Contents

the cost of installing LAN cabling and ease the task of relocation or otherwise modifying the network's structure. When Wireless LAN (WLAN) was first introduced in the market, the cost per node was higher than the cost of its counterpart in the wired domain. However, as time progressed, the cost per node started dropping, making wireless LAN quite attractive. Slowly WLAN started becoming popular and many companies started offering products. The question of interoperability between different wireless LAN products became critical. IEEE Standards committee took the responsibility to form the standard for WLAN. As a result the IEEE 802.11 series of standards emerged.

WLAN uses the unlicensed Industrial, Scientific, and Medical (ISM) band that different products can use as long as they comply with certain regulatory rules. These rules cover characteristics such as radiated power and the manner in which modulation occurs. The ISM bands specified by the ITU-R are: 6.765–6.795 MHz, 13.553–13.567 MHz, 26.957–27.283 MHz, 40.66–40.70 MHz, 433.05–434.79 MHz, 902–928 MHz, 2.400–2.500 GHz, 5.725–5.875 GHz, 24.00–24.25 GHz, 61.00–61.5 GHz, 122–123 GHz, 244–246 GHz. WLAN uses 2.4 GHz and 5.8 GHz ISM bands. WLAN works both in infrastructure mode and ad hoc mode. WLAN is also known as Wireless Fidelity or WiFi in short. There are many products which use these unlicensed bands along with WLAN; examples could be cordless telephone, microwave oven, etc.

1.2.4 Evolution of Wireless PAN

Wireless technology offers convenience and flexibility. Some people will call this freedom from being entangled with the wire. The success of wireless technology in cellular telephones or Wireless MAN (Metropolitan Area Network) made people think of using the technique in Wireless LAN and Wireless Personal Area Network (WPAN). Techniques for WPANs are infrared and radio waves. Most of the laptop computers support communication through infrared, for which standards have been formulated by IrDA (Infrared Data Association—www.irda.org). Through WPAN, a PC can communicate with another IrDA device like another PC or a Personal Digital Assistant (PDA) or a Cellular phone.

The other best known PAN technology standard is Bluetooth. Bluetooth uses radio instead of infrared. It offers a peak over the air speed of about 2.1 Mbps over a short range of about 100 meters (power dependent). The advantage of radio wave is that unlike infrared it does not need a line of sight. WPAN works in ad hoc mode only.

1.3 Mobile Computing

Mobile computing can be defined as a computing environment of physical mobility. The user of a mobile computing environment will be able to access data, information, or other logical objects from any device in any network while on the move. A mobile computing system allows a user to perform a task from anywhere using a computing device in the public (the Web), corporate (business information) and personal information spaces (medical record, address book). While on the move, the preferred device will be a mobile device, while back at home or in the office the device could be a desktop computer. To make the mobile computing environment ubiquitous, it is necessary that the communication bearer is spread over both wired and wireless media. Be it for the mobile

workforce, holidayers, enterprises, or rural population, access to information and virtual objects through mobile computing is absolutely necessary for optimal use of resource and increased productivity.

Mobile computing is used in different contexts with different names. The most common names are:

- *Mobile Computing*: This computing environment moves along with the user. This is similar to the telephone number of a GSM (Global System for Mobile communication) phone, which moves with the phone. The offline (local) and real-time (remote) computing environment will move with the user. In real-time mode the user will be able to use all his remote data and services online.
- *Anywhere, Anytime Information*: This is the generic definition of ubiquity, where the information is available anywhere, all the time.
- *Virtual Home Environment*: Virtual Home Environment (VHE) is defined as an environment in a foreign network such that the mobile users can experience the same computing experience as they have in their home or corporate computing environment. For example, one would like to keep the room heater on when one has stepped outside for about 15 minutes.
- *Nomadic Computing*: The computing environment is nomadic and moves along with the mobile user. This is true for both local and remote services.
- *Pervasive Computing*: A computing environment, which is pervasive in nature and can be made available in any environment.
- *Ubiquitous Computing*: A (nobody will notice its presence) everyplace computing environment. The user will be able to use both local and remote services.
- *Global Service Portability*: Making a service portable and available in every environment. Any service of any environment will be available globally.
- *Wearable Computers*: Wearable computers can be worn by humans like a hat, shoe or clothes (these are wearable accessories). Wearable computers need to have some additional attributes compared to standard mobile devices. Wearable computers are always on; operational while on the move; hands-free, context-aware (with different types of sensors). Wearable computers need to be equipped with proactive attention and notifications. The ultimate wearable computers will have sensors implanted in the body and supposedly integrate with the human nervous system. These are part of a new discipline of research categorized by "Cyborg" (Cyber Organism).

1.3.1 Mobile Computing Functions

We can define a computing environment as mobile if it supports one or more of the following characteristics:

- *User Mobility*: The user should be able to move from one physical location to another and use the same service. The service could be in a home or remote network. For example, a user moves from London to New York and uses Internet to access the corporate application the same way the user uses it in the home office.
- *Network Mobility*: Network mobility deals with two types of use-cases. In one use-case, the user is moving from one network to another and uses the same service seamlessly. An example could be a user moving from a WiFi network within the university

3G network outside while using the same online service.

In other use-case of network mobility, the network itself is mobile like in a Mobile Ad hoc Network (MANET). In MANET, each node in the network is a combination of a host and a router. As the nodes move, the routers within the network also move changing the routing table structure. These types of networks are used in battlefields or sensor networks, where routers/nodes are constantly moving.

- *Bearer Mobility*: The user should be able to move from one bearer to another and use the same service. An example could be a user using a service through WAP bearer in his home network in Bangalore. He moves to Coimbatore where WAP is not supported and switches over to the voice or SMS (short message service) bearer to access the same application.
- *Device Mobility*: The user should be able to move from one device to another and use the same service. An example could be sales representatives using their desktop computer in their home office. During the day while they are on the street they would like to use their Palmtop to access the application.
- *Session Mobility*: A user session should be able to move from one user-agent environment to another. An example could be a user using his service through a CDMA (Code Division Multiple Access) 1X network. The user entered into the basement to park the car and got disconnected from his CDMA network. He goes to his home office and starts using the desktop. The unfinished session in the CDMA device moves from the mobile device to the desktop computer.
- *Agent Mobility*: The user-agent or the applications should be able to move from one node to another. Examples could be aglets, crawler software, or even a malicious worm or virus software that moves from one node to another. There is another use-case of mobile agent in the Cloud Computing paradigm, where applications will be moving from platform to platform and infrastructure to infrastructure depending on temporal and economic considerations. In Cloud Computing, there will not be any fixed association between the application and the host running it—software agents in the cloud will constantly be mobile.
- *Host Mobility*: The user device can be either a client or server. When it is a server or host, some of the complexities change. In case of host mobility, mobility of the IP needs to be taken care of.

The mobile computing functions can be logically divided into the following major segments (Fig. 1.1):

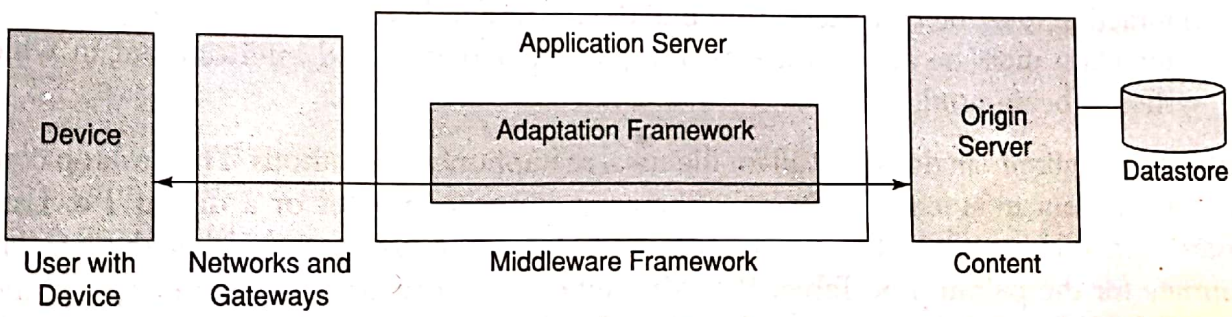

Figure 1.1 Mobile Computing Functions

1. *User with device*: This means that this could be a fixed device like a desktop computer in an office or a portable device like mobile phone. Example: laptop computers, desktop computers, fixed telephone, mobile phones, digital TV with set-top box, palmtop computers, pocket PCs, two-way pagers, handheld terminals, etc.
2. *Network*: Whenever a user is mobile, he will use different networks at different locations at different times. Example: GSM, CDMA, iMode, Ethernet, Wireless LAN, Bluetooth, etc.
3. *Gateway*: This acts as an interface between different transport bearers. These gateways convert one specific transport bearer to another. Example: From a fixed phone (with voice interface) we access a service by pressing different keys on the telephone. These keys generate DTMF (Dual Tone Multi Frequency) signals. These analog signals are converted into digital data by the IVR (Interactive Voice Response) gateway to interface with a computer application. Other examples will be WAP gateway, SMS gateway, etc.
4. *Middleware*: This is more of a function rather than a separate visible node. In the present context, middleware handles the presentation and rendering of the content on a particular device. It may optionally also handle the security and personalization for different users.
5. *Content*: This is the domain where the origin server and content is. This could be an application, system, or even an aggregation of systems. The content can be mass market, personal or corporate content. The origin server will have some means of accessing the database and storage devices.

1.3.2 Mobile Computing Devices

The device for mobile computing can be either a computing or a communication device. In the computing device category it can be a desktop, laptop, or a palmtop computer. On the communication device side it can be a fixed line telephone, a mobile telephone or a digital TV. Usage of these devices are becoming more and more integrated into a task flow where fixed and mobile, computing and communication functions are used together. The device is a combination of hardware and software; the hardware is technically called the User Equipment (UE) with software inside, which functions as an agent to connect to the remote service—this software is called a User Agent (UA). One of the most common UA today is a Web browser. When computing technology is embedded into equipment, Human-Computer Interaction (HCI) plays a critical role in effectiveness, efficiency, and user experience. This is particularly true as mobile information and communication devices are becoming smaller and more restricted with respect to information presentation, data entry and dialogue control. The human computer interface challenges are:

1. Interaction must be consistent from one device to another.
2. Interaction must be appropriate for the particular device and environment in which the system is being used.

Note: The requirement does not call for identical metaphors and methods. The desktop computer allows for different interaction techniques than a palmtop computer or a digital TV. Using the keyboard and a mouse may be appropriate for the desktop computer. Using the pen may be appropriate for the palmtop or Tablet PC. Microphones and speakers may be appropriate for a fixed or mobile phone. A remote control on the other hand will be more desirable for a digital TV.

1.4 DIALOGUE CONTROL

In any communication there are two types of user dialogues. These are long session-oriented transactions and short sessionless transactions. An example of a session-oriented transaction is: Reading a few pages from one chapter of a book at a time. Going to a particular page directly through an index and reading a particular topic can be considered a short sessionless transaction. Selection of the transaction mode will depend on the type of device we use. A session may be helpful in case of services offered through computers with large screens and mouse. For devices with limited input/output like SMS for instance, short sessionless transactions may be desired.

For example, consider enquiring about your bank balance over the Internet. In case of Internet banking through a desktop computer, the user has to go through the following minimum dialogues:

1. Enter the URL of the bank site.
2. Enter the account number/password and login into the application.
3. Select the balance enquiry dialogue and see the balance.
4. Logout from Internet banking.

This example is a session-oriented transaction. Using short sessionless transactions, the same objective can be met through a single dialogue. In a short sessionless transaction, the user sends an SMS message, 'mybal' to the system and receives the information on balance. The application services all the five dialogue steps as one dialogue. In this case steps like authentication and selection of transactions need to be performed in smarter ways. For example, user authentication will be done through the user's mobile number. It can be assumed that mobile devices are personal, therefore, authenticating the mobile phone implies authenticating the user account.

1.5 NETWORKS

Mobile computing will use different types of networks. These can be fixed telephone networks, GSM, GPRS, ATM (Asynchronous Transfer Mode), Frame Relay, ISDN (Integrated Service Digital Network), CDMA, CDPD (Cellular Digital Packet Data), DSL (Digital Subscriber Loop), Dial-up, WiFi (Wireless Fidelity), 802.11, Bluetooth, Ethernet, Broadband, etc.

1.5.1 Wireline Networks

This is a network, which is designed over wire or tangible conductors. This network is called fixedline or wireline network. Fixed telephone networks over copper and fiber-optic will be part of this network family. Broadband networks over Digital Subscriber Line (DSL) or cable will also be part of wireline networks. Wireline networks are generally public networks and cover wide areas. Though microwave or satellite networks do not use wire, when a telephone network uses microwave or satellite as part of its longhaul transmission infrastructure, it is considered part of wireline networks. When we connect to Internet Service Providers (ISP), it is generally a wireline network. The Internet backbone is a wireline network as well.

1.5.2 Wireless Networks

Mobile networks are called wireless network. These include wireless networks used by radio taxis, one-way and two-way pager, cellular phones. Examples will be PCS (Personal Cellular System), AMPS (Advanced Mobile Phone System), GSM, CDMA, DoCoMo, GPRS, etc. WiLL (Wireless in Local Loop) networks using different types of technologies are part of wireless networks as well. In a wireless network the last mile is wireless and works over radio interface. In a wireless network, other than the radio interface, rest of the network is wireline and is generally called the PLMN (Public Land Mobile Network).

1.5.3 Ad hoc Networks

In Latin, *ad hoc* means "for this purpose only". An ad hoc (or spontaneous) network is a small area network, especially one with wireless or temporary plug-in connections. In these networks some of the devices are part of the network only for the duration of a communication session. An ad hoc network is also formed when mobile or portable devices operate in close proximity to each other or with the rest of the network. When we beam a business card from our PDA (Personal Digital Assistant) to another, or use an IrDA port to print documents from our laptop, we have formed an ad hoc network. The term ad hoc has been applied to networks in which new devices can be quickly added using, for example, Bluetooth or wireless LAN (802.11). In these networks, devices communicate with the computer and other devices through wireless transmission. Typically based on short-range wireless technology, these networks don't require subscription services or carrier networks.

1.5.4 Bearers

For different type of networks, there are different types of transport bearers. These can be TCP/IP, HTTP, protocols or dial-up connection. For GSM it could be SMS, USSD (Unstructured Supplementary Service Data) or WAP. For mobile or fixed phone, it will be Voice.

2.4 ARCHITECTURE FOR MOBILE COMPUTING

In mainframe computers many mission critical systems use a Transaction Processing (TP) environment. At the core of a TP system, there is a TP monitor software. In a TP system, all the terminals—VDU (Visual Display Terminal), POS (Point of Sale Terminal), printers, etc., are terminal resources (objects). There are different processing tasks, which process different transactions or messages; these are processing resources (objects). Finally, there are database resources. A TP monitor manages terminal resources, database objects and coordinates with the user to pick up the right processing task to service business transactions. The TP monitor manages all these objects and connects them through policies and rules. A TP monitor also provides functions such as queuing, application execution, database staging, and journaling. When the world moved from large expensive centralized mainframes to economic distributed systems, technology moved towards two-tier conventional client/server architecture. With growth in cheaper computing power and penetration of Internet-based networked systems, technology is moving back to centralized server-based architecture. The TP monitor architecture is having a reincarnation in the form of three-tier software architecture.

In the early days of mainframes, the TP monitor and many other interfaces were proprietary. Even the networked interfaces to different terminals were vendor-specific and proprietary. The most successful early TP system was the reservation system for the American Airlines. This was over a Univac computer using U100 protocol. For IBM TP environment, which runs on OS/390 known as CICS (Customer Information Control System), the network interface was through SNA. In India DoT (Department of Telecommunication; currently BSNL and MTNL) launched the 197 telephone directory enquiry system in 1986, it used TPMS (Transaction Processing Management System) on ICL mainframe running VME operating system. The network interface was over X.25 interface.

The network-centric mobile computing architecture uses three-tier architecture as shown in Figure 2.1. In the three-tier architecture, the first layer is the User Interface or Presentation Tier. This layer deals with user facing device handling and rendering. This tier includes a user system interface where user services (such as session, text input, dialog and display management) reside. The second tier is the Process Management or Application Tier. This layer is for application programs or process management where business logic and rules are executed. This layer is capable of accommodating hundreds of users. In addition, the middle process management tier controls transactions and asynchronous queuing to ensure reliable completion of transactions. The third and final tier is the Database Management or Data Tier. This layer is for database access and management. The three-tier architecture is better suited for an effective networked client/server design. It provides increased *performance*, *flexibility*, *maintainability*, *reusability*, and *scalability*, while hiding the complexity of distributed processing from the user. All these characteristics have made three-tier architectures a popular choice for Internet applications and net-centric information systems. Centralized process logic makes administration and change management easier by localizing changes in a central place and using them throughout the system.

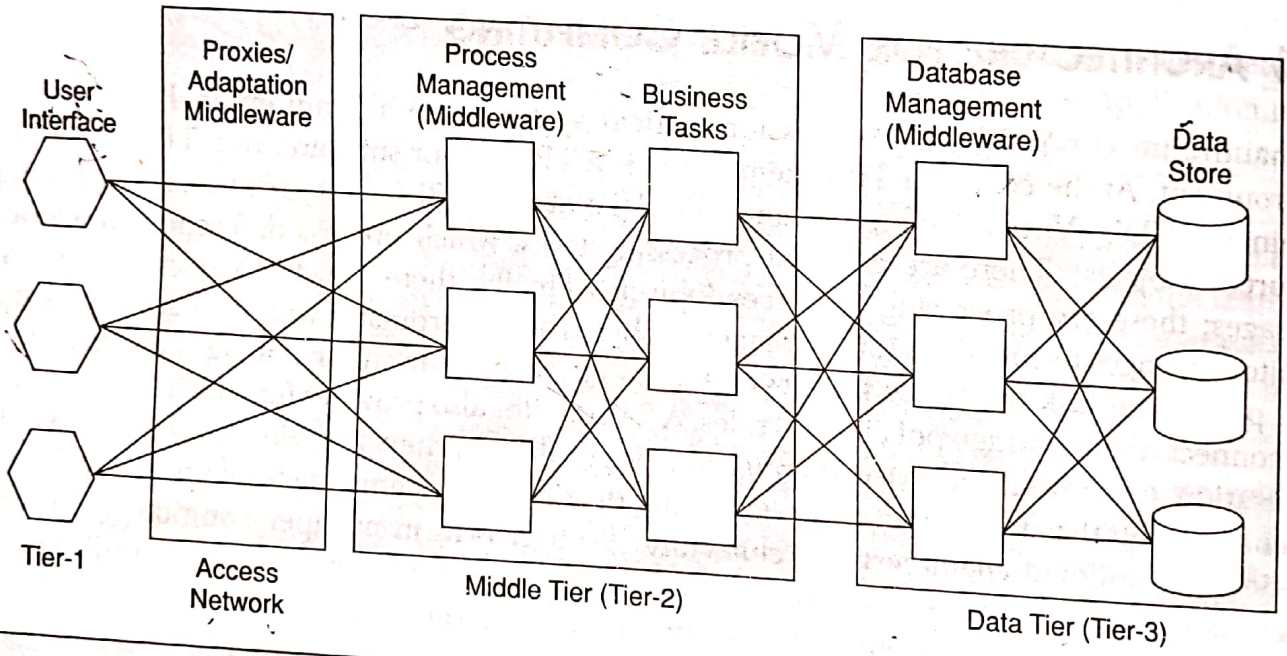

Figure 2.1 Three-tier Architecture for Mobile Computing

2.5 THREE-TIER ARCHITECTURE

To design a system for mobile computing, we need to keep in mind that the system will be used through any network, bearer, agent and device. To have universal access, it is desirable that the server is connected to a ubiquitous network like the Internet. To have access from any device, a web browser is desirable. The reason is simple; web browsers are ubiquitous, they are present in any computer. The browser agent can be Internet Explorer or Netscape Navigator or Mozilla or any other standard agent. Also, the system should preferably be context aware. We will discuss context awareness later.

We have introduced the concept of three-tier architecture. We have also discussed why it is necessary to go for Internet and three-tier architecture for mobile computing. The important question is what a mobile three-tier application actually should consist of. Figure 2.2 depicts a three-tier architecture for a mobile computing environment. These tiers are presentation tier, application tier and data tier. Depending upon the situation, these layers can be further sublayered.

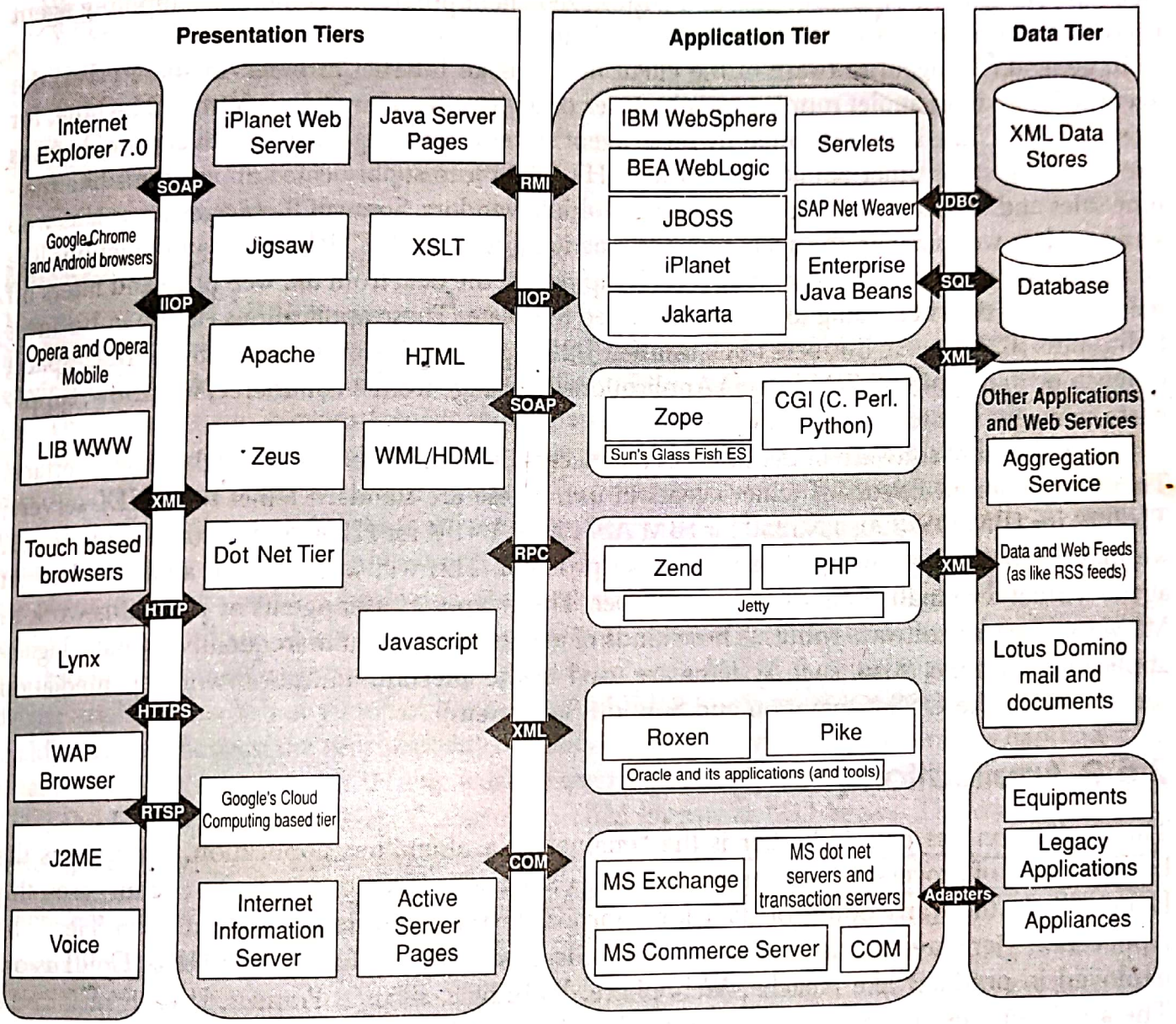

Figure 2.2 The Mobile Computing Architecture

2.5.1 Presentation (Tier-1)

This is the user facing system in the first tier. This is the layer of agent applications and systems. These applications run on the client device and offer all the user interfaces. This tier is responsible for presenting the information to the end user. Humans generally use visual and audio means to

receive information from machines (with some exceptions like vibrator in mobile phones). Humans also use keyboard (laptop computers, cell phones), pen (tablet PC, palmtops), touch screen (kiosks), or Voice (telephone) to feed the data to the system. In the case of the visual, the presentation of information will be through a screen. Therefore, the visual presentation will relate to rendering on a screen. 'Presentation Tier' includes web browsers (like Mozilla, Lynx, Internet Explorer and Netscape Navigator), WAP browsers and customized client programs. A mobile computing agent needs to be context-aware and device independent.

In general, the agent software in the client device is an Internet browser. In some cases, the agent software is an applet running on a browser or a virtual machine (Java Virtual Machine, for example). The functions performed by these agent systems can range from relatively simple tasks like accessing some other application through HTTP API, to sophisticated applications like real-time sales and inventory management across multiple vendors. Some of these agents work as web scrapers. In a web scraper, the agent embeds functionality of the HTTP browser and functions like an automated web browser. The scraper picks up part of the data from the web page and filters off the remaining data according to some predefined template. These applications can be in Business to Business (B2B) space, Business to Consumer (B2C) space or Business to Employee (B2E) space, or machine to machine (M2M) space. Applications can range from e-commerce, workflow, supply chain management to legacy applications.

There are agent software in the Internet that access the remote service through telnet interface. There are different flavors of telnet agents in use. These are standard telnet for UNIX servers; TN3270 for IBM OS/390; TN5250 for IBM AS/400 or VT3K for HP3000. For some applications, we may need an agent with embedded telnet protocol. This will work like an automated telnet agent (virtual terminal) similar to a web scraper. These types of user agents or programs work as M2M interface or software robots. These kinds of agents are used quite frequently to make legacy applications mobile. Also, such systems are used in the telecommunication world as mediation servers within the OSS (Operation and Support Subsystem).

2.5.2 Application (Tier-2)

The application tier or middle tier is the "engine" of a ubiquitous application. It performs the business logic of processing user input, obtaining data, and making decisions. In certain cases, this layer will do the transcoding of data for appropriate rendering in the Presentation Tier. The Application Tier may include technology like CGIs, Java, JSP, .NET services, PHP or ColdFusion, deployed in products like Apache, WebSphere, WebLogic, iPlanet, Pramati, JBOSS or ZEND. The application tier is presentation and database-independent.

In a mobile computing environment, in addition to the business logic there are quite a few additional management functions that need to be performed. These functions relate to decisions on rendering, network management, security, datastore access, etc. Most of these functions are implemented using different middleware software. A middleware framework is defined as a layer of software, which sits in the middle between the operating system and the user facing software. Stimulated by the growth of network-based applications and systems, middleware technologies are gaining increasing importance in net-centric computing. In case of net-centric architecture, a middleware framework sits between an agent and business logic. Middleware covers a wide range of software systems, including distributed objects and components, message-oriented

communication, database connectors, mobile application support, transaction drivers, etc. Middleware can also be considered as a software gateway connecting two independent open objects.

It is very difficult to define how many types of middleware are there. A very good description of middleware is available in Carnegie Mellon University Software Engineering Institute (http://www.sei.cmu.edu/str/descriptions/middleware.html), which readers can refer to.

We can group middleware into the following major categories:

1. Message-oriented Middleware.
2. Transaction Processing Middleware.
3. Database Middleware.
4. Communication Middleware.
5. Distributed Object and Components.
6. Transcoding Middleware.

Message-oriented Middleware (MOM)

Message-oriented Middleware is a middleware framework that loosely connects different applications through asynchronous exchange of messages. A MOM works over a networked environment without having to know what platform or processor the other application is resident on. The message can contain formatted data, requests for action, or unsolicited response. The MOM system provides a message queue between any two interoperating applications. If the destination process is out of service or busy, the message is held in a temporary storage location until it can be processed. MOM is generally asynchronous, peer-to-peer, and works in publish/subscribe fashion. In the publish/subscriber mode one or many objects subscribe to an event. As the event occurs, it will be published by the loosely coupled asynchronous object. The MOM will notify the subscribers about this event. However, most implementations of MOM support synchronous (request/response) message passing as well. MOM is most appropriate for event-driven applications. When an event occurs, the publisher application hands on to the messaging middleware application the responsibility of notifying subscribers that the event has happened. In a net-centric environment, MOM can work as the integration platform for different applications. An example of MOM is Message Queue from IBM known as MQ Series. The equivalent from Java is JMS (Java Message Service).

Transaction Processing (TP) Middleware

Transaction Processing Middleware provides tools and an environment for developing transaction-based distributed applications. An ideal TP system will be able to input data into the system at the point of information source and the output of the system is delivered at the point of information sink. In an ideal TP system, the device for input and output can potentially be different (Fig. 2.3). Also, the output can be an unsolicited message for a device. TP is used in data management, network access, security systems, delivery order processing, airline reservations, customer service, etc., to name a few. TP systems are generally capable of providing services to thousands of clients in a distributed client/server environment. CICS (Customer Information Control System) is one of the early TP application systems on IBM mainframe computers.

TP middleware maps numerous client requests through application-service routines to different application tasks. In addition to these processing tasks, TP middleware includes numerous management features, such as restarting failed processes, dynamic load balancing and ensuring

consistency of distributed data. TP middleware is independent of the database architecture. TP middleware optimizes the use of resources by multiplexing many client functions on to a much smaller set of application-service routines. This also helps in reducing the response time. TP middleware provides a highly active system that includes services for delivery-order processing, terminal and forms management, data management, network access, authorization, and security. In the Java world and net-centric systems, transaction processing is done through the J2EE application server with the help of entity and session beans.

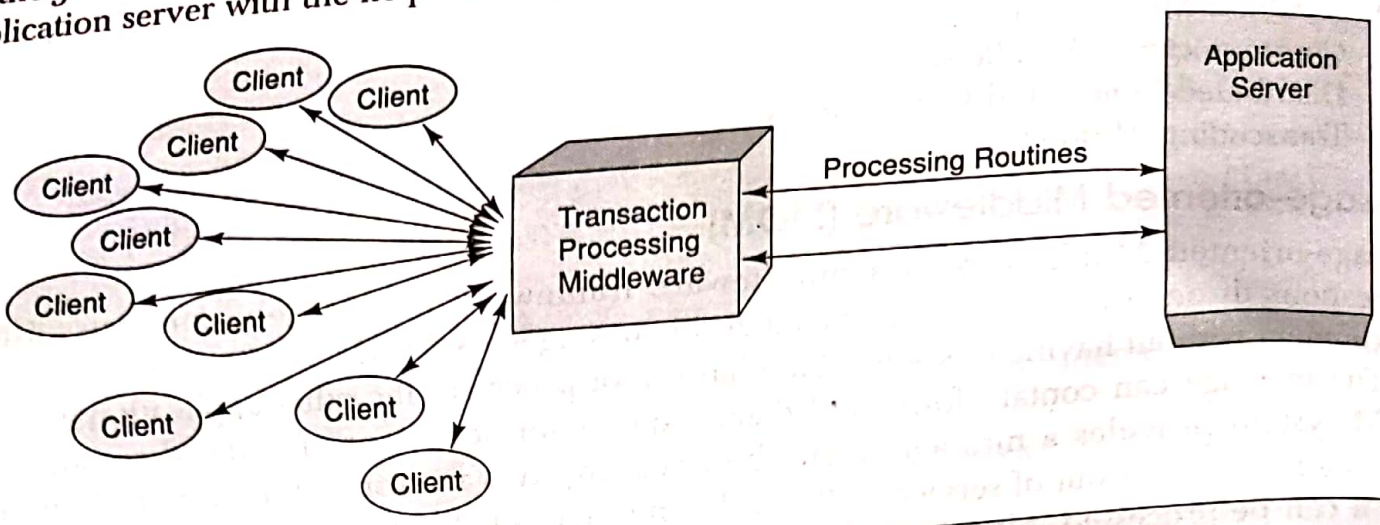

Figure 2.3 Transaction Processing Middleware

Model View Controller (MVC): Java uses the MVC architectural pattern which is an example of transaction processing system. It splits an application into separate layers, viz., presentation, domain logic, and data access. *Model* is the domain-specific representation of the information on which the application operates. Domain logic manipulates and adds meaning to the raw data. MVC does not specifically mention the data access layer because it is assumed to be encapsulated by the model. *View* is responsible for rendering the model into a form suitable for interaction and understood by the user, typically a user interface element. *Controller* manages processes and responds to events, typically user actions, and may invoke changes on the model. In the context of Web applications and J2EE, the MVC pattern is widely used. In Web applications, where the view is the actual HTML page, and the controller is the code which gathers dynamic data and generates the content within the HTML, the model is represented by the actual content, usually stored in a database.

Communication Middleware

Communication Middleware is used to connect one application to another through some communication middleware, like connecting one application to another through telnet. These types of middleware are quite useful in the telecommunication world. There are many elements in the core telecommunication network where the user interface is through telnet. A mediation server automates the telnet protocol to communicate with these nodes in the network. Another example could be to integrate legacy applications through proprietary communication protocols like TN5250 or TN3270.

Distributed Object and Components

An example of distributed objects and components is CORBA (Common Object Request Broker Architecture). CORBA is an open distributed object computing infrastructure being standardized by the Object Management Group (http://www.omg.org). CORBA simplifies many common network programming tasks used in a net-centric application environment. These are object registration, object location, and activation; request demultiplexing; framing and error-handling; parameter marshalling and demarshalling; and operation dispatching. CORBA is vendor-independent infrastructure. A CORBA-based program from any vendor on almost any computer, operating system, programming language and network, can interoperate with a CORBA-based program from the same or another vendor, on almost any other computer, operating system, programming language and network. CORBA is useful in many situations because of the easy way that CORBA integrates machines from so many vendors, with sizes ranging from mainframes through minis and desktops to hand-helds and embedded systems. One of its most important, as well as the most frequent uses is in servers that must handle a large number of clients, at high hit rates, with high reliability.

Transcoding Middleware

Transcoding Middleware is used to transcode one format of data to another to suit the need of the client. For example, if we want to access a web site through a mobile phone supporting WAP, we need to transcode the HTML page to WML page so that the mobile phone can access it. Another example could be accessing a map from a PDA. The same map, which can be shown in a computer, needs to be reduced in size to fit the PDA screen. Technically transcoding is used for content adaptation to fit the need of the device. Content adaptation is also required to meet the network bandwidth needs. For example, some frames in a video clip need to be dropped for a low bandwidth network. Content adaptation used to be done through proprietary protocols. To allow interoperability, IETF has accepted the Internet Content Adaptation Protocol (ICAP). ICAP is now standardized and described in RFC3507.

Internet Content Adaptation Protocol (ICAP)

Popular web servers are required to deliver content to millions of users connected at ever-increasing bandwidths. Progressively, content is being accessed through different devices and agents. A majority of these services have been designed keeping the desktop user in mind. Some of them are also available for other types of protocols. For example, there are a few sites that offer contents in HTML and WML to service desktop and WAP phones. However, the model of centralized services that are responsible for all aspects of every client's request seems to be reaching the end of its useful life. ICAP, the Internet Content Adaptation Protocol, is a protocol aimed at providing simple object-based content vectoring for HTTP services. ICAP is a lightweight protocol to do transcoding on HTTP messages. This is similar to executing a "remote procedure call" on a HTTP request. The protocol allows ICAP clients to pass HTTP messages to ICAP servers for some sort of transformation. The server executes its transformation service on messages and sends back responses to the client, usually with modified messages. The adapted messages may be either HTTP requests or HTTP responses. For example, before a document is displayed in the agent, it is checked for virus.

There are two major components in ICAP architecture:

1. What are the semantics for the transformation? How do I ask for content adaptation?

2. How is policy of the transformation managed? What kind of adaptation do I ask for and from where? How do I define and manage the adaptation?

ICAP works at the edge part of the network as depicted in Figure 2.4. It is difficult, if not impossible, to define the devices users may like to use to access content from within the Internet. Customized edge delivery of Internet content will help to improve user experience. When applications are delivered from an edge device, end users find that the applications execute more quickly and are more reliable. Typical data flow in an ICAP environment is depicted in Figure 2.4 and described here.

1. A user agent makes a request to an ICAP-capable surrogate (ICAP client) for an object on an origin server.
2. The surrogate sends the request to the ICAP server.
3. The ICAP server executes the ICAP resource's service on the request and sends the possibly modified request, or a response to the request back to the ICAP client.
4. The surrogate sends the request, possibly different from the original client's request, to the origin server.
5. The origin server responds to the request.
6. The surrogate sends the reply (from either the ICAP or the origin server) to the client.

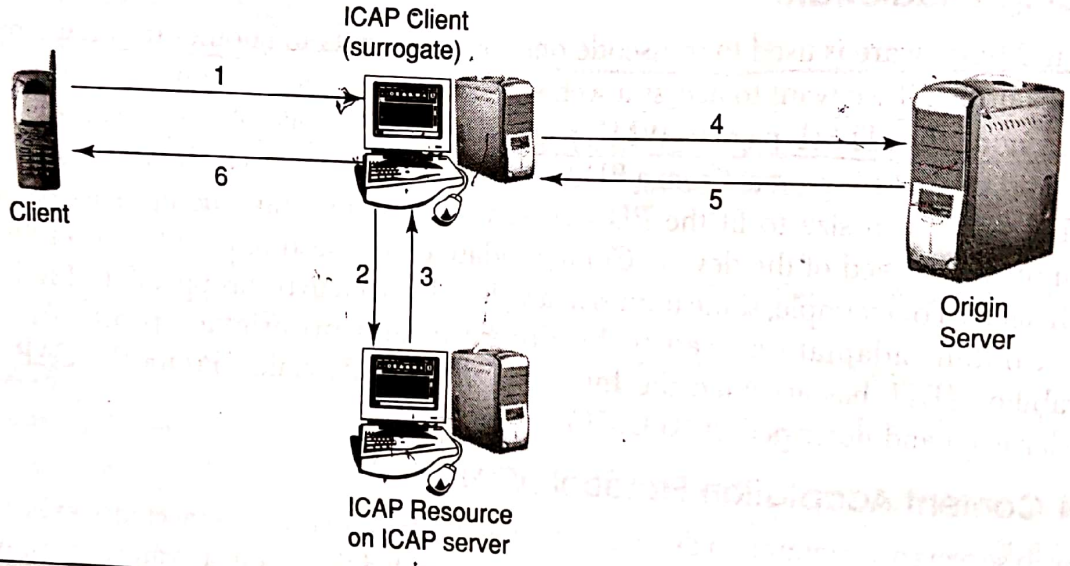

Figure 2.4 Typical Data Flow in an ICAP Environment

It is envisioned that in future, ICAP servers may be available to provide some of the following services:

- Suit content delivery based on network bandwidth.
- Suit content delivery based on device characteristics.
- Language translation based on the user's preference.
- Virus checking for the requested content.
- Content filtering based on sensor rating like PG (parental guidance), R (restricted).
- Local real-time advertisement insertion like television.
- Local real-time advertisement elimination for premium subscribers.

- Wireless protocol translation.
- Anonymous Web usage profiling for a dating service.
- Transcoding or image enhancement.
- Image magnification for the elderly.
- Image size reduction based on device display characteristics.
- Intelligent video condensation by dropping frames.
- Digest production/batch download of Web content.
- Content filtering based on copyright or digital signature.
- Peer-to-Peer compression and encryption of data.

Web Services

As the need for peer-to-peer, application-to-application communication and interoperability grows, the use of Web services on the Internet will also grow. Web services provide a standard means of communication and information exchange among different software applications, running on a variety of platforms or frameworks. Web service is a software system identified by a URI, whose public interfaces and bindings are defined using XML (eXtensible Markup Language). Its definition can be discovered by other software systems connected to the network. Using XML-based messages these systems may then interact with the Web service in a manner prescribed by its definition.

The basic architecture includes Web service technologies capable of:

- Exchanging messages.
- Describing Web services.
- Publishing and discovering Web service descriptions.

The Web services architecture defines the standards for exchange of messages between the service requester and service provider. Service providers are responsible for publishing a description of the services they provide. Requesters must be able to find and discover descriptions of the services.

Software agents in the basic architecture can take on one or all of the following roles:

- Service requester–requests the execution of a Web service.
- Service provider–processes a Web service request.
- Discovery agency–agency through which a Web service description is published and made discoverable.

The interactions involve the publish, find and bind operations. A service is invoked after the description is found, since the service description is required to establish a binding.

2.5.3 Data (Tier-3)

The Data Tier is used to store data needed by the application and acts as a repository for both temporary and permanent data. The data can be stored in any form of datastore or database. These can range from sophisticated relational database, legacy hierarchical database, to even simple text files. The data can also be stored in XML format for interoperability with other systems and datasources. A legacy application can also be considered as a data source or a document through a communication middleware.

Database Middleware

We have discussed that for a mobile computing environment, the business logic should be independent of the device capability. Likewise, though not essential, it is advised that business logic should be independent of the database. Database independence helps in maintenance of the system better. Database middleware allows the business logic to be independent and transparent of the database technology and the database vendor. Database middleware runs between the application program and the database. These are sometimes called database connectors as well. Examples of such middleware will be ODBC, JDBC, etc. Using these middleware, the application will be able to access data from any data source. Data sources can be text files, flat files, spreadsheets, or a network, relational, indexed, hierarchical, XML database, object database, etc., from vendors like Oracle, SQL, Sybase, etc.

SyncML

SyncML protocol is an emerging standard for synchronization of data access from different nodes. When we moved from the conventional client/server model of computing to the net-centric model of computing, we moved from distributed computing to centralized computing with networked access. The greatest benefit of this model is that resources are managed at a centralized level. All the popular mobile devices like handheld computers, mobile phones, pagers and laptops work in an occasionally connected computing mode and access these centralized resources from time to time. In an occasionally connected mode, some data are cached in the local device and accessed frequently. The ability to access and update information on the fly is key to the pervasive nature of mobile computing. Examples are emails and personal information like appointments, address book, calendar, diary, etc. Storing and accessing phone numbers of people from the phone address book is more user-friendly compared to accessing the same from a server. However, managing the appointments database is easier in a server, though caching the same on the mobile client is critical. Users will cache emails into the device for reference. We take notes or draft a mail in the mobile device. For workflow applications, data synchronization plays a significant role. The data in the mobile device and server need to be synchronized. Today vendors use proprietary technology for performing data synchronization. SyncML protocol is the emerging standard for synchronization of data across different nodes. SyncML is a new industry initiative to develop and promote a single, common data synchronization protocol that can be used industry-wide.

The ability to use applications and information on a mobile device, then to synchronize any updates with the applications and information back at the office or on the network, is key to the utility and popularity of mobile computing. The SyncML protocol supports naming and identification of records and common protocol commands to synchronize local and network data. It supports identification and resolution of synchronization conflicts. The protocol works over all networks used by mobile devices, both wireless and wireline. Since wireless networks employ different transport protocols and media, a SyncML will work smoothly and efficiently over:

- HTTP 1.1 (i.e., the Internet).
- WSP (the Wireless Session Protocol, part of the WAP protocol suite).
- OBEX (Object Exchange Protocol, i.e., Bluetooth, IrDA and other local connectivity).
- SMTP, POP3 and IMAP.
- Pure TCP/IP networks.
- Proprietary wireless communication protocols.

4.4 WIRELESS BROADBAND (WIMAX)

Wireless technologies are proliferating in a major way into the first-mile (as computer people call it) or last-mile (as communication people call it) subscriber access, as opposed to twisted-pair local loop. These technologies are generally referred to as (WLL–wireless local loop) or WiLL (wireless in local loop). Wireless local loop is also known as fixed-wireless system. The world is moving towards

a convergence of voice, data and video. This convergence will demand interoperability and high data rate. Keeping this in mind, the IEEE 802 committee set up the 802.16 working group in 1999 to develop wireless broadband or WirelessMAN (wireless metropolitan area network) standards. WirelessMAN offers an alternative to high bandwidth wired access networks like fiber optic, cable modems and DSL (Digital Subscriber Line). WirelessMAN is popularly known as WiMAX (Worldwide Interoperability for Microwave Access). WiMAX provides wireless transmission of data using a variety of transmission modes, from point-to-multipoint links to portable and fully mobile Internet access. This technology provides up to 10 Mbps bandwidth without the need for cables. Figure 4.4 illustrates the WiMAX Architecture; whereas, Fig. 4.5 illustrates a typical WirelessMAN deployment scenario.

The release of WirelessMAN (IEEE 802.16) standards in April 2002 has paved the way for the entry of broadband wireless access as a new bearer to link homes and businesses with core telecommunications networks. WirelessMAN provides network access to buildings through exterior antennas communicating with radio base stations. The technology is expected to provide less expensive access with more ubiquitous broadband access with integrated data, voice and video services. One of the most attractive aspects of wireless broadband technology is that networks can be created in just weeks by deploying a small number of base stations on buildings or poles to create high-capacity wireless access systems. In a wired set up, one physical wire will connect the device with the network. Also, we need to keep many wires reserved for future growth. Therefore, the initial investment in wired infrastructure is very high. Wireless network can grow as the demand increases. At any point in time the number of active users are always a fraction of the number of subscribers. In a wireless environment the number of channels is always low compared to the number of subscribers. This makes wireless technologies very attractive to the service providers.

IEEE 802.16 standardizes the air interface and related functions associated with WLL. Three working groups have been chartered to produce the following standards:

- IEEE 802.16.1–Air interface for 10 to 66 GHz.
- IEEE 802.16.2–Coexistence of broadband wireless access systems.
- IEEE 802.16.3–Air interface for licensed frequencies, 2 to 11 GHz.
- Extensive radio spectrum is available in frequency bands from 10 to 66 GHz worldwide. In a business scenario, 802.16 can serve as a backbone for 802.11 networks. Other possibilities are using 802.16 within the enterprise along with 802.11a, 802.11b or 802.11g.

IEEE 802.16 standards are concerned with the air interface between a subscriber's transceiver station and a base transceiver station. The 802.16 standards are organized into a three-layer architecture.

- The physical layer: This layer specifies the frequency band, the modulation scheme, error-correction techniques, synchronization between transmitter and receiver, data rate and the multiplexing structure.
- The MAC (Media Access Control) layer: This layer is responsible for transmitting data in frames and controlling access to the shared wireless medium through media access control (MAC) layer. The MAC protocol defines how and when a base station or subscriber station may initiate transmission on the channel.
- Above the MAC layer is a convergence layer that provides functions specific to the service being provided. For IEEE 802.16.1, bearer services include digital audio/video multicast, digital telephony, ATM, Internet access, wireless trunks in telephone networks and frame relay.

802.16

IEEE 802.16 standards define how wireless traffic will move between subscribers and core networks.

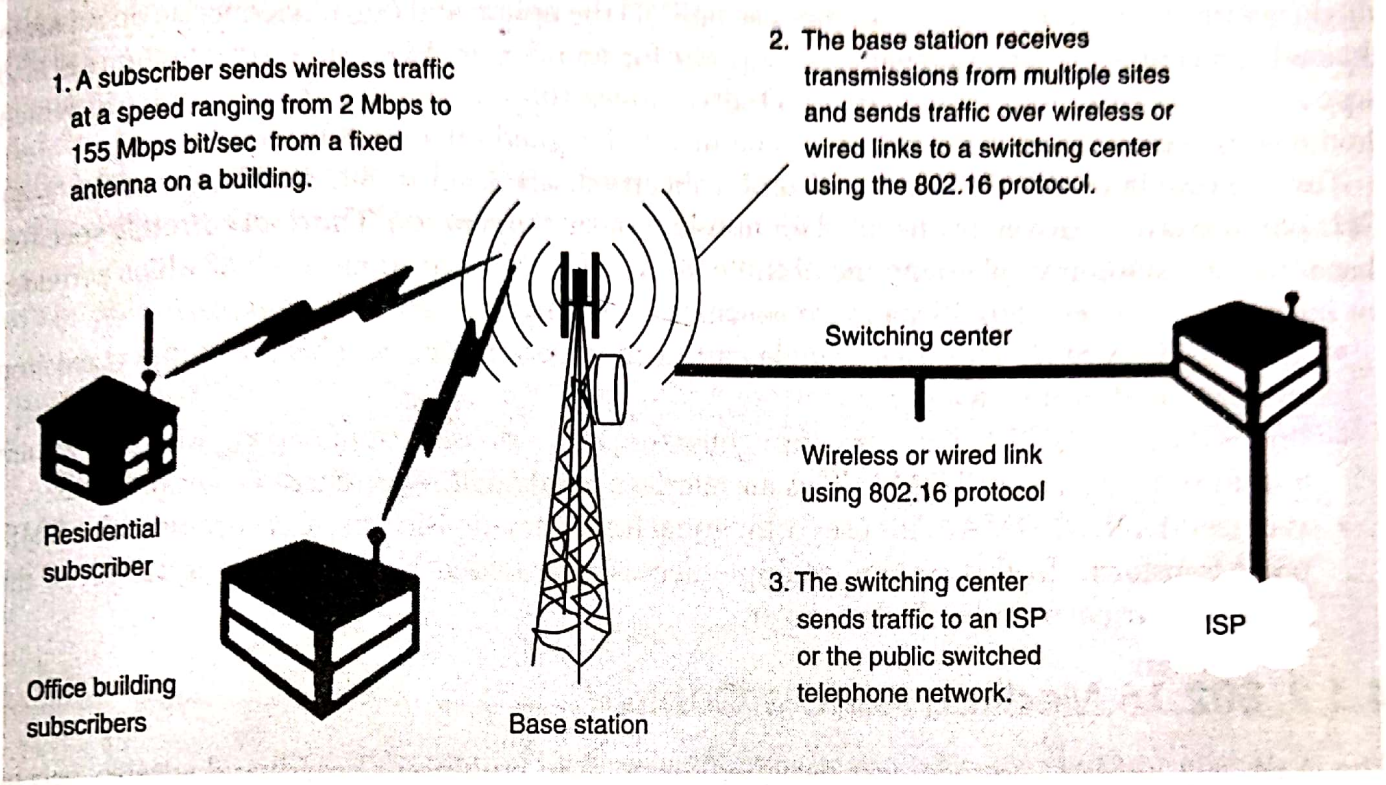

Figure 4.4(a) WiMAX (Wireless MAN) Deployment Architecture

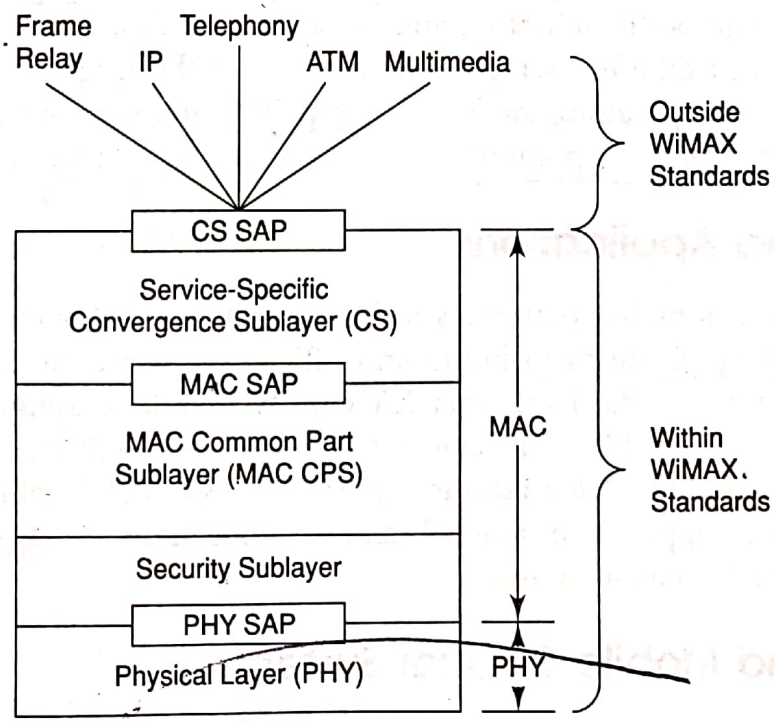

Figure 4.4(b) WiMAX Protocol Stack

4.4.1 Physical Layer

To support duplexing, 802.16 adapted a burst design that allows both time-division duplexing (TDD) and frequency-division duplexing (FDD). In TDD the uplink and downlink share a channel but do not transmit simultaneously. In the case of FDD the uplink and downlink operate on separate channels and sometimes simultaneously. Support for half-duplex FDD subscriber stations is also supported in 802.16. Both TDD and FDD alternatives support adaptive burst profiles in which modulation and coding options may be dynamically assigned on a burst-by-burst basis.

The 2–11 GHz bands, both licensed and unlicensed, are used in 802.16. Design of the 2–11 GHz physical layer is driven by the need for non-line-of-sight operation. The draft currently specifies that compliant systems implement one of three air interface specifications, each of which provides for interoperability. The 802.16 standard specifies three physical layers for services:

- WirelessMAN-SC2: This uses a single-carrier modulation format. This is to support existing networks and protocols.
- WirelessMAN-OFDM: This uses orthogonal frequency-division multiplexing with a 256-point transform. Access is by TDMA. This air interface is mandatory for license-exempt bands.
- WirelessMAN-OFDMA: This uses orthogonal frequency-division multiple access with a 2048-point transform. In this system, multiple access is provided by addressing a sub-set of the multiple carriers to individual receivers.

4.4.2 802.16 Medium Access Control

The IEEE 802.16 MAC protocol was designed for point-to-multipoint broadband wireless access. It addresses the need for very high bit rates, both uplink (to the base station) and downlink (from the base station). To support, a variety of services like multimedia and voice, the 802.16 MAC is equipped to accommodate both continuous and bursty traffic. To facilitate the more demanding physical environment and different service requirements of the frequencies between 2 and 11 GHz, the 802.16 project is upgrading the MAC to provide automatic repeat request (ARQ) and support for mesh, rather than only point-to-multipoint, network architectures.

4.4.3 Broadband Applications

Wireless broadband allows higher data rates in homes, offices, and even mobile environment. Therefore, all the user applications in home and offices are potential candidates for wireless broadband. These include standard Ethernet LAN or WiFi indoor using 802.16d and outdoor mobile using 802.16e. The IEEE 802.16 Broadband Wireless Metropolitan Area Network Standards can be found at IEEE site (http://standards.ieee.org/getieee802/802.16.html).

Along with the existing applications a new brand of applications are also being thought about. One such system is mobile cellular system.

4.4.4 Broadband Mobile Cellular System

During different discussions on systems and architecture of mobile computing, we talked about mobility with the network being static. In mobile cellular system the cellular network itself will be mobile. A cellular system like 3G can provide high data rate. WirelessMAN is also geared up to support high

data rate. However, these high data rates are possible with low speed mobility. Scientists are now thinking in terms of high-speed mobility specially designed for high-speed telematics application.

Figure 4.5 depicts one such mobile communication system to support high-speed mobility. This is achieved by installing moving base stations and fixed radio ports uniformly distributed along the median of the roadway. The moving base stations allow communication links to be established between the mobile units traveling on the roadway and a fixed communication network through the fixed radio ports. The small-cell (picocell) architecture of the proposed system enables the use of extremely lightweight low-power mobile units that can be used almost anywhere. In this architecture the picocell will move in the direction of the moving vehicle so that the relative speed between them is low. This proposed infrastructure is suitable for high-speed multilane highways in cities. The proposed system will be able to communicate to devices traveling at speeds up to and in excess of 150 kmph.

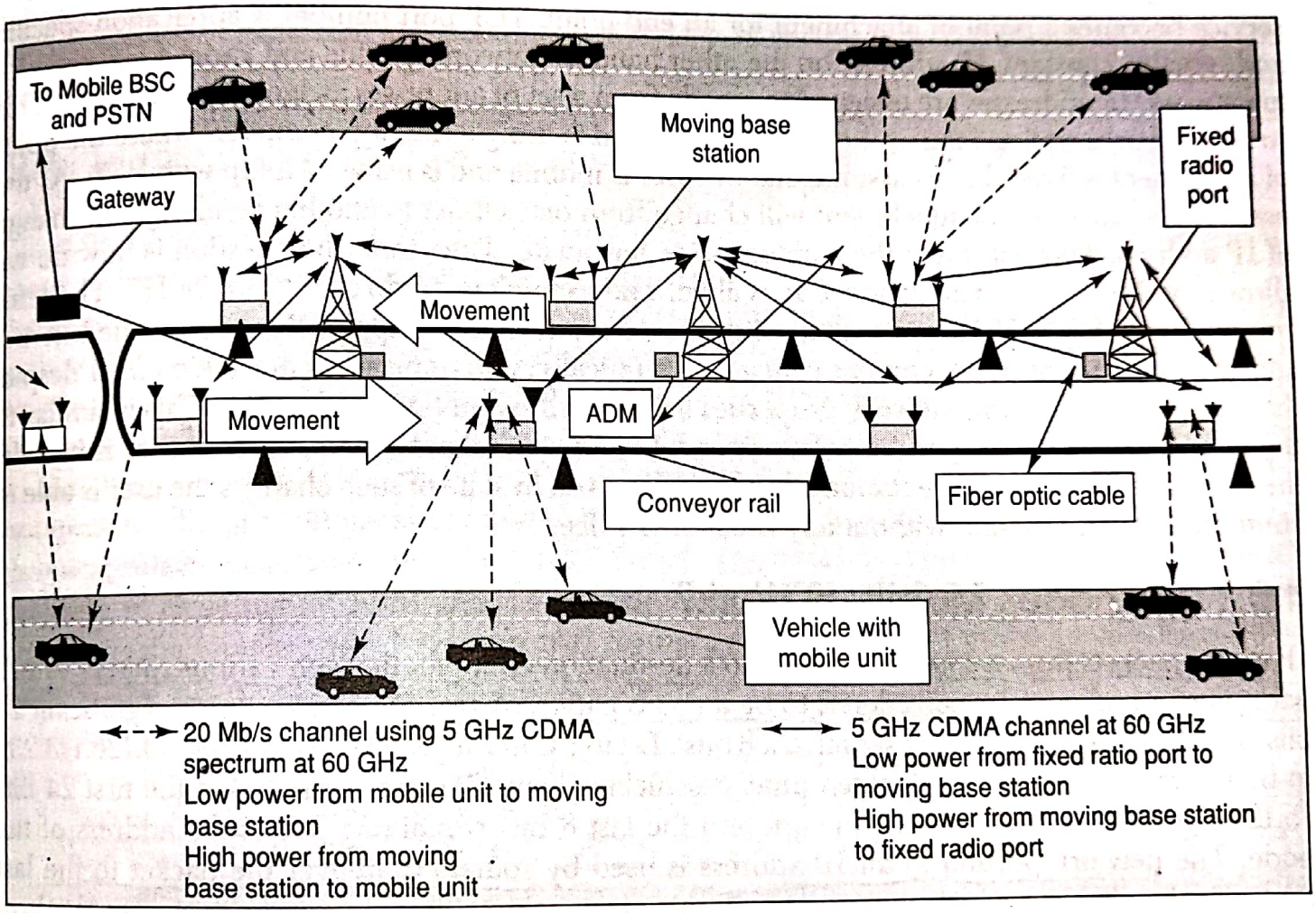

Figure 4.5 Mobile Broadband Communication System with Moving BTS

4.5 MOBILE IP

Mobile computing in the true sense will be able to provide an environment where a user will be able to continuously access data and services in a state of mobility. Mobile computing should not be confused with portable computing. In a portable computing environment, we move with the computing device from one location to another and use the network while stationary. For example,

while I am in office with my laptop computer, I use the company Ethernet LAN; and, when I am back home, I use the broadband at home. In this portable computing environment I use the network only when stationary and disconnect from one network before movement. Mobile computing on the other hand offers seamless computing and data networking facility even if the user is in a state of mobility and changes the network. Mobility Management (MM) deals with a situation where the user is at a vehicular state and accessing the network. Vehicular state generally means moving at a speed 60 kmph or higher. We will discuss the mobility management for voice network in Chapter 5. Here in Mobile IP, we will discuss the mobility management in TCP/IP data networks; Mobile IP standards are specified in RFC3344.

A data connection between two end-points through TCP/IP network requires a source IP address, source TCP port and a target IP address with a target TCP port. The combination of the IP address of the node (client or server device) system combined with the TCP port as the identification of a service becomes a point of attachment for an end-point. TCP port number is application-specific and remains constant. IP address, on the other hand, is network-specific and varies from network to network. IP addresses are assigned to a node from a set of addresses assigned to a network. This structure works well as long as the client is static and is using a desktop computer where the point of attachment is fixed. Let us assume that the user is mobile and is using a laptop with WiFi. As the user moves, the point of attachment will change from one subnet to another resulting in a change of IP address. This will force the connection to terminate. Therefore, the question is how do we allow mobility while a data connection is alive. The technology to do so is "Mobile IP". The term "mobile" in "Mobile IP" signifies that, while a user is connected to applications across the Internet and the user's point of attachment changes dynamically, all connections are maintained despite the change in underlying network properties including the point of attachments. This is similar to the handoff/roaming scenario in cellular networks. In a cellular network, when a user is mobile, the point of attachment (base station) changes. However, in spite of such changes the user is able to continue the conversation without any break in service.

4.5.1 How does Mobile IP Work?

IP routes packets from a source endpoint to a destination endpoint through various routers. An IP address of a node can be considered to be a combination of network address (most significant 24 bits) and the node address (least significant 8 bits). Let us assume a "C" class IP address 75.126.113.230 to be the mail server of Geschickten (mail.geschickten.com). We can assume that the first 24 bits 75.126.113 is the address of the network and the last 8 bits containing 230 is the address of the node. The network portion of an IP address is used by routers to deliver the packet to the last router in the chain to which the target computer is attached. This last router then uses the host portion (230 in this example) of the IP address to deliver the IP packet to the destination computer. In addition to the IP addresses of the nodes, for meaningful communication we need the TCP or UDP (User Datagram Protocal) port of the applications. The port number is used by the host to deliver the packet to the appropriate application.

A TCP connection is identified by a quadruplet that contains the IP address and port number of the sender endpoint along with the IP address and port number of the receiving endpoint. To ensure that an active TCP connection is not terminated while the user is mobile, it is essential that all of these four identities remain constant—physically or virtually. The TCP ports are application specific and generally constant—they do not change after an end-to-end connection is established.

However, the IP address will change when a node moves from one subnet to another. Therefore, to fix this problem mobile IP allows the mobile node to use two IP addresses. These IP addresses are called home address and care-of address. Home address is the original static IP address of the node and known to everybody as the identity of the node. The care-of address changes at each new point of attachment and can be thought of as the mobile node's location specific address. These are similar to MSISDN (Mobile Station ISDN) number and the MSRN (Mobile Station Roaming Number) respectively as in GSM network (see Chapter 5).

In addition to home address and care-of address there are two network elements in Mobile IP that play a very significant role in routing of the packets as part of mobility management; these are home agent and foreign agent. A home agent is a router on a mobile node's home network which forwards datagrams for delivery to the mobile node through a tunnel when it is away from home. The home agent also maintains current location information of the mobile node. In contrast, a foreign agent is a router on a mobile node's visited network which provides routing services to the mobile node while registered. The foreign agent detunnels and delivers datagrams to the mobile node that were tunneled by the mobile node's home agent. For datagrams sent by a mobile node, the foreign agent may serve as a default router for registered mobile nodes. This is similar to the concept of HLR (Home Location Register) and VLR (Visitor Location Register) in cellular networks (Chapter 5).

When the mobile node is located on its home network, it operates without mobility services. When the mobile node detects that it has moved to a foreign network, it registers with the foreign agent and obtains a care-of address on the foreign network. The care-of address can either be determined from a foreign agent's advertisements, or by some external assignment mechanism such as DHCP. The mobile node registers its new care-of address with its home agent informing its new location and new care-of address. The home agent forwards all incoming data packet to the foreign network using the care-of address. The delivery requires that the packet header is modified so that the care-of address becomes the destination IP address. This new header (Fig. 4.8) encapsulates the original packet, causing the mobile node's home address to have no impact on the encapsulated packet's routing. Figure 4.6 shows in general terms how Mobile IP deals with the problem of dynamic IP addresses. On returning to its home network from being registered elsewhere, the mobile node deregisters with its foreign agent, through exchange of a Registration Request and Registration Reply message.

Let us take an example of IP datagrams being exchanged over a TCP connection between the mobile node (A) and another host (server X in Fig. 4.6). The following steps occur:

1. Server X wants to transmit an IP datagram to node A. The home address of A is advertised and known to X. X does not know whether A is in the home network or somewhere else. Therefore, X sends the packet to A with A's home address as the destination IP address in the IP header. The IP datagram is routed to A's home network.
2. At the A's home network, the incoming IP datagram is intercepted by the home agent. The home agent discovers that A is in a foreign network. A care-of address has been allocated to A by this foreign network and available with the home agent. The home agent encapsulates the entire datagram inside a new IP datagram, with A's care-of address in the IP header. This new datagram with the care-of address as the destination address is retransmitted by the home agent.
3. At the foreign network, the incoming IP datagram is intercepted by the foreign agent. The foreign agent is the counterpart of the home agent in the foreign network. The foreign agent strips off the outer IP header, and delivers the original datagram to A.

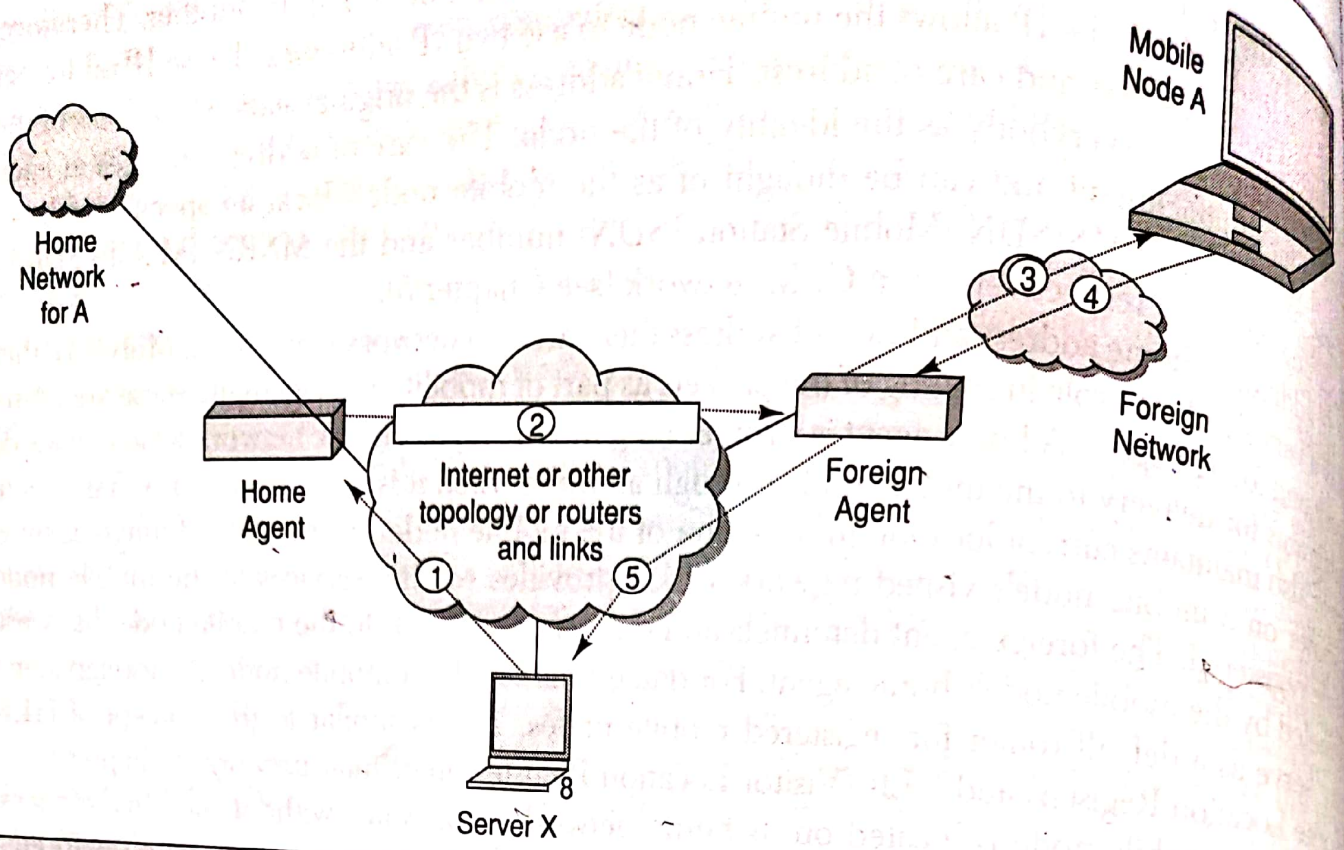

Figure 4.6 Mobile IP Architecture

4. A intends to respond to this message and sends traffic to X. In this example, X is not mobile; therefore X has a fixed IP address. For routing A's IP datagram to X, each datagram is sent to some router in the foreign network. Typically, this router is the foreign agent. A uses X's IP static address as the destination address in the IP header.
5. The IP datagram from A to X travels directly across the network, using X's IP address as the destination address.

To support the operations illustrated in fig. 4.6, mobile IP needs to support three basic capabilities:

- *Discovery*: A mobile node uses a discovery procedure to identify prospective home and foreign agents.
- *Registration*: A mobile node uses a registration procedure to inform its home agent of its care-of address.
- *Tunneling*: Tunneling procedure is used to forward IP datagrams from a home address to a care-of address.

4.5.2 Discovery

Agent advertisements are transmitted by both home and foreign agents to advertise their services on a link. Mobile nodes use these advertisements to determine their current point of attachment to the Internet. The Mobile IP discovery procedure has been built on top of an existing ICMP router discovery, router advertisement, and router solicitation procedure as specified for ICMP Router Discovery in RFC 1256. Mobile IP uses control messages that are sent to and from UDP port number 434. Mobile IP needs extensions to current messages formats. Extensions to ICMP router Discovery include:

0 One-byte Padding (encoded with no Length nor Data field);
16 Mobility Agent Advertisement; and
19 Prefix-Lengths.

Mobile IP control messages, however, include extensions like:

1 Registration Request;
3 Registration Reply;
32 Mobile-Home Authentication;
33 Mobile-Foreign Authentication; and
34 Foreign-Home Authentication.

Using the discovery procedure, the mobile node determines whether it is in a foreign network. For the purpose of discovery, a router or an agent periodically issues a router advertisement ICMP message. The mobile node on receiving this advertisement packet compares the network portion of the router IP address with the network portion of its own IP address allocated by the home network (home address). If these network portions do not match, then the mobile node knows that it is in a foreign network. A router advertisement can carry information about default routers and information about one or more care-of addresses. If a mobile node needs a care-of address without waiting for the agent advertisement, the mobile node can broadcast a solicitation that will be answered by any foreign agent.

4.5.3 Registration

Once a mobile node obtained a care-of address from the foreign network, the same needs to be registered with the home agent. The mobile node sends a registration request to the home agent with the care-of address information. When the home agent receives this request, it updates its routing table and sends a registration reply back to the mobile node.

Authentication: As a part of registration, the mobile node needs to be authenticated. Each mobile node, foreign agent, and home agent support a mobility security association (SA) for mobile entities, indexed by their security parameters index (SPI) and IP address. In the case of the mobile node, this must be its Home Address. Registration messages between a mobile node and its home agent MUST be authenticated with an authorization-enabling extension, e.g., the Mobile-Home Authentication Extension. This extension MUST be the first authentication extension; other foreign agent-specific extensions MAY be added to the message after the mobile node computes the authentication. Using 128-bit secret key and the HMAC-MD5 hashing algorithm, a digital signature is generated. Each mobile node and home agent shares a common secret. This secret makes the digital signature unique and allows the agent to authenticate the mobile node. At the end of the registration a triplet containing the home address, care-of address and registration lifetime is maintained in the home agent. This is called a binding for the mobile node. The home agent maintains this association until the registration life expires. The registration process involves the following four steps:

- The mobile node requests for forwarding service from the foreign network by sending a registration request to the foreign agent.
- The foreign agent relays this registration request to the home agent of that mobile node.
- The home agent either accepts or rejects the request and sends a registration reply to the foreign agent.
- The foreign agent relays this reply to the mobile node.

We have assumed that the foreign agent will allocate the care-of address. However, it is possible that a mobile node moves to a network that has no foreign agents or on which all foreign agents are busy. It is also possible that the care-of address is dynamically acquired as a temporary address by the mobile node such as through DHCP (Dynamic Host Configuration Protocol) as explained in RFC2131, or may be owned by the mobile node as a long-term address for its use only while visiting some foreign network. As an alternative therefore, the mobile node may act as its own foreign agent by using a co-located care-of address. A co-located care-of address is an IP address obtained by the mobile node that is associated with the foreign network. If the mobile node is using a co-located care-of address, then the registration happens directly with its home agent.

4.5.4 Tunneling

Figure 4.7 shows the tunneling operations in Mobile IP. In the mobile IP, an IP-within-IP encapsulation mechanism is used. Using IP-within-IP, the home agent, adds a new IP header called tunnel header. The new tunnel header uses the mobile node's care-of address as the tunnel destination IP address. The tunnel source IP address is the home agent's IP address. The tunnel header uses 4 as the protocol number (Fig. 4.8), indicating that the next protocol header is again an

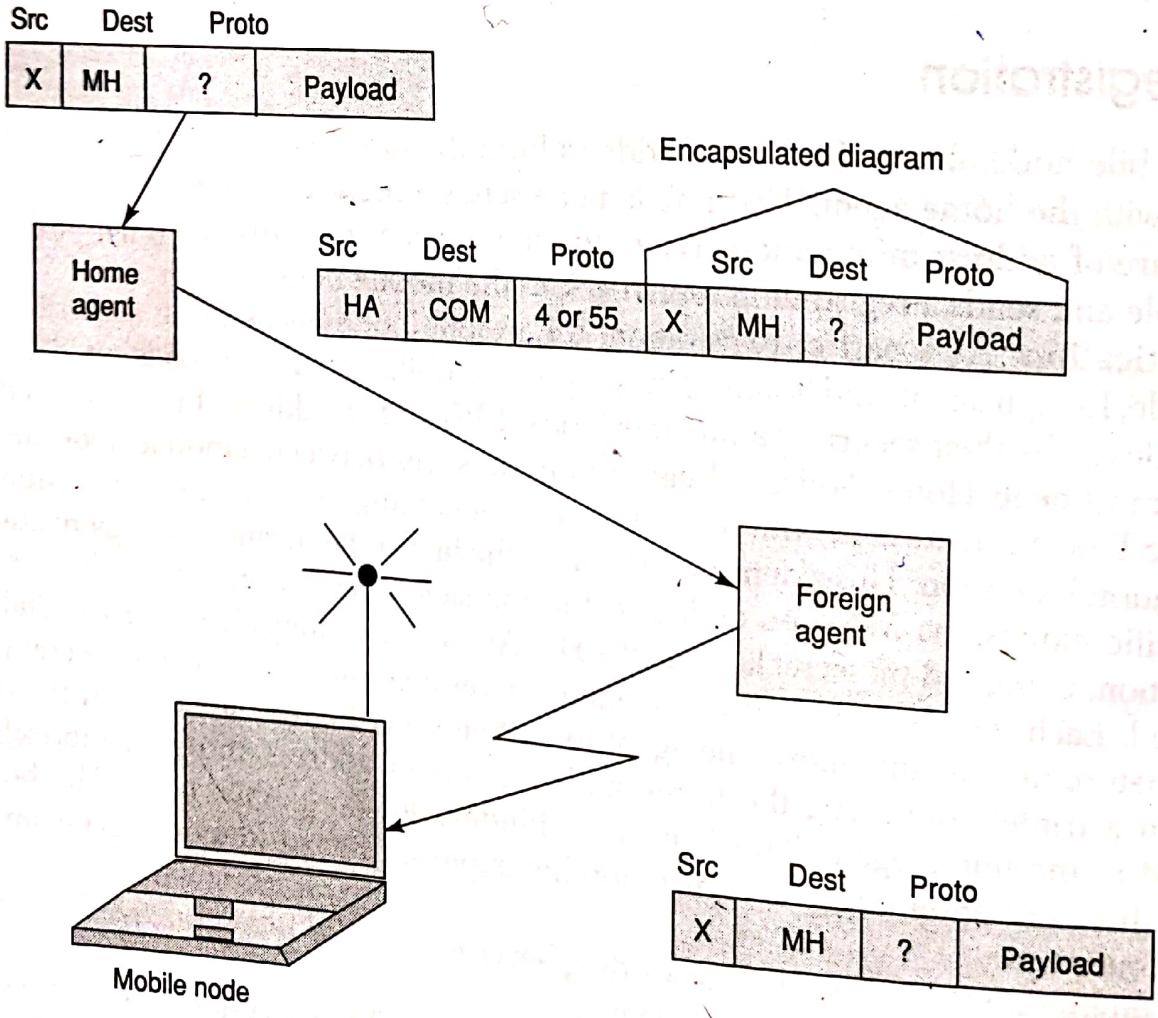

Figure 4.7 Tunneling

IP header. In IP-within-IP, the entire original IP header is preserved as the first part of the payload of the tunnel header. The foreign agent after receiving the packet, drops the tunnel header and delivers the rest to the mobile node.

When a mobile node is roaming in a foreign network, the home agent must be able to intercept all IP datagram packets sent to the mobile node so that these datagrams can be forwarded via tunneling. The home agent, therefore, needs to inform other nodes in the home network that all IP datagrams with the destination address of the mobile node should be delivered to the home agent. In essence, the home agent steals the identity of the mobile node in order to capture packets destined for that node that are transmitted across the home network. For this purpose ARP (Address Resolution Protocol) is used to notify all nodes in the home network.

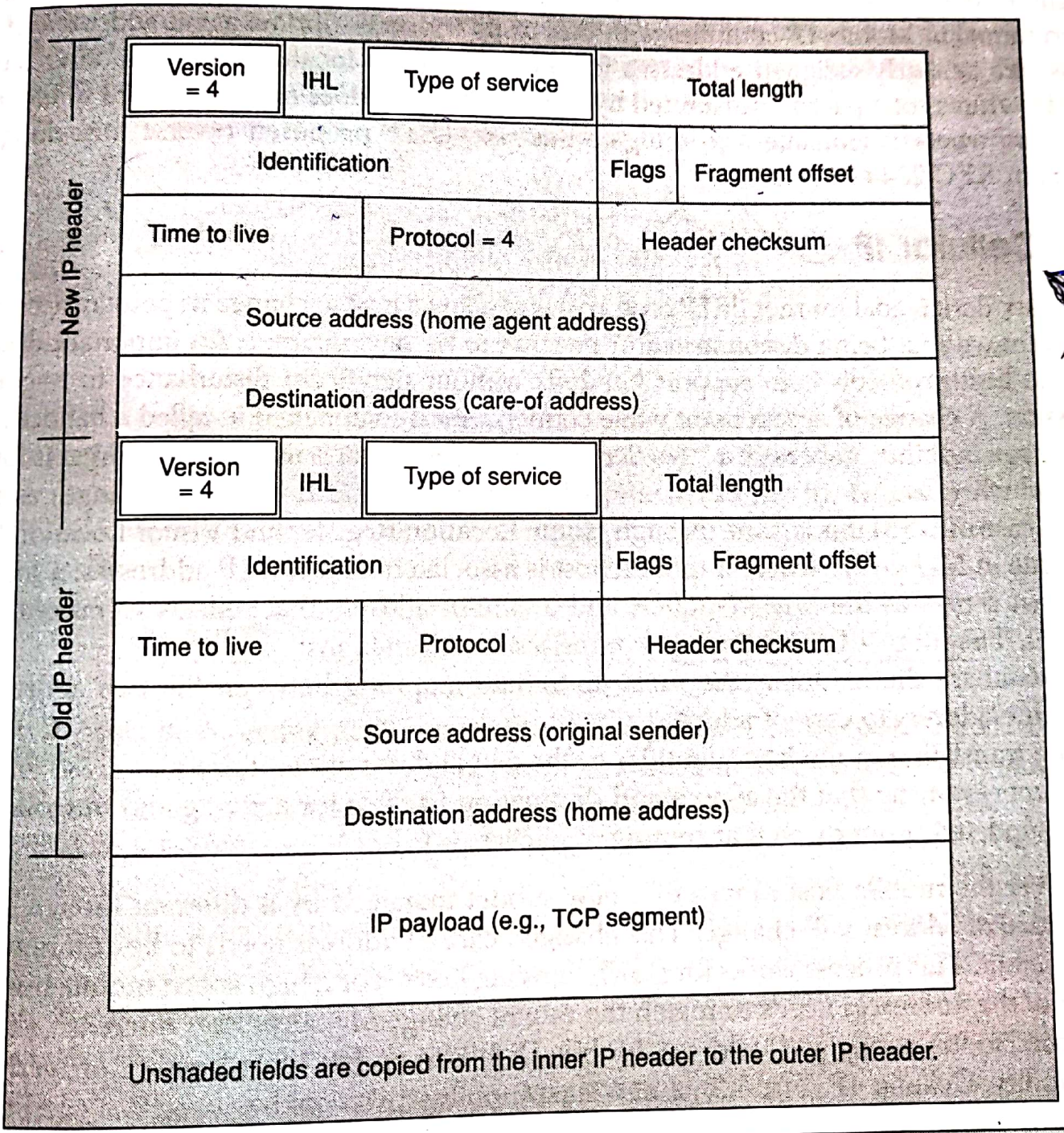

Figure 4.8 The IP Headers in Mobile IP (IP Encapsulation)

Let us take the example of Fig. 4.6. The original IP datagram from X to A has a source address as IP address of X and a destination address as the home IP address of A. The datagram is routed through the Internet to A's home network, where it is intercepted by the home agent. The home agent encapsulates the incoming datagram with an outer IP header. This outer header includes a source address same as the IP address of the home agent and a destination address equal to the care-of address. As the care-of address has the network portion of the foreign network, the packet will find its way directly to the mobile host. When this new datagram reaches the host in the foreign network, it strips off the outer IP header to extract the original datagram. From this stripped off packet it also finds out the original sender. This is necessary for the host to know who has sent the packet so that the response reaches the right destination.

In any IP data packet, the source and destination IP address must be topologically correct. The forward tunnel in Mobile IP complies with this, as its endpoints (home agent address and care-of address) are properly assigned addresses for their respective locations. On the other hand, the source IP address of a packet transmitted by the mobile node does not correspond to the network prefix from where it emanates. To mitigate this risk, IETF proposed reverse tunnelling that is specified in RFC 2344.

4.5.5 Cellular IP

The primary design goal for mobile IP protocols is to allow a host to change its point of access during data transfer without being disconnected or needing to be reconfigured. An important design goal for mobile host protocols is to support handoffs without significant disturbance to ongoing data transmission. A change of access point while connectivity is maintained is called a handoff.

To manage mobility, generally a "two tier addressing" scheme is used. One address is for a fixed location which is known to all; other one is for a dynamic location which changes as the user moves. In case of GSM this is done through Home Location Register and Visitor Location Register. Same is true in Mobile IP, where a mobile host is associated with two IP addresses: a fixed home address that serves as the host-identifier; and a care-of address that reflects its current point of attachment. The mobile IP architecture comprises three functions:

1. A database that contains the most up-to-date mapping between the two address spaces (home address to care-of address).
2. The translation of the host identifier to the actual destination address.
3. Agents ensuring that the source and destination packets for arriving and outgoing packets are updated properly so that routing of packets is proper.

Whenever the mobile host moves to a new subnet managed by a different foreign agent, the dynamic care-of-address will change. This changed care-of address needs to be communicated to the home agent. This process works for slowly moving hosts. For a high speed mobile host, the rate of update of the addresses needs to match the rate of change of addresses. Otherwise, packets will be forwarded to the wrong (old) address. Mobile IP fails to update the addressed properly for high speed mobility. Cellular IP (Fig. 4.9), a new host mobility protocol has been designed to address this issue.

In a 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 packets. This limited routing information

(referred to as mapping) is local to the node and does not assume that nodes have any knowledge of the wireless network topology. Mappings are created and updated based on the packets transmitted by mobile hosts.

Cellular IP uses two parallel structures of mappings through Paging Caches (PC) and Routing Caches (RC). PCs maintain mappings for stationary and idle (not in data communication state) hosts; whereas, RC maintains mappings for mobile hosts. Mapping entries in PC have a large timeout interval, in the order of seconds or minutes. RCs maintain mappings for mobile hosts currently receiving data or expecting to receive data. For RC mappings, the timeout is in the packet time scale. Figure 4.10 illustrates the relationship between PCs and RCs. While idle at location 1, the mobile host X keeps PCs up-to-date by transmitting dummy packets at a low frequency (Step 1 in Fig. 4.10). Let us assume that the host is mobile and moved to location 2 without transacting any data. The PC mapping for X now points to location 2. While at location 2, there are data packets to be routed to the mobile host X, the PC mappings are used to find the host (Step 2). As there is data transmission, the mapping database to be used will be the RC. As long as data packets keep arriving, the host maintains RC mappings, either by its outgoing data packets or through the transmission of dummy packets (Step 3).

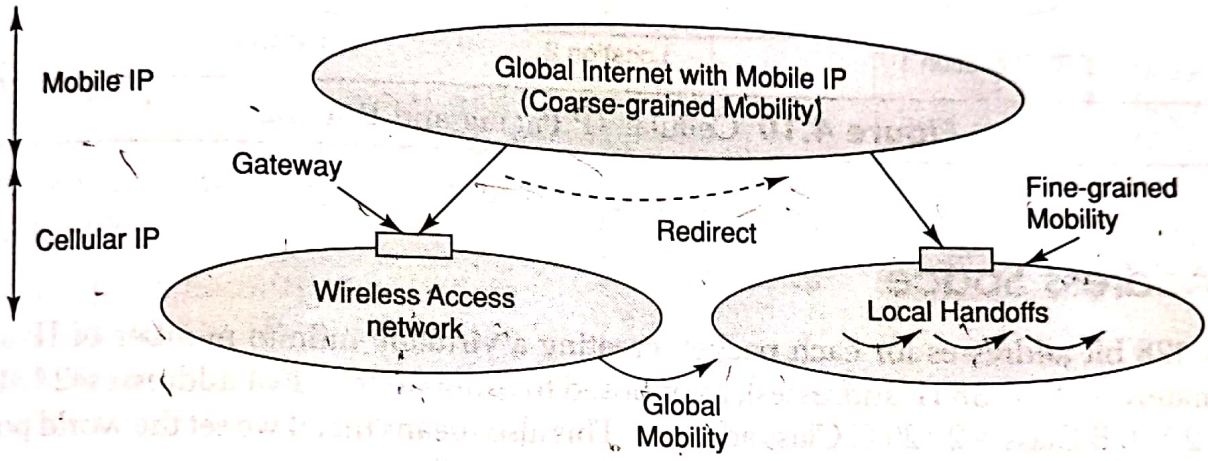

Figure 4.9 Relationships between Mobile IP and Cellular IP

Idle mobile hosts periodically generate short control packets, called paging-update packets. These are sent to the nearest available base station. The paging-update packets travel in the access network from the base station toward the gateway router, on a hop-by-hop basis. Handoff in Cellular IP is always initiated by the mobile host. As the host approaches a new base station, it redirects its data packets from the old to the new base station. First few redirected packets will automatically configure a new path of RC mappings for the host to the new base station. For a time equal to the timeout of RC mappings, packets addressed to the mobile host will be delivered at both old and new base stations.