CHAPTER 1

INTRODUCTION

The domain of information technology is growing more rapidly than any other domain in the world. Computers are used in almost all applications. Mobile technology has undoubtedly been replacing traditional computing technology. The use of mobile services and applications is quite common due to the evolution of connectivity and mobility.

Mobile computing is the process of performing computation with mobile devices. Mobile computing offers mobility with computing power. It provides distributed computations on a wide range of devices, systems, and networks which are mobile, synchronized and interconnected via mobile communication standards and protocols. The features of mobile computing systems such as their wireless network connectivity, their small size, the mobile nature of their use, their power sources and their functionalities make them different from the stationary systems. Because of these aspects, mobile computing applications differ from the applications written for stationary computing system. The unique features of the mobile computing systems are remarkably suited to the mobile users.

After the invention of first transportable voice radios of 1920s, radio technology matured in the subsequent decades with broadcast radio and television. In 1940s cellular technology was conceived with the ability to divide radio frequency service areas into cells in order to reduce interference and increase capacity. This is the basis for today’s wide area, voice and wireless local area networking technologies. In 1956, AT & T Bell system began to offer manual radio telephone service. In 1964,
Bell system introduced Improved Mobile Telephone Service (IMTS). In 1970s, the Federal Communication Commission (FCC) allotted spectrum space for mobile telephones. During the period from 1987 to 1995, a new air interface protocols such as TDMA (Time Division Multiple Access), CDMA (Code Division Multiple Access) were introduced. During 1992, GSM now called Global System for Mobile communication was standardized. In the subsequent years, IEEE 802.11, UMTS (Universal Mobile Telecommunication System), Bluetooth, GPRS (General Packet Radio Service), and IMT-d-2000 (International Mobile Telecommunications) were developed.

The recent advancements in portable devices, wireless network technology and satellite services have led to the development of mobile computing environment [1]. This environment will give the user the information accessing capability regardless of the location of the user or the information. With the evolution of PCS (Personal Communication System) and GSM (Global System for Mobile communication), advanced wireless communication services are being offered to the mobile users. Mobile Database System is a distributed client/server system based on PCS or GSM in which clients can move around freely while performing their data processing activities in connected or disconnected mode [2].

Mobile users can use their cellular phone to check e-mail and browse internet. Travellers with portable computers can surf the internet from airports, railway stations and other public locations. Tourists can use Global Positioning System (GPS) terminals installed inside rental cars to locate driving maps and tourist attractions. Researchers can exchange files and other information by connecting portable computers via wireless LANs while attending conferences. At home, users can synchronize data and transfer files between portable devices and desktops. Mobile
devices are not only getting smaller, cheaper, more convenient and more powerful, they also run more applications and network services, commonly fuelling the explosive growth of mobile computing equipment market.

1.1 Motivation

Mobile devices such as cell phones, PDA (Personal Digital Assistant), MP3 music players etc., have become popular devices for information retrieval by users. Using these portable computing devices with wireless interface, the desired information from the database can be retrieved from anywhere at any point of time. However, in this environment, the information can flow only from server to users. This limitation does not allow users to query or manipulate the database. Consequently users have to be satisfied with what server sends them which may not always be accurate or up to date. In database terminology, these systems are not capable of managing transactional activities. Recent research works focus on developing an information management system on a mobile computing environment that provides transaction management and database management solutions from anywhere and anytime.

Mobile users can access data in the fixed network via transactions. However, a transaction definition in this environment varies because of centralized and distributed environment. In addition, limitations of wireless and nomadic environment pose a new challenge to implementing efficient transaction processing using classical transaction models. Disconnection is the major obstacle in the mobile environment. Mobility and portability bring forth new challenges to Mobile Database management and distributed computing [3]. In a mobile computing environment, links between nodes in a network change dynamically. Thus we cannot rely on a fixed network
structure. A single site cannot play the role of a coordinator as in a centralized system. The Mobile Host and Fixed Host also differ in computational power and memory. Limited battery power of a Mobile Host can cause site failures. The above limitations lead to the identification of the proposed research work.

1.2 Mobile Computing Environment

Mobile computing environment consists of Fixed Hosts (FHs), Mobile Hosts (MHs) and Base stations or Mobile Support Stations (MSSs). MH is connected to the Fixed network through MSS via wireless channels. The Geographical area covered by a MSS is called a cell. Mobile Hosts are portable computers which move around in a cell. When a MH enters into a new cell hand-off or hand-over takes place. MH communicates only with the MSS responsible for its cell. Transactions and data management functions are done using the data base servers installed at MSS.

1.2.1 Mobile Host

Mobile Host is also known as Mobile Unit (MU) or Mobile Station (MS). It is a wireless device which consists of an antenna, transceiver and user interface. The antenna is used to capture the signal. The transceiver receives and sends the signal. The user interface supports user interaction by means of graphics display and texts. It also includes audio interface to handle voice conversation.

1.2.2 Types of Mobility

Mobile computing framework consists of wired and wireless components and mobile users. The wireless component is responsible for implementing terminal mobility and personal mobility which are distinguished as discussed below.
• **Terminal mobility** allows a Mobile Unit to access the desired services from any location while in motion or stationary irrespective of who is carrying the unit. It is the responsibility of wireless network to identify the communication device. In terminal mobility, the connection is established between two points and not between two persons calling each other.

• **Personal mobility** allows a mobile user to establish a connection with another user using a common device available in his location. He does not need to carry any communication device with him. Before getting connection, a user is required to verify his identity using an identification scheme.

1.2.3 Modes of Operation of a Mobile Host

In mobile computing, an MH can have four modes of operations [4, 64] due to the restriction on their resources: strong connection mode, disconnected mode, weak connection mode, and doze mode. Based on the bandwidth availability an MH can go from strong to weak connectivity.

• **Strong connection mode:** In a strong connection mode, the mobile computer is continually connected to a server. In this mode hand-over protocols would be required if a cellular structure is used and when mobiles move from one cell to a different cell. The hand-over protocol may involve a new communication link between the mobile unit and the new server, and the process of saving and transferring the states from the old to the new server. The communication hand-over to a new cell should be transparent to both users and applications not specifically involved with the hand-over process.

• **Disconnected mode:** An MH can change to a disconnected mode in order to minimize connection expenses or because no connection is possible. One
possible solution to deal with such disconnections would be the use of a proxy for the mobile computer [5]. This would ensure the continuity of the running of the query even when the mobile component is disconnected, and the mobile unit may request an update from its proxy when it is reconnected. In addition, mobile computers may voluntarily move into this disconnected mode when idle or low on battery power, to free up bandwidth resources and extend battery life [108]. While operating in disconnected mode, any applications that had used the communication link before the disconnection would be required to save their current communication state, and where possible continue with its other processes. Upon reconnection and depending on the saved communication states, applications may resume transmission or reception, or retransmit a request to begin the communication once again.

- **Weak connection mode:** In this mode, the mobile unit is connected to the rest of the network through low or intermittent bandwidth. The degree to which the communication bandwidth is available may vary from marginally less than full bandwidth availability to almost no connectivity. Partial or weak connection mode may occur when the mobile units are in shadow areas within or on the edge of a cell, where reception is poor. The partial-connection protocol would then be required to allow the mobile client to limit its communications to the network. Applications that use the communication link may then experience longer transmission time and may be required to extend timeouts to anticipate a lengthy response time.

- **Doze mode:** MH can go to energy conservation mode called doze mode. In this mode, clock speed is reduced and no user computation is performed. The CPU of the MH will be working at a lower rate. It can examine messages from
other MHs. The communication device can receive signals. So the MH can be awakened by a message from other MHs [86].

1.2.4 Mobile Networks

In the realm of telecommunication there are numerous ways and mediums through which effective communication and information exchange may be achieved. The nature of the data that is to be transmitted may also vary enormously and thus in order to accommodate these variations, there are diverse telecommunication networks that range from fixed networks to mobile networks, data and voice networks and/or analog and digital networks.

As compared to fixed networks, in which user’s equipment is connected to the telecom network using some type of physical medium, mobile network infrastructure varies in the last mile connection to the user. In mobile networks, the communication is typically achieved via radio transmission techniques and technology. As the service is mostly based on a wireless technology, the user has a huge advantage of added and improved mobility in his local area and/or wider area. However, this extent of mobility depends greatly on the nature of the mobile network.

For long distance communication, a commonly employed strategy is to use wireless communication embedded or as a part of a fixed network in order to extend and enhance the communication services to the user. An added advantage of this technique is that the need to install cumbersome physical wires or bulky cables is eliminated. Moreover, setting up a successful and effective mobile network infrastructure is no easy task as it involves overcoming a large number of problems which may arise due to signal degradation, noise and other factors. With the ever-growing trend towards smaller, sleeker and sophisticated equipments, the chief
problem being faced by telecom engineers is the size of the radio equipment that the user is expected to carry around.

As the primary objective of the research is concerned with mobile database transactions, only a brief overview of mobile networks and technology is presented in the following sections.

**Analog Mobile Networks**

When mobile phone networks were first introduced in the late 1970s, they comprised of a system that was based on analog communication and an inferior quality of service. As analog networks did not support data transmission, the user had to buy modems in order to achieve some form of data transfer. Initially this was limited to send faxes and similar data types. Moreover the modems were expensive and the data transfer rate was quite slow. The initial mobile phones were often bulky and had poor battery life.

**Digital Mobile Networks**

During the 1980s, development of the second generation of communication technology began. These Second-generation (2G) mobile phone systems were based on digital means of communication and hence the quality of service generally improved. Many advances were made in all aspects of mobile phone technology. The handsets became considerably smaller, lighter weight and thus easier to carry. Furthermore the battery life also improved and mobile communication became more reliable. The next phase in the evolution is Third Generation or 3G wireless technology which uses Spread Spectrum techniques for media access and encoding.
Mobile Ad-hoc Networks (MANET)

Recent advancements such as Bluetooth introduced a new type of wireless system known as Mobile Ad-hoc Networks (MANET). It is a network composed of mobile nodes mainly characterized by the absence of any centralized coordination or fixed infrastructure, which makes any node in the network act as a potential router. MANETs are also characterized by a dynamic, random and rapidly changing topology. This makes the classical routing algorithms fail to perform correctly since they are not robust enough to accommodate such a changing environment. Consequently, more and more research is being conducted to find optimal routing algorithms that would be able to accommodate for such networks. Mobile ad-hoc networks are suitable for emergency situations like natural or human-induced disasters, military conflicts, emergency medical situations etc.

MANETs are self-forming, self-maintained and self-healing, allowing for extreme network flexibility. While MANETs can be completely self-contained, they can also be tied to an IP-based global or local network (e.g. Internet or private networks). These are referred to as Hybrid MANETs.

In MANETs communication between mobile nodes always requires routing over multi-hop paths. Since no infrastructure exists and node mobility may cause frequent link failure, it is a great challenge to design an effective and adaptive routing protocol. Many restrictions should be well-considered, such as limited power and bandwidth. Some of the routing protocols designed for MANETs are: Destination-Sequenced Distance Vector routing protocol (DSDV), Ad-hoc On Demand Distance Vector (AODV), Dynamic Source Routing (DSR), Temporarily Ordered Routing Algorithm (TORA) and Associative Based Routing (ABR).
Current trends in Wireless technology

Few innovative wireless technologies which evolved during the past few years dominating wireless networking paradigm are listed below.

- **Wi-Fi (Wireless Fidelity):** The first wireless LAN technology based on IEEE 802.11 is Wi-Fi. It not only covers office based LANs, but also home based LANs which create hot spots, i.e., an area around central antenna in which people can wirelessly share information or connect to the internet with a properly equipped laptop.

- **WiMax:** It is similar to Wi-Fi in that it also creates hot spots. It is based on IEEE 802.16 spec. The range of WiMax network is from few 10’s of kms to an entire metropolitan city.

- **Mobile-Fi:** It is similar to mobile version of WiMax. It is based on IEEE 802.20 spec.

- **ZigBee:** It is used to deal with low data rate over relatively short distances.

- **Ultrawide band:** It is used to handle the movements of massive files at high data rates over short distances.

Wired vs. Wireless networks

Mobile Hosts communicate with other hosts via wireless networks. Radio Frequencies (RF) are used to establish communication link. The telecommunication system has to perform a number of steps and go through a number of wired segments for setting up a wireless communication session between two users. Compared to wired networks, wireless networks are characterized by lower bandwidth, unstable and disconnections [6]. The characteristics of the wireless networks are described as follows:
• **Low band width:** Low bandwidth of a wireless network results in more overhead for transaction processing

• **Data transmission errors:** High error rates during data transmission necessitate the repeated retransmission of the same data packages which increases communication overhead.

• **Disconnection:** Disconnection is a major problem in wireless networks. The execution of transactions may be interrupted, delayed or aborted due to disconnections in communication. Mobile Host will not be able to share data with other hosts in the disconnected mode. If the mobile host holds vital shared data, it can block transaction processes on other hosts. Furthermore, the duration of a disconnected period of a mobile host is not always as planned, i.e., it can be longer than expected. Caching the needed data is essential to continue transaction execution while the mobile host is being disconnected from the database servers.

### 1.3 Mobile Applications and Services

Data and information through mobile computing services are required by all people regardless of the fact that they are mobile or not. Mobile users will include people like mobile executives, sales people and service engineers. Some of the mobile applications which can be grouped into different categories are given below.

- Vehicles
- Emergencies
- Business
- Mobile Banking
- Location dependent services
1.3.1 Vehicles

Mobility aware applications are useful while travelling in a vehicle. Some of them are:

(i) It allows us to receive music, news, road conditions, weather reports and other broadcast information via Digital Audio Broadcasting (DAB)

(ii) Universal Mobile Telecommunication System (UMTS) phone might be available offering voice and data connectivity.

(iii) GPS is used for getting the current position of the car.

(iv) Today, car driving in the same area builds a local ad-hoc network and helps to exchange information in emergency situations.

(v) In future, a car will also inform other cars about accident via the ad-hoc network to help them slow down in time, even before a driver can recognize an accident.

1.3.2 Emergencies

If there is a high quality wireless connection between a hospital and an ambulance, the vital information about injured persons can be sent to the hospital from the scene of an accident. Hence, it will help specialists to get ready with the instruments for the early diagnosis.

1.3.3 Business

Generally, a travelling salesman needs an instant access to the company’s database. He may have details in his laptop, but it won’t reflect the current situations. With wireless access, the laptop can be turned into a mobile office but efficient and
powerful synchronization mechanisms are needed to ensure data consistency. These settings can be achieved by using WLAN, GSM and UMTS etc.

1.3.4 Mobile Banking

Mobile banking is a term used for performing balance checks, account transactions, payments, credit applications and other banking transactions through a mobile device such as a mobile phone or Personal Digital Assistant (PDA). The earliest mobile banking services were offered over SMS. With the introduction of the first primitive smart phones with WAP support enabling the use of the mobile web in 1999, the European banks for the first time started mobile banking on this platform to their customers.

1.3.5 Location Dependent Services

Users can access services depending on their actual locations. For example, a user might query the local wireless network to find the whereabouts of a nearby restaurant, or his current geographical location. The local network itself might advertise such data which the user can access.

1.4 Challenges in Mobile Applications

The characteristics of computing in the mobile environment are high communication latency, intermittent wireless connectivity, limited battery life and changing client location. Latency is caused by the processes such as coding data for wireless transfer, and tracking and filtering wireless signals at the receiver which are unique to the wireless medium. Battery life is determined by the battery size. Intermittent connectivity may happen intentionally or unintentionally. Unintentional disconnections happen in areas wireless signals cannot reach, e.g., subway tunnels.
Intentional disconnections may be due to user intent, e.g., when the mobile device is powered down. Mobile applications should include these characteristics in their design since they have a strong impact on data management.

From a data management standpoint, mobile computing may be considered as a variation of distributed computing. Mobile databases can be distributed under two possible scenarios: (a) The entire database is distributed mainly among the wired components, possibly with full or partial replication. A base station or fixed host manages its own database with a DBMS-like functionality, with additional functionality for locating mobile units and additional query and transaction management features to meet the requirements of mobile environments. (b) The database is distributed among wired and wireless components. Data management responsibility is shared among base stations or fixed hosts and mobile units. Hence, the distributed data management issues can also be applied to mobile databases with some modifications.

The characteristics of mobile computing environment and distributed nature of mobile databases pose many challenges in designing mobile applications as discussed below.

- **Scalability:** The problem of scalability arises due to the latency involved in wireless communication. Because latency due to wireless communications increases the time to service each client request, the server can handle fewer clients. One way servers relieve this problem is by broadcasting data whenever possible. Broadcast takes advantage of a natural characteristic of radio communications, and it is scalable because a single broadcast of a data item can satisfy all outstanding requests for it. For example, instead of sending
weather information to all clients in a cell, a server can simply broadcast it periodically. Broadcast also reduces the load on the server, as clients do not have to maintain active connections to it.

- **Client mobility:** In the case of client mobility, servers must keep track of client locations in order to route messages efficiently to them. Also, client data should be stored in the network location that minimizes the traffic necessary to access it. Keeping data in a fixed location increases access latency if the client moves “far away” from it. Finally, the act of moving between cells must be transparent to the client. The server must be able to divert the shipment of data gracefully from one base station to another without the client noticing.

- **Data distribution and replication:** Data is unevenly distributed among the base stations and mobile units. The ability to replicate the data objects is essential in mobile computing to increase availability and performance. Mobility of users and services and its impact on data replication and migration will be one of the main technical problems that are to be resolved. In [49], caching of data in mobile hosts and the cost of maintaining consistency among replicated data copies are discussed.

- **Cache consistency:** Caching of frequently accessed data plays a vital role in mobile computing because of its ability to alleviate the performance and the availability limitations during weak-connections and disconnections. Caching is useful during frequent relocation and connection to different DBS. In wireless computing, caching of frequently accessed data items is an important technique that will reduce contention on the small bandwidth wireless network. Cache consistency is severely hampered by both the disconnection and the mobility of clients since a server may be unaware of the current
locations and the connection status of clients. The server can solve this problem by periodically broadcasting the actual data, invalidation report, or even control information such as lock tables or logs. In [86], a number of cache invalidation schemes have been proposed.

- **Transaction models:** Issues of fault tolerance and correctness of transactions are aggravated in the mobile environment. A mobile transaction is executed sequentially through several base stations and possibly on multiple data sets depending upon the movement of the mobile unit. Central coordination of transaction execution is lacking. Moreover, a mobile transaction is expected to be long-lived because of disconnection in mobile units. Hence, traditional ACID properties of transactions may need to be modified and new transaction models must be defined.

- **Query processing:** Awareness of where the data is located is important and affects the cost/benefit analysis of query processing. Query optimization is more complicated because of mobility and rapid resource changes of mobile units. The query response needs to be returned to mobile units that may be in transit or may cross cell boundaries yet must receive complete and correct query results. A query processing facility suitable for mobile database applications is presented in [120].

- **Recovery and fault tolerance:** The mobile database environment must deal with site, media, transaction, and communication failures. Site failure at an MH is frequently due to limited battery power. If an MH has a voluntary shutdown, it should not be treated as a failure. Transaction failures are more frequent during handoff when an MH crosses cells. MH failure causes a network partitioning and affects routing algorithms.
• **Mobile database design:** The global name resolution problem for handling queries is compounded because of mobility and frequent shutdown. Mobile database design must consider several issues of metadata management like the constant updating of location information.

• **Location based service:** The fact that clients in a mobile environment can change locations enables the possibility of answering queries in a way that is dependent on the current position of the client [123]. In mobile environment, the location of a mobile client can be regarded as data item whose value changes with every move. In the mobile computing, the location management is a data management problem. Important issues here are how to know the current position of the MH? Where to store the location information and who should be responsible for determining and updating of information. As clients move location-dependent cache information may become stale. Eviction techniques are important in this case. Furthermore, updating location dependent queries, then applying these (spatial) queries in order to refresh the cache also poses a problem. For example, consider an electronic valet application that can tell a user the location of the nearest restaurant. Clearly, “nearest” is relative the client’s current position and movement can invalidate any data that are currently cached. Upon movement, the client must efficiently invalidate parts of its cache and request updated data from the database.

• **Division of labour:** Certain characteristics of the mobile environment force a change in the division of labour in query processing. In some cases, the client must function independent of the server. However, what are the consequences of allowing full independent access to replicated data? The relationship between client capabilities and their consequences have yet to be developed.
• **Security**: Mobile data is less secure than that which is left at the fixed location. Proper techniques for managing and authorizing access to critical data become more important in this environment. Data are also more volatile, and techniques must be able to compensate for its loss.

### 1.5 Organization of the Thesis

This thesis comprises eight chapters. A chapter by chapter breakdown is as follows.

Chapter 1 starts with the arguments which give the motivation for taking up this research work. It discusses briefly about the mobile computing environment, mobile applications and challenges in mobile applications.

In chapter 2, an overview of transaction management is presented. First, it introduces the transaction concept and concurrency control mechanisms. This is followed by a brief discussion on distributed transactions.

Chapter 3 discusses the Mobile Database System with the focus on mobile transactions. The characteristics and issues of mobile database transactions are also discussed in this chapter.

Chapter 4 is meant for literature review in which a survey of existing traditional transaction models and mobile transaction models is carried out. It also discusses the existing cache invalidation techniques in mobile environment. Finally it discusses the transaction management approaches used in MANETs.

Chapter 5 presents the proposed architecture called “Agent Based Direct transaction Architecture (ADTA)” in which two schemes one based on Fixed Agents (FADTA) and another based on Mobile Agents (MADTA) are explained. The
response time analyses for the two schemes are carried out and the results are discussed.

Chapter 6 presents the proposed architecture called “Agent Based Multi-hop Transaction Architecture (AMTA)” in which two schemes one based on Fixed Agents (FAMTA) and another based on Mobile Agents (MAMTA) are explained. This architecture makes use of multi-hop wireless networks. The response time analyses for the two schemes are carried out and the results are discussed.

In Chapter 7 the performances of four proposed schemes are compared and the results are analysed.

Chapter 8 presents the conclusion of this research work and future enhancements.