Chapter 1

Introduction

Mobile communication technology has given its user the freedom to roam and yet be able to communicate with the remote users. Eventually, it has allowed the mobile users not only to have voice communication but also be connected with the entire world through text and videos using more sophisticated devices (PDAs, Laptops etc.). It has resulted in mobile computing platform that gives the flexibility to the roaming user to perform the computation on the remote systems. High speed Internet is the backbone for such systems. Users’ mobility, in such an environment, needs to be managed in an effective manner. To handle the user’s mobility, two solutions for the mobility management have been proposed [1]. One is for macro level of mobility called Mobile IP protocol and the other is for micro level of mobility called Cellular IP protocol. It has been observed that these two protocols are good enough for the mobility management. The researchers are thriving to improve upon the abilities of these two protocols in terms of better Quality of Service (QoS).

Mobile computing experiences two kinds of traffic; real-time traffic such as multimedia traffic and non-real-time traffic e.g. e-mail or paging. For real time traffic, data transmission should ensure that no disconnection, no packet loss, or no substantial delay occurs whereas non real-time traffic may tolerate these up to some extent. Resource reservation, e.g. bandwidth, in both types of networks (Mobile IP networks and Cellular IP networks) can improve the QoS especially for the real-time traffic that requires wider attention of the resources. Many other resources such as
buffer, routers etc. affect the performance of the system, and therefore the reservation of these may also help in improving the QoS of the system [2, 3].

This chapter briefs about mobile computing and the challenges in mobile computing. It also explores the research issues in these networks. To improve QoS, resource reservation has been identified to be the better technique. The chapter introduces the QoS in wireless networks and the techniques used for improving QoS in Cellular IP network. Finally, the organization of the thesis is listed.

1.1 Mobile Computing

Mobile computing is the term used to describe technologies that enable people to access network services at anyplace, anytime, and anywhere. It is the computing environment over the physical mobility. Mobile computing is also referred as ubiquitous computing and nomadic computing. Information access via a mobile device is plagued by low available bandwidth, poor connection maintenance, and addressing problems. Mobile computing system allows the user to perform a task, such as surfing the net, anywhere. Other synonyms for mobile computing is “Anywhere, Anytime Information”, “Global Service Portability”, etc. [4, 5].

1.1.1 Mobile Computing Functions

The computing environment is considered as mobile computing environment if it supports the following functions.

User Mobility

Mobility management is an important issue for any mobile computing environment. Mobile computing environment must support the mobility of the user from one physical location to another. The same service a user gets at home (in fixed) must be available while the user is on move.
Introduction

**Network Mobility**

Network mobility means that the user must be able to roam from one network to another without service disturbance. Mobile IP is an example of network mobility function in which the users can handoff between two networks.

**Service Mobility**

Service mobility implies the ability of the user to avail different services while on move. The user should be allowed to use Internet browser to surf the net on move, in addition to use other necessary services (e.g. computation, etc.)

1.1.2 Challenges in Mobile Computing

Though the mobile computing environment offers many services, it has many challenges also. These challenges are essentially included in various layers. Mobility, in mobile computing environment has the main effect on many layers of the protocol including data link layer (packet loss due to wireless medium), network layer (addressing and routing issues), and transport layer (end-to-end QoS). Figure 1.1 shows the associated challenges in mobile computing environment in each layer of the protocol. The technical challenges that mobile computing must overcome are mobility, wireless communication and portability [4, 5].

1.1.2.1 Mobility

Mobility is the ability of a mobile user to change its physical location while it is connected to the network. Mobility causes many problems in the network such as packet loss due to the route change. Some of the challenges inducted by mobility are:

**Address migration**

When a user hands off the session to another network, it gets different address. It registers with the new address in the new network. This address will be stored (cached) with long expiration time.
Chapter 1

Introduction

Location-dependent information

Another challenge for mobility management is the information that depends on location of the user. This information must be maintained and a mechanism for obtaining them must be provided.

Privacy

When a user moves from one physical location to another physical location, it needs to keep its position private (unknown) for the other users in the network. Privacy means, some sensitive information for a particular user should be hidden from the others. The challenge of mobile computing is to allow more flexible access to this information without violating privacy.

<table>
<thead>
<tr>
<th>Application Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource Discovery</td>
</tr>
<tr>
<td>Profile Management</td>
</tr>
<tr>
<td>New Multimedia Applications</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transport Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow and Congestion Control</td>
</tr>
<tr>
<td>End-to-end QoS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addressing and Routing</td>
</tr>
<tr>
<td>Location Management</td>
</tr>
<tr>
<td>QoS Control Handoff</td>
</tr>
<tr>
<td>Authentication</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Physical/Link Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal Modulation</td>
</tr>
<tr>
<td>Encryption/Compression</td>
</tr>
<tr>
<td>Power Control</td>
</tr>
<tr>
<td>Channel Access</td>
</tr>
</tbody>
</table>

Fig. 1.1 Challenges in Mobile Computing
1.1.2.2 Wireless Communications

It's obvious that wireless communication is necessity at present. The use of wireless communication implies several design challenges. Some of these challenges are as follows [6].

**Disconnection**

Disconnection refers to the interruption of the services for the users in the network. Network failure and the loss of connectivity are big challenges in any mobile computing system due to the sensitive nature of wireless users towards disconnection. Designers of the wireless communication systems must decide to use the available resources either to prevent disconnection or to enable the systems to face the disconnectivity. Autonomous system is the system which is able to tolerate disconnection [6].

**Low Bandwidth**

Bandwidth consumption is the most important issue in any wireless system due to low bandwidth availability. Therefore, mobile computing designs need to reflect a great concern for bandwidth consumption. Network capacity can be improved by installing more wireless cells in order to serve more users. There are two approaches to achieve this purpose: one is to overlap cells on different wavelengths. Second is to reduce transmission ranges so that more cells fit in a given area. First approach is not scalable because it needs software for bandwidth allocation. The second one is preferred as it reduces power requirements also.

**Heterogeneous networks**

Unlike stationary computers connected through wires, mobile computers encounter more heterogeneous network connections in different ways such as moving from a range of one base station to the range of another in the same network or moving from one network to another. The heterogeneity makes mobile networks more complex than traditional networks.
Chapter 1

Introduction

1.1.2.3 Portability

Mobile elements must be light and small to be easily carried around. Such considerations in conjunction with a given cost and level of technology will keep mobile elements having less resources than static elements including memory, screen size and disk capacity. This justifies the asymmetry between static and mobile elements. Mobile elements rely on battery and even with the advances in battery technology; this concern does not cease to exist. Concern for power consumption must span various levels in hardware and software design. Mobile elements are easier to be accidentally damaged, stolen, or lost. Thus, they are less secure and reliable than static elements. Some of the challenges faced for portability are as follows [7].

Risks to data

The risk resulting from making the computer mobile is that the data will be prone to loss, theft or physically damaged. Minimizing the essential data carried away in these computers reduces these risks.

Small storage capacity

Due to the size limitation in the portable computers, the storage size is limited and rather small. Desktops (fixed computers) provide bigger storage spaces than the portable ones. In a mobile computer, the disk drives consume more power than the disk drives in the static devices.

1.1.3 Mobile IP and Cellular IP

The Internet Protocol (IP) represents today’s standard in Internet networking. The IP technology has recently been improved with the introduction of new services such as Voice over IP (VoIP) and QoS enhancements which represent key features of wireless networks. Thus, the ability to handle data and voice services and mobility management for the users make extreme attraction for finding good solutions based on extensions of the standard IP protocol. Mobile IP and Cellular IP are the two protocols suggested to handle the mobility of users in micro and macro levels [8].
Chapter 1

Introduction

Mobile IP is the solution for user's mobility management at the macro level. Mobile IP protocol allows location-independent routing of IP datagrams on the Internet. Each mobile node is identified by its home address. While away from its home network, a mobile node is associated with a Care-Of-Address (COA) which identifies its current location and its home address is associated with the local endpoint of a tunnel to its Home Agent (HA). Mobile IP specifies how a mobile node registers with its home agent and how the home agent routes datagrams to the mobile node through the tunnel. Mobile IP provides an efficient, scalable mechanism for roaming within the Internet. Using Mobile IP, nodes may change their point-of-attachment to the Internet without changing their home IP address. This allows them to maintain transport and higher-layer connections while roaming. Node mobility is realized without the need to propagate host-specific routes throughout the Internet routing fabric.

Cellular IP protocol is used for mobility management of mobile users across the cells within one network (micro level). Whenever the mobile host moves to a new network managed by a different foreign agent, the dynamic Care-Of-Address will change. This change needs to be communicated to the home agent. This process works for slowly moving hosts. For high speed mobile hosts, the rate of update of the address needs to match the rate of change of address. Otherwise, packets will be forwarded to the wrong (old) address. Mobile IP fails to update the address properly for high speed mobility. Cellular IP has been designed to address this issue. In Cellular IP, none of the nodes know the exact location of a mobile host. Packets addressed to a mobile host are routed to its current base station on a hop-by-hop basis where each node only needs to know on which of its outgoing ports to forward the packets. This limited routing information is local to the node and does not presume the nodes having any knowledge of the wireless network topology. Figure 1.2 shows the relationship between Mobile IP and Cellular IP.
1.2 Research Issues in Mobile Computing

The development in wireless networks and access technologies is quite enthralling and the most interesting research field. It is easy nowadays to access Internet through the laptops or other mobile devices. These connections, most of the time, face the problems of bandwidth limitation and reliability. These two issues limit the nature of the applications in addition to the number of users. Other often encountered problem, in any mobile computing system, is disconnection. Disconnection is probable and frequent in mobile systems. Mobile computing needs to be studied and effort is required to develop robust system architecture. A set of protocols and models to enhance the performance of mobile computing systems are required for this purpose.

There are many interesting research issues to be studied in mobile computing systems. These issues are explained as follows. [8]

1.2.1 Bandwidth and Disconnection

Wireless networks have lower bandwidth than wired networks. Bandwidth utilization can be improved by compression of data before transmission. When a mobile user is
Introduction

connected to a new access network, the bandwidth might vary according to availability of the bandwidth. This variation in bandwidth may cause failure to some running applications. Therefore, these applications should be designed accordingly. They can be designed to run at full bandwidth or to minimum bandwidth or to adapt their features automatically with the available bandwidth.

The common difficulty in wireless networks is that they are more prone to failure due to disconnection. When the computer is more autonomous it can deal with disconnection. Disconnection might be caused by the lack of bandwidth in the network, therefore; the challenge and the research issue which tries to manage the bandwidth should take into consideration the most important goal of bandwidth management so as to reduce or prevent disconnection.

1.2.2 Location Intelligence

As a mobile computer moves it connects to networks with different features and will be able to continue their functions as if they are still in the home network. A mobile computer must be able to switch from one mode to another such as switching from cellular mode of operation to satellite mode when it moves to another network. Mobile computers should be able to figure out the information in the network they are connected to such as server name, shared printers, etc. automatically and intelligently.

1.2.3 Security Risks

The ease of connecting to the wireless link and data transmission in large areas introduces the security issue. Stored data in mobile computers are prone to risk due to theft or physical damage. To prevent unauthorized disclosure of information, data stored on disks and removable memory cards should be encrypted. To prevent data loss, backup copies must be immediately made on a stationary server.

1.2.4 Network Layer Mobility

Mobility is essentially an address translation problem that can be best resolved at the network layer. Mobile systems frequently change their point of attachment to the
network. To ensure that mobile systems run without disruption, an internetworking infrastructure is needed. Routing in wireless networks is an important research issue and many protocols have been developed to improve the routing operation in the mobile computing systems.

1.2.5 Power Consumption

Batteries are the primary power sources for Mobile Computers. Batteries should be ideally as light as possible but at the same time they should be capable of longer operational times. Battery life should be kept as long as possible using some power consumption schemes. Normally the clock in the mobile computers chip causes the power consumption. It is stated that with every MHz of clock rate in a 386 laptop the power consumption increases by approximately 0.2 watts [164].

Power management is necessary where individual components can be powered down when they are idle. Applications should be redesigned to make them less computation and communication intensive. Idle mobile hosts in Cellular IP networks are grouped in areas called paging areas. These mobile hosts are idle and don’t spend much power.

1.3 QoS in Wireless Networks

The word quality is defined by ISO 8402 is “the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs”. ISO 9000 defines quality as the degree to which a set of inherent characteristics fulfils requirements. ITU basically defines Quality of Service (QoS) as “the collective effect of service performance which determines the degree of satisfaction to the user of the service”. IETF considers QoS as the ability to segment traffic or differentiate between traffic types in order for the network to treat certain traffic flows differently from others. QoS defines both the service categorization and the overall performance of the network for each category [9].

From network point of view, QoS is the ability of a network element (e.g. an application, host or router) to have some level of assurance that its traffic and service requirements will be satisfied. QoS manages bandwidth according to application
Introduction

demands and network management settings. The term “QoS” is used in different meanings, ranging from the users’ perception of the service to a set of connection parameters necessary to achieve particular service quality. The meaning of the “QoS” changes depending on the application field and on the scientific scope. There are three types of QoS: intrinsic, perceived and assessed [10]. Intrinsic QoS is directly supported by the network and may be described in terms of objective parameters, e.g. loss and delay. Perceived QoS (P-QoS) is the quality perceived by the users. It heavily depends on the network performance but is measured by the “average opinion” of the users. Mean Opinion Score (MOS) methods are often used to perform the measure of the quality. Users assign an MOS rating to the application they are evaluating that may be as follows: 1 – bad, 2 – poor, 3 – fair, 4 – good, 5 – excellent. The MOS is the arithmetic mean of all the individual scores, and can range from 1 (worst) to 5 (best) [MOS]. Assessed QoS is referred as the will of a user to continue using a specific service. It is related to P-QoS and also depends on the pricing mechanism, level of assistance of the provider and other marketing and commercial aspects. For example, a performance decrease may be surely tolerated by a user if the service is free, but the same decrease will raise criticism if the user is paying for it.

When a stream of packets is sent from source to destination, it is called flow. All the packets in a flow follow the same route in connection-oriented networks. In connectionless networks, the packets may take different routes. The main parameters characterizing the flow are: delay, jitter, reliability and bandwidth. These parameters together determine the Quality of Service (QoS). QoS Requirements depend on the application and its nature. Real-time applications such as VoIP and videoconferencing have strict delay requirements. These applications are also extremely sensitive to jitter. If a user is watching video over the network and the frames are delayed by exactly 2 msec, then the harm will be acceptable. But if the transmission time varies randomly between 1 to 2 seconds, the result will be terrible. Applications differ in their bandwidth needs. With text e-mail, the bandwidth is not needed so much but video in all forms needs a great amount of bandwidth.
1.3.1 Techniques for Achieving Good Quality of Service

Definition of QoS mentions the nature of QoS and the parameters affecting it. Moreover, it mentions the relations between the different applications and the QoS. It doesn’t mention or discuss a way to improve or achieve good QoS. In fact, there is no single way in which achieving good QoS is possible. Many techniques have been developed for this purpose. Following is an unended and brief description of the techniques used for achieving good QoS [11].

Overprovisioning

It seems to be the easiest solution. It means providing much router capacity, buffer space and bandwidth for the flow of packets. With the limitations in resources especially in wireless environment, this solution is impractical as it is quite expensive.

Buffering

Buffering the packets of a flow doesn’t affect the bandwidth needed for a type of traffic as well it doesn’t affect the reliability. Buffering solution smoothes out the jitter, where for video and audio demand jitter is the main concern. The solution can be found in our daily life in some websites which have some video or audio advertising. These websites use players that buffer for around 10 seconds before starting to play.

Traffic Shaping

When a server outputs the data packets in an irregular manner, the traffic congestion may occur. This non-uniform output is common if the server is handling many streams at the same time, and also allows other actions. This problem introduces traffic shaping which smoothes out the traffic on the server side rather than on the client side. Traffic shaping implies regulating the average rate of data transmission. Traffic shaping reduces congestion and thus helps in improving QoS for real-time traffic such as video and audio data which, as mentioned before, require more attention than other types of traffic.
Resource Reservation

Once the route for the flow of packets is specified, it is possible to reserve resources along that route to make sure that the needed capacity is available. The main resources which can be reserved are Bandwidth, Buffer space and CPU cycles. Bandwidth is the most important resource in any network to be reserved. Reserving bandwidth doesn't mean oversubscribing any output line. For example, if a flow requires 1 Mbps of bandwidth and the line capacity is 2 Mbps, this flow will go well. But trying to direct three flows through that line will not work. The second resource which is often in short supply is buffer space. When a packet arrives to a router, the router must copy this packet to a buffer and queue it for further transmission on the outgoing line. If there is no buffer space available, the packet has to be dropped. For a good service, the buffer space must be reserved for real-time traffic which doesn't have to compete for buffer with other types of traffic. Using buffer reservation, there will always be buffer availability when the flow needs one, up to maximum limit. Another important resource is the CPU cycles. A router has finite capacity, and according to this capacity it can process a specific number of packets per second. CPU needs every cycle in its work to ensure that its work is done. Loosing a few cycles may put the CPU in a blockage. Reserving CPU cycles, needed to process the real-time packets, helps in improving the performance of the CPU and as a result the CPU doesn't go in idle situation. Overall, this can improve the quality of service in the network [11].

In recent years, a variety of resource reservation algorithms have been proposed to improve QoS in cellular networks [12], [13], [14], [15], [16], [17], [18], [19], [20], [21] and [22]. Some of them suggest reserving some bandwidth in the target cells during handoff operation and the neighboring cells at the same time improves the performance [16]. Levine et al. presented a scheme to reserve resources with neighboring cells by exchanging information related to the movement pattern and position [17]. In [18], a scheme based on max−min fairness protocol to provide QoS in wireless multimedia network is proposed. In spite of potentially improving both the CBP and the CDP in this scheme, the users might be subjected to significant
Chapter 1

Introduction

bandwidth fluctuations. Lee et. al., presented a handoff management scheme using simultaneous multiple bindings that reduces packet loss and generates negligible delays due to handoff in IP-based third-generation cellular systems [19]. The CDP is probably reduced whereas the bandwidth levels of ongoing multimedia traffic are also degraded. Kuo et. al. took use of the knowledge of staying time, available time, and the class of the MH to develop a resource semi-reservation scenario and it turns out to be ideal since the speed of the MH is difficult to detect accurately [20].

1.4 Evolutionary Algorithms as a Tool

In artificial intelligence, an evolutionary algorithm (EA) is a subset of evolutionary computation, a generic population-based metaheuristic optimization algorithm. An EA uses some mechanisms inspired by biological evolution: reproduction, mutation, recombination, and selection. Candidate solutions, to the optimization problem play the role of individuals in a population and the fitness function determines the environment within which the solutions "live". Evolution of the population then takes place after the repeated application of the above mentioned operators.

Evolutionary algorithms often perform well approximating solutions to all types of problems because they ideally do not make any assumption about the underlying fitness landscape. This generality is shown by successes in fields as diverse as engineering, biology, economics, marketing, genetics and many others. Apart from their use as a mathematical optimizer, evolutionary computation and algorithms have also been used as an experimental framework within which to validate theories about biological evolution and natural selection, particularly through work in the field of artificial life.

In recent past the applications of EAs, useful search procedures, have attracted attention of many researches of many fields as a problem solving tool. Many algorithms are classified under the category of evolutionary algorithms such as Genetic Algorithm and Particle Swarm Optimization. These two algorithms have been applied widely and successfully for solving optimization problems. Researchers of
mobile computing have used EAs for bandwidth reservation [23]. Other applications
used EA for resource reservation in Cellular IP networks [24].

EAs start with an initial population. The population, a potential solution to the
problem, consists of individuals also called chromosomes at the implementation level.
A new set of chromosomes is created, in every generation, using information from the
individuals of the previous generations. Generating the initial populations in GA and
PSO is usually done randomly. Evolution reflects the possibility of generating new
offspring which is likely to be better than the parents in some features. More details
about GA and PSO are presented in the next chapter section 2.7.

1.5 Concluding remarks

The challenges in mobile computing include mobility management, disconnections,
weak connectivity, adaptivity, upward compatibility with existing and legacy
applications and support of variable types of mobile units.

Researchers realized that either application or system adaptation is a key requirement
in any mobile computing system. Researchers also realized that they must identify,
and systematically eliminate, the limitations of this new environment through various
optimizations. Mobile computing (including mobile client-server information access)
is a rapidly changing research field that depends on a rapidly evolving set of
technologies. Researchers, however, should monitor these advances closely and
should adapt and direct their research based on the parameters of the latest
technology. This is important because some of the strong assumptions regarding the
limitations of the wireless network or the mobile computer are being relaxed or
nullified by newer technological developments. Many problems still remain to be
understood and solved. At this time, it is difficult to quantitatively evaluate and
compare the various proposed models and techniques because of the lack of available
application experiences and samples. Experimentation with these models should be
the critical next step in mobile computing research.

There are several technological problems that are yet to be solved to exploit the full
potential of mobile computing; many of these issues are being studied in the
laboratories. Software as well as hardware engineers need to be acquainted with those techniques to successfully meet the challenges of the future. There is a need to develop educational materials and curricula that incorporate newly created fundamental engineering and computer and information science knowledge in mobile computing. This will help to enhance the education and careers of future engineers and scientists by enabling them to compete in the global environment.

In this chapter we have given a review about mobile computing system and the most important challenges and the research issues which are well studied by researchers and must be addressed in the field of mobile computing in order to produce new models and improve the performance of mobile computing systems.

QoS, the main concern of this thesis, is addressed in this chapter along with the techniques that can achieve better QoS. Resource reservation, as the main principle adopted in this work, is discussed. The thesis elaborates optimizing the resources of the Cellular IP network for improving QoS. Evolutionary algorithms, the tool applied to manage QoS parameters such as bandwidth, buffer and Router CPU time, has been introduced in this chapter.

1.6 Organization of the Thesis

The remaining of the thesis has been organized as follows.

In chapter two, Quality of Service issue along with the parameters affecting QoS have been discussed. The problem, addressed in the thesis, has also been explained in this chapter. Genetic Algorithm and Particle Swarm Optimization, the evolutionary techniques used to solve the problem, are elaborated in this chapter. Related models that deal with the same problem have been listed in chapter two.

The very first resource for better QoS that has been considered is bandwidth. Chapter three proposes two models for bandwidth management in Cellular IP networks. First is the model for bandwidth management using PSO and the second one is using GA.
Chapter 1

Introduction

A comparison between the two models has been done. The models have been simulated for performance studies and observations have been derived.

Next resource, addressed in this thesis, is buffer. Two models for buffer management in Cellular IP networks have been proposed in chapter four. Theses two models use PSO and GA for buffer management. Simulation has been done for performance evaluation of both the models alongwith the comparative study in chapter four.

The models for Router CPU cycles, the third resource affecting QoS in Cellular IP network, have been proposed and designed in chapter five. The models consider a flow of real-time packets and a number of routers in the route. In these models both GA and PSO are applied in order to find the best processing time in the route and optimize this time. Comparison between GA based model and PSO based model in reducing processing time has been done in this chapter. All the experiments have been conducted to check the efficacy of the models.

In chapter six, the conclusion for our research work towards this thesis is drawn along with the direction for the future work.