Wireless and Mobile Computing

7.1 Introduction

Wireless networks utilize radio waves and/or microwaves to maintain communication channels between computers. Wireless networking is a more modern alternative to wired networking that relies on copper and/or fiber optic cabling between network devices. Wireless is rapidly gaining in popularity for both home and business networking. Wireless technology continues to improve, and the cost of wireless products continues to decrease. Popular wireless local area networking (WLAN) products conform to the 802.11 "Wi-Fi" standards. The gear a person needs to build wireless networks include network adapters (NICs), access points (APs), and routers.

The basic wireless LANs, or WLANs, that are the most commonly used today typically run half-duplex communication—everyone is sharing the same bandwidth and only one user is communicating at a time. Transmitting a signal using the typical 802.11 specifications works a lot like it does with a basic Ethernet hub: They're both two-way forms of communication, and they both use the same frequency to both transmit and receive, often referred to as half-duplex and mentioned earlier. Wireless LANs (WLANs) use radio frequencies (RFs) that are radiated into the air from an antenna that creates radio waves.

These waves can be absorbed, refracted, or reflected by walls, water, and metal surfaces, resulting in low signal strength. So because of this innate vulnerability to surrounding environmental factors, it's pretty apparent that wireless will never offer us the same robustness as a wired network can, but that still doesn't mean wireless has no advantages. A wireless network offers advantages and disadvantages compared to a wired network. Advantages of wireless include mobility and elimination of unsightly cables. Disadvantages of wireless include the potential for radio interference due to weather, other wireless devices, or obstructions like walls.

In wireless networking, a dead zone (sometimes called "dead spot") is an indoor or outdoor area where wireless devices are unable to connect to a network. Wireless

dead zones are common inside homes, especially larger ones with brick or plaster walls. Dead zones also exist outdoors due to physical obstructions like trees and large buildings, in gaps between the coverage areas of public Wi-Fi hotspots or anyplace signal interference from other sources is prevalent. While often just an annoyance, dead zones become a serious safety issue when depending on wireless connectivity for emergency calls.

7.2 The 802.11 Wireless Protocol

802.11 is a set of technology standards for wireless network devices. These standards are determined by the IEEE (Institute of Electrical and Electronic Engineers), and they basically govern how different wireless devices are designed and how they communicate with each other. You'll see 802.11 mentioned when you are looking to buy a wireless-enabled device or a piece of wireless hardware. When researching what netbook to buy, for example, you may see some advertised as communicating wirelessly at "ultra-high" 802.11n speeds (in fact, Apple touts its use of 802.11n technology in its latest computers and devices). The 802.11 standard is also mentioned in descriptions of wireless networks themselves; for example, if you want to connect to a public wireless hotspot, you may be told that it is an 802.11g network.

7.2.1 What do the letters mean?

The letter after "802.11" indicates an amendment to the original 802.11 standard. Wireless technology for consumers/the general public has progressed from 802.11a to 802.11b to 802.11g to, most recently, 802.11n. (Yes, the other letters, "c" and "m," for example, also exist in the 802.11 spectrum, but they are only primarily relevant to IT engineers or other specialized groups of people.)

802.11b was the first standard to be widely used in WLANs. The 802.11a standard is faster but more expensive than 802.11b; 802.11a is more commonly found in business networks. 802.11g, attempts to combine the best of both 802.11a and 802.11b, though it too is more a more expensive home networking option.

Important to note is the fact that the 802.11 specifications were developed so that there would be no licensing required in most countries—to ensure the user the freedom to install and operate without any licensing or operating fees. This means

that any manufacturer can create products and sell them at a local computer store or wherever. It also means that all our computers should be able to communicate wirelessly without configuring much, if anything at all.

Without getting into more detailed distinctions between 802.11a, b, g, and n networks, we can just generalize that each new version of 802.11 offers improved wireless network performance, compared to prior versions, in terms of:

- ***Data rate*:** maximum data transfer speed (i.e., how fast information can travel over the wireless network)
- ***Range*:** the distance the wireless signals can reach or how broad an area the wireless signals cover (i.e., how far you can be from the wireless signal source and still maintain a reliable connection)

We can increase the transmitting power and gain a greater transmitting distance, but doing so can create some nasty distortion, so it has to be done carefully. By using higher frequencies, we can attain higher data rates, but this is, unfortunately, at the cost of decreased transmitting distances. And if we use lower frequencies, we get to transmit greater distances but at lower data rates. This should make it pretty clear to you that understanding all the various types of WLANs you can implement is imperative to creating the LAN solution that best meets the specific requirements of the unique situation you're dealing with.

7.2.2 802.11

In 1997, the Institute of Electrical and Electronics Engineers (IEEE) created the first WLAN standard. They called it 802.11 after the name of the group formed to oversee its development. Unfortunately, 802.11 only supported a maximum network bandwidth of 2 Mbps - too slow for most applications. For this reason, ordinary 802.11 wireless products are no longer manufactured.

7.2.3 802.11b

IEEE expanded on the original 802.11 standard in July 1999, creating the 802.11b specification. 802.11b supports bandwidth up to 11 Mbps, comparable to traditional Ethernet. 802.11b uses the same unregulated radio signaling frequency (2.4 GHz) as the original 802.11 standard. Vendors often prefer using these frequencies to lower their production costs. Being unregulated, 802.11b gear can

incur interference from microwave ovens, cordless phones, and other appliances using the same 2.4 GHz range. However, by installing 802.11b gear a reasonable distance from other appliances, interference can easily be avoided.

- **Advantages:** lowest cost; signal range is good and not easily obstructed
- **Disadvantages:** slowest maximum speed; home appliances may interfere on the unregulated frequency band

7.24 802.11a

While 802.11b was in development, IEEE created a second extension to the original 802.11 standard called 802.11a. Because 802.11b gained in popularity much faster than did 802.11a, some folks believe that 802.11a was created after 802.11b. In fact, 802.11a was created at the same time. Due to its higher cost, 802.11a is usually found on business networks whereas 802.11b better serves the home market.

802.11a supports bandwidth up to 54 Mbps and signals in a regulated frequency spectrum around 5 GHz. This higher frequency compared to 802.11b shortens the range of 802.11a networks. The higher frequency also means 802.11a signals have more difficulty penetrating walls and other obstructions. Because 802.11a and 802.11b utilize different frequencies, the two technologies are incompatible with each other. Some vendors offer hybrid 802.11a/b network gear, but these products merely implement the two standards side by side (each connected devices must use one or the other).

- **Advantages:** fast maximum speed; regulated frequencies prevent signal interference from other devices
- **Disadvantages** - highest cost; shorter range signal that is more easily obstructed

7.2.5 802.11g

In 2002 and 2003, WLAN products supporting a newer standard called 802.11g emerged on the market. 802.11g attempts to combine the best of both 802.11a and 802.11b. 802.11g supports bandwidth up to 54 Mbps, and it uses the 2.4 Ghz frequency for greater range. 802.11g is backwards compatible with 802.11b,

meaning that 802.11g access points will work with 802.11b wireless network adapters and vice versa.

- **Advantages: -** fast maximum speed; signal range is good and not easily obstructed
- **Disadvantages: -** costs more than 802.11b; appliances may interfere on the unregulated signal frequency

7.2.6 802.11n

802.11n (also known as "Wireless-N"), being the latest wireless protocol, offers the fastest maximum data rate today and better signal ranges than the prior technologies. In fact, demonstrated speeds for 802.11n products have been 7 times faster than 802.11g; at 300 or more Mbps (megabits per second) in real world usage, 802.11n is the first wireless protocol to seriously challenge wired 100 Mbps Ethernet setups.

Wireless-N products are also designed to perform better at greater distances, so that a laptop can be 300 feet away from the wireless access point signal and still maintain that high data transmission speed. By contrast, with the older protocols, your data speed and connection tend to be weakened when you are that far away from the wireless access point.

- **Advantages: -** fastest maximum speed and best signal range; more resistant to signal interference from outside sources
- **Disadvantages: -** standard is not yet finalized; costs more than 802.11g; the use of multiple signals may greatly interfere with nearby 802.11b/g based networks.

The increased performance benefits of 802.11n are definitely worth a look, but keep in mind the following caveats/tips when deciding whether to stick with the more widely used 802.11g protocol or invest in 802.11n now:

- Network performance will be greatest when each of the devices on the wireless network are using the 802.11n technology. On the flip side, if an older device using 802.11g or 802.11b connects to your 802.11n-based router, the speed and data rate of all the devices on the network will decrease. One

way to get around this issue for your home wireless network is to get what's called a dual-band router or access point. This will allow older devices to run over one frequency band (2.4 GHz) and newer 802.11n-based devices to use the other frequency band (5 GHz).

- Look for recently manufactured network devices, which will have a greater likelihood of conforming to the ratified 802.11n standard. Definitely avoid "pre-n" or "draft n" products.
- Also look out for products that are certified by the Wi-Fi Alliance (they will have the Wi-Fi CERTIFIED logo on their packaging), as these products are tested for compatibility and interoperability.
- Finally, keep in mind that most public wireless hotspots and wireless networks in general are more likely to be running 802.11g or even b. Although your newer 802.11n device is backwards-compatible with (i.e., can work on) these networks, it'll do so at the slower g or b speed.

7.3 Wireless Local Area Networks and Satellite-Based Networks

7.3.1 Introduction

The Wireless Local Area Network (WLAN) utilizes and uses the electromagnetic waves which are spread by spectrum technology and are based on radio waves to transfer information in form of signals between devices. WLANS are of two types infrastructure WLANs Independent WLANs.

The Infrastructure WLANs are used where the wireless network is linked to a wired network or with the cables, is more commonly used today. In an infrastructure WLAN, the wireless network is usually connected to a wired network such as Ethernet, via access points, which possesses both Ethernet links and antennas to send and receive signals as well as making them powerful to transmit. These signals span micro cells, or circular coverage areas (depending on walls and other physical obstructions) or buildings in the way of access points and transmitter, in which devices can communicate with the access points, and through these, with the wired network. In a wireless LANs the devices can move within and between coverage areas without experiencing disruption or obstruction in connectivity as long as they stay within range of an access point or extension point which is similar to an access point at all times and occasions.

The simplest configuration is an Independent (or peer-to-peer) WLAN that connects a group of PCs with wireless adapters. Any time, two or more wireless adapters are within range of each other, they can set up an ad hoc independent network. These on-demand networks typically require no administration or pre-configuration. Basically this is a network which links two or more computers without using wires or cables. WLAN utilizes spread-spectrum technology based on radio waves to enable communication between devices in a limited area. This gives users the mobility to move around within a broad coverage area and still be connected to the network. As regards the home user, the wireless has become most popular and economical due to ease of installation, and location freedom with the gaining popularity of laptops and PDAs.

7.3.2 The Benefits of Wireless LANs

- ***Convenience:*** The wireless nature of such networks allows users to access network resources from nearly any convenient location within their primary networking environment (a home or office). With the increasing saturation of laptop-style computers, this is particularly relevant.
- ***Mobility:*** With the emergence of public wireless networks, users can access the internet even outside their normal work environment. Most coffee shops, for example, offer their customers a wireless connection to the internet at little or no cost.
- ***Productivity:*** Users connected to a wireless network can maintain a nearly constant affiliation with their desired network as they move from place to place. For a business, this implies that an employee can potentially be more productive as his or her work can be accomplished.
- ***Deployment:*** Initial setup of an infrastructure-based wireless network requires little more than a single access point. Wired networks, on the other hand, have the additional cost and complexity of actual physical cables being run to numerous locations (which can even be impossible for hard-to-reach locations within a building).
- ***Expandability:*** Wireless networks can serve a suddenly-increased number of clients with the existing equipment. In a wired network, additional clients would require additional wiring.

- ***Cost:*** Wireless networking hardware is at worst a modest increase from wired counterparts. This potentially increased cost is almost always more than outweighed by the savings in cost and labor associated to running physical cables.

7.3.3 Drawbacks of Wireless LANs

The Wireless LAN technology, while filled with the conveniences and advantages which are described above. It has its share of downfalls or demerits also. If you think of a networking solution, the wireless LANs may not be desirable for a number of reasons. As most of these have to do with the inherent limitations of the technology because they are sometimes dependent devices...

- ***Security:*** The Wireless LAN transceivers are designed to serve computers throughout a structure with uninterrupted service using radio frequencies. Due this space and cost, the antennas typically present on wireless networking cards in the end computers are generally relatively poor. To receive signals properly using such limited antennas or devices throughout even a modest area, the wireless LAN transceiver utilizes a fairly considerable amount of power. It means that not only can the wireless packets be intercepted by a nearby adversary's poorly-equipped computer, but more importantly, a user willing to spend a small amount of money on a good quality antenna can pick up packets at a remarkable distance as the case may be . Perhaps, the people in hundreds of time the radius as the typical user so they are also vulnerable as they are not secure data transmitters.
- ***Range:*** As regards the typical range of a common 802.11n network with standard equipment is on the order of tens of meters. While sufficient for a typical home, it will be insufficient in a larger structure. To obtain additional range, repeaters or additional access points will have to be purchased. Costs for these items can add up quickly. Other technologies are in the development phase, however, which feature increased range, hoping to render this disadvantage irrelevant.
- ***Reliability:*** Like any radio frequency transmission, wireless networking signals are subject to a wide variety of interference, as well as complex propagation effects (such as multi path fading) that are beyond the control of the network

administrator. As a result, important network resources such as servers are rarely connected wirelessly.

- ***Speed:*** The speed on most wireless networks (typically 54 Mbps) is far slower than even the slowest common wired networks (100Mbps up to several Gbps). Newer standards such as 802.11n are addressing this limitation and will support peak throughputs in the range of 100-200 Mbps.

7.3.4 Wireless LAN architecture using an Infrastructure BSS

WLAN can be connected through a cell phone card via GSM network, satellite hardware from your satellite company, or the most common a 802.11 router and either a network card for PCI/PCI express slot on a desktop, or PCMCIA card for a laptop. Some laptops already come prepared with built-in wireless. Satellite communications have distinct benefits over terrestrial alternatives:

- ***Universal:*** Satellite communications are available virtually everywhere. A small constellation of satellites can cover the Earth's entire surface. And even the reach of a single satellite is far more extensive than what any terrestrial network can achieve.
- ***Versatile:*** Satellites can support all of today's communications needs - transactional and multimedia applications, video, voice, cellular networks, entertainment and even breaking news.
    - Bring broadband to the last mile of residences and businesses.
    - Overcome regulatory issues that make alternative carriers dependent on incumbents.
    - Deliver a communications infrastructure to areas where terrestrial alternatives are unavailable, unreliable or simply too expensive.
- ***Reliable:*** Satellite is a proven medium for supporting a company's communications needs. Whereas terrestrial IP networks are often a mix of different networks and topologies, with different level of congestion and latency. Satellite networks are extremely predictable allowing constant and uniform quality of service to hundreds of locations, regardless of geography.
- ***Seamless:*** Satellite's inherent strength as a broadcast medium makes it ideal for the simultaneous distribution of bandwidth-intensive information to hundreds or thousands of locations.

- ***Fast:*** Unlike most terrestrial alternatives, satellite networks can be rolled out quickly and inexpensively to hundreds or thousands of locations, connecting cities or remote locations across a large landmass, where copper or fiber is cost prohibitive. Since satellite networks can be set up quickly, companies can be fast-to-market with new services.
- ***Expandable:*** Satellite networks are easily scalable, allowing users to expand their communications networks and their available bandwidth easily. In coordination with local vendors, expanding a network on the ground requires the ordering of new terminal components and the commissioning of increased bandwidth at each site.
- ***Flexible:*** Satellites can be easily integrated to complement, augment or extend any communications network, helping overcome geographical barriers, terrestrial network limitations and other constraining infrastructure issues.

7.4 Mobile Internet Protocol

7.4.1 Introduction

The Internet Protocol (IP) is the most successful network layer protocol in computing due to its strengths, but it also has some weaknesses, most of which have become more important as networks have evolved over time. While mobile devices can certainly use IP, the way that devices are addressed and datagrams routed causes a problem when they are moved from one network to another. At the time IP was developed, computers were large and rarely moved. Today, we have millions of notebook computers and smaller devices, some of which even use wireless networking to connect to the wired network. The importance of providing full IP capabilities for these mobile devices has grown dramatically. To support IP in a mobile environment, a new protocol called IP Mobility Support, or more simply, Mobile IP, was developed.

7.4.2 The Problem with Mobile Nodes in TCP/IP

IP addresses are fundamentally divided into two portions: a network identifier (network ID) and a host identifier (host ID). The network ID specifies which network a host is on, and the host ID uniquely specifies hosts within a network. This structure is fundamental to datagram routing, because devices use the network ID portion of the destination address of a datagram to determine if the

recipient is on a local network or a remote one, and routers use it to determine how to route the datagram.

This is a great system, but it has one critical flaw: the IP address is tied tightly to the network where the device is located. Most devices never (or at least rarely) change their attachment point to the network, so this is not a problem, but it is certainly an issue for a mobile device. When the mobile device travels away from its home location, the system of routing based on IP address “breaks”. Mobile IP solves this problem by giving mobile devices and routers the capability to forward datagrams from one location to another.

7.4.2 Mobile IP Overview: "Address Forwarding" for the Internet

Mobile IP works in a manner very similar to a mail forwarding system. The mobile node is normally resident on its home network, which is the one that is indicated by the network ID in its IP address. Devices on the internetwork always route using this address, so the pieces of “mail” (datagrams) always arrive at a router at the device's “home”. When the device “travels” to another network, the home router (“post office”) intercepts these datagrams and forwards them to the device's current address. It may send them straight to the device using a new, temporary address, or it may send them to a router on the device's current network (the “other post office) for final delivery. An overview of Mobile IP operation can be seen in Figure 7-1.

7.4.3 Mobile IP Addressing: Home and "Care-Of" Addresses

Just as most of us have only a single address used for our post mail, most IP devices have only a single address. The mobile device, however, needs to have two addresses; a normal one and one that is used while it is away.

Home Address: The “normal”, permanent IP address assigned to the mobile node. This is the address used by the device on its home network, and the one to which datagrams intended for the mobile node are always sent.

Care-Of Address: A secondary, temporary address used by a mobile node while it is 'traveling” away from its home network. It is a normal 32-bit IP address in most respects, but is used only by Mobile IP for forwarding IP datagrams and for

administrative functions. Higher layers never use it, nor do regular IP devices when creating datagrams.

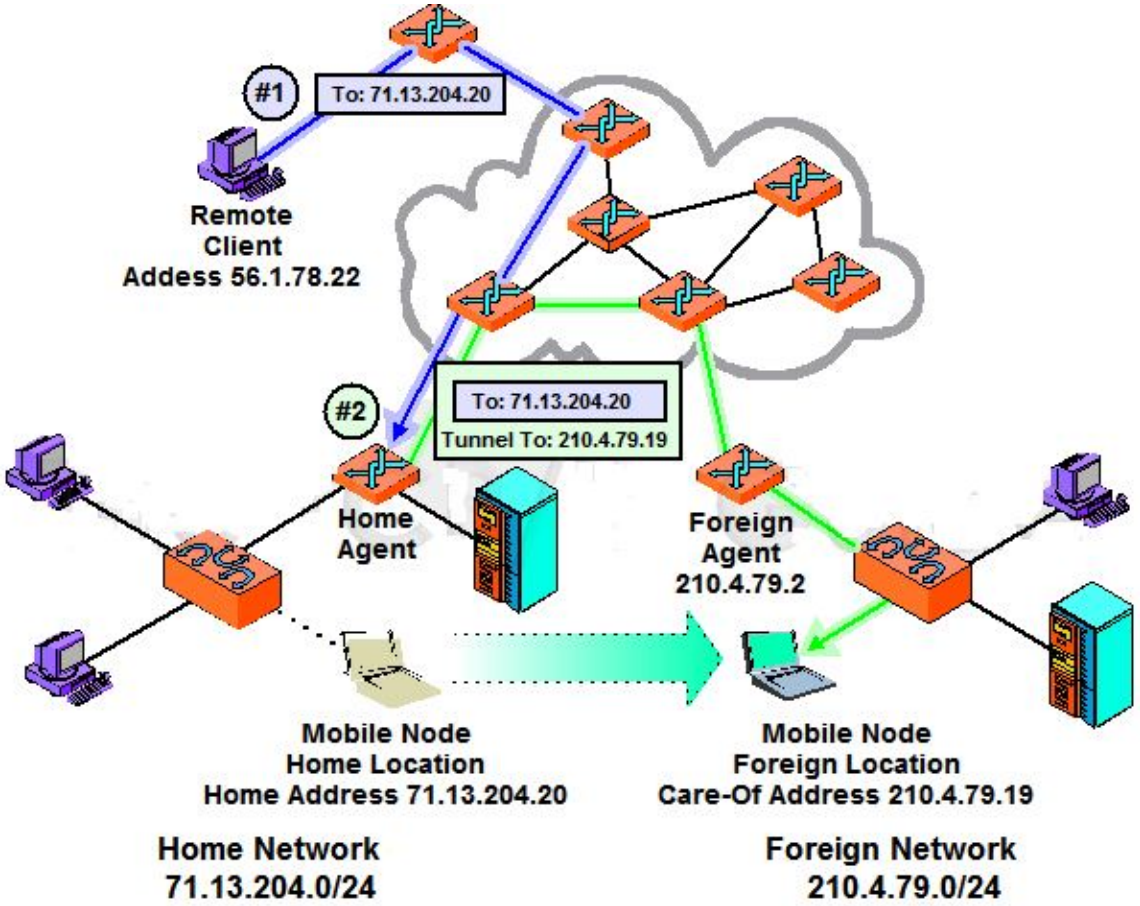

Figure 7-1: General Operation of the Mobile IP Protocol

7.4.4 Mobile IP Care-Of Address Types

The care-of address is a slightly tricky concept. There are two different types, which correspond to two distinctly different methods of forwarding datagrams from the home agent router.

a) Foreign Agent Care-Of Address

This is a care-of address provided by a foreign agent in its Agent Advertisement message. It is, in fact, the IP address of the foreign agent itself. When this type of care-of address is used, all datagrams captured by the home agent are not relayed

directly to the mobile node, but indirectly to the foreign agent, which is responsible for final delivery. This arrangement is illustrated in Figure 7-2.

In normal mail delivery analogy, this type of care-of address is like forwarding from the one post office to another. The personnel would take a letter sent to the first post office, and repackage it for delivery to "the recipient, care of the second post office". The second post office (or the recipient) would need to worry about the last leg of the delivery.

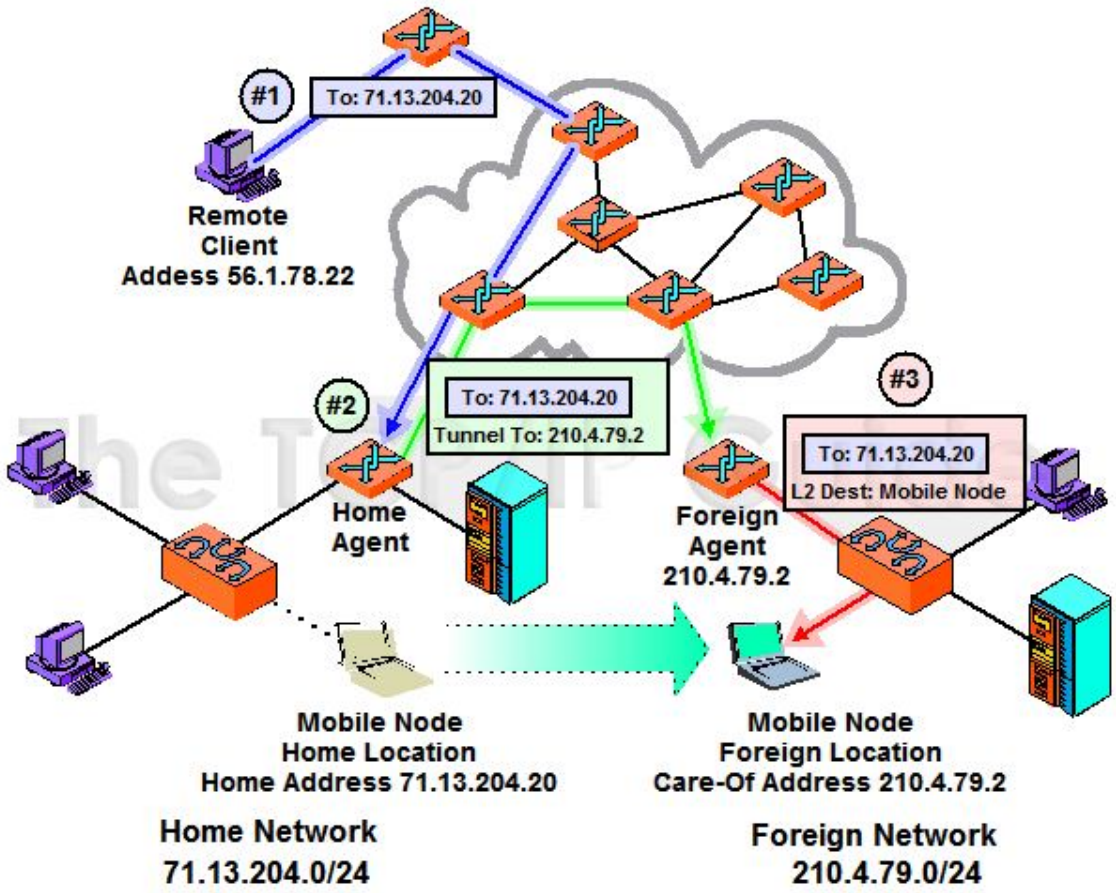

Figure 7-2: Mobile IP Operation with a Foreign Agent "Care-Of" Address

b) Co-Located Care-Of Address

This is a care-of address assigned directly to the mobile node using some means external to Mobile IP. For example, it may be assigned on the foreign network

manually, or automatically using DHCP. In this situation, the care-of address is used to forward traffic from the home agent directly to the mobile node. This was the type of address shown in Figure 7-1.

In normal mail delivery analogy, this is like the recipient obtaining a temporary address for his use while in the new location. The home post office would forward directly to his new address. They would not specifically send it to the second post office (though of course that PO would handle the mail at some point).

Key Concept: In Mobile IP, each mobile device uses a temporary, care-of address while on a foreign network. A co-located care-of address is one that is assigned directly to the mobile node, and enables direct delivery of datagrams to the node. The alternative is to use a foreign agent care-of address. In this situation the mobile node actually uses the IP address of the foreign agent; datagrams are sent to the foreign agent, which delivers them to the mobile node.

7.5 Client-Server Computing in Mobile Environments

7.5.1 Introduction

Recent advances in wireless data networking and portable information appliances have engendered a new paradigm of computing, called mobile computing, in which users carrying portable devices have access to data and information services through a shared infrastructure regardless of their physical location or movement behavior.

Such a new environment introduces new technical challenges in the area of information access. Traditional techniques for information access are based on the assumptions that the location of hosts in distributed systems does not change and the connection among hosts also does not change during the computation. In a mobile environment, however, these assumptions are rarely valid or appropriate. Mobile computing is distinguished from classical, fixed-connection computing due to the mobility of nomadic users and their computers and the mobile resource constraints such as limited wireless bandwidth and limited battery life. The mobility of nomadic users implies that the users might connect from different access points through wireless links and might want to stay connected while on the move, despite possible intermittent disconnection.

Wireless links are relatively unreliable and currently are two to three orders of magnitude slower than wireline networks. Moreover, mobile hosts powered by batteries suffer from limited battery life constraints. These limitations and constraints leave much work to be done before mobile computing is fully enabled. This remains true despite the recent advances in wireless data communication networks and hand-held device technologies.

There has been a recent proliferation of research addressing issues of mobile systems and applications, especially for the purpose of mobile information access.

In this section, we provide a concrete framework and categorization of the various ways of supporting mobile client-server computing for information access. We examine characteristics of mobility that distinguish mobile client-server computing from its traditional counterpart. We provide a comprehensive analysis of new paradigms and enabler concepts for mobile client-server computing, including mobile-aware adaptation, extended client-server model, and mobile data access. A comparative and detailed review of major research prototypes for mobile information access is also presented.

7.5.2 Paradigms of Mobile Client-Server Computing

Existing research on mobile client-server computing can be categorized into the following three paradigms: (1) mobile-aware adaptation, (2) extended client-server model, and (3) mobile data access.

a) Mobile-aware Adaptation

The dynamics of mobile environments and the limitations of mobile computing resources make adaptation a necessary technique when building mobile systems and applications. The paradigm of mobile-aware adaptation covers various strategies and techniques in how systems and applications respond to the environmental changes and the re- source requirements. It also suggests the necessary system services that could be utilized by mobile-aware applications.

b) Extended Client-Server Model:

The extended client-server model facilitates mobile client-server information access. One distinguishing feature is the dynamic partitioning of client-server

functionality and responsibilities. The extended client-server model provides a way to support the adaptation of mobile systems and applications. The paradigm of the extended client-server model includes various client-server computing architectures that enable the functional partitioning of applications between clients and servers.

c) Mobile Data Access:

Mobile data access addresses issues such as how server data can be delivered to client hosts, how data over wireless and mobile networks is structured, and how the consistency of client cache is ensured effectively. The adaptive strategies for mobile data access depend largely on the type of communication links, the connectivity of mobile hosts, and the consistency requirements of applications.

Mobile data access provides another way to characterize the impact of mobile computing constraints on information access. It should be noted that these new paradigms are closely related to each other. For example, the implementation for data delivery strategies and extended client-server architectures may involve the use of adaptation solutions. Extended client-server architectures might be needed to take advantage of new data delivery strategies. The categorization of new paradigms in this section provides a comprehensive way to understand and analyze various proposed techniques in building mobile client-server information systems.

7.6 Mobile-Aware Adaptation

Mobile clients could face wide variations and rapid changes in network conditions and local resource availability when accessing remote data. In order to enable applications and systems to continue to operate in such dynamic environments, the mobile client-server system must react by dynamically adjusting the functionality of computation between the mobile and stationary hosts. In other words, the computation of clients and servers has to be adaptive in response to the changes in mobile environments.

The range of strategies for application and system adaptation is identified, as shown in Figure 7-3. The range is delimited by two extremes. At one extreme, adaptation is entirely the responsibility of individual applications. This approach,

called laisse-faire adaptation, avoids the need for system support. The other extreme, called application-transparent adaptation, places the entire responsibility for adaptation on the system. A typical case of this approach is to use proxies to perform adaptation on behalf of applications. Between these two extremes lies a spectrum of possibilities that are referred to as application-aware adaptation. This approach supports collaborative adaptation between the applications and the system. That is, the applications can decide how to best adapt to the changing environment while the system provides support through the monitoring of resources and the enforcing of resource allocation decisions.

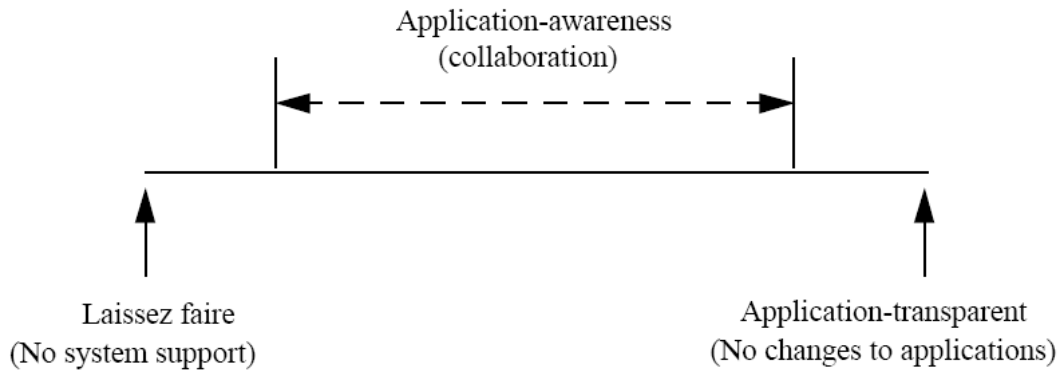

Figure 7-3: Range of adaptation strategies

7.6.1 Application-Transparent Adaptation

Many existing client-server applications are built around the assumption that the environment of a client does not change. These applications are usually unaware of the mobility and make certain assumptions about the resource availability. The approach of application- transparent adaptation attempts to make these applications work with no modification in mobile environments.

This is done by having the system shield or hide the differences between the stationary and mobile environments from applications. A local proxy runs on the mobile host and provides an interface for regular server services to the applications. The proxy attempts to mitigate any adverse effects of mobile environments as shown in Figure 7-4.

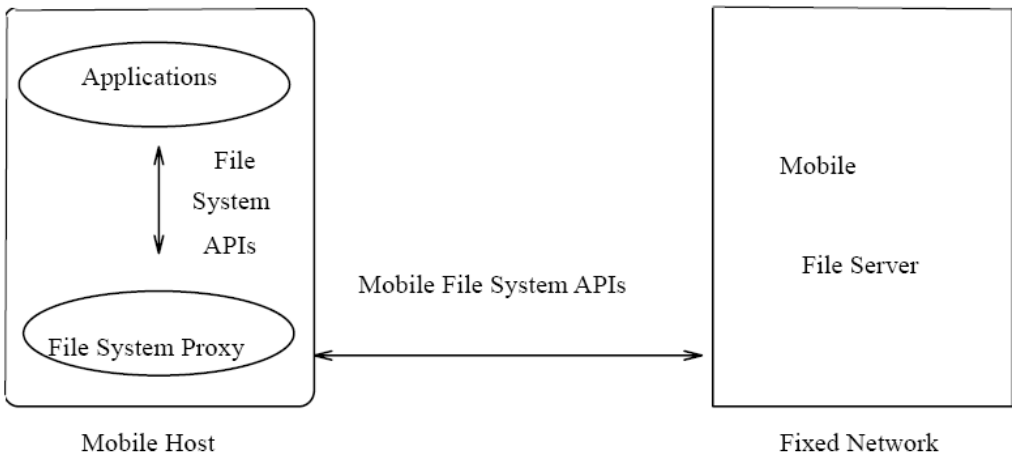

Figure 7-4: File System Proxy

7.6.2 Application-Aware Adaptation

The approach of application-transparent adaptation does not require changing existing applications to run in mobile environments. However, it could sacrifice functionality and performance. As applications are shielded from dealing with mobility, it might be very hard for the system to make adaptation decisions that meet the needs of different and diverse applications. As a result, it may have to require some manual intervention by the user (e.g., having the user indicate which data to pre-fetch onto the mobile device) to make applications run smoothly. Such user-administered manual actions could be less agile to adapt to the changing environment.

To address these problems, application-aware adaptation has been developed. Application-aware adaptation allows applications or their extensions to react to the mobile resource changes. One way to realize the application-aware adaptation is through the collaboration between the system and individual applications. The system monitors resource levels, notifies applications of relevant changes, and enforces resource allocation decisions. Each application independently decides how best to adapt when notified. In a video player application, for example, such adaptation allows the video player system to scale back quality (and resource consumption) when application performance is poor and to attempt to discover additional resources by optimistically scaling up usage.

Depending on where the adaptive application logic resides, the approaches of application-aware adaptation can be divided into the following categories: client-based application adaptation, client-server application adaptation, and proxy-based application adaptation. The client-based adaptation allows the applications on mobile clients to react to the environmental changes, while client-server adaptation might have applications on both client and server to adapt to the changes. The proxy-based adaptation supports application-specific adaptation on the proxy server in the fixed networks.

7.7 Extended Client-Server Model

Another way to characterize the client-server computing in mobile environments is to examine the effect of mobility on the client-server computing model. In a client-server information system, a server is any machine that holds a complete copy of one or more databases. A client is able to access data residing on any server with which it can communicate. Classic client-server systems assume that the location of client and server hosts does not change and the connection among them also does not change. As a result, the functionality between client and server is statically partitioned.

In a mobile environment, however, the distinction between clients and servers may have to be temporarily blurred, resulting in an extended client-server model shown in Figure 7-5. The resource limitations of clients may require certain operations normally performed on clients to be performed on resource-rich servers. Conversely, the need to cope with uncertain connectivity requires clients to sometimes emulate the functions of a server. One extreme case is called the thin client architecture and the other extreme case is the full client architecture.

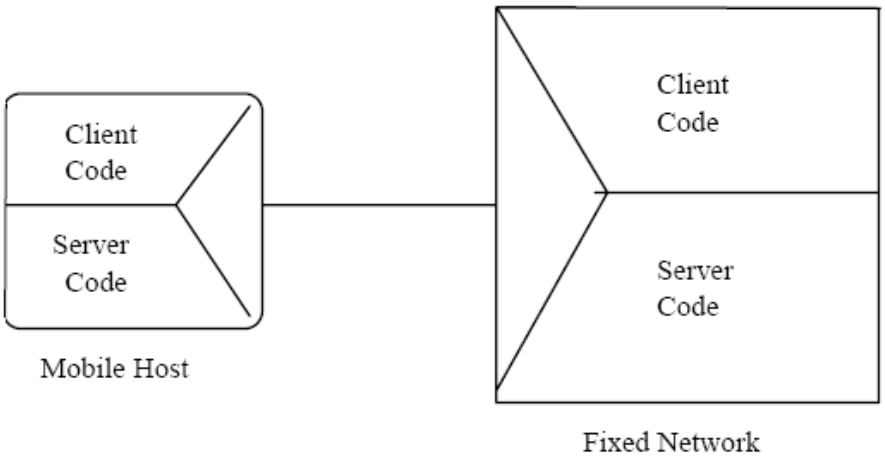

Figure 7-5: Extended Client-Server Model

7.7.1 Thin Client Architecture

The thin client architecture offloads most application logic and functionality from clients to stationary servers. In the thin client architecture, applications in stationary servers are usually mobile-aware and optimized for mobile client devices. The thin client architecture is especially suitable for small PDA applications. A thin client can refer to either a software program or to an actual computer that relies heavily on another computer to do most of its work. A thin client is part of a network, and the client software or computer acts as an interface, while the network server computer does all the real work. In the case of a computer, a thin client is unable to perform many functions on its own. A thin client computer may be a machine designed only for online use, sending and receiving email, and surfing the net. A thin client computer may also be part of a larger network, at a company or school for example.

The thin client computer contains enough information to start up and connect to a more powerful network server, and the server computer provides the rest of the computing horsepower. The thin server may not even have a hard drive. If the thin client computer needs to use a program or save a file, it will connect to the network server computer to do so. In software terms, a thin client is a program which is mostly interface. The user of the thin client software sees all the data, tools, and features they would on a normal piece of software, but another program running on a remote server does all the work. The reasons someone might use a

thin client, both hardware and software versions, include reduced cost, ease of maintenance, ease of use, and security.

A thin client is much more simple than a complete computer. In a situation in which many people need to perform a similar task, it is more cost effective to have one network server computer and many inexpensive thin client computers, than to have many complete computers. Because thin clients are relatively simple, it is much easier to diagnose problems and repair them. A standard computer has a lot of parts, and a thin client only has a few. Fewer parts mean fewer things can go wrong.

People who are not as computer literate will have an easier time using a thin client than a standard computer or software program. With fewer features and functions, a thin client means a person has fewer things that they need to learn about. Thin clients are also relatively easy to secure. A thin client user has restricted access to programs or functions that could breach security. Restricting all the real computing power to a single network server also means that all the security can be focused in one place. While a thin client is not the right tool for every job, it has many valuable uses. If a user only needs to perform select tasks and does not need all the functionality of a standard computer or program, a thin client may be the right tool for the job.

7.7.2 Full Client Architecture

The full client architecture emulates server functions on the client devices and, therefore, is able to minimize the uncertainty of connectivity and communications. Mobile clients must be able to use networks with rather unpleasant characteristics: intermittence, low bandwidth, high latency, or high expense. The connectivity with one or more of these properties is referred to as weak connectivity.

In the extreme case, mobile clients will be forced to work under the disconnected mode. The ability to operate in disconnected mode can be useful even when connectivity is available. For example, disconnected operations can extend battery life by avoiding wireless transmission and reception. It can reduce network charges, an important feature when charge rates are high. It allows radio silence to be maintained, a vital capability in military applications.

Finally, it is a viable fallback position when network characteristics degrade beyond usability. A full client architecture can be used to effectively support the disconnected or weakly connected clients. Compared to a thin client architecture, the full client architecture is at the other extreme of the range of extended client-server model. The full client architecture supports the emulation of functions of servers at the client host so that applications can be executed without fully connecting to remote servers.

7.7.3 Flexible Client-Server Architecture

Flexible client-server architecture generalizes both thin client and full client architectures in that the roles of clients and servers and application logic can be dynamically relocated and performed on mobile and stationary hosts (see Figure 7-6). In the flexible architecture, the distinction between clients and servers may be temporarily blurred for purposes of performance and availability. Furthermore, the connection between clients and servers can be dynamically established during the execution of applications.

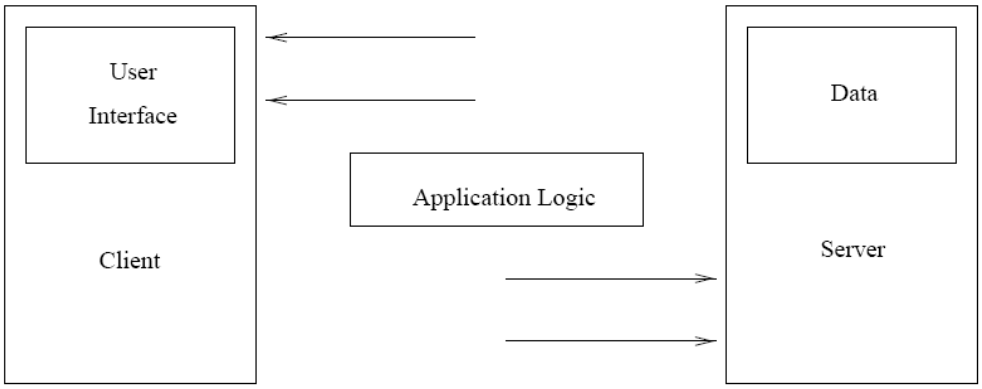

Figure 7-6: A flexible client-server computing

7.8 Mobile Data Access

Mobile data access enables the delivery of server data and the maintenance of client-server data consistency in a mobile and wireless environment. Efficient and consistent data access in mobile environments is a challenge research area because of weak connectivity and resource constraints. The data access strategies in a mobile information system can be characterized by delivery modes, data organizations, and consistency requirements, etc. The mode for server data

delivery can be server-push, client-pull, or hybrid. The server-push delivery is initiated by server functions that push data from the server to the clients. The client-pull delivery is initiated by client functions which send requests to a server and "pull" data from the server in order to provide data to locally running applications. The hybrid delivery uses both server-push and client- pull delivery. The data organizations include mobility-specific data organizations like mobile database fragments in the server storage and data multiplexing and indexing in the broadcast disk. The consistency requirements range from weak consistency to strong consistency. Figure 7-7 illustrates the paradigm of mobile data access for mobile information access. In this section, we will examine various proposed approaches that offer the new paradigm of data access in mobile client-server information systems.

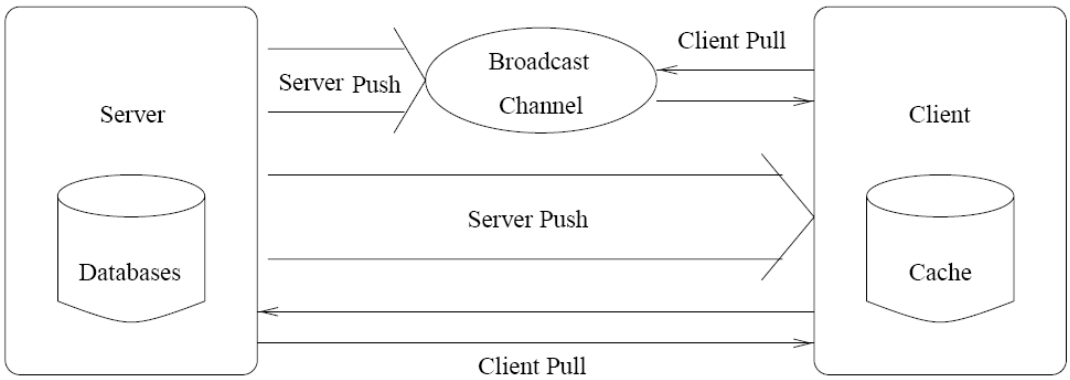

Figure 7-7: A Mobile Data Access Paradigm

7.8.1 Server Data Dissemination

In many applications (e.g., Web access), the downstream data volume from servers to clients is much greater than the upstream data volume from clients back to servers. The unbalanced communications are referred to as asymmetrical communications between clients and servers. A challenge problem in supporting applications with asymmetrical communications is how to deliver server data and information to a large number of clients. To address this scalability problem, a new information system architecture that exploits broadcast-based dissemination capability of communications has been proposed. The central idea is that the servers exploit the downstream communication capacity in bandwidth by broadcasting data to multiple clients. This arrangement is called a push-based

architecture where data is pushed from the server to the clients. In contrast, most traditional client-server information systems use pull-based data delivery to provide data to locally running applications.

7.8.2 Client Cache Management

Caching of frequently-accessed data items is an important technique that reduces contention and improves query response times on narrow bandwidth wireless links. The cached data can also support disconnected or intermitted connected operations. However, cache pre-fetching and consistency strategies can be greatly affected by the disconnection or weak connectivity of mobile clients. The weak connectivity makes cache coherence expensive due to communication latency and intermittent failures. Pre-fetching (or hoarding) data into the client cache prior to disconnection is a difficult challenge in mobile client-server computing. This subsection describes an automated hoarding approach and two cache validation mechanisms.

a) Automated Hoarding

A useful solution to support disconnected operations is hoarding, in which nonlocal files are cached on the client cache prior to disconnection. The difficult issue for hoarding is which files should be selected and stored locally. Possible solutions include choosing the most recently referenced files or asking the user to participate at least peripherally in managing hoard contents. The former approach might be wasteful of scarce hoard space, while the latter requires more expertise and involvement that most users are willing to offer. The automated predictive hoarding is based on the idea that a system can observe user behavior, make inferences about the semantic relationships between files, and use those inferences to aid the user.

b) Varied Granularity of Cache Coherence

Consistency methods in traditional client-server architecture can be divided into two categories: (1) callback approach when servers send invalidation messages directly to the clients that have cached the data items to be updated and (2) detection approach when clients send queries to servers to validate cached data. The difficulty of using these traditional methods in mobile environments is due to

the disconnection and weak connectivity of clients. Frequently disconnected clients make it very ineffective to use the detection approach. On the other hand, the classic callback approach may also be very expensive, due to network latency or intermittent failures. After a long disconnection, many data items at the server side may have been updated. In this case, the time for the callback invalidation of each data item can be substantial on a low network.

c) Cache Invalidation Reports

A dissemination-based approach to the problem of invalidating caches in wireless environments is by utilizing wireless broadcast channels. In this approach, a server periodically broadcasts an invalidation report that reports data items which have been changed. Rather than querying a server directly regarding the validation of cached copies, clients listen to these invalidation reports over broadcast channels. This approach is attractive because a server does not need to know the location and connection status of its clients, and the clients do not need to establish an "uplink" connection to the server to invalidate their caches.

7.9 Performance Issues

With the availability of wireless interface cards, mobile users are no longer required to remain confined within the static network premises to get network access. A host that can move while retaining its network connection is a mobile host. Mobility is not the same as wirelessness. A mobile host is the one who has the ability to communicate anytime, anywhere. On the other hand a wireless host is one which is physically free from a communication link, which is a capability of the physical media in use. We now turn our attention to the limitations and challenges of mobile wireless networks. The technical challenges that mobile computing must overcome to achieve its potential are far from trivial. The following are the limitations and challenges:

- Wireless bandwidth
- Unreliable wireless link
- Mobility of hosts
- Available storage on Mobile computers
- Disconnections