will require direct surveying of our features [15].

Table Map Accuracy Standards				
Horizontal Accuracy		No more Than 10% of the points tested shall be in error of more than		
Representative Fraction	Feet/Inch	Inches on the Map	Feet on Earth	Meters on Earth
1:240,000	20,000	0.020	400.00	121.92
1:120,000	10,000	0.020	200.00	60.96
1:24,000	2,000	0.020	40.00	12.19
1:12,000	1,000	0.033	33.33	10.16
1:2,400	200	0.033	6.67	2.03
1:1,200	100	0.033	3.33	1.02
1:600	50	0.033	1.67	0.51
Vertical Accuracy. Regardless of scale, no more than 10% of the elevations tested shall be in error of more than one half the contour interval.				

Source: USCS National Mapping Program - National Map Accuracy Standards. rocykweb.cr.usgs.gov/nmpstds/nmas647.html

Table 4.6 Table Map Accuracy Standards

4.2.8.2 Precision

After accuracy we need to take into account the precision of the features in our data development strategy. Precision is the ability of a measurement to be reproduced and the number of significant digits to the right of a decimal place a feature can be reliably measured to. This issue of precision is frequently misused in GIS. It is very easy for the user of a system to change configuration settings and have the computer provide location and distance measurements that can vary from no decimal places to two decimal places to eight places when measuring the same feature. This does not mean that the feature's precision has change, only that the user has altered the way the computer represents the precision. What determines the precision of a feature are the methods used to collect and input the feature into the system? For example, a surveyor or inspector could measure the foundation of a house after it has been built to determine its size. Measurements could be taken to the one-hundredth of a foot and recorded in a field book. When this person returns to the office and draws the foundation into the system, it can be drawn using coordinate geometry (COGO). If properly measured and recorded in the field, the feature would close and create a polygon representing the foundation when the measurements were checked. This would indicate a high level of relative accuracy. When in the field, the worker could also record three (or more) GPS points to determine the foundation's location in a standard coordinate system. These points could then be used to locate the foundation in the GIS. The precision of this foundation and any of its sides would be one-hundredth of a foot, but the precision of any measurements taken from this foundation to any other feature measured and located in the same manner would be based on the precision of the GPS locations take to locate the corners. If these locations were accurate to 1 m, the precision of measurements between two foundations would-be ±2 meters. But if the corners were located to within one-hundredth of an inch, the measurements between foundations would also be that precise.

4.2.8.3 Completeness

The third related issue is the need for completeness within a data layer. We should design standards for absolute accuracy and precision but also for completeness. Completeness is defined as locating all the features represented in the data layer or some stated percentage of those features. Location accuracy standards are stated as the percentage of visible features that are within a certain distance of their reallocation, and completeness is a statement of the percentage of the actual features that are captured in the database. It is important to state exactly what is the basis for measuring completeness? We would certainly expect to capture all land parcels delimited on the parcel maps for a community as of a certain date, but can we be certain that those maps contain depictions of all the land parcels in the community? There is a probably apocryphal tale told in GIS of a community that paid for its GIS by discovering many land parcels that had not been correctly entered into the taxation database or were missing entirely. The back taxes collected on those parcels were said to have covered a significant percentage of the GIS development cost. It would be nice to be able to document this story, but it probably is not true. Specifying the desired level of completeness is particularly important when data are to be captured from aerial photography. It is common for features to be obscured by shadow and branch overhang, and although we might wish to collect every catch basin and utility pole, that may not be possible at the scale we have selected. Measuring the delivered positional accuracy and completeness of delivered data are quality control tasks that we need to design in as we are specifying our spatial data. Determining that a set of digital data actually meets a certain map accuracy standard or that we actually have captured 95 percent of the features we expected to capture is a task that is done after completing the data by field work with higher accuracy measurement systems and careful field checking of the presence or absence of visible features. In the design phase we specify our desired levels of accuracy, precision, and completeness, and in the implementation phase we document whether or not we have met these standards.

4.2.9 Accuracy Concerns—Global Positioning Systems

GPS is another very common method for obtaining spatial data to include in a GIS. So it is important in the design phase to specify the standards we expect for data derived from GPS. GPS technology was originally developed by the U.S.Department of Defense and allows the user to determine exactly where they are on the face of the earth by triangulating data from a series of satellites that orbit the earth. Although the military applications drove the construction of the system, civilian use far outstrips military use of GPS around the world today. The user has a GPS receiver that receives the signals from a constellation of these satellites, processes them, and then calculates the latitude and longitude of the receiver at the time the data were collected. The accuracy of the data that can be acquired from this process depends on the type of receiver, the methods that are used to collect the data, the amount of time that data are collected, and what, if any, post processing is done on the data with more advanced software. If we are trying to get meter-level accuracy, there are eight major areas in the system that can generate error; getting more accurate than that, the list gets even longer. Error in GPS is a complex topic, and we only introduce it here in general terms. The primary components of the system are the satellites, ground stations, receivers, and software used to correct known errors in the data. The satellites make up a worldwide network of 24 satellites that orbit the earth every 12 hours. With this number of satellites moving at that speed there are almost always enough satellites in direct line of sight of a receiver that determining location is possible. The signal broadcast by the satellites contains information about exactly when the signal was sent, and information that tells the receiver which satellite it is listening to. By comparing the time it receives the signal to its internal clock, the receiver can calculate its distance to the satellite. If the earth were not nearly spherical, it would require this distance information from a minimum of three satellites to determine location in latitude and longitude. Let's say we measure a distance of 10,000 km from the first satellite (the satellites actually orbit at about 11,000 nautical miles, so these numbers are just for illustration).This first distance defines a sphere with a radius of 10,000,000 km and the receiver must be on the surface of the sphere. Now, the distance from the second satellite, let's say 12,000 km, locates the receiver on the surface of a second sphere with a radius of 12,000 km. The intersection of these two spheres defines a series of locations where the sphere must be. Calculate the distance to the third satellite and, if the clocks were perfectly in synchronization, the number of possible locations is narrowed to two points where the receiver may be. One of those points is on the face of the earth (assuming that is where the receiver is), and the other is usually some where in space. To resolve this conflict and because it is not possible to have clocks telling exactly the same time on both the satellites and the receivers, a fourth satellite is required to eliminate the incorrect location and correct for the clock differential. Because it takes between .05 and .07 seconds for the signal to travel the 11,000 nautical miles, accurate clocks are critical, and the clocks on the satellites are large and accurate atomic clocks. Because we cannot (financially or physically) place anatomic clock on a receiver, a mathematical correction is made in the receiver. If the clocks were perfectly synchronized, the four spheres would intersect at a single point, the location of the receiver. Because of clock error they intersect in a three dimensional surface, and the receiver applies a correction to its clock until the points all intersect at a single location, which it reports to the user. All this happens almost instantly at the push of a button after the receiver has locked on to at least four satellites. This makes GPS receivers very accurate clocks as well. The United States military, which controls the GPS satellite system, can degrade the signal so that only military receivers will get accurate locations. This is known as selective availability, and when it is on, the signal is degraded. Since May2000 it is now always turned off, so there is no difference between positional accuracy from military and civilian receivers. The chief of the U.S. Geodetic Survey has related this to a football field; if selective availability is off, we can only know that we are on or off the field, but if it is on, we can tell at which yard marker we are located. Another important factor to making GPS work is knowing the exact location of the satellites at all times. For low-end or hand-held GPS receivers such as used by hikers and boaters this is not important, but for survey-grade receivers, those typically used to collect GIS data, this is important. The U.S. Department of Defense monitors these satellites very exactly using radar and records the position, altitude, and speed of the satellites. Along with the pseudo-code mentioned above, the satellite also broadcasts this information, which is known as the ephemeris errors. The errors themselves are usually pretty small, but if we want exact measurements, they need to be corrected for. All of this discussion is based on the premise that we can easily calculate the distance from a receiver to a satellite. The problem is that the signal itself has to travel from the satellite, through the atmosphere, and a number of factors can affect the speed of the signal, including the temperature of the atmosphere and how much moisture is in it. Now, if all of these factors could be known, they could easily be accounted for with mathematics, but the reality is that these factors are difficult if not impossible to track. The way these are accounted for is by tracking two different signals from the same satellite and calculating their relative speed.

For this reason, the receivers that are used to perform accurate measurements need to be dual frequency where the information is carried twice on different frequencies. A dual-frequency receiver can measure two signals from the same satellite, know the time that each signal left the satellite, and calculate correction factors to apply to the results to reduce error in measurements. These types of receivers cost more but provide much improved accuracy in measurements. Ground stations are permanent locations, also known as the control segment, that are in constant communication with the satellites. These control stations transmit signals to the satellites to check their exact position in space and assure that they are operating properly. The signals sent to them are then used by the satellites to adjust the position and the signals that they send out to the GPS receivers on the surface of the earth. Five ground stations exist and are located in Hawaii, Ascension Island, Diego Garcia, Kwajalein, and Colorado Springs [13].

4.2.9.1 Differential Processing

Another type of ground station, much more widely distributed, is known as a base station. These are locations with highly accurate and expensive GPS receivers that constantly monitor their location as determined by the satellite. Because of all the possible errors in the system, the location changes slightly depending on atmospheric conditions, satellite errors, and other factors. But the location of the base station is know by surveying methods to a very high degree of accuracy, so it is possible to mathematically model the deviations from the known location compared to the location estimated at the same time our receiver was taking measurements. Of course, we want to select a base station near to where we are taking our measurements, and there is a widespread and growing network of base stations around the world. Both receivers, the base station and our receiver, are communicating with the satellite at the same time. And this processing is known as differential GPS (DGPS).This processing can occur either in real time by having radio links between our receiver and the base station or on a computer after we have taken our measurements, known as post processing DGPS. Many federal and state agencies have established permanent base stations that transmit their locational data constantly and store the files on servers tied to the World Wide Web for post processing. Using DGPS and post processing we can anticipate measurement results within a couple of meters for moving applications where few points are taken at each location. The results can be within sub centimeters if we use dual-frequency real-time DGPS and stay at one location for an extended time.GPS has not yet evolved to the point where we can stand on top of a feature and push a button and in one or two seconds receive locational information accurate enough to locate that feature within a centimeter of where it is, but the systems get closer to that goal all the time.

4.2.9.2 Real-Time Kinematic GPS

Real-time Kinematic (RTK) GPS is an advanced type of differential GPS .With these units a combination of a stationary base station located at a known coordinate location and a roving RTK receiver, usually in a backpack, are used at the location where we want to calculate new coordinates. The difference between a DGPS and an RTK unit is that the RTK roving unit communicates back to the stationary unit using radio waves, and the unit can calculate its location much faster and more precisely. These type of receivers are typically much more expensive, but because of their efficiency, they are often used to collect data for a GIS where a large number of points need to be collected; the positional accuracy is very high (less than a meter).A typical application of this type of RTK GPS is utility infrastructure mapping such as water valves or utility manholes and catch basins. With hand-held units and no differential processing we can achieve accuracies between 3 to 5 m, and 10 to15 ft, which approaches map accuracy values at the 1-inch-to-200-ft scale. Differential processing and hand-held units that allow we to process the data this way can now achieve accuracies in the 2- to 5-meter (6-to15-ft) range. Submeter accuracy using RTK GPS is very expensive and requires sophisticated equipment, real-time information from multiple base stations, dual frequency radio, and complex mathematical processing, but submeter accuracy is getting more affordable over time.

4.2.9.3 Accuracy across Layers

The largest concern relating to accuracy, however, is not determining how accurate one layer of a set of features may be but how to assess accuracy when there are multiple layers in our GIS, each with differing levels of accuracy. This has been stumping the profession for years. What is commonly done is to take the most accurate data and use that data source to realign the inaccurate information. For example, often in GISs that uses aerial photography as a backdrop or data source that is the most accurate information in the database. Features from other sources can be shifted to fit visible features on the digital photograph. This process of conflating one data set to fit another improves the visual fit of data Designing Spatial Data Choosing a Coordinate System and Map Projection layers, but we still cannot state that the layer is necessarily more accurate than it was before. It probably is, but all we can say about it is that we performed that function on the data to improve the fit. In the end, the level of accuracy we require is determined by the uses to which we plan to put the GIS. For many uses (e.g., regional planning over multiple states), available data at a scale of 1:24,000 (accuracy ±40 ft) is adequate. For city planning in dense areas we may need data accurate to a map scale of 1:1,200(±3.3 ft). The issue of accuracy often comes down to differences of opinion between engineers, who have been trained to strive for the highest accuracy possible, and others who would be satisfied with the lowest level of accuracy that keeps them from making bad locational decisions. Rather they are intended to be a model of the real world. The real world is a messy place, and sometimes a high level of locational accuracy is not only unnecessary but also unattainable.

4.2.10 Choosing a Coordinate System and Map Projection

A human being, as an object, has the properties of height and weight, and we can measure those properties in different measurement systems such as inches, meters, pounds, and kilograms. Similarly, geographic objects have the property of location, and we can measure that property in different coordinate systems and map projections. The correct choices of coordinate system and projection are important because this is how we will present our data to the users and how we will take measurements from the data. It may seem that the question. Although there are many terms specific to map projections, Table 4.7 is a simple glossary of the terms use in this sect

Developable surface	A three-dimensional feature that can be broken apart and which will lie flat; a sphere is not one, whereas a cylinder is.
Conformal	Property of a map projection such that relative local angles around every point on the map are shown correctly.
Equal-area	An area on the map covers the same area on the earth no matter at all locations on the map
Geoid	The shape that the earth would take if it were all measured at mean sea level. Newer projections use a mathematical representation of the geoid rather than any particular spheroid.
Ellipsoid or spheroid	The earth is nearly spherical but not quite; it is elliptical and slightly flattened at the poles. There are over 25 spheroids with different semi major and semi minor axes used in different parts of the world and at different times. Ground-measured spheroids are still in wide use for specific areas although there are several satellite-measured spheroids designed for use around the entire globe.
Semi major axis	Equatorial radius.
Semi minor axis	Polar radius; always smaller because of flattening.
Flattening	Difference between the two axes expressed as a percentage of the semi major axis.
Datum	A set of parameters that define a coordinate system. A spheroid is part of a datum, but many datums could be based on single spheroid. Local datums, such as the North American Datum of 1927 (NAD27), have a central point, and all locations are measured relative to that point. Local datums are aligned to the spheroid to fit locations in a particular area. A geocentric datum such as the World Geodetic System of 1984 (VVCS84) uses the center of the earth's mass as the reference point.
NAD27	A local datum based on the Clarke 1866 ellipsoid; defined for used in North America with a reference point at Meade's Ranch in Kansas. It is being phased out and replaced with the North American Datum of 1983 (NAD83).
NAD83	A geocentric datum, both satellite and ground-based, using an ellipsoid developed from the Geodetic Reference System of 1980 (GRS80). Gradually replacing NAD27 as the standard datum in North America.
UTM	Universal Transverse Mercator. A set of transverse mercator projections adopted by the U.S. Army in 1947 for global large-scale military maps. Between N 84 degrees and S 80 degrees the globe is divided into 6-degree segments numbered 1 to 60 and 8-degree latitudinal segments using letters. The measurement units are meters; the central meridian is the midpoint meridian between the bounding meridians and is given a false easting of 500,000 m. N/S coordinates are measured from the equator with any value of 0 for the northern hemisphere and 10,000,000 for the southern. Numbers increase going east and north, avoiding negative numbers.
Standard Parallel	In conic projections. The one or two parallels (lines of latitude) along which scale is true; at latitudes off the standard parallels map scale varies.
Central Median	The line of longitude in the middle of the projection; x locations are measured from this line.
Latitude of origin	Line of latitude representing the 0 or base line for measuring distance in the y dimension.
False easting	Arbitrary number given to the central meridian to avoid negative

	numbers
False northing	Arbitrary number given lo the latitude of origin to avoid negative numbers.

Table 4.7 Some basic projection technology

4.2.11 Characteristics of Map Projections

Some practitioners believe that all spatial data should be stored in decimal latitude and longitude because it facilitates transfer and can be converted to any other projection using standard transformation tools. Most, however, store their data in particular projection and in measurement units such as feet or meters. Map projections exist because the Earth is essentially a sphere, a three-dimensional object, and a paper map or computer screen is a two-dimensional object. Some three dimension objects, such as a cube, will unfold and lie perfectly flat. Three-dimensional objects (cones and cylinders) that can do this are known as developable surfaces. A sphere or spheroid is not one of these surfaces, so map projections are the mathematical transformation of the sphere to a flat surface. Distortions or compromises of various sorts occur because the sphere will not lie perfectly flat. Because this cannot be done without distortion, the choice of map projection is fundamentally a choice of which distortions our applications can survive. For large-scale work (small areas) the choice of map projection is not as important. Ata scale of 1:24,000 on a paper map more distortion is introduced by expansion and contraction of the paper due to humidity than by choice of map projection [16]. For small-scale work (large areas) the choice of projection becomes more important. The characteristics we need to consider when choosing a projection are as follows:

Area

Some projections are equal-area, which means that an area of a certain size on the map covers the same amount of area on the earth's surface no matter where on the map it is. This forces some distortion in the other map properties.

Shape

Some projections are conformal, which means that relative local angles around all points on the map are correct. On conformal maps lines of latitude and longitude cross each other at right angles, but areas are enlarged or reduced except along certain lines in some projections. All the large-scale maps used by the USGS and most other agencies of the U.S. federal government involved in mapping or GIS are conformal projections, and all state plane coordinate projections in the United States are conformal. No map or set of geographic data can be both equal area and conformal.

Scale

Map projections do not show scale correctly throughout the map, which is why we will get different measurements between the same two points depending on the projection used. There are usually one or two lines along which scale is correct, and if we choose those correctly, we can minimize the scale variation across the data or the map. There is a special class of projections, equidistant, where scale is correct to all points on the map from one or two specified points on the map or along all lines of longitude.

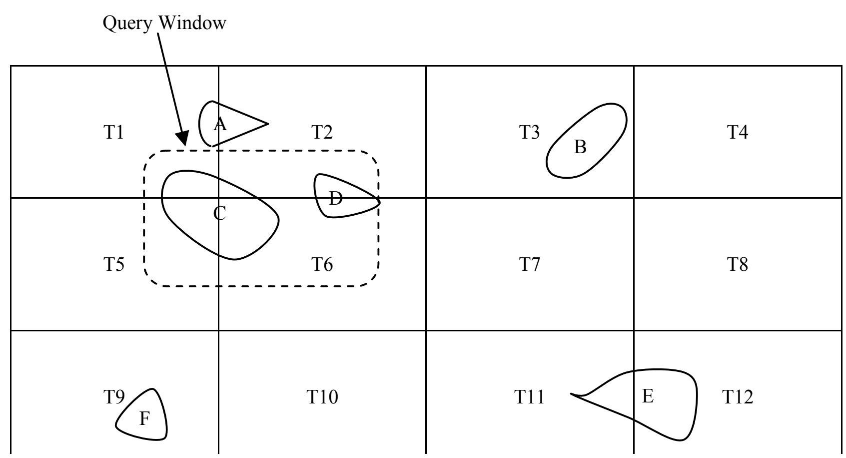
Direction

On a group of map projections known as azimuthally the directions of all points on the map are correct with respect to one location at the center of the map. On other map projections direction is distorted.

4.2.12 Spatial Indexing

Another design concern for spatial data is indexing of the data. Indexing allows rapid retrieval of features from large data sets, and without it a GIS may operate so slowly that users will not use the system. GISs automatically index the spatial information, so it is really not necessary to understand it unless our database is so large that the default indexing scheme is not adequate or super-fast access is critical to our applications. In those cases we need to set up the spatial indexes with considerable care. Efficient spatial indexing has been an area of research and development in computer science for almost 30 years, and new techniques for indexing spatial data are regularly proposed. Until they get worked into the existing GISs and RBDMSs, though, they remain essentially research tools. The spatial indexing process discussed here is typical of those implemented currently in commercially available GISs.

As with an attribute-indexed query, a spatially indexed query for features reads from a structured index table or tables and then, through relational joins between the index tables and the feature table, pulls out only the features that have been requested. Almost every interaction with the data in the map view is a spatial query. Panning the map from one location to another is the equivalent of SELECT ALL FEATURES FROM THESE_LAYERS AT THIS_SCALEWITHIN A RECTANGLE OF THIS_X_DIMENSION AND THIS_Y_DIMENSION CENTERED ON THIS X_Y_LOCATION. Every time we pan the map view, that query is sent again to the database with a different value for the center of the view. Zooming in or out is a spatial query as is selecting features inside a circle drawn on the map view .All these queries require reads against the feature database, which is not ordered geographically, and by reading against the indexed tables first, we are able to reach directly into the feature table to retrieve the requested geographic features. The simplest kind of spatial index is created with a set of tiles that exclusively and exhaustively cover the area of the geographic features. Exclusively means that all parts of the region of interest are covered by only one tile and exhaustively means that no area of the region of interest is uncovered by a tile. This is called a tessellation. In Figure 4.5 the tiles are numbered T1 through T12.The spatial query being executed here is a so-called window query that seeks to find and display all the features inside the window we have drawn on the screen. The search determines which tile(s) touch the query window (in this case,T1,T2,T5, and T6). It then goes into the index table to find out which geographic121features are inside these tiles and retrieves only those features from the data set (i.e., features A, B and C).All these features are retrieved from the T1 record in the index table; the other records,T2,T5, and T6, would have to be read, but no new features would be added to the already selected set. Because tile number organizes the index table, that first search of the index table is actually an indexed search on the tile number. With no index the entire feature table would have to be read and each feature evaluated to see whether it is inside the query window or not. There are only five features in this simple example, but what if there were several hundred thousand and we were interested only in a small area? One feature retrieved, A, is actually not inside the query window but would be retrieved anyway because it, or at least part of it, is inside tile 1. Slightly more sophisticated indexing systems will execute a second query at this time of selecting from this set (A, B, C) only those features that intersect the query window. This will result in A being thrown away and not displayed and B and C being the only features returned. One GIS system, ArcInfo, uses a three-step indexed search where each feature is a record in a feature envelope table that contains the x/x pairs of the corner of a rectangle that completely encloses the feature. The first stage of the search pulls the feature envelopes that touch the selected tiles; the second stage eliminates those feature envelopes on one of the selected tiles but outside the query window; the third stage compares the actual features that remain to identify those that are inside the query window. Because the simplest indexing scheme would have retrieved a feature that is not in the requested set, it is clear that choice of the size of the tile is important. Inmost cases the default tile size will result in adequate results. If we make the tiles too large, as in this case, we retrieve excess features, but if we make them too small, we can create a much larger index table; searching that table slows retrieval down. More complex indexing systems support three levels of tile sizes where there are three index tables and the one searched is determined by how far out we are zoomed in the display. So if we draw a large query window with the entire area of interest in view, the largest tile is searched, but if we draw a small window, the search accesses the index table based on the smallest of the tiles. If we have literally hundreds of thousands of features to search and typically are zoomed into a small set of x/x ranges (zoom levels), we could set up the tiling system to approximate the extent of the areas displayed at these zoom levels as our three levels of tiling. Setting up the spatial indexes and correctly locating the various files that need access for queries is referred to as tuning the database. Tuning is a craft that database professionals excel at, and the addition of a spatial index to already existing indexing schemes is something they will quickly understand and implement. Locating the files that are read frequently in the center of the storage medium (disk) and those that are read less frequently on the periphery will also speed access, as will partitioning or splitting very large feature tables onto several different disk storage devices, so long as each with its own disk controlling hardware and software. Indexing often the size of the database is so large that access is not determined as much by the speed of the CPU or the amount of RAM we have as by the physical/mechanical limitation of disk access. The setup and tuning of very large databases (VLDBs) is a specialized craft and beyond the scope of this book, so seek professional help when the number of features we need to query regularly and quickly extends into the hundreds of thousands. Multi tile indexing, careful file placement and partitioning of large tables onto several disks are necessary for rapid retrieval in that situation[12].



INDEX TABLE

T1	A,C,D
T2	A,C,D
T3	B
T5	C
T6	C,D
T9	F
T11	E
T12	E

FEATURED TABLE

A	X1,Y1,X2,Y2,….X1,Y1
B	X1,Y1,X2,Y2,….X1,Y1
C	X1,Y1,X2,Y2,….X1,Y1
D	X1,Y1,X2,Y2,….X1,Y1
E	X1,Y1,X2,Y2,….X1,Y1
F	X1,Y1,X2,Y2,….X1,Y1

Figure 4.5 Spatial index

4.3 Designing the GIS Database Schema

Database designers use the word schema to refer to the diagram and documents that lay out the structure of the database and the relationships that exist between elements of the database. A schema is like a blueprint for a database that tells a knowledgeable builder exactly how to construct it. Naturally, designers spend a lot of time thinking about the schema. This work comes before worrying too much about the exact content of tables and even before design concerns for the spatial data. Rushing into building a database without laying out our schema is like trying to build a house without a set of plans; it might stand up for a while, but it will not be as useful as it could be. The tools that assist in the construction of these schemas are called computer assisted software engineering (CASE) tools. These same tools are used to design the structure of complex computer programs as well as databases, and most programmers know how to use them. Many in the GIS world do not, but it is usually possible to design our database with paper and pencil, and some database designers still work this way. The ability to erase entire tables, delete relationships, add relationships, and soon, is sometimes easier with pencil and paper or on a whiteboard than mastering anew set of tools. One of the problems with GIS is that it appears to force us to develop areas of specialization and skill that we didn't have before. Sometimes it just takes too long to learn the new tools, so feel free to use simpler ones we have mastered instead of new tools that do basically the same thing[11].

Elements of a Schema

A schema at its simplest consists of an arrangement of tables and the relationships between them. Because organizations differ so widely in the kind of work they do and the types of data they need to do this work, it is impossible to provide a cookbook schema for every application. The task of customizing an existing schema in a complex organization that is planning to construct a large geographic database that integrates most of its existing data and incorporating new data tables into that schema ought to be done by highly skilled database administrators and designers. From the perspective of the users of the geographic data, who may be a minority of the total set of users, it is important that our geographic feature tables be correctly linked to the other tables we need to do our jobs.

Data Dictionary

Data Dictionary is a sub system of DBMS to keep track of the definition of all the data items in the database. It is valuable source of information for end users and developers to find out what data is available, what the data means and where and how to get it. The data stored in the DD are also called metadata or data about data. It also keeps track of a relationship that exists between various data structures.

Tables and Relationships

The second critical part of a database schema, and actually the one we create first, is a diagram that shows the relationships among the various tables in the database, as shown in Figure 3.1. Relationships have a property called cardinality that describes the type of relationship.

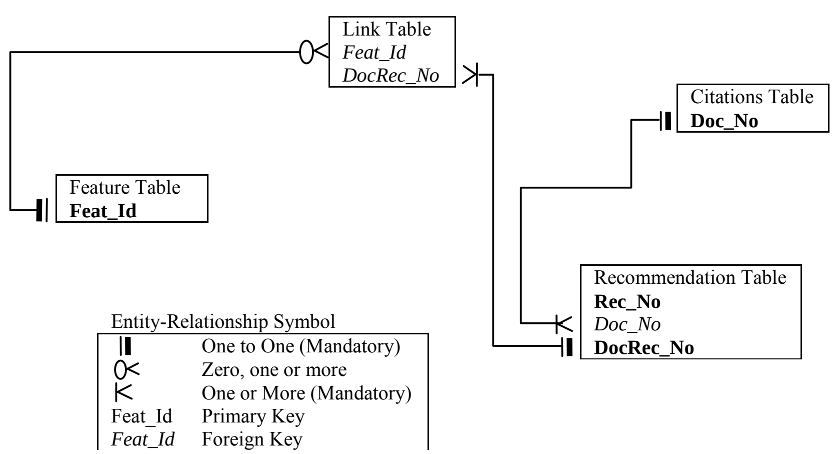
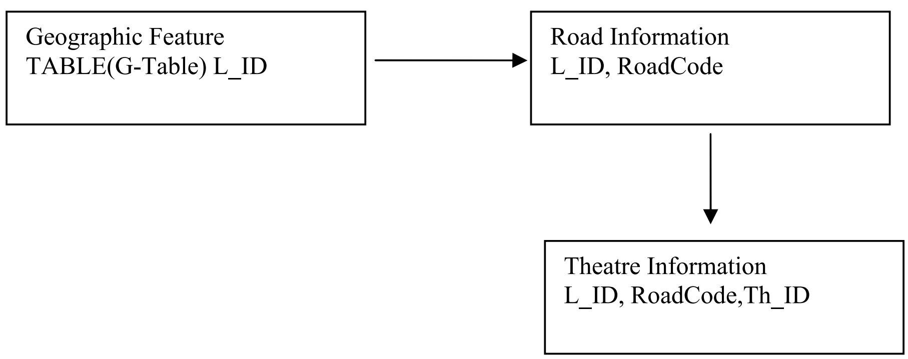


Figure 4.6 Schema diagram

The possibilities for relationships are one to one, one to many, and many to many. Additionally, relationships may have the property of being required (mandatory) or optional. Database programs relate tables one to one and one to many directly between the tables, but a many-to-many relationship requires the construction of an intermediate table. By establishing one-to-many relationships between each data table (the Feature and Recommendations tables) and the link or composite table, we create a many-to-many relationship between the two data tables. The one-to-many relationship between the Feature table and Link table allows features to exist in the Feature table that have not been discussed in a document and for which there are no recommendations .This means that geographic features that might be important for identifying where we are not linked to any recommendations. Another way to handle this would be to create a dummy document and dummy recommendation that said “No recommendation has been made on this feature,” possibly in the DESCRIPT field of the Recommendations table. Then the Feature table-to-Link table relationship would be a one-or-many (mandatory) relationship. We have the ability to design it either way. If a zero relationship is possible, clicking on the feature in the data (map) view will produce nothing. If we have created the dummy recommendation and citation records, the text “No recommendation has been made” would appear. If we plan it one way and change our mind, it is always possible to modify the schema, but it is better to think through questions like that at the beginning of the design process. The reason that a schema diagram is important and not an optional element in designing a GIS should be clear from the preceding paragraphs. It is possible to document the relationships with words and descriptions, but the graphic picture of how the relationships flow is much clearer once we understand the symbols. With the tables and relationships in this schema it would be possible to click on feature on the computer screen—a point (well), line (section of stream), or polygon (wetland)—and immediately know at least all of the following: Who made this recommendation and when was it made.

- ❖ If this feature has had any water quality recommendations made on it and when.

- ❖ If this feature is recommended for preservation in any document. Of more interest are the queries that this structure makes possible. For example, we could create a query that would show all features:
- ❖ For which a recommendation related to fish habitat was made between1990 and 1995
- ❖ For which recommendations were made in a particular document
- ❖ That have a recommendation pertaining to wetlands and are recommended for preservation
- ❖ That have conflicting recommendations made for them in different documents



Schema Example

Metadata

Metadata is usually described as data about the data. It is the information we need to document our data set sufficiently so that an outsider can understand all the key issues involved in the construction of the data set, what the various values in the data set mean, what projection we are using, and so on. One analogy is to a catalogue card for a book in a library, although most metadata is considerably more involved than that. The product of the process of creating metadata is a file that describes our data set, or pieces of our database. The mountains of information available on how to produce it, what to include, how to check it against a standard, and publish it is huge, but the actual product is rather small. The likelihood that this chapter will be out of date by the time it hits print is high because there is a lot of work being done around the world to implement standards and also because the techniques to disseminate metadata are expanding rapidly. Early implementations of geographic metadata were simple text files that could be read by word processors. The National Spatial Data Infrastructure (NSDI) initiative of the U.S. government led to a standard for implementing metadata using hypertext markup language (HTML). Currently, the push is tousing extensions to HTML for producing metadata, Simple Graphics Markup Language (SGML) and Extensible Markup Language (XML). It is almost certain that there will be new high-level languages developed out of these that will have additional advantages for implementing metadata. Since the 1990s there has been a proliferation of metadata standards, but in recent years national organizations have begun to cooperate on a set of international standards that may eventually make it easier to document and share geographic information across the world. The United States Federal Geographic Data Committee (FGDC) standards were an early version of metadata standards, and now there are groups in Europe, Australia and New Zealand (anzlic.org.au/asdi/metaelem.htm), and internationally that are working on geographic metadata.

4.4 Managing World Wide Web-Based Interfaces

Interaction of viewer with GIS databases is occurring through Internet and intranet applications. Users gain access to these applications completely through their Internet browsing software (e.g., Netscape or Microsoft Internet Explorer) and require no GIS functionality on the client side of the connection. The software that constructs the maps and processes the requests from the client's browser to the map and data server, however, must have specialized software that facilitates the serving of the GIS database to the internal intranet or external Internet. In this type of interaction, the interface design is of extreme importance, and decisions about what functionality to place on it and how to design it so that users get the information they need must be made at relatively high levels in the organization because of the face that it presents to the public. There is a lot of work going on both by the GIS core software providers and third parties around Web-based interfaces to geographic data.

Serving static graphic images of maps across the Internet was a very early use of the technology and is still an important means of distributing geographic information. Taking a paper map, scanning it into digital form, and then making that digital file available across the Internet is an effective way to distribute already existing maps. This has been especially useful in making historic maps available to many viewers who otherwise would not be able to access these documents easily. More recently organizations have taken to placing some or their entire GIS database on the World Wide Web for users both inside and outside the organization to query. And people are using these databases all the time. The map-making and direction-finding sites on the Internet such as Mapquest.com and MapBlast are among the most frequently visited sites on the Web. These sites are network based GISs that, due to their efficient indexing and system architecture, can rapidly determine least-cost paths from any addressable point in the United States to any other addressable point and create a set of driving directions complete with required turns and distances in less time[7].

People get lost with the directions and sometimes they appear not to be the optimum path to people who use similar paths daily, but they have become an important location resource. The design of these interfaces is always much simpler than the professional desktop interfaces that come with commercial GIS and look more like customized, simplified application interfaces. They make use of standard tools such as (pan the map to a new location, (zoom into a small area of the map view), (zoom out to a wider view), and (display attribute information about the feature. The goal is to have the tools and buttons so self-explanatory that users can start getting information from the database almost immediately with no training. Testing and serious user feedback is vital in designing these interfaces. There is a temptation to include too much space on the screen for identifying information about the organization (e.g., large names, logos, disclaimers, etc.), and too little space for actual map and data content. No matter what programming language is being used for these Web interfaces or what GIS core software or database structure they are built around, they all work in fundamentally the same way:

When the user accesses the Web page through a browser, a default template page and the server sends a graphic map image to the client browser. This page contains the software coding necessary to communicate with the GIS data server.

The user clicks on the available options to modify the map and/or access tools to query the map. There is often a list of layers like a legend that can be clicked on and off. Usually the available layers are predetermined for the user, and there is no opportunity to browse for new layers. The software in the local downloaded page takes the information and creates the query in a format the server GIS data serving software will understand and be able to execute.

The browser then sends that query command to the GIS data serving software, which executes the query and sends back a new map image if the request was for a new map or a file containing attribute information if the request was for data. Every zoom, pan, or request for modification or data is a separate query and must be sent across the network to the server, which returns either a new map graphic or data. There are some applications now that actually download spatial and attribute data and make the query on the local client machine. These applications are becoming more and more common, particularly with organizations that have legal obligations to provide certain information to the public. Organizations that implement these Web applications for routine queries notice significant drop-offs in the number of clients who come directly to the organization for the same information over a counter. The service is often provided free, but sometimes the organization will charge monthly or yearly access fees and protect access to the Web site in that way. If we are considering Web access for our data, it is important that we consider the hardware implications, particularly the server and networking issues for intranets, in the early design phase. The interface design can wait until the database is up and going and typically will go through its own development cycle.

4.4.1 GIS Interaction and the Organization

User roles, interfaces, control, and access must be handled at a detailed level, but looking at interaction with the GIS from the point of view of the organization as a whole, we can expect that the interaction will differ as a function of where in the organization the user is. Front-line users have the most direct interaction; as we move through management levels, interaction becomes less direct, and people begin to interact with the output from the system rather than the system itself. Top management is concerned with the output of ordered information to help them make decisions, and how they get that information is not a major concern. Midlevel management is a bit more directly connected with the GIS, but the connection is tighter with management issues than with the data and applications themselves. Midlevel managers in GIS-using departments have interaction concerns that center on how they can get applications that will help their people do find out the proper information [12].

4.5 Prototype development

To visualize the Location Based Application, following prototype was developed in Java Environment to simulate availability of tickets of a certain movie in a particular location. The named of prototype is Smart Ticket. The Smart Ticket sample contains a J2EE component for the enterprise back end and a J2ME component for the mobile front end. Running the application requires a J2EE application server and a MIDP v2.0-compatible device (or emulator) with Internet connectivity. Prototype has been tested under both the J2EE server and MIDP emulator on the same computer as follows

1. Set up the following environment variables:
`JAVA_HOME`: The JDK installation directory (JDK v1.4.1 and above required)
`J2EE_HOME`: The J2EE reference implementation (RI) installation directory (J2EE v1.3.1 and above required)
`J2MEWTK_HOME`: The J2ME Wireless ToolKit installation directory (J2ME WTK v2.0 and above required)

2. Start the J2EE server using the following two commands:

3. `J2EE_HOME\bin\cloudscape -start`

4. `J2EE_HOME\bin\j2ee -verbose`

5. Deploy the J2EE application using the following script. The setup script in turn invokes the corresponding ANT task in `setup.xml`:

```
setup deploy
```

6. Point the browser to `http://localhost:8000/smartticket` and click on the populate database link to import the mock theater and movie data to the Smart Ticket database.

7. Start the J2ME WTK v2.0 and run MIDlet `smart_ticket-client.jad`.
Now, it is ready to order some movie tickets from the phone emulator!
Note: To run the MIDP client on an actual smart phone device, the J2EE server must run on an Internet-accessible computer and we must configure the MIDP client for the correct server IP address[23].

4.5.1 Smart Ticket in Action

The Smart Ticket client allows to manage user preferences, browse movie schedules, order tickets, rate watched movies, and pre-download the schedule. The overuse of design patterns results in unnecessary abstraction layers and produces slow and large applications.

4.5.1.1 Manage User Preferences

When the user starts the MIDP client for the first time, he/she will be asked to create a profile. The profile includes two types of information:

Account credentials: The username, password, and optional credit card numbers.

User preferences: The theater search zip codes, favorite day of the week, and preferred seating. Figure 4.7 shows how to manage user preferences.

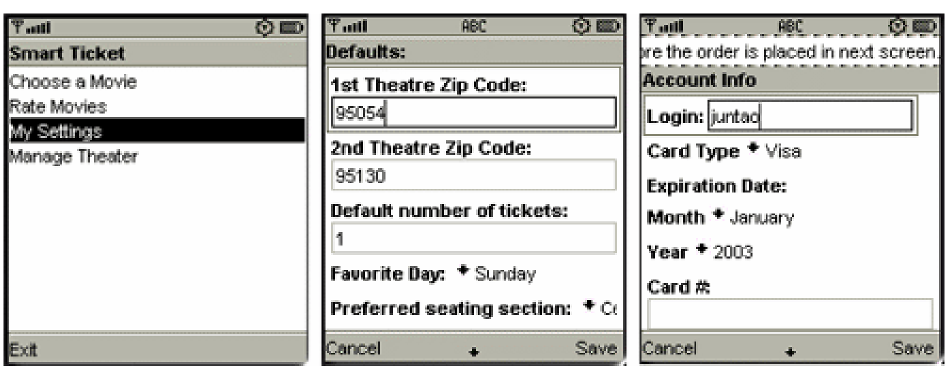


Figure 4.7 Managing User Preferences

After the user submits the profile, a corresponding user account is created on the J2EE server. The preference information is cached on the device. The user could configure the MIDP client to cache the account credentials so that he/she does not need to manually sign in every time she wants to purchase tickets or submit movie ratings. User preferences can be modified at any time through the MIDP UI [22].

4.5.1.2 Search and Purchase Tickets

Once logged in, the user can browse theaters, movies, and show times in his/her zip code areas. This process involves a series of real-time queries to the J2EE server. Once he/she selects a show, he/she will be asked to select currently available seats from an interactive seating map to make reservations. The reservation is persisted to the server database. Figure 4.8 illustrates the browsing and reserving process from the user's perspective.

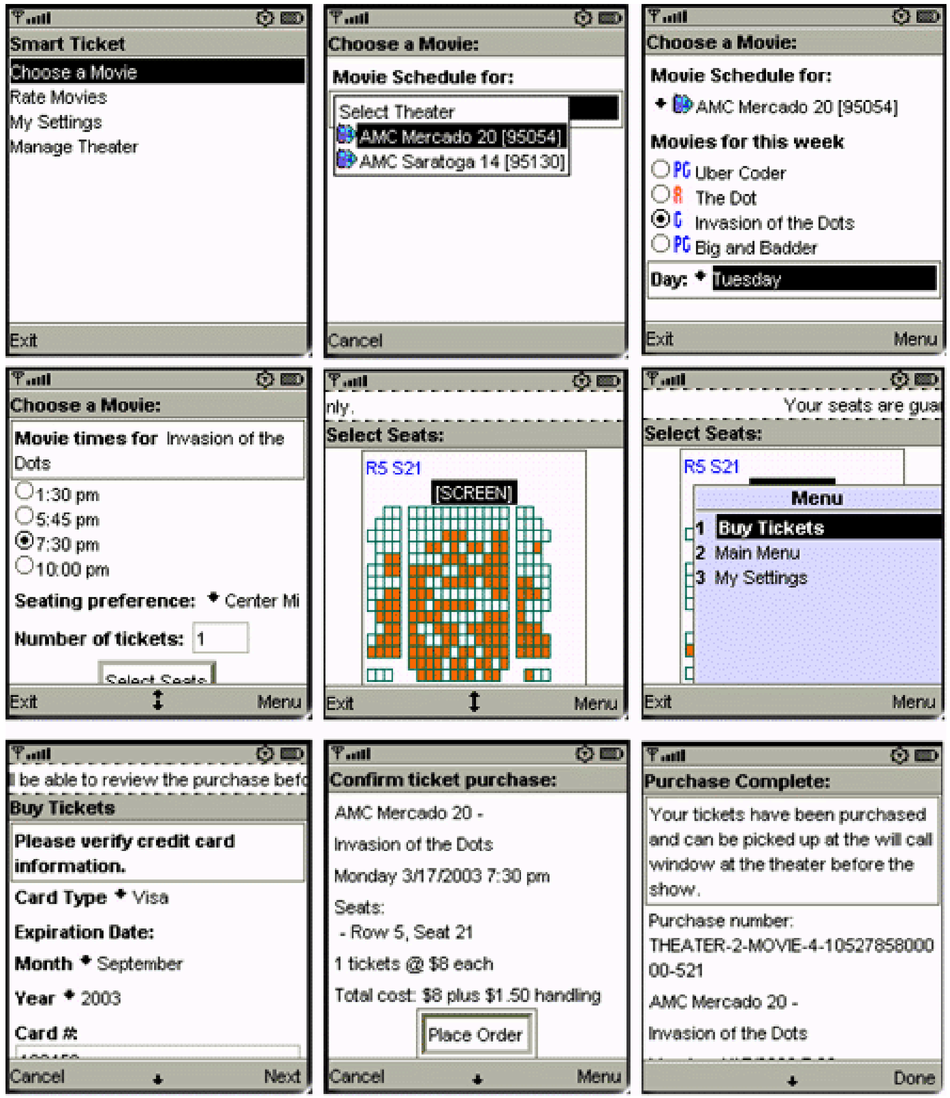


Figure 4.8 Browse and purchase movie tickets.

The interactive seating map allows the user to move a flashing cursor using the phone's navigation pad. The currently reserved and available seats are differentiated by colors. It demonstrates the rich UI capabilities of the MIDP.

4.5.1.3 Rate Movies

The user can rate movies he/she has seen (Figure 4.9). The ratings are not immediately submitted to the server rather they are cached on device and can be synchronized to the server upon the user request. That allows the user to rate movies even when the phone is out of network range (for example, in a shielded cinema building!). The synchronization agent is smart: When the same user rates the same movie multiple times, it resolves the issue by keeping only the most recent rating in the backend database. The user can only rate movies for which she has purchased tickets via the Smart Ticket system. This is necessary to ensure the integrity of the rating system. It also simplifies the UI design.

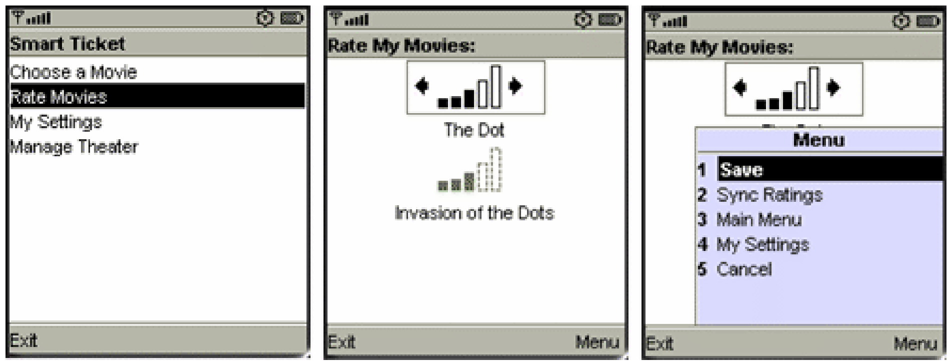


Figure 4.9 Rate watched movies

4.5.1.4 Cache Theater Schedules

Smart Ticket allows the user to download a theater's schedule to the mobile client. The cached schedule enables offline browsing and improves the performance by reducing network round trips. The user can delete or re-download the schedule as needed (Figure 4.10). The cached schedules have a different icon. That helps the user to avoid downloading the same schedule multiple times. It also helps the user to make quick choices of theaters in the browsing mode (Figure 4.8).

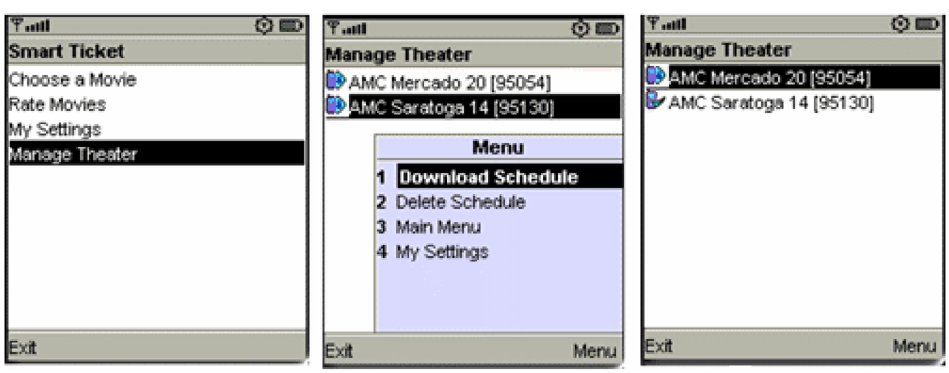


Figure 4.10 Download movie schedules into the on-device cache.

4.5.2 Important Architectural Patterns

Smart Ticket utilizes several architectural design patterns, which are commonly used by enterprise architects. It is important for enterprise mobile developers to understand those patterns.

4.5.2.1 The Overall Model View Controller Pattern

The overall architecture of the Smart Ticket application follows the Model-View-Controller pattern. According to Martin Fowler in his Patterns of Enterprise Application Architecture, the MVC pattern "splits user interface interaction into three distinct roles." In an MVC application, the view and controller components work together as the UI, which primarily concerns how to present the information to the user through a series of displays and interactions. The model component represents the domain model. The model's primary concern is business logic and machine-to-machine interactions (e.g., database access). The use of the MVC patterns brings some important benefits:

Since the presentation and the underlying data model are separated, a number of persons can be employed (experts) to work on different parts of the code. For example, a UI expert can design the views while a database expert optimizes the database connections at the same time.

MVC allows us to develop multiple views for the same model. For example, the Smart Ticket.

Non visual objects in the model layer are easier to test using automatic tools than are the UI components.

In the Smart Ticket prototype, the MVC pattern is implemented as follows:

- ❖ Model: Classes in the model layer contain all the business logic. In fact, the entire J2EE server component, on-device caches, and communication classes all belong to the model layer. The most notable design pattern in the model layer is the facades

- ❖ View: Each interactive screen is represented by a view class. There are 17 view classes in the Smart Ticket prototype client. Once the user generates a UI event (e.g., by pressing a button or selecting an item from a list), the view class's event handler captures the event and passes it to the controller class. Listing 4.1 demonstrates the UI classes for the screen to confirm ticket purchase.

- ❖ Controller: The controller class knows all the possible interactions between the user and the program. In the Smart Ticket application, the controller is the `UIController` class (Listing 4.2). It has one method for each possible action (e.g., `purchaseRequested()`). The action method often starts two new threads: one to perform the action in the background and the other to display a progress bar for the user. The background action thread is represented by the `EventDispatcher` class. The `EventDispatcher.run()` method contains a long list of `switch` statements that invoke the corresponding methods in the model layer to perform the action. When the model method returns, the controller displays the next UI screen using the appropriate view class.

The overall end-to-end architecture of the Smart Ticket application is illustrated in Figure 4.11 The model layer is by far the most sophisticated. Let us visualize how the model is assembled around a series of facades.

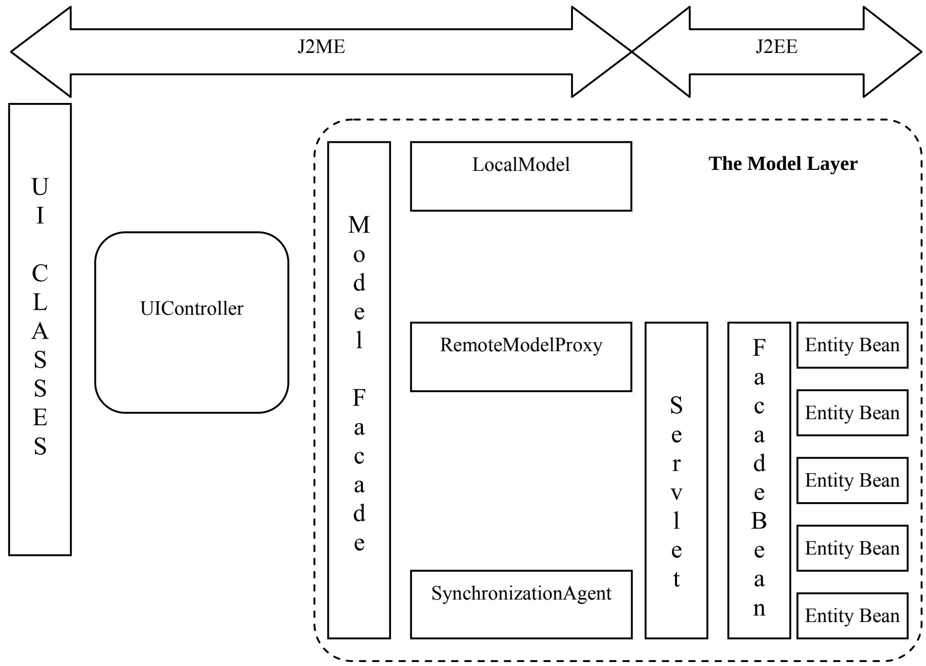


Figure 4.11 The overall architecture of the Smart Ticket application

4.5.2.2 The Clientside Facade

The facade pattern is a structural pattern that provides a simple interface for complex subsystems. In the Smart Ticket prototype, the clientside subsystems in the model layer, such as the `LocalModel`, `RemoteModelProxy`, and `SynchronizationAgent` classes, are behind the facade class `ModelFacade` (Listing 4.3), which is the entry point from the controller to the model. The `ModelFacade` class contains one method for each action in the model layer.

- The `LocalModel` class handles actions that access the local on-device storage. For example, the `purchaseTickets()` method adds the purchased movie to the on-device rating list. The `addMovieRating()` action method in the `LocalModel` class is called.

- The `RemoteModelProxy` class, which implements the `RemoteModel` interface, handles actions that require access to the remote J2EE server. For example, if the user decides to purchase tickets (`reserveSeats()` and `purchaseTickets()`), the transaction has to be done on the server side and be persisted to the database through the `RemoteModelProxy`. Action methods in the `RemoteModelProxy` class invoke remote procedure calls (RPC) to the remote facade on the server side. The details of the remote facade and the RPC format are discussed later.

- The `SynchronizationAgent` class handles all synchronization actions from the local data storage to the remote server. In the case of the Smart Ticket application, it handles only the movie ratings synchronization. It has two action methods: The `synchronizeMovieRatings()` method synchronizes the ratings; the `commitMovieRatings()` method commits the resolved synchronization requests to the back end and updates the content of the local store.

4.5.2.3 The Serverside Facade

One of the most important benefits of the facade pattern is that it reduces network round trips between remote systems. A properly designed facade allows us to use fine-grained objects in the subsystems yet still have a coarsegrained, simple network interface. It is especially important for mobile applications, since the wireless network is very slow.

When an RPC is made from the `RemoteModelProxy` to the server side, the HTTP servlet `SmartTicketServlet` (Listing 4.4) invokes the corresponding action method in Session EJB `SmartTicketFacadeBean` (Listing 4.6) through a business delegate object `SmartTicketBD` (Listing 4.5). Depending on the nature of the action, it is further delegated to either `TicketingBean` or `SynchronizingBean`, both of which are session EJBs too. The application data on the server side is persisted to the relational database through an array of Container Managed Persistence (CMP) v2.0 entity EJBs.

4.5.3. Implementation Techniques

The MVC and facade patterns define the overall architecture of the application. In addition, Smart Ticket showcases some important behavioral patterns and implementation techniques.

4.5.3.1 Chain of Handlers

On the J2ME device side, the RemoteModelProxy class (see Listing 4.3) further delegates the action to a chain of handler classes that transparently work out the dirty plumbing of the RMS and HTTP serialization. The chained handlers are based on the RequestHandler interface and the RemoteModelRequestHandler abstract class (Listing 4.7), which implements the former. Concrete handler classes extend the RemoteModelRequestHandler class. A chain of handlers is established through nested constructors. Two handler classes are available in the Smart Ticket prototype: the RMSCacheHandler and HTTPCommunicationHandler classes. Listing 4.8 illustrates how the chain is assembled and used (e.g., getMovie()) in the RemoteModelProxy class.

A handler can selectively implement any action methods in the RemoteModel interface. There are two possibilities:

If a RemoteModelProxy class calls an action method not implemented by the first handler class in the chain, the default implementation in the base class RemoteModelRequestHandler ensures that the call is passed to the next handler in the chain.

If a handler in a chain decides that it has finished processing an action, it returns directly. Otherwise, it can invoke the same action method in the base class to pass it to the next handler in the chain.

The following code snippets (Listing 4.9 and 4.10) illustrate how to implement the getMovie() method in the two handlers. The RMSCacheHandler looks up the on-device cache for the requested movie. If the requested movie is not cached, RMSCacheHandler calls its base class's getMovie() method, which passes the control to the next handler in the chain: the HTTPCommunicationHandler class. The getMovie() method in HTTPCommunicationHandler performs some network tasks to retrieve the movie object from the J2EE back end. To understand the inner workings of the HTTPCommunicationHandler class, we need to read on to the next section.

4.5.3.2 Binary RPC over HTTP

In the model layer, the HTTPCommunicationHandler class in the RemoteModelProxy class invokes remote procedures on the J2EE server side through a binary RPC protocol over the HTTP.

All RPC requests from the client to the server follow the same basic pattern: The first byte in the HTTP request data stream specifies the action method to be executed on the serverside session facade EJB. The RPC request code constants are defined in the MessageConstants class (Listing 4.11). The second byte to the end of the request stream encodes a sequence of UTF strings that represent the parameters to be passed to the remote method. The response HTTP stream contains the RPC return value. The format is unique to each method, and we have to look at the source code for each method to figure out the exact format. The two code snippets below demonstrate the entire RPC round trip to get a list of theaters using a zip code. The RPC request is assembled in HTTPCommunicationHandler's action method getTheaters() (Listing 4.12), and the response array is unmarshaled by the shared model object Theater (Listing 4.13).

The SmartTicketServlet first determines the RPC action code from the first byte in the request stream. It then dispatches the RPC to the corresponding action method through the facade and passes all the RPC parameters remaining in the stream. In the Smart Ticket prototype, the client and server are tightly coupled. This approach can improve network efficiency, since each RPC exchange can be specially designed and optimized. However, the trade-off is development speed and robustness. If we make small changes to the server, the protocol and the parsing code on the client are likely to need to change too. It is necessary to keep track of and update the code in multiple places, which could prove error prone. It is also needed to often need to recompile and redistribute clients.

4.5.3.3 The Clientside Thread Model

The Smart Ticket prototype uses a sophisticated threading model on the mobile client side. During a prolonged background task, another thread displays a moving gauge to the user indicating the progress (Figure 4.12). The gauge screen could also provide a button for the user to cancel the long action if she does not want to wait.

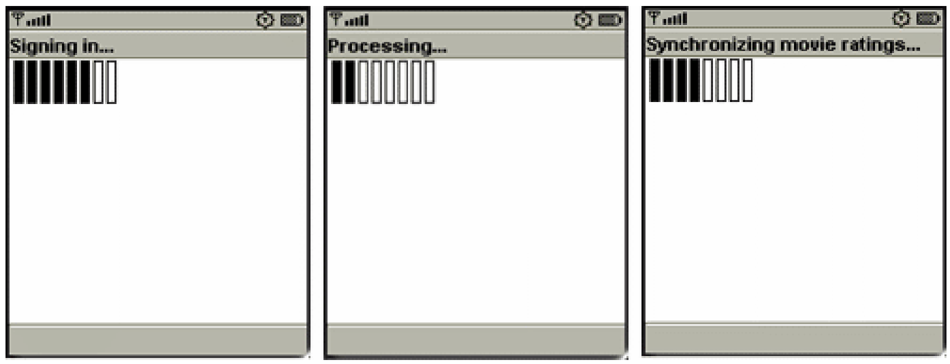


Figure 4.12 The progress gauge

As it has been seen, action methods in the UIController class are simply wrappers of the runWithProgress() method (Listing 4.14), which sets the display to ProgressObserverUI and starts the EventDispatcher thread. The ProgressObserverUI screen displays a gauge and an optional Stop button, which is monitored by the main MIDlet system UI thread. As described in The Overall MVC Pattern, the EventDispatcher thread eventually delegates the action to methods in the model layer. The model action method calls the ProgressObserverUI's updateProgress() method (Listing 4.15) at certain stages over the execution to update the gauge and inform the user of the progress (see Listing 4.12).

4.5.4 Summary

The prototype demonstrates the use of several important end-to-end application design patterns, including

- ❖ The Model-View-Controller pattern in rich client and J2EE application server settings.

- The clientside and serverside facade patterns.
- The chain of responsibility pattern for transparent network and persistence support on the client side.
- The binary RPC protocol for tight integration between the mobile client and the J2EE server.
- The worker thread pattern for non-blocking background tasks and user notification.

5. GIS Data Distribution through the World Wide Web

Increasingly, organizations that take GIS Data Distribution approach are using the World Wide Web to make their data available. There are hundreds of Web sites, mostly state and local governmental units that make their data available to all users this way. It still requires staff time to set up the distribution Web pages and to replace old information with updates, but it is a very efficient way to distribute the data [15]. If it is necessary to make the data available either on a fee basis or at no cost, there are some concerns we must need to deal with:

5.1 Metadata

Metadata should always be available for any information that is placed on the Web. Ideally it should be incorporated directly into the data or the file that the users will download rather than having a separate button they have to click to get it. Distributing metadata along with the GIS data is, by itself, a good thing to do, but the real reason that it should do this is to save the time it might take to answer the questions Good metadata should describe the data so well that there is no need to e-mail or telephone the distributor about the data.

5.2 Disclaimers

There are always liability issues around data released by organizations. This concern should not stop from distributing the data to people who need them so long as the users understand the data and what the appropriate uses of the data are. Disclaimers such as these have their origins in the disclaimers that organizations have been putting on maps for generations. They serve as a warning to the users in sufficiently legal terms that they are receiving an as-is product with no warranty or guarantee about its accuracy and, most importantly, does not place the organization at any risk for use of the information.

5.3 Update frequency

Distributing information via the World Wide Web, it is not going to be possible to have real-time, at the moment, access to the data. Periodically it is necessary to make a copy of the data we want to distribute and place it on the Web site. It is a good idea to plan for the frequency of the updates and explicit state the currency of the data in the metadata. If we need to refresh the Web site data frequently, it makes sense to construct specific applications that will automate the process. But outsiders will probably be working with information that is not as up to date as the data to which the workers in the organization have access.

5.4 Data formats

As anyone who has worked in GIS for some time knows, there is a bewildering array of data and file structures that different software systems use for GIS data. There is no way any organization could, or should, make their Web-based data available in all the possible formats for the occasional user who might need them. Generally, Web-based data access is provided for knowledgeable GIS users who can transform the data from whatever format(s) have been select into the format that meets their needs, and that should be their responsibility, not users. Casual browsers will need map-like interfaces rather than data access. Generally, it is easier to use exchange formats that take an entire layer of data and deconstruct it into a single file. It is also need to consider the abilities of other GISs to be able to read or import the data. This is why the Ungen format of ESRI [20], which is structurally the simplest format in the list, is not recommended for Web posting; users will need the more expensive core GIS software from ESRI to be able to read and use the data. If it is intended to provide data in a format that will require the least processing on the user's part to be able to view and use them.

5.5 Interface design

Organizations are usually quite sensitive about how they present themselves to the world through the Web, and a well designed interface is important. For the liability-conscious it is not difficult to place disclaimers in the interface so that the user has to click to get by the disclaimer, thereby registering and documenting that they have seen the disclaimer. This is just like the many license and use agreements that are on the Web for software and other services; most people don't read them and just click by. They are the Web equivalent of the fine-print product information on packages and are a good idea.

5.6 Provide an image of the data or not

It is common, but not universal, for providers to supply a graphic image of what the data will look like once they are imported into a GIS or mapping program. This is a nice service, but it adds another layer to the interface. For example, if we choose to make metadata available and two different formats for the data and a graphic image of the mapped data, it will need at least four separate files for each layer. If it is update frequently, this can lead to significant work.

5.7 Compression technology

Fortunately, file compression has become so standardized and the software much better that there are not many decisions to make. Because most GIS layers tend to get very large, it is almost essential to compress the files before placing them on the Web. Compression also allows us to assemble multiple files, if our format requires it, and the image and metadata into a single file so that when the user clicks on the download button, the entire package will be transmitted. With the widespread diffusion of the WinZip compression utility, most PC users are using that compression program to zip, or compress, files and folders. Unix users have a slightly different set of compression utilities, but the WinZip decompression program recognizes and can deal with that type of compression. Compression is necessary to move large amounts of data around the Web, but frequently a user will be using a particular Internet provider that has restrictions of the size of files it will allow us to receive.

5.8 Alternate means of distribution

Many public organizations provide the data at no cost on the Web but charge distribution costs for data that must be prepared on CD, disk, or tape. This is also the process adopted by several commercial distributors of data; people who wish to wait and do the work themselves can do it free, but if we want the company to copy and send the data, there will be a charge. Compact disk is becoming the standard format for that kind of distribution because the technology to copy disks has become so inexpensive. However our organization chooses to deal with these issues of Web deployment of our GIS data, the up-front setup time will quickly pay off in decreased time for staff to deal with data distribution requests. Organizations that do not wish to be as free and available with their data, of course, will not adopt Web technology or will control access to the Web distribution, which is an in-between way to control access. With this hybrid approach we can license or sell the rights to access the data pages and protect that access with a password that will let us into the Web interface.

5.9 Access Controls

Organizations differ radically in how they control access to the geographic data and applications they have created in their GISs. Some control access very tightly for reasons of confidentiality, security, and/or market concerns, and others open their entire databases and applications to everyone in the organization and even the outside world. This variation in control extends within the organization as well, with some units being very restrictive and others very open. So questions of access control and data security need to be addressed. Traditionally, the GIS community has been very open. Partly as a result of that tendency and to make it easier for people to use it, GIS software as it comes off the shelf has little security built in, other than from improper use of the software. One widely used software system allows us to put a password on a geographic data layer, but that password is written into a text file that describes the project and is not encrypted, almost no security at all. Because there is little security functionality in the core GIS software, we have to deal with access, control, and security from within our computer operating system and the RDBMS we are using to manage our enterprise GIS. Fortunately, both of these components of our system come with strong ability to manage access and security. Through the operating system we can assign permissions to folders, directories, and individual files. Through the RDBMS we can control access to data tables and can lock people out from even viewing selected fields in individual tables.

5.10 Control through the RDBMS

If the data are stored on a centralized server, control access is achieved through permissions granted to users and classes or groups of users when we set up their profiles. Because a GIS may have dozens of users and people come and go within an organization, setting up access inside defined roles or groups within the RDBMS is a simpler method of control than controlling access on a person-by-person basis. When a new user is entered into the system, the administrator will attach the roles that person may assume. When people leave the organization, removing them as users strips them of all roles and access ability. RDBMSs even have the ability to temporarily lock out users without deleting them from the entire system. In a local government it might be wanted to define a role of Property_Viewer, which would allow users to view, in map and table form, selected fields of a complex property database but not allow any modification or even querying of data other than to locate a particular property through an address or PIN query. A simple interface attached to that role would allow only those queries, but the

Property_Viewer could not query the database to find property based on the name of the property owner.That functionality could be reserved to a role of Property_Query. Roles are usually hierarchical. If, within the category of users that will be dealing with property data we have View, Query, Selective_Modifer, and Developer roles, each role will need the access privileges of the roles beneath it in the hierarchy. The setup and modification of roles within a database are tasks that database administrators are trained to do; it is really no different for geographic data access than for any other kind of data access within a database. Setting up the roles and the access privileges is the responsibility of the GIS database administrator working with the overall database administrator. Managers of the operational units that need the GIS will be helpful in defining the parameters of the roles that will support their units, but the details of role creation are best left to database administrators rather than line managers. 235

5.11 Control through the Operating System

Access control through the definition of roles can be backstopped by access controls at the operating system level. These types of controls were present on the early mainframe versions of GISs, became easier to manage with the advent of Unix operating systems, and are now present in the versions of the various Windows operating systems. The notion of administrator at this level relates to the hardware rather than the database. Roles and permissions work within the enterprise database, whereas operating system controls work within the pieces of hardware, (i.e., the disk drives on the system). Although database administrators control access to the database, machine administrators control access to the hardware on which the database resides. Obviously, these people need to work with each other. Although role assignment is a good way to control data access, permissions assignment is a good way to control application access. If we locate the applications that run our GIS in folders where only certain individual users have permissions to execute programs, only those users may run the applications. By denying them the right to modify (write to) those files or directories, we can disable their ability to change the application; that permission belongs to other users. This kind of control allows us to permit access to certain users during the testing of a new application and then open it to a wider set of users after we have worked out the problems. If the data and applications reside on central servers, the control is also centralized, but when we distribute data to users' individual computers, control becomes more difficult. All GIS software contains functionality to take in a data layer, perform some kind of query on it, and then export that to be a new data layer. If users access a data layer to which they have viewing access on a server and then create a copy of that data in a folder to which they have write access, they can then work on that copy of the layer and do whatever they want to with it. Removing that functionality from all the users could be difficult if there are lots of users in the organization. It is like a copying machine; we can control the initial access to the document, but unless we control access to the copying machine, it is difficult to stop users from copying the data and then distributing them to people who should not have access. The only solution to this problem is completely limiting access to the user's desktop machine or removing the functionality to create copies of data within the core GIS software. Data are going to get out. Copies and partial copies of data sets are going to be all over the place, and there is a real possibility that people will be making decisions and developing output from the copies instead of using the centralized, controlled data. If data is distributed across a network or through some physical medium like a tape or compact disk rather than serve it centrally, this problem will be worse, but it exists even though data is control by a central server. Very sophisticated techniques are necessary to completely control the ability of a user to copy data to a different location, but it can be done.

5.12 Controlling Public Access

For organizations that used public money to develop their GIS, the issue of how to allow and control access by the public to the database and applications is important. This aspect of GIS data access frequently receives a lot of attention outside the organization. Managers and administrators can tightly control who has access to what within the organization and have a responsibility to do that to maintain the integrity and quality of the data. It is not in anyone's interest to allow unqualified people to modify data. Controlling the public's access to the data is more difficult. This discussion is separate from the question of what we can charge for the information and deals only with what we have to make available and how we do it. Any governmental organization has dealt with the concerns before, and the implementation of a GIS is only another step in this process of making data accessible. All government's agencies usually seek legal advice before establishing public access policies, and if they do not, they certainly should. There are a number of questions the legal department might ask the managers of the line units using the GIS about public access. A simpler request, which is almost impossible for a public agency to deny, is the request to view the data set, and local and state governments are moving rapidly to provide map interfaces into their GIS databases across the World Wide Web. Because this is becoming so common, the question frequently arises early in the design phase, but it properly belongs in the management phase of the process. If the database is well designed, it can be made available in an Internet or intranet application. Web-based access through browser software is the thinnest client we can use to view the data. It can put information in the hands of a lot of people rather quickly. Private corporations, of course, have no legal requirement to make viewing of their data available to the public because the public did not bear the cost of constructing the data sets. The decision of whether and how to allow public access is a business decision except around the information that state and federal regulations require for publicly held companies. This information deals with the company's structure and financial well being and does not include proprietary information in databases the company uses to create profit (i.e., a GIS database). The management of access to a GIS database is a balancing act. On the one hand, we want to have as much data available to users as possible because the simple ability to view a wide range of spatial information may suggest new ideas. Pre-GIS processes were stingy with spatial data because of the time and cost of making the maps. When we can put it up quickly on a computer screen, people almost immediately want to see more information and more layers because they come up with new questions to ask. Generally this is a positive thing for an organization, but we can go a little overboard and add so much information to a screen that the clutter makes it very difficult to focus on the key issues. On the other hand, we can restrict access to only that information that was needed to do the job before the GIS, having the GIS reproduce the exact process used before. If we take that approach we lose the ability to consider new processes that might work better, but it is able to gain the simplicity of an interface that relates directly to the task. And users vary widely in their ability to absorb and use spatial information. Some people do not relate well to maps and want to see their information presented only in tabular form, whereas others enjoy and need to see the where as well as the what. Our applications need to be sensitive to the different ways people process information. Most, but not all, GIS practitioners are visual learners and are almost messianic in diffusing their way of looking at the world. Some users will resist this as well as the design of access to the database needs to meet their needs. Balancing between too little and too much information and meeting the needs of the visual users and the non-visual users creates interesting challenges, and how well we meet those challenges is a key determinant to how successful our GIS will be.

5.13 Managing the System—The Maintenance Plan

Before implementation one should have prepared a strategic plan, a management overview of how and where GIS would fit in the organization, and an implementation plan, a detailed process of how one going to do it. It is needed to prepare those plans, in that order, before implementation. The third of the three plans, a maintenance plan, is no less important but can wait until after the GIS is functioning. Table 5.1 shows the principal issues we need to treat in maintenance plan for the GIS. Applications often need updating and maintenance as we get feedback from users about problems with the way it works or does not work. Additionally, if the implementation is phased, there will be new applications to develop. An agreed upon schedule for development of applications in new units and the updating or maintenance of existing applications is important so that all units in the organizations feel that the GIS will work for them and no unit feels left out. Large GIS implementations typically will have dedicated staff for this task, and smaller ones may outsource it to consultants. As users in GIS-using units get more familiar and skilled with it, they will be able to make concrete suggestions for improving the application and may even be able to do it themselves. Additionally, it will discover applications that is originally did not plan to implement in the GIS but later realize that it makes sense to do so. In the first few years after implementation this aspect of maintenance will be particularly important, and one should budget adequate resources.

GIS Component	Maintenance Issues	How do we accomplish it?
Applications	Updating existing applications; creating new ones	Feedback from users, in-house or contract application development
People	Upgrading skills of existing users; training new users	In-house or contract training and workshops
Software	Upgrades and technical support, licenses for new users	Maintenance contracts with vendors
Hardware	Replacement cycle, hardware for new users	User feedback, keeping current with improvements in hardware
Data	Periodic large-scale data replacement	Periodic update of entire service area versus rotating updates of selected portions
Evaluation	Serious assessment of CIS utility and acceptance	Structured user feedback

Table 5.1 Maintenance plan for GIS

Maintaining the skills of the users of the system is also key. The implementation plan will contain a description and timeline for the necessary training, and the maintenance plan should contain the same for upgrading those skills and training new users. Training budgets in most organizations are sometimes considered luxuries and are often early sacrifices in tough financial times. GIS consultants especially recognize the steepness of the learning curve for GIS and are eager to offer their services in this area. Most organizations with GIS will have at least a few very skilled users, but they may be poor trainers, unable to transfer their skills and knowledge base to others effectively. The major software vendors have programs to certify trainers, and these people can provide a valuable service at a reasonable cost. Local universities with GIS programs are also good places to look for training resources if our organization lacks them. Whether we do all the training and skills development in-house or contract with outsiders, make sure any training we engage does the following:

a. Involve our organization's data, avoid training with canned databases that are unlike those we need to work with.

b. Be geared to the application needs and specifically involve our applications.

c. Be paced appropriately for the staff. The software vendors prefer to run multiday training sessions either at workplace or at their facility because it minimizes travel and support costs for the trainers

5.14 Data Dissemination

Ownership of GIS layers and feature data sets inside the organization is determined by who needs to use the information and what kinds of roles and privileges they need to do their jobs; it is internal to the organization and driven by its needs. When ownership is examined from outside the organization, legal issues of copyright, licensing, and access arise. If we separate issues of management from those of design and implementation, legal concerns are a management issue, complex and closely linked with legal issues around all kinds of information. But in the design process the legal concerns are simpler. If some state or local statute requires that certain information be kept about geographic features, our database design needs to include space for that information. Highway and engineering departments are required to keep track of the road signage, where it is, what kind of sign it is, when it was installed, and so on. Accident liability can be a significant cost for a government. These legal concerns in the design process, however, involve the details of the table and not organizational design.

5.15 Inside the Organization

One organizational issue that does have substantial legal implications is the establishment of procedures to disseminate the GIS data, both within the organization and outside it. Within the organization, decision-making users need to have timely information that is adequate for their needs. If the organization fails to deliver that information and a poor decision is made, there can be significant consequences for the organization. Decisions on how to make the information available to all users within the organization need to come early in the design process because they have significant implications on network and database design and on hardware. There are many options for how to do this, but they all come down to sub options or mixes of the following options:

a. Continuous, real-time interaction between the user and the database, which is maintained centrally.

b. Regular updates of the entire database or portions of the database delivered to the users to reside on their computers. Delivery can be through a network or by some storage medium such as a CD or portable disk drive.

5.16 Outside the Organization

It does not matter whether a private business, a governmental organization, or a nonprofit; once we create a GIS database, someone else will want a copy of all of it or pieces of it, and we need to be prepared for how the organization will deal with these requests. One will spend a lot of time thinking about this and doing it once we have figured out the process. This is actually part of the design process that can wait because it does not become a concern until the database is implemented. After all, we cannot distribute data the we don't have. If, as a private business, one need to develop geographic data sets that help us do our job and make profits, we do not have to provide copies of that data to anyone if we do not wish to. The license agreement will specify what the licensee may and may not do with the data and what the responsibilities are of the licensor with respect to the data (e.g., delivery of updates, answering technical questions about the data).The data are actually delivered, and the license buyer has a copy of them, but the use of the data is governed by the license agreement and, theoretically, if the user violates that agreement, there are penalties, the data must be returned, and so on. The GIS data of a private firm is a product, and the firm may market it as it does any product. In the public sector data distribution gets more complicated because public money was the funding source for the development of the data. :

- ❖ Profit making. Only private businesses may make money selling their data. Governmental units are not allowed to make a profit on data sales.
- ❖ Cost recovery. This costing model requires the provider to calculate the costs of building the data set and estimate the number of potential users over the life of the data set, and by dividing the costs by the estimated users arrive at a cost per user. This would be what we would charge the buyer.
- ❖ Cost of duplication. It is clear that governments may charge for duplication costs and have been doing so with information for many years.

6. Location Management

The growth of mobility aspects in cellular networks occurs at three different levels. First, there is the spatial level, that is, users desire to roam with a mobile terminal. Second, growth occurs from the penetration rate of mobile radio access lines. Third, the traffic generated by each wireless user is constantly growing. On one hand (e.g., cellular, ad hoc)[2] subscribers use their mobile terminals; on the other hand, more capacity-greedy services (e.g., Internet accesses, multimedia services) arrive one after another. From these considerations, the generalized mobility features will have serious impacts on the wireless telecommunications networks. Mobility can be categorized into two areas:

a. Radio mobility, which mainly consists of the handover process

b. Network mobility, which mainly consists of location management (location updating and paging).

6.1 Location Updating and Paging

The main task of location management is to keep track of the user's current location, so that an incoming message (call) can be routed to his or her

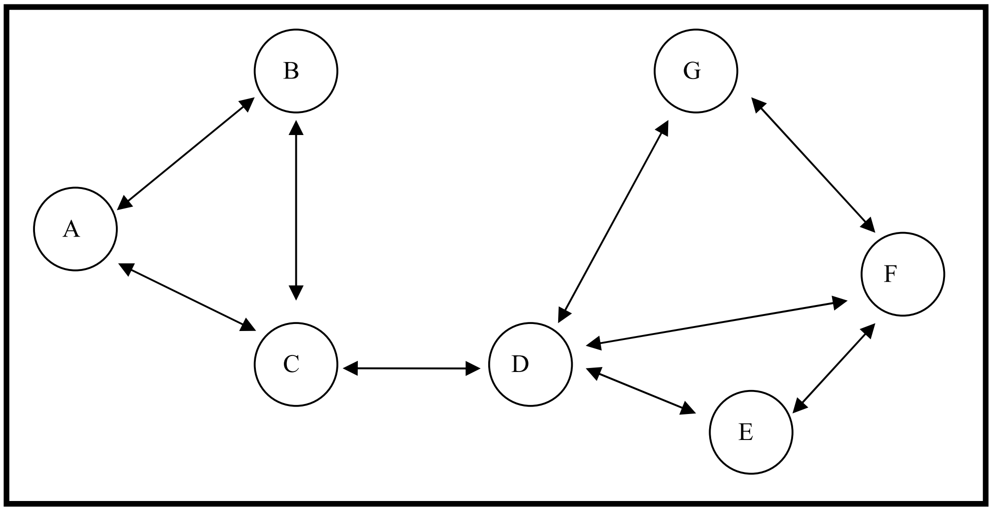


Figure 6.1 Basic structure of an ad hoc network.

mobile station (MS). Location management schemes are essentially based on users' mobility and incoming call rate characteristics[1]. The network mobility process has to face strong antagonism between its two basic procedures: (1) updating (or registration), the process by which a mobile endpoint initiates a change in the location database according to its new location; and (2) finding (or paging), the process by which the network initiates a query for an endpoint's location (which may also result in an update to the location database). The location updating procedure allows the system to keep the user's location knowledge, more or less accurately, in order to be able to find him or her, in case of an incoming call, for example. Location updating is also used to bring the user's service profile near its location and allows the network to rapidly provide the user with his or her services. The paging process achieved by the system consists of sending paging messages in all cells where the mobile terminal could be located. Most location management techniques use a combination of updating and finding in an effort to select the best trade-off between update overhead and delay incurred in finding. Specifically, updates are not usually sent every time an endpoint enters a new cell, but rather are sent according to a predefined strategy such that the finding operation can be restricted to a specific area. There is also a trade-off, analyzed formally, between the update and paging costs. For this purpose, the MS frequently sends location update messages to its current MSC. If the MS seldom sends updates, its location (e.g., its current cell) is not known exactly and paging is necessary for each downlink packet, resulting in a significant delivery delay. On the other hand, if location updates happen very often, the MS's location is well known to the network, and the data packets can be delivered without any additional paging delay. Quite a lot of uplink radio capacity and battery power, however, is consumed for mobility management in this case. Thus, a good location management strategy must be a compromise between these two extreme methods.

6.2 Mobility Models

Three mobility models, namely, the fluid flow model, the random-walk model, and the gravity model, are addressed. The fluid flow model considers traffic flow as the flow of a fluid, modeling macroscopic movement behavior. The random-walk model (also known as Markovian model) describes individual movement behavior in any cellular network. The gravity model has also been used to model human movement behavior. It is also applied to regions of varying sizes, from city mobility models to national and international mobility models. Mobility traces indicate current movement behavior of users and are more realistic than mobility models. However, mobility traces for large population sizes and large geographical areas have been categorized into a hierarchy by three different scales: Metropolitan Mobility Model, National Mobility Model, and International Mobility Model.

6.3 Location Tracking

In a cellular network, location-tracking mechanisms may be perceived as updating and querying a distributed database (the location database) of endpoint identifier-to-address mappings. In this context, location tracking has two components:

(1) determining when and how a change in a location database entry should be initiated, and

(2) organizing and maintaining the location database. In cellular networks, endpoint mobility within a cell is transparent to the network, and hence location tracking is only required when an endpoint moves from one cell to another. The location-tracking methods are broadly classified into two groups. The first group includes all methods based on algorithms and network architecture, mainly on the processing capabilities of the system. The second group contains the methods based on learning processes, which require the collection of statistics on subscribers' mobility behavior, for instance. This type of method emphasizes the information capabilities of the network.

6.4 Radio Resource Management

The problem of radio resource management is one important issue for good network performance. The radio resource management problem depends on the three key allocation decisions that are concerned with waveforms (channels), access ports (or base stations), and with the transmitter powers. Both channel derivation and allocation methods will influence the performance. The use of TDMA and CDMA are alternatives to FDMA used in the first-generation systems. With TDMA, the usage of each radio channel is partitioned into multiple timeslots, and each user is assigned a specific frequency and timeslot combination. Thus, only a single mobile in a given cell is using a given frequency at any particular time. With CDMA (which uses direct sequence spreading), multiple mobiles in a given cell use a frequency channel simultaneously, and the signals are distinguished by spreading them with different codes. The channel allocation is an essential feature in cellular networks and impacts the network performance.

6.5 Wireless Routing Techniques

A network must retain information about the locations of endpoints in the network, in order to route traffic to the correct destinations. Location tracking (also referred to as mobility tracking or mobility management) is the set of mechanisms by which location information is updated in response to endpoint mobility[8]. In location tracking, it is important to differentiate between the identifier of an endpoint (i.e., what the endpoint is called) and its address (i.e., where the endpoint is located). Mechanisms for location tracking provide a time varying mapping between the identifier and the address of each endpoint. In any communication network, procedures for route selection and traffic forwarding require accurate information about the current state of the network (e.g., node interconnectivity, link quality, traffic rate, endpoint locations) in order to direct traffic along paths that are consistent with the requirements of the session and the service restrictions of the network. Traffic sessions in wireline networks usually employ the same route throughout the session, and the route is calculated once for each session (normally, prior to the beginning of the session). Traffic sessions in mobile wireless networks, however, may require frequent rerouting because of network and session state changes. The degree of dynamism in route selection depends on several factors, such as (1) the type and frequency of changes in network and session state; (2) the limitations on response delay imposed in assembling, propagating, and acting upon this state information; (3) the amount of network resources available for these functions; and (4) the expected performance degradation resulting from a mismatch between selected routes and the actual network and session state. For instance, if the interval of time between successive state changes is shorter than the minimum possible response delay of the routing system, better performance may actually be achieved by not attempting to reroute for every state change. Moreover, the routing system can decrease its sensitivity to small state changes while continuing to select feasible routes, by capturing statistical characterizations of the session and network state and by selecting routes according to these characterizations. If a state change is large enough to significantly affect the quality of service provided along the route for a session, the routing system attempts to adapt its route to account for this change, in order to minimize the degradation in service to that session. As in stationary networks, the types of route selection and forwarding procedures employed in mobile networks depend partially upon whether the underlying switching technology is circuit-based or packet-based, and in part on whether the switches themselves are stationary or mobile. In most cellular networks, routes are computed by an off-line procedure, and calls are forwarded along circuits set up along these routes. Handoff procedures enable a call to continue when a mobile endpoint moves from cell to cell. In most mobile ad hoc networks, the mobile hosts themselves compute routes, and traffic is forwarded hop-by-hop at each switch along the route. The mobile hosts individually adjust routes according to perceived changes in network topology resulting from host movement. In mobile networks with stationary infrastructure (i.e., cellular networks), the main component of route selection for mobile endpoints is handoff. In mobile networks with mobile infrastructure (i.e., mobile ad hoc networks), the hosts not only need to keep track of the locations of other mobile endpoints but also need to keep track of each other's location and interconnectivity as they move. Route selection requires information about the interconnectivity and services provided by the hosts as well as information about the service requirements for the session and the locations of the session endpoints. This is a difficult task, however, in such a highly dynamic environment, since the topology update information needs to be propagated frequently throughout the network. In an ad hoc network, where network topology changes frequently and where transmission and channel capacity is scarce, the procedures for distributing routing information and selecting routes must be designed to consume a minimum amount of network resources and must be able to quickly adapt to changes in network topology . In cellular wireless networks, there are a number of centralized entities to perform the function of coordination and control. In ad hoc networks, since there is no preexisting infrastructure, these centralized entities do not exist. Thus, lack of these entities in the ad hoc networks requires distributed algorithms to perform equivalent functions. Designing a proper medium access control and routing scheme in this context is a challenging task which will be discussed in detail in subsequent chapters. Portable computers and communication devices with wireless connection to the network are changing the way people think about and use computing and communication. These wireless devices can communicate with each other even though the user is mobile. People carrying a mobile computer will, therefore, be able to access information regardless of time and current position. For example, they will be able to receive and send e-mail from any location or receive current information about local traffic, bus, and train services. But, location management will be an important problem in these situations because wireless devices can change location while connected to the network. New strategies must be introduced to deal with the dynamic changes of a mobile device's network address. A detailed description of the means and techniques for user location management in present cellular networks is addressed in this chapter. The ability to change locations while connected to the network creates a dynamic environment. This means that data, which is static for stationary computing, becomes dynamic for mobile computing. A stationary computer, for example, is permanently attached to the nearest server, while mobile computers need a mechanism to determine which server to use. There are a few questions that must be answered when looking at a location management scheme. What happens when a mobile user changes location? Who should know about the change? How can we contact a mobile host? Should we search the whole network or does anyone know about the mobile users moves? Location management schemes are essentially based on users' mobility and incoming call rate characteristics. Two basic mechanisms to determine a mobile terminal's current location are: location update (or registration) and paging. The network mobility process has to find a balance between these two basic procedures. The location update procedure allows the system to keep the user's location knowledge, more or less accurately, in order to be able to find him or her, in case of an incoming call. Location registration is also used to bring the user's service profile near its location and allows the network to provide the user with his or her services rapidly. The paging process achieved by the system consists of sending paging messages in all cells where the MT could be located. Location management methods are broadly classified into two groups. The first group includes all methods based on algorithms and network architecture, mainly on the processing capabilities of the system. The second group contains the methods based on learning processes, which require the collection of statistics on subscribers' mobility behavior, for instance. For location management purposes, cells in a cellular network are usually grouped together into location areas (LAs) and paging areas (PAs). An LA is a set of cells, normally (but not necessarily) contiguous, over which an MS may roam without needing any further location updates. In effect, an LA is the smallest geographical scale at which the location of the MS is known. A PA is the set of cells over which a paging message is sent to inform a user of an incoming call. A network must retain information about the locations of endpoints in the network in order to route traffic to the correct destinations. In cellular networks, endpoint mobility within a cell is transparent to the network, and hence location tracking is only required when an endpoint moves from one cell to another. In location management, it is important to differentiate between the identifier of an endpoint (i.e., what the endpoint is called) and its address (i.e., where the endpoint is located). Mechanisms for location management provide a time varying mapping between the identifier and the address of each endpoint. We have already introduced that location management typically consists of two operations [14]:

(1) updating (or registration), the process by which a mobile endpoint initiates a change in the location database according to its new location; and

(2) finding (or paging), the process by which the network initiates a query for an endpoint's location (which may also result in an update to the location database). Most location management techniques use a combination of updating and paging in an effort to select the best trade-off between update overhead and latency involved in paging. Specifically, updates are not usually sent every time an endpoint enters a new cell, but rather are sent according to a predefined strategy such that the finding operation can be restricted to a specific area. There is also a trade-off, analyzed formally , between the update and paging costs. Location management methods as adapted in current cellular networks, such as GSM, Interim Standard 54 (IS-54), and IS-95, may be perceived as updating and querying a distributed database (the location register database) of endpoint identifier-to-address mappings. So it is important to determine when and how a change in a location register database entry should be initiated and how to organize and maintain the location register database

7. Location Management Policies

The problem of location management (LM) is relevant in parallel and distributed systems, where objects dynamically relocate. Techniques for managing location, i.e., Location Management Policies (LMP), describe the rules which are used to find objects and the actions to be taken when objects migrate to the new locations in the network. A LMP should provide efficient implementations for move and find operations on objects. Efficiency in the context of LMP is defined in terms of

(1) Communication,

(2) Computation overheads, and

(3) Response time.

However, for the same LMP, it is possible that optimizations of one operation will deteriorate the performance of another and vise versa. This is illustrated by the following two extreme strategies.

The “full-information" strategy requires up-to-date information about all objects for efficient find operations, but then the cost for performing move is high (all nodes have to be up- dated).

On the other hand the “no-information" strategy does not require location updates. Consequently, the find operation is very expensive, its cost is almost equivalent to a global search. The impact of location management policies on performance of parallel and distributed applications that require object migration. The LMPs we evaluate combine existing experience of location management in Parallel Distributed Computations (PDC), mobile communication networks, and mobile agents computing. This evaluation is the first (to the best of our knowledge) comprehensive evaluation of LMPs in PDC. One of the objectives of this study is to classify existing location management approaches in terms of their impact on the overall performance of parallel and distributed computation applications[2].

7.1 Parallel Distributed Computations

Running programs or parts of programs concurrently and in an organized manner on different processors has long fascinated researchers with its promise of a quantum leap in computing power. Parallel computation presents problems which are either nonexistent or trivial in the context of sequential computation. Parallel computing involves the study of the ways in which the potential power of thousands of connected computers can be harnessed to solve challenging and very computationally time-consuming tasks. Research in the parallel and distributed systems group focuses on developing and analysing parallel algorithms for solving problems in computational geometry such as path planning for mobile robot and in image processing

7.2 Mobile Communication Networks

During the last years we have been observing constant development of wireless and cellular communication technologies. Different kinds of Public Land Mobile Networks are becoming more and more ubiquitous. Cellular phones, palm-top computers, laptops with wireless network cards, i.e., mobile terminals (MT), are not attached to a single stable physical location, but roam around. This creates the need for special techniques to handle such movement in order to guarantee communication between MTs. In this section we describe standard location management procedures in cellular communication networks and survey some of the proposed modifications. Typical structure of a network infrastructure supporting cellular wireless communication is depicted on figure 7.1. The geographical area is divided into location areas (LA). Each LA can contain one or more cells. A mobile support station (MSS) is assigned to

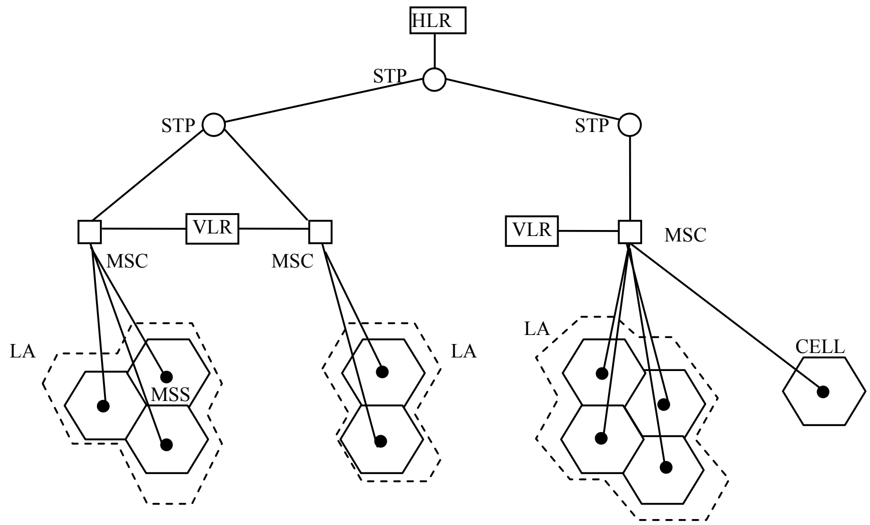


Figure 7.1 Example architecture of a mobile communication network

every cell to handle all network traffic directed from a MT located within the cell. Up in the hierarchy a Mobile Switching Center (MSC) governs one or more LAs and maintains a database with MTs locations. Multiple MSCs are connected together by a fixed backbone and/or intelligent network through a number of Signal Transfer Points (STPs). In cellular networks mobile users are tracked using two-tier scheme [8, 40] (as defined in IS-41 and GSM [17] standards). A location database, called Home Location Register (HLR), is predefined for each MT. Another database, a Visitor Location Register (VLR), is associated with one or more LAs (see figure 7.1).

Two procedures, governed by the standard, define what happens when MT moves from one LA to a different one, and how a call recipient can be found. Following is the brief description of location procedures as implemented in the current mobile networks. There are two possible scenarios when a MT moves from one coverage area to another. If the new area shares the local database (VLR) with the original one, that VLR is simply updated with the new location of a user. If VLRs are different, home registry has to be updated with the user's new location. The MT requests to remove its record from the old VLR and registers with the new one. If there is a need to locate a particular MT when another MT makes a call from some cell, the request is first sent to the local VLR of the caller. No further actions are required if that VLR possesses information about the recipient's location. Otherwise, a query is propagated to the callee's HLR. The up-to-date location information (which HLR always has) is sent back to the caller's support station. The support station covers the whole location area consisting of multiple cells. The actual cell where the recipient is located is determined by polling, or paging, within the LA. The search request is broadcast to all cells of the LA, and the recipient reports its location cell upon receiving this request. At that point connection between the two MTs is finally established. Most of the research about location management in Mobile Communication Networks has been concerned with the costs of updating the HLR. Some studies were trying to keep and improve the centralized nature of the scheme, while others were attempting to distribute the process of location. Interestingly enough, all of the described techniques are just proposals. They have been evaluated using theoretical analysis, simulations and traces, but none is a part of the existing standards. The nature of mobile network communication is usually unpredictable, but the infrastructure should support any particular pattern. Schemes that can adapt to the communication and migration characteristics of MTs are advantageous. When a home database is queried for a specific MT, the response is stored locally, so that the subsequent call to the same MT may not require communication with the HLR. It has been shown, that if CMR is high, caching performs very well. Another proposed improvement is based on user profile replication. A profile represents the set of mobile users, whose location information is always kept up-to-date at the local VLR. This enables quick location of the most popular users. Forwarding technique eliminates the update operation by keeping a pointer to the new location of a migrated MT at the source LA VLR. When a request to locate that MT arrives, it will be forwarded to that new location. Forwarding techniques decrease the load on HLR, but have high overheads if forwarding chains become long. If Call-to-Mobility Ratio (CMR, the number of calls issued to the user over the number of times it changes location) is lower than 0.5 and forwarding chains are at most 5 hops long, forwarding reduces user location costs network overheads by 20-60%. A conceptually different approach uses distributed database architecture instead of a centralized HLR. This technique takes advantage of the fact, that in most cases back- bone/intelligent network architecture has hierarchical tree structure. This allows distributing the load of location management among the non-leaf nodes of the tree.

Partitioning of the coverage area into zones, among which MT moves infrequently, is yet another modification which reduces the number of LSs and query time for certain call/migration patterns. A partition consists of location areas, which are represented by the dedicated location server. That representative LS is not aware of the exact MT location, but knows its current partition. This technique reduces update-induced communication. Summarizing, hierarchical location schemes eliminate the need for centralized HLR at the cost of increased general complexity of location management and increased storage requirements at the intermediate LSs. Hierarchical techniques support locality of communication and migration of MTs. Compared with LM in PDC, mobile network systems have a number of distinct properties. In PDC a local location directory is associated with each processor. This directory is analogous to VLR in combination with supporting stations, which communicate with MTs local to a location area. However, in PDC location directory is always aware of all the objects local to the process address space. This eliminates requirement for paging. In PDC applications, similarly to MNC, communication and migration patterns are not predictable in general case. At the same time, in MNC the migration options for a MT are limited by neighboring areas, while in PDC object can migrate to any of the processors regardless of their geographical location. Of course, PDC application does not include unpredictability of human character, present in cellular phone networks. Another difference of PDC is the time required for a mobile object to change its location. In MNC there are strict limitations on maximum travel speed for cellular phone users; the sizes of communication cells are also predefined. PDC applications can possibly move hundreds of objects in few seconds between geographically distant locations. There are also differences in the system architecture. Mobile networks in most cases have hierarchical structure. There can be dedicated location servers on non-leaf nodes of the hierarchy. In PDC applications computation is done either on a COW, or a collection of clusters. All nodes have equal functions, and the application can rarely take advantage of the underlying network routing, as it is handled by the low-level protocols. The PDC model assumes fine-grained object mobility. The number of objects may be large. We argue though, that the number of objects in mobile communication networks, e.g., in cellular networks, is generally much higher (hundreds of millions). This puts certain size and memory limitations on MNC LM algorithms.

7.3 Mobile Agents Computing

Mobile agent computing is a relatively new area of distributed computing, which is gaining more popularity with the development and growth of the Internet. Mobile Agent (MA) is an independent piece of code and data. It can be taken from the execution context on one host, migrated to a different machine, and continue execution there after migration completes. Recent progress in developing platform-independent environments (e.g., Java Virtual Machine) addressed many technical difficulties, inherent to the implementation of mobile agents, which also contributed to the growing popularity of the model. The spectrum of applications, which can take advantage of mobile agents, includes e-commerce, distributed collaboration environments, information search and dissemination, network management and monitoring. Some of the MA applications require support for communication between the agents. In such cases location management techniques play very important role. Nevertheless, being important, location management is not the major research issue in mobile agent computing: the main challenges in MA systems are support for mobility, security, naming services and fault-tolerance. Home server location algorithm associates a specific host with each mobile agent. Every time a mobile agent changes its location, home server is updated with the new location. A message addressed to the mobile agent is sent to its home server, which forwards that message to the agent.

8. Future Trends and Challenges

Location based services has generated a lot of interest in recent years, as a new source for mobile operators to enhance their service offerings, thus potentially increasing revenues. Prediction of LBS usage have generated lot of interest and attracted many new players developing and offering numerous application and services. It has been the regulator that forced operators to be able to provide subscribers' location data for safety application. The lack of standards and preferred technologies in area such as Position determination Technology, services platforms, mobile terminals, graphical presentation opens the opportunities for many newer technologies. Figure 8.1, illustrates the various entities involved in making LBS work. But still, the partnership between these entities would still need to be worked out.

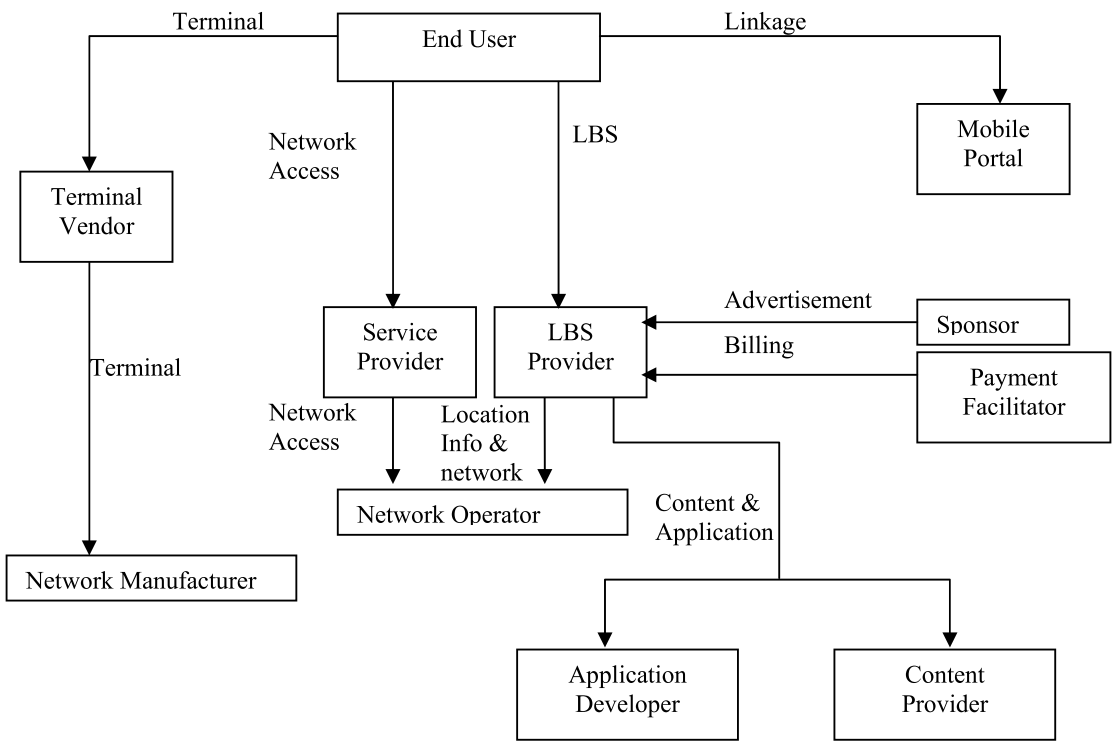


Figure 8.1 Today's Location Based Services

8.1 Key Trend in Location Based Services

To offer LBS is tough challenge for mobile operator. Many different areas exist where thorough consideration is required, that might have a crucial effect on the bottom line. Challenges include the choice of Position determination technology, the use of a third party platform vs. the in-house design and implementation of an application platform, the selection of services to offer and selection of business models to use etc. As with any new, and somewhat hype driven business field, operator are trying to be innovative, and yet keep an eye on the completion to make sure they are not left behind. It is quite interesting to note that marketing effort for LBS are usually quite low. Some of key trends in Mobile communication industry for LBS are:

8.1.1 Launch off LBS

In global market, launch of LBS is mainly driven by E-911 directives. It is mainly driven by competition, leading to earlier deployment of LBS. Most operators in are already offering some related services and are planning to launch more services in the future period on their GSM, as well as next generation networks. The most common services for the consumer market at the moment include 'yellow pages' and 'point of interest'. These services allow users to locate the nearest hotel, cinema etc., based on their current location. Other common services include navigation and traffic information. The next planned steps for consumer application include improved functionality of the existing services, and offering entertainment services such as community application and games.

8.1.2 Business Model for LBS

In order to enable launch of attractive services for LBS, mobile operator seems to recognize the need to share the LBS revenues with other players. The business model used by application developer in LBS market include a combination of one time set-up fees, revenue sharing and monthly payment for additional services such as technical support and upgrades, customer care, etc.

On the subscriber side, operators are also experimenting various charging schemes, including charging per additional time/traffic, charging for premium services and a monthly subscription. Due to low use of LBS today, and the low revenues associated with it, it is impossible for the different players to rely on revenue sharing with the operators. This trend is likely to change in future if and when revenues from LBS actually increase as predicted [18].

8.1.3 Implementation of LBS Location Based Services

The key trend in the Nepal at the moment is E-OTD for GSM/GPRS operators and A-GPS for CDMA/TDMA operators. But since GSM/GPRS is gaining momentum in Nepal, E-OTD has chance of becoming a dominant technology[14].

8.2 Opportunities for New Companies

IT professionals and software developers should note that cost of entry for developing content applications is relatively low. Higher-level application development tools available from major players, such as AirFlash/Webraska, IBM, Nokia, Openwave, Oracle and Sun, are making it easier to develop location-specific content for various industries - hospitality (hotels, restaurants), gas stations, movie halls and theatres. There are a number of business opportunities for new companies. These can be classified as follows:

Location-specific portals - Besides the large national and multinational portals, there is room for creating local portals. The major players are expected to franchise this to each geographical area because they cannot go to individual cities and enroll local merchants.

Content providers - local, national and global administrative, management and customer billing tools that apply to this industry specifically - this applies to other segments of Telecom industry as well

Unique location-specific applications - largest scope lies in this space. There is opportunity to market these applications through carriers for other locations [18].

8.3 Key areas of consideration

It is believed that in coming few years the competitive pressure in general will increase, forcing operators to offer even more innovative services, including LBS. Operator who have started with simple information services will gradually start to offer more attractive and wider range of LBS. Current Positioning services technologies will replace with more accurate technologies, enabling new services. In most LBS will not necessarily be a service category in itself, but rather an added feature to existing services, increasing their usability and values to the others. Some need to b considered and will have crucial effect on this development are:

8.3.1 New network technologies

The availability of new network technologies including 2.5G and 3G technologies will increase the use of data services. The 'always on' data connection, the higher data transfer rates, and the charging per volumes and per user-values, will enable LBS to benefit from these technologies. The ability to push data to users based on their location and preferences, in a seamless and inexpensive manner, is likely to help LBS services to proliferate [11].

8.3.2 Standardization

Much effort is put in standardizing LBS, both on network side and application side. Main forces are the 3G-partnership program (3GPP) defining mainly the addition of LBS capabilities to future releases of 3G networks. The Location Interoperability Forum (LIF), formed by vendors and interested parties are developing and promoting the common and ubiquitous solutions for LBS that is network and Position Determination Technology independent. The outcome of these efforts will have an enormous effect on the success of LBS, affecting the technology choice operators will make, the required investment to launch or upgrade existing LBS., as well as on the actual availability, usability, and cost of services.

8.3.3 Availability off attractive services

LBS will not take off unless there are attractive and easy to use services. Some of these future services are likely to benefit from higher accuracy positioning services technology. The ability to offer such services requires tight cooperation between mobile operators, application developers and equipment vendors. This requires the understanding of subscriber's preferences and usage habits as well technology expertise. Standardization is likely to facilitate the development and launch of services, but the key is still attracting the subscribers.

8.3.4 User acceptance

A key question remains whether subscribers will be willing to pay additional fees to use these services. User acceptance surveys provide different answers, most of them debatable. General usage figure based on past experience with other services show that the answer lies in the usability and value services brought to users. This adds further dimension to the attractive services mentioned before – services should be tailored and offered to specific user segment, maximizing their values from such services. Operators are in a key position to define and package such services and tailor them to the need of their different subscriber segment.

8.4 GIS Data Distribution through the Mobile

It is always necessary to make the decision to give the data away to anyone who asks. Increasingly, organizations that take this approach to outside distribution of their data are using the mobile to make their data available. There is a few mobile applications, mostly finding location, that make their data available to all users this way. It still requires staff time to set up the distribution LB application and to replace old information with updates, but it is a very efficient way to distribute the data. If we choose to make our data available either on a fee basis or at no cost, there are some concerns we need to deal with:

a. Metadata

Metadata should always be available for any information we place on the Mobile. Ideally it should be incorporated directly into the data or the file that the users will download rather than having a separate button they have to click to get it. Distributing metadata along with the GIS data is, by itself, a good thing to do, but the real reason we should do this is to save the time it might take to answer the questions 245 people will have about the data. Good metadata should describe the data so well that there is no need to e-mail or telephone the distributor about the data.

b. Disclaimers

There are always liability issues around data released by organizations. This concern should not stop us from distributing our data to people who need them so long as the users understand the data and what the appropriate uses of the data are. Disclaimers such as these have their origins in the disclaimers that organizations have been putting on maps for generations. They serve as a warning to the users in sufficiently legal terms that they are receiving an as-is product with no warranty or guarantee about its accuracy and, most importantly, does not place the organization at any risk for use of the information.

c. Update frequency

If it is distributing via the World Wide Web, it is not going to be possible to have real-time, at the moment, access to the data. Periodically one will make a copy of the data he/she want to distribute and place it on the Web site. It is a good idea to plan for the frequency of the updates and explicit state the currency of the data in the metadata. If we are going to refresh the Web site data frequently, it makes sense to construct specific applications that will automate the process. But outsiders will probably be working with information that is not as up to date as the data to which the workers in the organization have access.

d. Data formats

As anyone who has worked in GIS for some time knows, there is a bewildering array of data and file structures that different software systems use for GIS data. There is no way any organization could, or should, make their Web-based data available in all the possible formats for the occasional user who might need them. Generally, Web-based data access is provided for knowledgeable GIS users who can transform the data from whatever format(s) we select into the format that meets their needs, and that should be their responsibility, not ours. Casual Application will need map-like interfaces rather than data access. Generally, it is easier to use exchange formats that take an entire layer of data and deconstruct it into a single file. We also need to consider the abilities of other GISs to be able to read or import the data. This is why the Ungen format of ESRI, which is structurally the simplest format in the list, is not recommended for Web posting; users will need the more expensive core GIS software from ESRI to be able to read and use the data. We want to provide data in a format that will require the least processing on the user's part to be able to view and use them.

8.4.1 Interface design

If we have ever cruised the mobile application looking for geographic information for a project, we will appreciate the different interfaces people have developed for data access. Some clearly work better than others, and some seemed designed to hide the data rather than make it easy to find. Organizations are usually quite sensitive about how they process data in the world through the Web, and a well designed application is important. For the liability-conscious it is not difficult to place disclaimers in the application so that the user has to click to get by the disclaimer, thereby registering and documenting that they have seen the disclaimer. This is just like the many license and use agreements that are on the Web for software and other services; most people don't read them and just click by. They are the Web equivalent of the fine-print product information on packages and are a good idea.

8.4.2 Provide an image of the data or not

It is common, but not universal, for providers to supply a graphic image of what the data will look like once they are imported into a GIS or mapping program. This is a nice service, but it adds another layer to the interface. For example, if we choose to make metadata available and two different formats for the data and a graphic image of the mapped data, we will need at least four separate files for each layer. If we update frequently, this can lead to significant work.

8.4.3 Compression technology

Fortunately, file compression has become so standardized and the software much better that there are not many decisions to make. Because most GIS layers tend to get very large, it is almost essential to compress the files before placing them on the Web. Compression also allows us to assemble multiple files, if our format requires it, and the image and metadata into a single file so that when the user clicks on the download button, the entire package will be transmitted. With the widespread diffusion of the WinZip compression utility, most PC users are using that compression program to zip, or compress, files and folders. Unix users have a slightly different set of compression utilities, but the WinZip decompression program recognizes and can deal with that type of compression. Compression is necessary to move large amounts of data around the Web, but frequently a user will be using a particular Internet provider that has restrictions of the size of files it will allow us to receive.

8.4.4 Alternate means of distribution

Some people who will need our data may not have Internet access or their access is not adequate to deal with the volume of data they may need. For those people we will need an alternate means of distribution. Many public organizations provide the data at no cost on the Web but charge distribution costs for data that must be prepared on multimedia cards etc or in built-in application. This is also the process adopted by several commercial distributors of data; people who wish to wait and do the work themselves can do it free, but if we want the company to copy and send us the data, there will be a charge. Compact Disk is becoming the standard format for that kind of distribution because the technology to copy disks has become so inexpensive [12].

8.5 Discussion of future and ethics for location-based services

Are location-based services a threat against our personal integrity? Do I want the authorities or my family to know every move I make? Is it not possible to be anonymous and unknown for a short while? Can my boss always count on me to be able to work 24-hours a day? Are there no more peaceful moments in the weekend cottage? However lets not take to many terrifying examples and instead concentrate on how the personal integrity is maintained today. If we withdraw money on an ATM a camera registers us and the transaction is also saved digitally. The telephone calls we make can be listed or even bugged. The sites we visits on the Internet are registered on the www servers. Each of these actions is not so terrifying (except the bugging!) on its own but if all such things are put together, a pattern on our movements and lifestyle can arise. It can be very insulting for a person if someone supervises ones habits. The big solution is trust. If we can rely on that the bank office only uses the video from the camera in case of a crime, we can usually accept being filmed. The following six bullets could make people trust the location-based services.

- Laws. It would be calming for the users to know that there is a law that regulates what companies might do and not do with the locations. Currently no such law exists. In the Enhanced 911 decision it is clearly stated that the operators do not have permission to use the position except for emergency situations. It is not clearly stated on how the operators receive the users' permission to use position in other cases. Regulations about collecting and saving positions are also not stated yet.
- Clear Information. Information about if the locations are stored and if it is going to be sold to another company must be clearly stated when a user joins a service. A company can make directed advertisements if they know a users daily movement routine. Today it is natural for companies to ask if they can sell e-mail addresses etc.
- High mobile usage. Even if location-based services currently are considered to be a corner stone in mobile Internet and wireless services it cannot be the only part. People need to find other applications and devices that are interesting too. A high mobile penetration will benefit the entire business.
- Give high-quality services. To make applications that people use and like is maybe the most important task. Many people do their bank affairs through the Internet today even though the security mechanisms are not fool-proof. Dishonest crackers can destroy the day for a person. But on the other hand the service is easy to use, fast and highly recommended from the

bank. Humans can live with a little more unsafety if they see that they can take benefit of it.

- ❖ Ability to turn positioning off. The companies are probably going to develop services that are both terminal and network based. The problem with network based services is that the positioning cannot be turned off, compared to when using a GPS receiver. Anyway a user must rely on the operator that whenever they want they can switch off the positioning. Today profiles can be set for the ring signal on the mobile phone. Profiles should also be possible to specify for positioning services. One profile is for work; in that case every allowed work colleague may position the user. On the weekend a profile that only lets the family position the user is selected. One tricky question is if it should be legal to locate a suspected criminal who has turned off his positioning service?

- ❖ Decreasing stress. The goal is to make life easier for people. If someone positions a user on the tennis court they probably understand that it is not the right time to disturb during the game. The business must nevertheless watch out for not increasing the feel of stress. More efficient applications will sure creates higher demands on the humans both in the home and at work.

9. Thesis outcomes

9.1 Findings of Research Work

The objective of this research is to investigate and develop an Object-Oriented Application Framework to improve the software reusability of the wireless communication framework for Location Based Services applications for commercial aspects. LBS incorporation in Commercial platform increases efficacy for the consumers in many aspects. LBS has made foray into the market. Attractive location-based services covering all walks of life are here to stay hampered only by mans imaginative limits. We must, however, understand that there is no single universal solution to LBS. According to our research we have find out that Location based service work effectively with GIS database providing spatial based DECISION SUPPORT SYSTEM in handheld device to the consumers. J2ME supported handheld device can be effectively used as LBS Application Deployment. LBS application can be well designed and can be interpolated with most of the commercial system that needs positioning, tracking and communication system. LBS system works fine with 2.5 G Wireless architecture and its performance can be increased with new wireless technologies like 3G. LBS application can make a good use of geographic database system for commercial approach.

The control center is another important element in LBS. It is the center of data management and dissemination of location services. The location data, geo-coded data or map data need to be managed and manipulated to provide the various services based on the location to the user.

Retrieving and using Geo-coding is of prime importance as it determines the latitude and longitude of an address. The technical requirements of geo-coding services can vary from trivial to demanding. A trivial approach could mean getting the approximate location of an address in a street network.

LBS has a great future in these areas : Mobile Computation Services(Communication and computations using mobile devices) , Location Enabling Services (User Location Providing) and Location Aware Services (IP based applications) which when integrated with commercial application provides ease to customer and also results a good service oriented throughput.

The standards that have been used in this research work are primarily :

- TCP and IP
- WAP
- J2ME, CLDC, MIDP
- XML, XLS and XLD and other W3C recommendations
- OGC and ISO TC 211 recommendations
- XML Geo-Database

The finding of this research work that the prototype developed during the research works determines that:

- All devices work any where, any time from 2G to latest technology
- Devices from different companies (perhaps with different positioning systems) work with all Net Servers - they will give the same answer
- Services will be delivered on different wireless systems easily
- How will hand off from one to the other be conducted by specified methods
- Security can be ensured by owner encryption
- Services can be extended beyond retail goods and services

9.2 Conclusion

Efficiency, performance, and the costs of development and maintenance for scientific computing applications are directly dependent on the quality and capabilities of the runtime software. LBS application must have the three fundamental issues: correctness, performance, and ease- of-use. The second major contribution for LBS application is the survey, comparison and evaluation of location management techniques. LBS applications depend on multiple factors: number of nodes involved in the computation, properties of the communication network, migration and communication patterns of the application.

Location Based Service or LBS, is the ability to find the geographical location of the mobile device and provide services based on this location information. For an example a person at any location want to find the nearest theatre to watch the movie with economy budget, he needs only movies names and addresses which are within his reach, say within one sq.km., out of the database of say few theatres in the city spread over few sq. km.

The foundation stone of Location Based Services was laid by the Federal Communications Commission of US (www.fcc.gov) ruling which required the network operators to provide emergency services by locating the user of the mobile device within 125 meters. It required wireless network operators to supply public emergency services with the caller's location and callback phone number. This lasts the emergence of new and dynamic field called LBS, where the service was based on the geographical location of the calling device. Further, the developments in the field of Positioning Systems, Communications and GIS, fueled the imagination of the industry people with regards to the LBS. This ability to provide the user a customized service depending upon his geographical location could be used by telecommunication companies to any commercial providers.

In the days to come, the LBS will be benefiting both the consumers and network operators. While the consumers will have greater personal safety, more personalized features and increased communication convenience, the network operators will address discrete market segments based on the different service portfolios.

Unfortunately when talking about LBS, most think in terms of static user-initiated scenarios like finding the nearest taxi rank or getting a map proximate to the user's position. Although such applications of location awareness are indeed useful and can bring revenue to many service providers, it is somehow limiting the domain and range of the possibilities arising from location aware computing, especially in the context of mobile phones. Granted that these are the early days and such services are limited and not widespread, nevertheless we need to think about LBS in terms of an enabling (infrastructure almost) technology, where location awareness is invisible and not an end in itself. In many cases, useful LBS can be achieved without high degrees of positioning accuracy or in fact availability. While for some other cases only accurate positioning will do. In this paper, a detailed description of Location Based Services has been provided. This paper initially begins with introducing the location based services and its significance to cellular operators and mobile phone user. Then widely used geodetic datum WSG-84 is discussed, which defines the standard size and shape of earth, origination and orientation of coordinate system. Further different technology available for location determination in a mobile network is detailed. Comparative analysis of these location determination technologies in terms of performance, implementation trends and cost factor is carried out. The general architecture of a location-enabled GSM/GPRS network is presented. The process of making location request and returning the appropriate information to the end user location-enabled GSM/GPRS network is explained. Further, end user application is presented which make use of location and show how location is a powerful enabling factor for any mobile application. Location based services are extensively used by companies such as NTT DoCoMo, Autodesk, Orange and Mapinfo is also mentioned. An overview of market perspective & recent development in location-based services is presented. Finally challenges and future trends for location based industry are also touched upon.

References

[1]. **AMITAVA MUKHERJEE, SOMPRAKASH BANDYOPADHYAY**, **DEBASHISH SHAH**. *Location Management And Routing In Mobile Wireless Network*., pp 11 , http:// www.artechhouse.com/

[2]. **AMITAVA MUKHERJEE, SOMPRAKASH BANDYOPADHYAY**, **DEBASHISH SHAH**. *Location Management And Routing In Mobile Wireless Network*. pp 13, http:// www.artechhouse.com/

[3] **ZHE LIU**, *A Java-Based Wireless Framework for Location-Based Services Applications, June 2002*. http://www.geomatics.ucalgary.ca/links/GradTheses.html

[4]. **JIANG BIN, ZIPF ALEXANDER**. *An Introduction to the Special Issue on LBS and GIS, Vol. 10, No. 2, December 2004*.

[5] **JOHN E. HARMON AND STEVEN J. ANDERSON**. *The Design and Implementation of Geographic Information Systems, 2003 , John Wiley & Sons*.

[6] **SHARAD CHANDRA AGRAWAL, SANDEEP AGRAWAL (CO-AUTHOR)**, *Location Based Services, September 2003*.

[7**] WANG FANGXIONG A, JIANG ZHIYONG**, *Research On A Distributed Architecture Of Mobile Gis Based On Wap*.

[8] **Wireless Developer Network**, *http://www.wirelessdevnet.com*

[9] **MANEESH PRASAD**, *Location Based Services, GIS Development, http://www.gisdevelopement.net*

[10] **MICHAEL JUNTAO YUAN**, *Enterprise J2ME: Developing Mobile Java Applications, Prentice Hall PTR , October 23, 2003*.

[11] **REENA SHUKLA**, *LBS, the ingredients and the alternatives, GIS Development*, http://www.gisdevelopement.net

[12] **JACEY-LYNN MINOI, DR. PETER GREEN, SYLVESTER ARNAB**, *Navigation Application with Mobile Telephony: Shortest-path, GIS Development*, http://www.gisdevelopement.net

[13] **ANAND RAMAMOORTHY**, *New approaches in GPS based location systems*

[14] **P. M. UDANI, R. K. GOEL**, *GPS enabled mobile GIS services*

[15] **YONG HE, FULIN BIAN, XICHUN WANG**, *The Research on Spatial Process Modeling in GIS*, http://www.gisdevelopment.net

[16] **PLEWE, B**., *GIS Online: Information, Mapping, and the Internet. Onward Press, 1997*

[17]. GSM ASSOCIATION, 2004, http://www.gsmworld.com

[18]. INTERNET GIS, http://map.sdsu.edu/gisbook/ch1.htm

[19]. OGC, Open GIS Consortium. http://www.opengis.org

[20]. ESRI, GIS and Mapping Software, 2006. http://www.esri.com

[21]. GIS Development, http://www.gisdeveloment.net

[22]. The Mobile Information Device Profile (MIDP), http://java.sun.com/products/midp

[23]. The J2ME Web Services API. http://www.jcp.org/en/jsr/detail?id=172

[24]. http://www.microsoft.com

Appendix A - Source Code Listing

Listing 4.1. The `ConfirmTicketUI` class represents the screen for confirming the ticket purchase

```
package com.sun.j2me.blueprints.smartticket.client.midp.ui
public class ConfirmTicketUI extends Form
           implements CommandListener, ItemCommandListener {
  private UIController uiController;
  private Command cancelCommand;
  private Command confirmCommand;
  private StringItem theater, movie, showTimeStr, seatsStr;
  private StringItem cost, totalCost, placeOrderBtn;
  public ConfirmTicketUI(UIController uiController) {
    super(uiController.getString(UIConstants.CONFIRM_TITLE));
    this.uiController = uiController;
    createItems();
    append(theater); append(movie); append(showTimeStr);
    append(seatsStr); append(cost); append(totalCost);
    append(placeOrderBtn);
    confirmCommand =
       new Command(uiController.getString(UIConstants.CONFIRM),
                    Command.OK, 5);
    cancelCommand =
        new Command(uiController.getString(UIConstants.CANCEL),
                    Command.EXIT, 5);
    addCommand(confirmCommand);
    addCommand(cancelCommand);
    setCommandListener(this);
    placeOrderBtn.setDefaultCommand(confirmCommand);
    placeOrderBtn.setItemCommandListener(this);
  }
  public void init(String theaterName, String movieName,
                          int[] showTime, Seat[] seats) {
    // Set the display strings to the correct values
  }
  // Command callback for UI events for text button
"placeOrderBtn"
  public void commandAction(Command command, Item item) {
    if (command == confirmCommand) {
      uiController.purchaseRequested();
    }
  }
  // Command callback for UI events on the command buttons
  public void commandAction(Command command, Displayable
displayable) {
    if (command == cancelCommand) {
        uiController.mainMenuRequested();
    } else if (command == confirmCommand) {
        uiController.purchaseRequested();
    }
  }
}
```

Listing 4.2. Process the `purchaseTickets` action in the `UIController` class in the controller layer

```
package com.sun.j2me.blueprints.smartticket.client.midp.ui;
public class UIController {
  // references to all UI classes
```

```
  public UIController(MIDlet midlet, ModelFacade model) {
    this.display = Display.getDisplay(midlet);
    this.model = model;
  }
    public void purchaseRequested() {
    runWithProgress(
       new EventDispatcher(EventIds.EVENT_ID_PURCHASEREQUESTED,
                           mainMenuUI),
       getString(UIConstants.PROCESSING), false);
  }
  class EventDispatcher extends Thread {
    private int taskId;
    private Displayable fallbackUI;
    EventDispatcher(int taskId, Displayable fallbackUI) {
      this.taskId = taskId;
      this.fallbackUI = fallbackUI;
      return;
    }
    public void run() {
      try {
        switch (taskId) {
          case EventIds.EVENT_ID_PURCHASEREQUESTED: {
            model.purchaseTickets(reservation);
            purchaseCompleteUI.init(reservation.getId(),
                                   selectedTheater.getName(),
                                   selectedMovie.getTitle(),
                                   selectedShowTime);
            display.setCurrent(purchaseCompleteUI);
            break;
          }
        }
      } catch (Exception exception) {
        // handle exceptions
      }
    } // end of run() method
  } // end of the EventDispatcher class
}
```

Listing 4.3. The `ModelFacade` class

```
package com.sun.j2me.blueprints.smartticket.client.midp.model;
public class ModelFacade {
  private SynchronizationAgent syncAgent;
  private RemoteModelProxy remoteModel;
  private LocalModel localModel;
  public Reservation reserveSeats(String theaterKey,
        String movieKey, int[] showTime, Seat[] seats)
                        throws ApplicationException {
    try {
      return remoteModel.reserveSeats(theaterKey,
                         movieKey, showTime, seats);
    } catch (ModelException me) {
   }
  }
  public void purchaseTickets(Reservation reservation)
                         throws ApplicationException {
    try {
      remoteModel.purchaseTickets(reservation.getId());
      // Purchased movies are eligible for rating.
      localModel.addMovieRating(
        new MovieRating(
```

```
          remoteModel.getMovie(reservation.getMovieId()),
                    reservation.getShowTime()));
    } catch (ModelException me) {
  }
  return;
 }
 public void synchronizeMovieRatings(
                    int conflictResolutionStrategyId)
                          throws ApplicationException {
    try {
syncAgent.synchronizeMovieRatings(conflictResolutionStrategyId)
;
    return;
    } catch (ModelException me) {
   }
  }
}
```

Listing 4.4. The gateway servlet `SmartTicketServlet`

```
package com.sun.j2me.blueprints.smartticket.server.web.midp;
public class SmartTicketServlet extends HttpServlet {
public static final String SESSION_ATTRIBUTE_SMART_TICKET_BD =
"com.sun.j2me.blueprints.smartticket.server.web.midp.SmartTicke
tBD";
 protected void doPost(HttpServletRequest request,
                        HttpServletResponse response)
                        throws ServletException, IOException {
     HttpSession session = request.getSession(true);
        SmartTicketBD smartTicketBD =
(SmartTicketBD)
session.getAttribute(SESSION_ATTRIBUTE_SMART_TICKET_BD);
     // Calls handleCall() method and encode the URL for
     // session tracking
  }
  public int handleCall(SmartTicketBD smartTicketBD,
InputStream in, OutputStream out) throws IOException,
ApplicationException {
    // Identifies the requested action method
// Execute the method through a list of switch – case satements
    switch (method) {
    case MessageConstants.OPERATION_GET_MOVIE:
         getMovie(smartTicketBD, call, successfulResult);
    break;
    }
  }
}
```

Listing 4.5. The business delegate class `SmartTicketBD`

```
package com.sun.j2me.blueprints.smartticket.server.web.midp;
public class SmartTicketBD implements RemoteModel {
public static final String EJB_REF_FACADE =
"ejb/SmartTicketFacade";
  private SmartTicketFacadeLocal facade;
  private ServletContext servletContext = null;
  public SmartTicketBD(ServletContext servletContext)
        throws ApplicationException {
    this.servletContext = servletContext;
    try {
      Context context =
         (Context) new nitialContext().lookup("java:comp/env");
```

```
      facade =
((SmartTicketFacadeLocalHome)context.lookup(EJB_REF_FACADE)).cr
eate();
      return;
    } catch (Exception e) {
      throw new ApplicationException(e);
    }
  }
  public Movie getMovie(String movieKey)
            throws ModelException, ApplicationException {
    try {
      MovieLocal movieLocal = facade.getMovie(movieKey);
      Movie movie = new Movie(movieLocal.getId(),
                              movieLocal.getTitle(),
                              movieLocal.getSummary(),
                              movieLocal.getRating());
      return movie;
    } catch (SmartTicketFacadeException stfe) {
      throw new
ModelException(ModelException.CAUSE_MOVIE_NOT_FOUND);
    } catch (Exception e) {
      throw new ApplicationException(e);
    }
  }
 // Other action methods in RemoteModel interface
}
```

Listing 4.6. The facade session bean **`SmartTicketFacadeBean`**

```
package com.sun.j2me.blueprints.smartticket.server.ejb;
public class SmartTicketFacadeBean implements SessionBean {
public void ejbCreate() throws CreateException {
Context context = (Context) new
InitialContext().lookup("java:comp/env");
    ticketingHome =
        (TicketingLocalHome) context.lookup(EJB_REF_TICKETING);
    synchronizingHome = (SynchronizingLocalHome)
context.lookup(EJB_REF_SYNCHRONIZING);
}
  public MovieLocal getMovie(String movieId)
          throws SmartTicketFacadeException {
    try {
      return movieHome.findByPrimaryKey(movieId);
    } catch (FinderException fe) {
      throw new SmartTicketFacadeException("No matching
movie.");
    }
  }
  public void purchaseTickets(String reservationId)
            throws SmartTicketFacadeException {
    if (ticketing != null) {
      ticketing.purchaseTickets(reservationId);
      return;
    }
    throw new SmartTicketFacadeException("User not logged
in.");

  }
  public MovieRatingData[] synchronizeMovieRatings(
```

```
                      MovieRatingData[] movieRatings,
                      int conflictResolutionStrategyId)
                      throws SmartTicketFacadeException {
    if (synchronizing != null) {
      return
synchronizing.synchronizeMovieRatings(movieRatings,
                    conflictResolutionStrategyId);
    }
    throw new SmartTicketFacadeException("User not logged
in.");
  }
}
```

Listing 4.7. The `RemoteModelRequestHandler` class

```
public interface RequestHandler {
    RequestHandler getNextHandler();
    void init() throws ApplicationException;
    void destroy() throws ApplicationException;
}
abstract public class RemoteModelRequestHandler
           implements RequestHandler, RemoteModel {
  private RemoteModelRequestHandler nextHandler;
  private Preferences preferences;
  protected static ProgressObserver progressObserver;
  public RemoteModelRequestHandler(
      RemoteModelRequestHandler nextHandler) {
    this.nextHandler = nextHandler;
  }
  public RequestHandler getNextHandler() {
    return nextHandler;
  }
  public void init() throws ApplicationException {
    if (nextHandler != null) {
      nextHandler.init();
    }
    return;
  }
  public void destroy() throws ApplicationException {
    if (nextHandler != null) {
      nextHandler.destroy();
    }
    return;
  }
  public void login(String userName, String password)
            throws ModelException, ApplicationException {
    getRemoteModelRequestHandler().login(userName, password);
    return;
  }
  public void createAccount(AccountInfo accountInfo)
            throws ModelException, ApplicationException {
    getRemoteModelRequestHandler().createAccount(accountInfo);
    return;
  }
  // Other action methods declared in RemoteModel
}
```

Listing 4.8. Assemble the handler chain in class `RemoteModelProxy`

```
public class RemoteModelProxy extends ModelObjectLoader
                                 implements RemoteModel {
```

```
  private RemoteModelRequestHandler requestHandlerChain;
  private Preferences preferences = null;
  private Hashtable movies = new Hashtable();
  public RemoteModelProxy(String serviceURL)
        throws ApplicationException {
    requestHandlerChain = new RMSCacheHandler(
            new HTTPCommunicationHandler(null, serviceURL));
    return;
  }
 // get a movie from the chain of handlers
  public Movie getMovie(String movieKey)
            throws ModelException, ApplicationException {
    Movie movie = (Movie) movies.get(movieKey);
    if (movie == null) {
      movie = requestHandlerChain.getMovie(movieKey);
      movies.put(movieKey, movie);
    }
    return movie;
  }
  // Other action methods etc.
}
```

Listing 4.9. The `getMovie()` method in `RMSCacheHandler`

```
public class RMSCacheHandler extends RemoteModelRequestHandler
{
  public Movie getMovie(String movieKey)
            throws ModelException, ApplicationException {
    IndexEntry indexEntry = rmsAdapter.getIndexEntry(movieKey,
                IndexEntry.TYPE_MOVIE, IndexEntry.MODE_ANY);
    if (indexEntry != null) {
      return rmsAdapter.loadMovie(indexEntry.getRecordId());
    }
    return super.getMovie(movieKey);
  }
}
```

Listing 4.10. The `getMovie()` method in `HTTPCommunicationHandler`

```
public class HTTPCommunicationHandler
            extends RemoteModelRequestHandler {
  public Movie getMovie(String movieKey)
          throws ModelException, ApplicationException {
    HttpConnection connection = null;
    DataOutputStream outputStream = null;
    DataInputStream inputStream = null;
    try {
      connection = openConnection();
      updateProgress();
      outputStream = openConnectionOutputStream(connection);
  outputStream.writeByte(MessageConstants.OPERATION_GET_MOVIE);
      outputStream.writeUTF(movieKey);
      outputStream.close();
      updateProgress();
      inputStream = openConnectionInputStream(connection);
      Movie movie = Movie.deserialize(inputStream);
      updateProgress();
      return movie;
    } catch (IOException ioe) {
      throw new
  ApplicationException(ErrorMessageCodes.ERROR_CANNOT_CONNECT);
  }
```

```
   finally {
      closeConnection(connection, outputStream, inputStream);
    }
  }
}
```

Listing 4.11. The RPC action codes in `MessageConstants`

```
package com.sun.j2me.blueprints.smartticket.shared.midp;
  public final class MessageConstants {
  public static final byte OPERATION_LOGIN_USER = 0;
  public static final byte OPERATION_CREATE_ACCOUNT = 1;
  public static final byte OPERATION_UPDATE_ACCOUNT = 2;
  public static final byte OPERATION_GET_THEATERS = 3;
  public static final byte OPERATION_GET_THEATER_SCHEDULE = 4;
  public static final byte OPERATION_GET_MOVIE = 5;
  public static final byte OPERATION_GET_MOVIE_POSTER = 6;
  public static final byte OPERATION_GET_MOVIE_SHOWTIMES = 7;
  public static final byte OPERATION_GET_SEATING_PLAN = 8;
  public static final byte OPERATION_RESERVE_SEATS = 9;
  public static final byte OPERATION_PURCHASE_TICKETS = 10;
public static final byte OPERATION_CANCEL_SEAT_RESERVATION =11;
  public static final byte OPERATION_GET_LOCALES = 12;
  public static final byte OPERATION_GET_RESOURCE_BUNDLE = 13;
public static final byte OPERATION_INITIATE_SYNCHRONIZATION=1
4;
public static final byte OPERATION_SYNCHRONIZE_MOVIE_RATINGS =
15;
  public static final byte OPERATION_COMMIT_MOVIE_RATINGS = 16;
  public static final byte ERROR_NONE = 0;
  public static final byte ERROR_UNKNOWN_OPERATION = 1;
  public static final byte ERROR_SERVER_ERROR = 2;
  public static final byte ERROR_MODEL_EXCEPTION = 3;
  public static final byte ERROR_REQUEST_FORMAT = 4;
  private MessageConstants() {}
}
```

Listing 4.12. The `HTTPCommunicationHandler` *class generates the RPC request in the handler chain*

```
package com.sun.j2me.blueprints.smartticket.client.midp.model;
public class HTTPCommunicationHandler
            extends RemoteModelRequestHandler {
public Theater[] getTheaters(String zipCode)
            throws ModelException, ApplicationException {
    HttpConnection connection = null;
    DataOutputStream outputStream = null;
    DataInputStream inputStream = null;
    try {
      connection = openConnection();
      updateProgress();
      outputStream = openConnectionOutputStream(connection);
outputStream.writeByte(MessageConstants.OPERATION_GET_THEATERS)
;
      outputStream.writeUTF(zipCode);
      outputStream.close();
      updateProgress();
      inputStream = openConnectionInputStream(connection);
      // The first number in the response stream indicates
      // the number of theater objects to follow.
      Theater[] theaters = new Theater[inputStream.readInt()];
```

```
// Iterate to unmarshal all theater objects in the response.
      for (int i = 0; i < theaters.length; i++) {
        theaters[i] = Theater.deserialize(inputStream);
      }
      updateProgress();
      return theaters;
    } catch (IOException ioe) {
      throw new
  ApplicationException(ErrorMessageCodes.ERROR_CANNOT_CONNECT);
    } finally {
        closeConnection(connection, outputStream, inputStream);
    }
  }
}
```

Listing 4.13. The Theater class in the J2ME model layer unmarshals the RPC response

```
package com.sun.j2me.blueprints.smartticket.shared.midp.model;
public class Theater {
  private String primaryKey;
  private String name;
  private String address;
  private String zipCode;
  public static Theater deserialize(DataInputStream dataStream)
throws ApplicationException {
    try {
      Theater theater = new Theater();
      theater.zipCode = dataStream.readUTF();
      theater.primaryKey = dataStream.readUTF();
      theater.name = dataStream.readUTF();
      theater.address = dataStream.readUTF();
      return theater;
    } catch (IOException ioe) {
      throw new ApplicationException(ioe);
    }
  }
}
```

Listing 4.14. The `runWithProgress()` method in the `UIController` class

```
public class UIController {
  public void chooseMovieRequested() {
    runWithProgress(
      new EventDispatcher(
        EventIds.EVENT_ID_CHOOSEMOVIEREQUESTED, mainMenuUI),
        getString(UIConstants.PROCESSING), false);
  }
  public void runWithProgress(Thread thread, String title,
                              boolean stoppable) {
    progressObserverUI.init(title, stoppable);
    getDisplay().setCurrent(progressObserverUI);
    thread.start();
  }
  class EventDispatcher extends Thread {
    public void run() {
    }
  }
}
```

Listing 4.15. The `ProgressObserverUI` class

```
public class ProgressObserverUI extends Form
     implements ProgressObserver, CommandListener {
  private UIController uiController;
  private static final int GAUGE_MAX = 8;
  private static final int GAUGE_LEVELS = 4;
  int current = 0;
  Gauge gauge;
  Command stopCommand;
  boolean stoppable;
  boolean stopped;
  public ProgressObserverUI(UIController uiController) {
      super("");
      gauge = new Gauge("", false, GAUGE_MAX, 0);
      stopCommand =
         new Command(uiController.getString(UIConstants.STOP),
                                Command.STOP, 10);
      append(gauge);
      setCommandListener(this);
  }
  public void init(String note, boolean stoppable) {
    gauge.setValue(0);
    setNote(note);
    setStoppable(stoppable);
    stopped = false;
  }
  public void setNote(String note) {
    setTitle(note);
  }
  public boolean isStoppable() {
    return stoppable;
  }
  public void setStoppable(boolean stoppable) {
    this.stoppable = stoppable;
    if (stoppable) {
      addCommand(stopCommand);
    } else {
      removeCommand(stopCommand);
    }
  }
  // Indicates whether the user has stopped the progress.
  // This message should be called before calling update.
  public boolean isStopped() {
    return stopped;
  }
  public void updateProgress() {
    current = (current + 1) % GAUGE_LEVELS;
    gauge.setValue(current * GAUGE_MAX / GAUGE_LEVELS);
  }
  public void commandAction(Command c, Displayable d) {
    if (c == stopCommand) {
      stopped = true;
    }
  }
}
```