UNIT 1

Introduction to Mobile Computing

1.1. MOBILE COMPUTING

1.1.1. Introduction

The rapidly expanding technology of cellular communication, wireless LANs, and satellite services will make information accessible anywhere and at any time. In the near future, tens of millions of people will carry a portable palmtop or laptop computer. Smaller units often called personal digital assistants or personal communications, will run on AA batteries and may have only a small memory; larger ones will be powerful laptop computers with larger memories and powerful processors.

Regardless of size, most mobile computers will be equipped with a wireless connection to the fixed part of the network, and perhaps, to other mobile computers. The resulting computing environment, which is often referred to as mobile computing, no longer requires users to maintain a fixed and universally known position in the network and enables almost unrestricted mobility. Mobility and portability will create an entire new class of applications and possibly, new massive markets combining personal computing and consumer electronics.

Mobile Computing is a technology that allows transmission of data, voice and video via a computer or any other wireless enabled device without having to be connected to a fixed physical link. The main concept involves:

1) Mobile Communication: The mobile communication in this case, refers to the infrastructure put in place to ensure that seamless and reliable communication goes on. These would include devices such as Protocols, Services, Bandwidth, and Portals necessary to facilitate and support of the stated services. The data format is also defined at this stage. This ensures that there is no collision with other existing systems which offer the same service.

   Since the media is unguided/unbounded, the overlaying infrastructure is more of radio wave oriented. That is, the signals are carried over the air to intended devices that are capable of receiving and sending similar kinds of signals.

2) Mobile Hardware: Mobile hardware includes mobile devices or device components that receive or access the service of mobility. They would range from Portable laptops, Smartphones, Tablet Pc's, Personal Digital Assistants.

   These devices will have receptor medium that are capable of sensing and receiving signals. These devices are configured to operate in full- duplex, whereby they are capable of sending and receiving signals at the same time. They don't have to wait until one device has finished communicating for the other device to initiate communications.

3) Mobile Software: Mobile software is the actual program that run on the mobile hardware. It deals with the characteristics and requirements of mobile applications. This is the engine of that mobile device. In other terms, it is the operating system of that appliance. It's the essential component that makes the mobile device operate.

   Since portability is the main factor, this type of computing ensures that users are not tied or pinned to a single physical location, but are able to operate from anywhere. It will incorporate all aspects of wireless communications.

1.1.2. History of Mobile Computing

Figure 1.1 shows a timeline of mobile computing development:

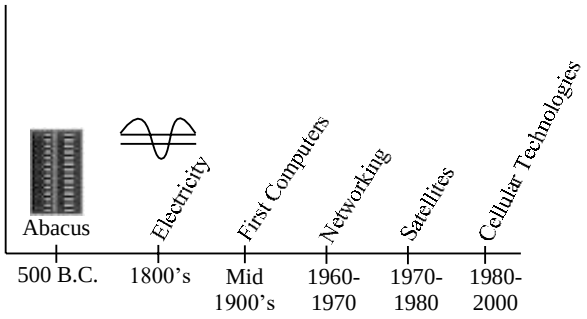


Figure 1.1: Timeline of Mobile Computing

One of the very first computing machines, the abacus, which was used as far back as 500 B.C., was, in effect, a mobile computing system because of its small size and portability. As technology progressed, the abacus evolved into the modern calculator. Most calculators today are made with an entire slew of mathematical functions while retaining their small size and portability. The abacus and calculators became important parts of technology not only because of their ability to compute but also because of their ease of use and portability. You can calculate the proceeds of a financial transaction anywhere as long as you had an abacus in 500 B.C. or have a calculator today. But, calculating numbers is only one part of computing.

Other aspects of computing, namely storage and interchange of information, do not date as far back as the abacus. Though writing has always been a way of storing information, we can hardly call a notebook a computing storage mechanism. The first mobile storage systems can be traced back only as far as the advent of the age of electronics.

A mobile computing system, as with any other type of computing system, can be connected to a network. Connectivity to the network, however, is not a prerequisite for being a mobile computing system. Dating from the late 1960s, networking allowed computers to talk to each other. Networking two or more computers together requires some medium that allows the signals to be exchanged among them. This was typically achieved through wired networks. Although wired networks remain the predominant method of connecting computers together, they are somewhat cumbersome for connecting mobile computing devices. Not only would network ports with always-available network connectivity have to be pervasive in a variety of physical locations, it would also not be possible to be connected to the network in real time if the device were moving. Therefore, providing connectivity through a wired system is virtually cost prohibitive. This is where wireless communication systems come to the rescue (figure 1.2):

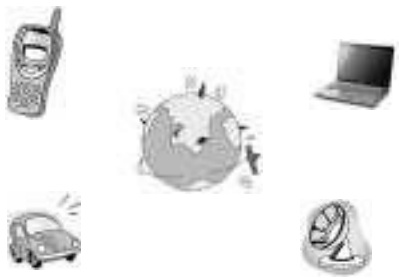


Figure 1.2: Wireless Communication Systems

By the 1960s, the military had been using various forms of wireless communications for years. Not only were wireless technologies used in a variety of voice communication systems, but the aviation and the space program had created great advances in wireless communication as well. First, the military developed wireless communication through line of sight – If there were no obstacles between point A and point B, you could send and receive electromagnetic waves. Then came techniques that allowed for wireless communication to encompass larger areas, such as using the atmosphere as a reflective mechanism. But, there were limitations on

how far a signal could reach and there were many problems with reliability and quality of transmission and reception.

By the 1970s, communication satellites began to be commercialised. With the new communication satellites, the quality of service and reliability improved enormously. Still, satellites are expensive to build, launch, and maintain. So the available bandwidth provided by a series of satellites was limited. In the 1980s cellular telephony technologies became commercially viable and the 1990s were witness to advances in cellular technologies that made wireless data communication financially feasible in a pervasive way.

Today, there are a plethora of wireless technologies that allow reliable communication at relatively high bandwidths. Of course, bandwidth, reliability, and all other qualitative and quantitative aspects of measuring wireless technologies are relative to time and people's expectations (as seems to be with everything else in life). Though most wireless networks today can transmit data at orders of magnitude faster speeds than just ten years ago, they are sure to seem archaically slow soon. It should, however, be noted that wired communication systems will almost certainly always offer us better reliability and higher data transmission bandwidths as long as electromagnetic communications is the primary means of data communications. The higher frequency sections of the electromagnetic spectrum are difficult to use for wireless communications because of natural noise, difficulty of directing the signal (and therefore high losses), and many other physical limitations.

Because the greatest advances in mobile communications originated in the military, it is no surprise that one of the first applications of wireless communication for mobile computing systems was in displaying terrain maps of the battlefield. From this, the Global Positioning System (GPS) evolved so that soldiers could know their locations at any given time. Portable military computers were provided to provide calculations, graphics, and other data in the field of battle. In recent years, wireless telephony has become the major provider of a revenue stream that is being invested into improving the infrastructure to support higher bandwidth data communications.

1.1.3. Dimensions of Mobile Computing

It should be obvious that any mobile computing system can also be stationary. If we stop moving it, it is stationary. So, we can say that mobile computing systems are a superset of stationary computing systems. Therefore, we need to look at those elements that are outside of the stationary computing subset. These added dimensions will help us to pick out variables that in turn allow us to divide and conquer the problems of mobile computing.

The dimension of mobility is the tools that allow us to qualify our problem of building mobile software applications and mobile computing systems. Although these dimensions of mobility are not completely orthogonal with respect to each other, they are separate enough in nature that we can distinguish them and appropriate them as orthogonal variables. Also, some of these dimensions are limitations; nevertheless, they are still added dimensions that need not be considered when dealing with the typical stationary application. These dimensions of mobility are as follows:

1) Location Awareness: A mobile device is not always at the same place – its location is constantly changing. The changing location of the mobile device and the mobile application presents the designers of the device and software applications with great difficulties. However, it also presents us with an opportunity of using the location and the change in location to enhance the application. These challenges and opportunities can be divided into two general categories:
    i) Localisation: It is the mere ability of the architecture of the mobile application to accommodate logic that allows the selection of different business logic, level of work flow, and interfaces based on a given set of location information commonly referred to as locales. Localisation is not exclusive to mobile applications but takes a much more prominent role in mobile applications. Localisation is often required in stationary applications where users at different geographical locations access a centralised system. For example, some Point of Sale (POS) systems and e-commerce websites are able to take into account the different taxation rules depending on the locale of the sale and the location of the purchase. Whereas localisation is something that stationary applications can have, location sensitivity is something fairly exclusive to mobile applications.

ii) Location Sensitivity: It is the ability of the device and the software application to first obtain location information while being used and then to take advantage of this location information in offering features and functionality. Location sensitivity may include more than just the absolute location of the device (if there is such a thing as absolute location). It may also include the location of the device relative to some starting point or a fixed point, some history of past locations, and a variety of calculated values that may be found from the location and the time such as speed and acceleration.

2) Network Connectivity Quality of Service (QOS): Whether wired or wireless connectivity is used, mobility means loss of network connectivity reliability. Moving from one physical location to another creates physical barriers that nearly guarantee some disconnected time from the network. If a mobile application is used on a wired mobile system, the mobile system must be disconnected between the times when it is connected to the wired docking ports to be moved. Of course, it is always a question whether a docking port is available when required let alone the quality and type of the available network connectivity at that docking port. In the case of wireless network connectivity, physical conditions can significantly affect the Quality of Service (QOS). For example, bad weather, solar flares, and a variety of other climate- related conditions can negatively affect the QOS. This unreliability in network connectivity has given rise to the QOS field and has led to a slew of accompanying products. QOS tools and products are typically used to quantify and qualify the reliability, or unreliability, of the connectivity to the network and are mostly used by network operators. Network operators control the physical layer of the network and provide the facilities, such as Internet Protocol (IP), for software application connectivity.
3) Limited Device Capabilities (Particularly Storage and CPU): No one wants to carry around a large device, so most useful mobile devices are small. This physical size limitation imposes boundaries on volatile storage, non-volatile storage, and CPU on mobile devices.
4) Limited Power Supply: The size constraints of the devices limit their storage capabilities and that their physical mobility affects network connectivity. For the same set of reasons that wireless is the predominant method of network connectivity for mobile devices, batteries are the primary power source for mobile devices. Batteries are improving every day and it is tough to find environments where suitable AC power is not available. The desirability of using batteries instead of an AC power source combined with the size constraints creates yet another constraint, namely a limited power supply. This constraint must be balanced with the processing power, storage, and size constraints; the battery is typically the largest single source of weight in the mobile device. The power supply has a direct or an indirect effect on everything in a mobile device.
5) Support for a Wide Variety of User Interfaces: A mobile application, based on its device support, the type of users using it, perhaps the biggest paradigm shift that designers and implementers of mobile applications must undergo is to understand the necessity of finding the best user interface(s) for the application, architecting the system to accommodate the suitable user interface(s), implementing them, and keeping in mind that a new user interface may be required at any time. User interfaces are difficult to design and implement for the following reasons:
   i) Designers have difficulties learning the user's tasks.
   ii) The tasks and domains are complex.
   iii) A balance must be achieved among the many different design aspects, such as standards, graphic design, technical writing, internationalisation, performance, multiple levels of detail, social factors, and implementation time.
   iv) The existing theories and guidelines are not sufficient.
   v) Iterative design is difficult.
   vi) There are real-time requirements for handling input events.
   vii) It is difficult to test user interface software.
   viii) Today's languages do not provide support for user interfaces.
   ix) Programmers report an added difficulty of modularisation of user interface software.
6) Platform Proliferation: It has very significant implications on the architecture, design, and development of mobile applications. Platform proliferation heightens the importance of designing and developing devices independent of the platform. Writing native code specific to the mobile device, unless absolutely necessary because of performance requirements, is not a recommended practice because of the proliferation of devices.
7) Active Transactions: These transactions are those transactions which are initiated by the system. Active transactions may be synchronous or asynchronous. All active transactions are initiated by the system.

1.1.4. Applications of Mobile Computing

1) For Estate Agents: Estate agents can work either at home or out in the field. With mobile computers they can be more productive. They can obtain current real estate information by accessing multiple listing services, which they can do from home, office or car when out with clients. They can provide clients with immediate feedback regarding specific homes or neighbourhoods, and with faster loan approvals, since applications can be submitted on the spot. Therefore, mobile computers allow them to devote more time to clients.
2) Emergency Services: Ability to receive information on the move is vital where the emergency services are involved. Information regarding the address, type and other details of an incident can be dispatched quickly, via a CDPD (Cellular Digital Packet Data) system using mobile computers, to one or several appropriate mobile units which are in the vicinity of the incident.
3) In Courts: Defence counsels can take mobile computers in court. When the opposing counsel references a case which they are not familiar, they can use the computer to get direct, real-time access to online legal database services, where they can gather information on the case and related precedents. Therefore mobile computers allow immediate access to a wealth of information, making people better informed and prepared.
4) In Companies: Managers can use mobile computers in, say, critical presentations to major customers. They can access the latest market share information. At a small recess, they can revise the presentation to take advantage of this information. They can communicate with the office about possible new offers and call meetings for discussing responds to the new proposals. Therefore, mobile computers can leverage competitive advantages.
5) Stock Information Collation/Control: In environments where access to stock is very limited, i.e., factory warehouses. The use of small portable electronic databases accessed via a mobile computer would be ideal.

   Data collated could be directly written to a central database, via a CDPD network, which holds all stock information hence the need for transfer of data to the central computer at a later date is not necessary. This ensures that from the time that a stock count is completed, there is no inconsistency between the data input on the portable computers and the central database.
6) Credit Card Verification: At Point of Sale (POS) terminals in shops and supermarkets, when customers use credit cards for transactions, the intercommunication required between the bank central computer and the POS terminal, in order to effect verification of the card usage, can take place quickly and securely over cellular channels using a mobile computer unit. This can speed up the transaction process and relieve congestion at the POS terminals.
7) Taxi/Truck Dispatch: Using the idea of a centrally controlled dispatcher with several mobile units (taxis), mobile computing allows the taxis to be given full details of the dispatched job as well as allowing the taxis to communicate information about their whereabouts back to the central dispatch office. This system is also extremely useful in secure deliveries, i.e., Securicor. This allows a central computer to be able to track and receive status information from all of its mobile secure delivery vans. Again, the security and reliability properties of the CDPD system shine through.
8) Electronic Mail/Paging: Usage of a mobile unit to send and read e-mails is a very useful asset for any business individual, as it allows him/her to keep in touch with any colleagues as well as any urgent developments that may affect their work. Access to the internet, using mobile computing technology, allows the individual to have vast arrays of knowledge at his/her fingertips.

   Paging is also achievable here, giving even more intercommunication capability between individuals, using a single mobile computer device.

1.1.5. Issues in Mobile Computing

The main issues of mobile computing environment are:

1) Communication: Mobile hosts connect with mobile support stations through wireless network. It is obvious that this wireless network does not have capacity as fixed wired network.
   i) The wireless bandwidth is very low, for example, cellular network has bandwidth in the order of 10Kbps or wireless local area network has bandwidth of 10Mbps.
   ii) Wireless network have higher error rate and frequent disconnection. The same network data package may require retransmission many times.

When MH is moving from one cell to another cell, the current connection with MSS will need to be changed to new connection. This process required two steps:

i) Disconnect from the current connection, and
ii) Establish new connection

The above disadvantage results in taking more time to transfer a same amount of data from MH to FH and vice-versa. Re-transmit data causes unnecessary processing power, which is already very limited on the mobile host. The situation is more complicated if two mobile hosts need to exchange data during cooperative work. Message cannot be delivered directly between two MH but via one or more MSS. Because of larger overhead in communication time, the longer time is required for mobile host to perform computation. Caching Mechanism is currently the major method to ease the problem.

2) Mobility: Mobility is the most frequent activity of a mobile host. When MH is moving from one cell to another cell in wireless network, the connection will need to be changed because one MSS can only support mobile hosts within its limited area. This causes frequent need of reconfiguration network topology and protocols. The more mobility, the more time spends on reestablishing communication between MH and MSS.

   Because the activities of MH need support from its MSS, therefore, location management is another problem caused by the mobility of MH. Mobile hosts need to track MSS in order to obtain data from the FH or other MH. In other words, MSS also needs to keep track on MH in order to transmit the result from the FH or to update the state of current mobile host profile.

   Mobility of MH raises the question on location dependent data. The same query will have different results depending on the location of MH. For example, bus time-table will depend on the location of the bus stop.

3) Portability: The availability of mobile devices depends on their power supply. A mobile phone can live up to five days but the laptop can only be for several hours. The more complicated application requires more processing power. Redefining computation into smaller partition (fined grain) or shifting heavy process from MH to FH for processing can save energy.

   Communication in mobile hosts requires a lot of power. Compressing data or data distilling before transmission can reduce communication time.

   Portability of mobile devices requires more sophisticated software application. MH has smaller user interface like display screen, keyboard. Many PDA support handwriting, therefore handwriting recognition software is required.

4) Heterogeneity: One MSS needs to support broad types of mobile devices which operate in its cell, identifying what kind of hardware of the MH is important. Different MH requires different applications. When MH requests communication with other MH, the heterogeneous problem needs to be taken into account.

   Different MSS are in different heterogeneous network and these MSS need to co-operate and communicate with each other for exchanging data. A standard interface is needed between MSS. Java technologies or a middleware like CORBA can be used to solve the heterogeneous problems.

5) Security Issues: Many authors have presented classifications of security issues in communication networks. There are five fundamental goals of security in information system.
   i) Confidentiality: preventing unauthorized users from gaining access to critical information of any particular user.
   ii) Integrity: It ensures unauthorized modification; destruction or creation of information cannot take place.
   iii) Availability: It ensuring authorized users getting the access they require.
   iv) Legitimate: ensuring that only authorized users have access to services.

v) Accountability: ensuring that the users are held responsible for there security related activities by arranging the user and his/her activities are linked if and when necessary.

The way these goals are achieved depends on the security policy adopted by the service providers.

1.1.6. Advantages of Mobile Computing

1) Locational Flexibility: The main benefit of mobile computers is that you do not have to bind yourself to a certain place. One can communicate with other people while sitting anywhere in the world. This will play an important role in the economy of the country and the world. One no longer need to stay plugged in to a specific location for performing computing activities. Mobile computing allows you unprecedented flexibility to move about and perform computing activities at the same time.
2) Saves Time: Mobile computing technology allows to use transit time more effectively.
3) Enhanced Productivity: Increased work flexibility is directly proportionate to enhanced work productivity - the fact that one can do work from any place you want, without waiting for, and making efforts to, get access to computing facility translates into people being able to do more work with greater flexibility. This is the reason why most companies these days offer home-computing access to employees. Suppose a national emergency is declared or any natural calamity occurs (or any other reason) due to which offices stay closed, work can still go on as people are no longer dependent upon office computing systems to get their work done!
4) Ease of Research: Mobile computing and the flexibility offered by it enable students as well as professionals to conduct in-depth research on just about any topic or subject even when on the go!
5) Entertainment: Various entertainment options available on mobile computing devices these days - games, movies, music, videos etc.

1.1.7. Disadvantages of Mobile Computing

Mobile computing systems are constrained in important ways, relative to static systems. These constraints are intrinsic to mobility, and are not just artifacts of current technology:

1) Mobile Elements are Resource-Poor Relative to Static Elements: At any given cost and level of technology, considerations of weight, power, size and ergonomics will render mobile elements less computationally capable than their static counterparts. While mobile elements will undoubtedly improve in absolute ability, they will always be at a disadvantage relative to static elements.
2) Mobile Elements are more Prone to Loss, Destruction, and Subversion than Static Elements: A Wall Street stockbroker is more likely to be mugged on the streets of Manhattan and have his or her laptop stolen than to have the workstation in a locked office be physically subverted. Even if security isn't a problem, portable computers are more vulnerable to loss or damage.
3) Mobile Elements must operate under a much Broader range of Networking Conditions: A desktop workstation can typically rely on LAN or WAN connectivity. A laptop in a hotel room may only have modem or ISDN connectivity. Outdoors, a laptop with a cellular modem may find itself in intermittent contact with its nearest cell.

1.2. WIRELESS COMMUNICATION

1.2.1. Overview

Wireless is a term used to describe telecommunications in which electromagnetic waves carry the signal over part or the entire communication path. In telecommunications, wireless communication is the transfer of information without the use of wires. The distances involved may be short (a few meters as in television remote control) or long.

The term wireless refers to technology that enables two or more computers to communicate using standard network protocols, but without network cabling.

Wireless technology is rapidly evolving, and is playing an increasing role in the lives of people throughout the world. In addition, ever-larger numbers of people are relying on the technology directly or indirectly.

The origin of wireless communication goes back to 1896, when Marconi invented the wireless telegraphy. His invention allowed two parties to communicate by sending each other alphanumeric characters encoded in an analog signal. Over the last century, advances in wireless technologies have led to the radio, the television, the mobile telephone, and communications satellites. Recently, a great deal of attention has been focused on satellite communications, wireless networking, and cellular technology.

Communication satellites were first launched in the 1960s. These first satellites could only handle 240 voice circuits. Today satellites carry about one-third of the voice traffic and all of the television signals between countries. Modern satellites typically introduce a quarter-second propagation delay to the signals they handle. Satellites in lower orbits with less inherent signal delay are being deployed to provide data services such as Internet access.

Wireless networking is allowing businesses to develop Wireless Area Networks, Metropolitan Area Networks, Local Area Networks without a cable plant. The IEEE has developed 802.11 as a standard for wireless LANs. The Bluetooth industry consortium is also working to provide a seamless wireless networking technology.

The cellular or wireless telephony is the modern equivalent of Marconi's wireless telegraph, offering two-party, two-way communication. The first generation of wireless telephony used analog technology, while the second generation used digital communication principles.

Digital networks can carry much more traffic and provide better reception and security than analog networks. The third generation of wireless telephony also uses digital technology and uses new frequency ranges at higher information rates.

The emergence of future generation networks has led to the convergence of various wireless communication devices that will allow the creation of global wireless network that will deliver a wide variety of services.

1.2.2. Elements of Wireless Communication System

Figure 1.1 is a block diagram of a complete wireless system.

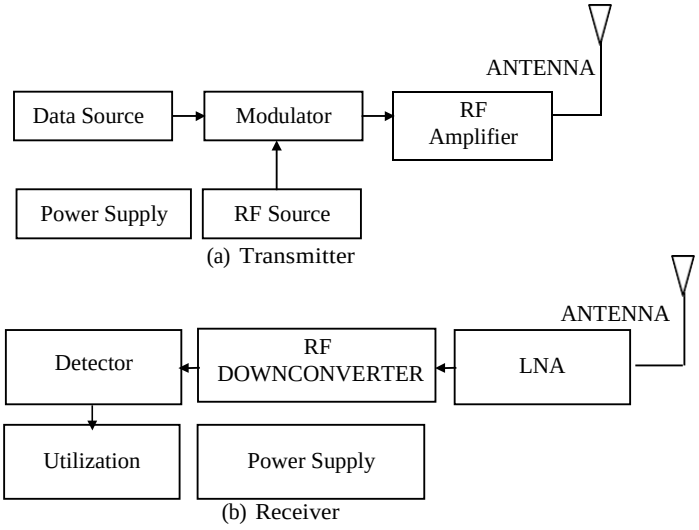


Figure 1.1: Wireless Communication System

The elements of wireless communication system are:

1) Data Source: This is the information to be conveyed from one side to the other. Data source may be analog or digital. A change of state of the data will cause a message frame to be modulated on a RF carrier wave. The format of frame is given below:

Address Bits	Data	Parity

Figure 1.2: Message Frame

2) Radio Frequency Generating Section: This part of the transmitter consists of an RF source (oscillator or synthesizer), a modulator, and an amplifier. In the simplest short-range devices, all three functions may be included in a circuit of only one transistor.

3) RF Conduction and Radiation: Practically all short-range devices have built-in antennas, so their transmission lines are relatively short and simple. However, particularly on the higher frequencies, their lengths are a high enough percentage of wavelengths to affect the transmission efficiency of the transmitter.

4) Radio Channel: Radio channel for short-range applications is short, and for a large part the equipment is used indoors. The allowed radio frequency power is relatively low and regulated by the telecommunication authorities.

5) Receivers: Receivers are similar to transmitters, but their operation is reversed. They have an antenna and transmission line, RF amplifiers, and use oscillators in their operation. The ultimate purpose of the receiver is to convert the data source that was implanted on the RF wave in the transmitter back to its original form.

6) Power Supplies: When size is limited, as it is in hand-operated remote control transmitters and security detectors, battery size and therefore energy is limited. The need to change batteries often is not only highly inconvenient but also expensive, and this is an impediment to more widespread use of radio in place of wires. Thus, low-current consumption is an important design aim for wireless devices.

1.2.3. Types of Wireless Systems

Wireless computing systems can be broadly classified into the following two categories:

1) Fixed Wireless Systems: These wireless computing systems support little or no mobility of the computing equipment associated with the wireless network. For example, a LAN can be set up using wireless technology to get rid of the hassles of laying cables. The LAN will work as a conventional wired LAN except for the difference that it does not need any cabling to be carried out.

2) Mobile Wireless Systems: These wireless computing systems support mobility of the computing equipment, which the users use to access the resources associated with the wireless network. These systems support mobility of users and allow the mobile users to access information from anywhere and at anytime. For example, the mobile wireless system includes smart phones, personal digital assistants (PDAs), and pagers with Internet access.

1.2.4. Advantages of Wireless Communication

1) Completes Access Technology Portfolio: Customers commonly uses more than one access technology to service various parts of their network and during the migration phase of their networks, when upgrading occurs on a scheduled basis. Wireless enables a fully comprehensive access technology portfolio to work with existing dial, cable, and DSL technologies.
2) Goes Where Cable and Fiber Cannot: The inherent nature of wireless is that it doesn't require wires or lines to accommodate the data/voice/video pipeline. As such, the system will carry information across geographical areas that are prohibitive in terms of distance, cost, access, or time.
3) Involves Reduced Time to Revenue: Companies can generate revenue in less time through the deployment of wireless solutions than with comparable access technologies because a wireless system can be assembled and brought online in as little as two to three hours.
4) Provides Broadband Access Extension: Wireless commonly both competes with and complements existing broadband access. Wireless technologies play a key role in extending the reach of cable, fiber, and DSL markets, and it does so quickly and reliably. It also commonly provides a competitive alternative to broadband wire line or provides access in geographies that don't qualify for loop access.

1.2.5. Disadvantages of Wireless Communication

1) Security: Wireless technology offers much less security compared to a wired network. Many wireless networks utilize encryption technologies, but not all of these are effective.
2) Range of Signal Problem: Common home network, using ordinary equipment, only sends a signal a few tens of yards. This works for most homes, but can be problematic in a large structure, or in a building where the signal is likely to be stopped by heavy walls. Additional range must be added using repeaters or more access points. This can increase your network costs significantly. However, wireless manufacturers are in the process of developing technology that will improve the range of transmission.
3) Unreliable: Sometimes, wireless networks are unreliable. Because they are based on radio transmissions, signals from wireless networking can suffer from interference or other problems beyond the administrator's control. This means that wireless networks can suffer unexpected downtime or interruption of signal. Important networking components, like servers, are almost never connected via wireless network.
4) Slower: Compared to wire networking, wireless LANs can be much slower (1-108 Mbit/s versus 100 Mbit/s to multiple Gbit/s). The methods used to send data over a wireless network can also cause performance problems.
5) Compatibility: Because wireless is still a fairly new technology, components which are not made by the same company may not work together. It may also require a lot of tinkering (tampering) to get them to communicate properly.

1.2.6. Applications of Wireless Communication

Many applications can benefit from wireless networks and mobile communications, particular application are:

1) Vehicles: Vehicles will comprise many wireless communication systems and mobility aware applications. Music, news, weather reports, and other broadcast information are received via wireless system in any location.
2) Emergencies: Possibilities of an ambulance with a high-quality wireless connection to a hospital. Vital information about injured persons can be sent to the hospital from the scene of the accident. All the necessary steps for this particular type of accident can be prepared and specialists can be consulted for an early diagnosis.

   Wireless networks are the only means of communication in the case of natural disasters such as hurricanes or earthquakes.
3) Business: A traveling salesman today needs instant access to the company's database: to ensure that files on his or her laptop reflect the current situation, to enable the company to keep track of all activities of their traveling employees, to keep databases consistent etc. With wireless access, the laptop can be turned into a true mobile office, but efficient and powerful synchronization mechanisms are needed to ensure data consistency.
4) Infotainment: Wireless networks can provide up-to-date information at any appropriate location. The travel guide might tells something about the history of a building (knowing via GPS, contact to a local base station, or triangulation where one is) downloading information about a concert in the building at the same evening via a local wireless network.
5) Location Dependent Services: In many cases, it is important for an application to 'know' something about the location or the user might need location information for further activities. Several services that might depend on the actual location can be distinguished:
   i) Follow-on Services: The function of forwarding calls to the current user location is well known from the good old telephone system. Wherever you are, just transmit your temporary phone number to your phone and it redirects incoming calls. Using mobile computers, a follow-on service could offer, for example, the" same desktop environment wherever you are in the world. All e-mail would automatically be forwarded and all changes to your desktop.
   ii) Location Aware Services: Imagine you wanted to print a document sitting in the lobby of a hotel using your laptop. If you drop the document over the printer icon, where would you expect the document to be printed? Certainly not by the printer in your office!

iii) Privacy: There might be locations and/or times when you want to exclude certain services from reaching you and you do not want to be disturbed. You want to utilize location dependent services, but you might not want the environment to know exactly who you are.

iv) Information Services: While walking around in a city you could always use your wireless travel guide to 'pull' information from a service, for example, 'Where is the nearest Pizza Corner.

v) Support Services: Many small additional mechanisms can be integrated to support a mobile device. Intermediate results of calculations, state information, or cache contents could 'follow' the mobile node through the fixed network. As soon as the mobile node reconnects, all information is available again. This helps to reduce access delay and traffic within the fixed network. Caching of data on the mobile device (standard for all desktop systems) is often not possible due to limited memory capacity. The alternative would be a central location for user information and a user accessing this information through the (possibly large and congested) network all the time as it is often done today.

6) Mobile and Wireless Devices: The following list gives some examples of mobile and wireless devices graded by increasing performance:

i) Sensor: A very simple wireless device is represented by a sensor transmitting state information. Sensor are used in many industrial and consumer applications, such as industrial process monitoring and control, machine health monitoring, and so on.

ii) Embedded Controllers: Many appliances already contain a simple or sometimes more complex controller. For example, the Keyboards, mice, headsets, washing machines, coffee machines, hair dryers and TV set. Hair dryer as a simple mobile and wireless device that is able to communicate with the mobile phone. Then the dryer would switch off as soon as the phone starts ringing.

iii) Pager: As a very simple receiver, a pager can only display short text messages, has a tiny display, and cannot send any messages. Pagers can even be integrated into watches. The tremendous success of mobile phones has made the pager virtually redundant in many countries. Short messages have replaced paging. The situation is somewhat different for emergency services where it may be necessary to page a larger number of users reliably within short time.

iv) Mobile Phones: The traditional mobile phone only had a simple black and white text display and could send/receive voice or short messages. Today, mobile phones migrate more and more toward PDAs. Mobile phones with full color graphic display, touch screen, and Internet browser are easily available.

v) Personal Digital Assistant: PDAs typically accompany a user and offer simple versions of office software (calendar, note-pad, mail). The typical input device is a pen, with built-in character recognition translating handwriting into characters. Web browsers and many other software packages are available for these devices.

vi) Pocket Computer: The next steps toward full computers are pocket computers offering tiny keyboards, color displays, and simple versions of programs found on desktop computers (text processing, spreadsheets etc.).

vii) Notebook/Laptop: Laptops offer more or less the same performance as standard desktop computers; they use the same software - the only technical difference being size, weight, and the ability to run on a battery. If operated mainly via a sensitive display (touch sensitive or electromagnetic), the devices are also known as notepads or tablet PCs.

1.2.7. Wireless Telephony

A system of telephone communication in which the action of the transmitter brings about fluctuations in electric waves which are radiated through space by a high frequency current. These fluctuations are reproduced at the distant station in the form of the original sounds.

Wireless telephony differs from wireless telegraphy in that the transmission of waves is continuous instead of interrupted. Wireless telephony has become a practical method of communication between ships at sea and between moving railroad trains. The distance to which sounds can be transmitted is practically limitless.

1.2.8. Cellular Concept

The design objective of early mobile radio systems was to achieve a large coverage area using a single, high powered transmitter with an antenna mounted on a tall tower. The cellular concept is a system-level idea which calls for replacing a single, high power transmitter (large cell) with many low power transmitter (small cells) each providing a coverage to only a small portion of the service area.

The need to operate and grow "indefinitely" within an allocation of hundreds of channels has been the primary driving force behind the evolution of the cellular concept.

The concept was developed in 1947 at AT&T Bell Laboratories. First tests were conducted in 1962 for commercial applications. FCC finally set aside new radio frequencies for land mobile communications in 1970. In 1970, AT&T proposed to build the first high capacity cellular phone system called "Advanced Mobile Phone Service (AMPS)".

A cellular System is a radio network distributed over land areas called cells, each served by at least one fixed-location transceiver known as a cell site or base station. When joined together these cells provide radio coverage over a wide geographic area. This enables a large number of portable transceivers (e.g., mobile phones, pagers, etc.) to communicate with each other and with fixed transceivers and telephones anywhere in the network, via base stations, even if some of the transceivers are moving through more than one cell during transmission. Adjacent cells are assigned different frequencies to avoid interference or crosstalk.

Cellular systems for mobile communications implement SDM (Space division multiplexing). Cell radii can vary from tens of meters in buildings, and hundreds of meters in cities, up to tens of kilometers in the countryside. The shape of cells are never perfect circles or hexagons as shown in figure 1.2, but depends on the environment (buildings, mountains, vallays), on weather conditions, and sometimes even on system load.

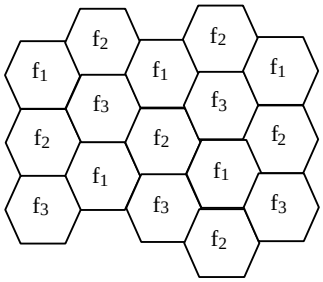


Figure 1.2: Cellular System with Three Cell Clusters

The first design decision to make is the shape of cells to cover an area. A matrix of square cells would be the simplest layout to define as shown in figure 1.3. This geometry is not ideal.

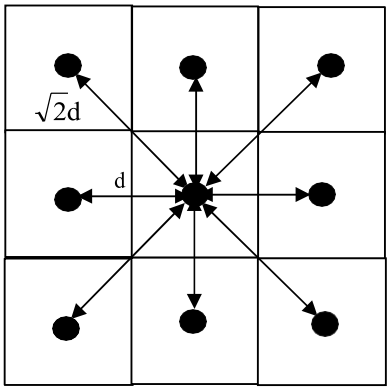


Figure 1.3: Square Pattern

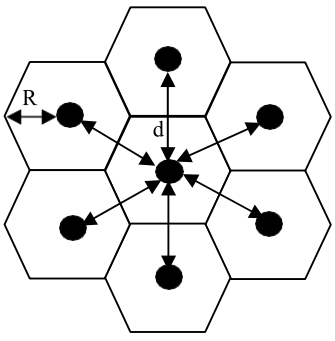


Figure 1.4: Hexagonal Pattern

In square pattern, if the width of a square cell is d, then a cell has four neighbors at a distance d and four neighbors at a distance $\sqrt{2}$d . As a mobile user within a cell moves toward the cell's boundaries, it is best if all of the adjacent antennas are equidistant.

In hexagonal pattern as shown in figure 1.4, provides for equidistant antennas. The radius of a hexagon is defined to be the radius of the circle that circumscribes it. For a cell radius R, the distance between the cell center and each adjacent cell center is d = R. $\sqrt{3}$

1.2.8.1. Operations of Cellular Systems

Figure 1.5 shows the principal elements of a cellular system. In the approximate center of each cell is a Base Station (BS).

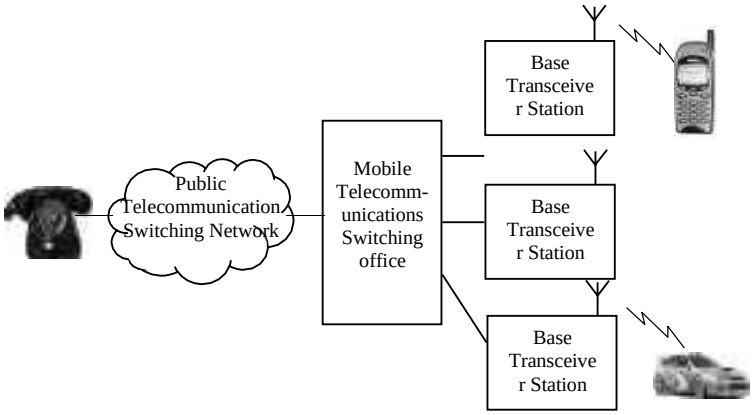


Figure 1.5: Overview of Cellular System

The BS includes an antenna, a controller, and a number of transceivers, for communicating on the channels assigned to that cell. Each BS is connected to a Mobile Telecommunications Switching Office (MTSO), with one MTSO serving multiple BSs.

Typically, the link between an MTSO and a BS is by a wire line, although a wireless link is also possible. The MTSO connects calls between mobile units. The MTSO is also connected to the public telephone or telecommunications network and can make a connection between a fixed subscriber to the public network and a mobile subscriber to the cellular network. The MTSO assigns the voice channel to each call, performs handoffs, and monitors the call for billing information.

There are two types of channels are available between the mobile unit and the Base Station (BS) as follows:

1) Control Channels: These channels are used to exchange information having to do with setting-up and maintaining calls and with establishing a relationship between a mobile unit and the nearest BS.

2) Traffic Channels: These channels carry a voice or data connection between users. Figure 1.6 illustrates the steps in a typical call between two mobile users within an area controlled by a single MTSO:

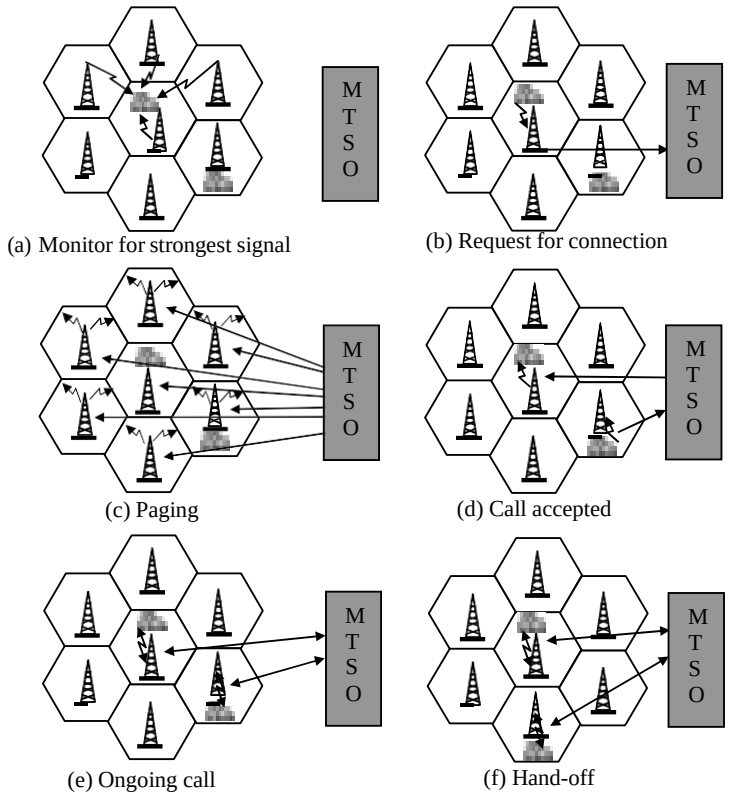


Figure 1.6: Example of Mobile Cellular Call

i) Mobile Unit Initialization: When the mobile unit is turned on, it scans and selects the strongest set-up control channel used for this system (figure 1.6(a)).

ii) Mobile-Originated Call: A mobile unit originates a call by sending the number of the called unit on the pre-selected set-up channel (figure 1.6(b)).

iii) Paging: The MTSO then attempts to complete the connection to the called unit. The MTSO sends a paging message to certain BSs depending on the called mobile number (figure 1.6(c)). Each BS transmits the paging signal on its own assigned set-up channel.

iv) Call Accepted: The called mobile unit recognizes its number on the set-up channel being monitored and responds to that BS, which sends the response to the MTSO. The MTSO sets-up a circuit between the calling and called BSs. At the same time, the MTSO selects an available traffic channel within each BS's cell and notifies each BS, which in turn notifies its mobile unit (figure 1.6(d)). The two mobile units tune to their respective assigned channels.

v) Ongoing Call: While the connection is maintained, the two mobile units exchange voice or data signals, going through their respective BSs and the MTSO (figure 1.6(e)).

vi) Hand-off: If a mobile unit moves out of range of one cell and into the range of another during a connection, the traffic channel has to change to one assigned to the BS in the new cell (figure 1.6(f)). The system makes this change without either interrupting the call or alerting the user.

1.2.8.2. Channel Allocation in Cellular Systems

Channel allocation deals with the allocation of channels to cells in a cellular network. Once the channels are allocated, cells may then allow users within the cell to communicate via the available channels. Channels in a wireless communication system typically consist of time slots, frequency bands and/or CDMA pseudo noise sequences, but in an abstract sense, they can represent any generic transmission resource.

In view of the increasing growth of mobile users in the present scenario with limited bandwidth, the bandwidth should be utilized in an optimal way so that more number of users may be serviced. This limitation means that frequency channels should be reused in an efficient way to support many users. Thus, evolved the concept of cellular architecture, which consists of collection of geometric areas called cells; each serviced by a base station.

The cellular concept was a major breakthrough in solving the problem of spectral congestion and user capacity. It offered very high capacity in a limited spectrum allocation without any major technological changes.

The cellular concept is a system-level idea which calls for replacing a single, high power transmitter (large cell) with many low power transmitters (small cells), each providing coverage to only a small portion of the service area. Each base station is allocated a portion of the total number of channels available to the entire system, and nearby base stations are assigned different groups of channels so that all the available channels are assigned to a relatively small number of neighboring base stations. Neighboring base stations are assigned different groups of channels so that the interference between base stations is minimized.

By systematically spacing base stations and their channel groups throughout the market, the available channels are distributed throughout the geographic region and may be reused as many times as necessary, as long as the interference between co-channel (same frequency channels) stations is kept below acceptable levels.

As the demand for service increases, the number of base stations may be increased thereby providing additional radio capacity with no additional increase in radio spectrum. This fundamental principle is the foundation for all modern wireless communication systems, since it enables a fixed number of channels to serve an arbitrarily large number of subscribers by reusing the channel throughout the coverage region.

As the frequency channels are a scarce resource in a cellular mobile system, various schemes of assigning the resources are prevailing such as:

1) Fixed Channel Allocation (FCA): Fixed Channel Allocation systems allocate specific channels to specific cells. This allocation is static and cannot be changed. For efficient operation, FCA systems typically allocate channels in a manner that maximizes frequency reuse. Thus, in a FCA system, the distance between cells using the same channel is the minimum reuse distance for that system.

   The problem with FCA systems is quite simple and occurs whenever the offered traffic to a network of base stations is not uniform. For example, consider a case in which two adjacent cells are allocated N channels each. There clearly can be situations in which one cell has a need for N+k channels while the adjacent cell only requires N-m channels (for positive integers k and m). In such a case, k users in the first cell would be blocked from making calls while m channels in the second cell would go unused. Clearly in this situation of non-uniform spatial offered traffic, the available channels are not being used efficiently.

2) Dynamic Channel Allocation (DCA): Dynamic Channel Allocation attempts to alleviate the problem mentioned for FCA systems when offered traffic is non-uniform. In DCA systems, no set relationship exists between channels and cells. Instead, channels are part of a pool of resources. Whenever a channel is needed by a cell, the channel is allocated under the constraint that frequency reuse requirements cannot be violated. There are two problems that typically occur with DCA based systems.
   i) First, DCA methods typically have a degree of randomness associated with them and this leads to the fact that frequency reuse is often not maximized unlike the case for FCA systems in which cells using the same channel are separated by the minimum reuse distance.
   ii) Secondly, DCA methods often involve complex algorithms for deciding which available channel is most efficient. These algorithms can be very computationally intensive and may require large computing resources in order to be real-time.

3) Hybrid Channel Allocation: The third category of channel allocation methods includes all systems that are hybrids of fixed and dynamic channel allocation systems. Several methods have been presented that fall within this category and in addition, a great deal of comparison has been made with corresponding simulations and analyses.

1.2.8.3. Frequency Re-Use

Cellular radio systems rely on an intelligent allocation and re-use of channels throughout a coverage region. Each cellular base station is allocated a group of radio channels to be used within a small geographic area called a cell. Base stations in adjacent cells are assigned channel groups which contain completely different channels than neighbouring cells. The base station antennas are designed to achieve the desired coverage within the particular cell. By limiting the coverage area to within the boundaries of a cell, the same group of channels may be used to cover different cells that are separated from one another by distances large enough to keep interference levels within tolerable limits. The design process of selecting and allocating channel groups for all of the cellular base stations within a system is called frequency re-use or frequency planning.

Figure 3.1 illustrates the concept of cellular frequency re-use, where cells labelled with the same letter use the same group of channels:

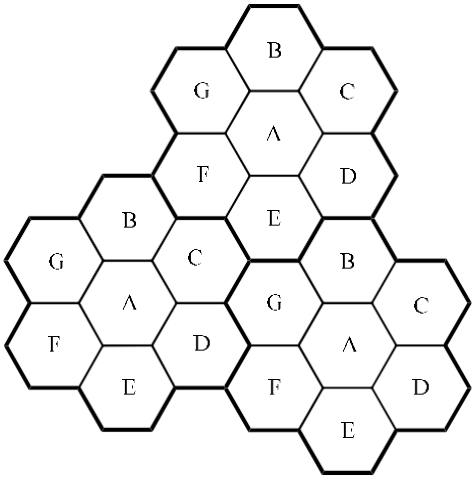


Figure 3.1: Illustration of the Cellular Frequency Re-Use Concept. Cells with the same letter use the same set of frequencies. A cell cluster is outlined in bold and replicated over the coverage area. In this example, the cluster size, N, is equal to seven, and the frequency re-use factor is 1/7 since each cell contains one-seventh of the total number of available channels.

The frequency re-use plan is overlaid upon a map to indicate where different frequency channels are used. The hexagonal cell shape shown in figure 3.1 is conceptual and is a simplistic model of the radio coverage for each base station, but it has been universally adopted since the hexagon permits easy and manageable analysis of a cellular system. The actual radio coverage of a cell is known as the footprint and is determined from field measurements or propagation prediction models. Although the real footprint is amorphous in nature, a regular cell shape is needed for systematic system design and adaptation for future growth. While it might seem natural to choose a circle to represent the coverage area of a base station, adjacent circles cannot be overlaid upon a map without leaving gaps or creating overlapping regions. Thus, when considering geometric shapes which cover an entire region without overlap and with equal area, there are three sensible choices – a square, an equilateral triangle, and a hexagon. A cell must be designed to serve the weakest mobiles within the footprint, and these are typically located at the edge of the cell. For a given distance between the centre of a polygon and its farthest perimeter points, the hexagon has the largest area of the three. Thus, by using the hexagon geometry, the fewest number of cells can cover a geographic region, and the hexagon closely approximates a circular radiation pattern which would occur for an omnidirectional base station antenna and free space propagation. Of course, the actual cellular footprint is determined by the contour in which a given transmitter serves the mobiles successfully.

When using hexagons to model coverage areas, base station transmitters are depicted as either being in the centre of the cell (centre-excited cells) or on three of the six cell vertices (edge-excited cells). Normally, omnidirectional antennas are used in centre-excited cells and sectored directional antennas are used in corner-excited cells. Practical considerations usually do not allow base stations to be placed exactly as they appear in the hexagonal layout. Most system designs permit a base station to be positioned up to one-fourth the cell radius away from the ideal location.

To understand the frequency re-use concept, consider a cellular system which has a total of S duplex channels available for use. If each cell is allocated a group of k channels ($k < S$), and if the S channels are divided among

N cells into unique and disjoint channel groups which each have the same number of channels, the total number of available radio channels can be expressed as:

$$S = kN \qquad \text{.....(3.1)}$$

The N cells which collectively use the complete set of available frequencies is called a cluster. If a cluster is replicated M times within the system, the total number of duplex channels, C, can be used as a measure of capacity and is given by

$$C = MkN = MS \qquad \text{.....(3.2)}$$

As seen from equation (3.2), the capacity of a cellular system is directly proportional to the number of times a cluster is replicated in a fixed service area. The factor N is called the cluster size and is typically equal to 4, 7, or 12. If the cluster size N is reduced while the cell size is kept constant, more clusters are required to cover a given area, and hence more capacity (a larger value of C) is achieved. A large cluster size indicates that the ratio between the cell radius and the distance between co-channel cells is small. Conversely, a small cluster size indicates that co-channel cells are located much closer together. The value for N is a function of how much interference a mobile or base station can tolerate while maintaining a sufficient quality of communications. From a design viewpoint, the smallest possible value of N is desirable in order to maximise capacity over a given coverage area (i.e., to maximise C in equation (3.2)). The frequency re-use factor of a cellular system is given by 1/N, since each cell within a cluster is only assigned 1/N of the total available channels in the system.

Due to the fact that the hexagonal geometry of figure 3.1 has exactly six equidistant neighbours and that the lines joining the centres of any cell and each of its neighbours are separated by multiples of 60 degrees, there are only certain cluster sizes and cell layouts which are possible. In order to tessellate – to connect without gaps between adjacent cells – the geometry of hexagons is such that the number of cells per cluster, N, can only have values which satisfy equation (3.3):

$$N = i^2 + ij + j^2 \qquad \text{.....(3.3)}$$

where i and j are non-negative integers. To find the nearest co-channel neighbours of a particular cell, one must do the following:

1) Move i cells along any chain of hexagons, and then
2) Turn 60 degrees counter-clockwise and move j cells.

This is illustrated in figure 3.2 for i = 3 and j = 2 (e.g., N = 19):

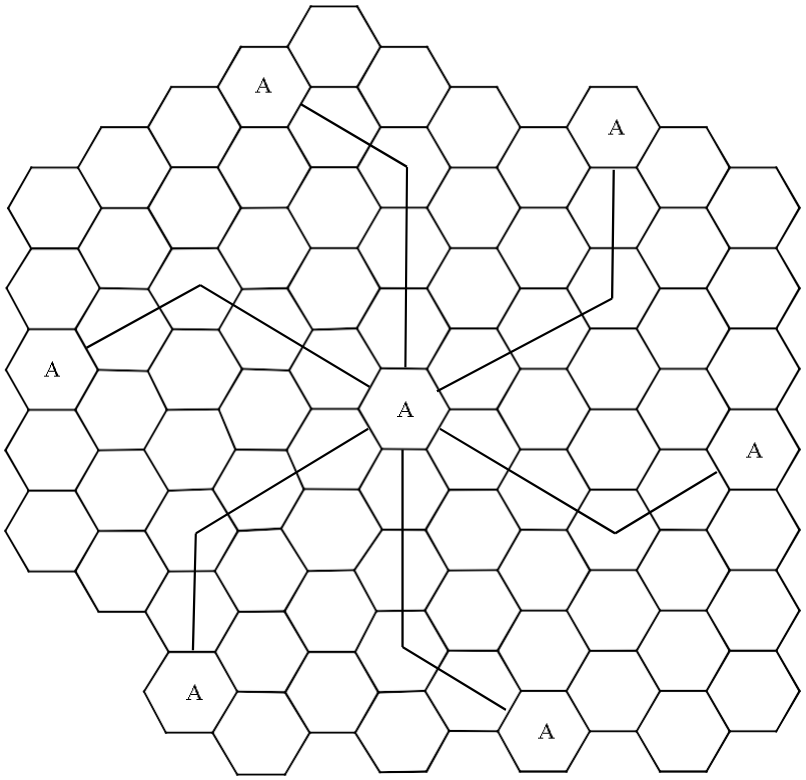


Figure 3.2: Method of Locating Co-Channel Cells in a Cellular System. In this Example, N = 19 (i.e., i = 3, j = 2)

1.3. GSM

1.3.1. Introduction

GSM (Global System for Mobile Communications) is a standard set developed by the European Telecommunications Standards Institute (ETSI) to describe protocols for second generation (2G) digital cellular networks used by mobile phones. It became the de facto global standard for mobile communications with over 80% market share.

GSM is an open, digital cellular technology used for transmitting mobile voice and data services.

GSM supports voice calls and data transfer speeds of up to 9.6 kbps, together with the transmission of SMS (Short Message Service).

GSM is a globally accepted standard for digital cellular communication. GSM is the name of a standardisation group established in 1982 to create a common European mobile telephone standard that would formulate specifications for a European mobile cellular radio system operating at 900MHz.

GSM provides recommendations, not requirements. The GSM specifications define the functions and interface requirements in detail but do not address the hardware. The reason for this is to limit the designers as little as possible but still to make it possible for the operators to buy equipment from different suppliers. The proposed GSM system had to meet certain business objectives:

1) Support for international roaming,
2) Good speech quality,
3) Ability to support handheld devices,
4) Low service cost,
5) Use of spectrum efficiently,
6) Support for a range of new services and facilities, and
7) ISDN compatibility.

1.3.2. Advantages of GSM

GSM is already used worldwide with over 450 million subscribers.

1) Roaming: International roaming permits subscribers to use one phone throughout.
2) Mature Technology: GSM is mature, having started in the mid-80s. This maturity means a more stable network with robust features. CDMA is still building its network.
3) Identity Mobility: GSM network uses SIM, or subscriber identity module, cards to identify users' phones. SIM cards are small chips that contain information like a subscriber's phone number, contacts, preferences and other data. Users can move a single SIM card from one phone to another, making it easy to transfer service between phones without losing important data.

1.3.3. Disadvantages of GSM

1) Lack of Visibility in United States: Another significant disadvantage to the GSM network is its relative lack of visibility in the United States. While US cell phone provider AT&T uses GSM technology, most other US cell phone companies still rely on CDMA. Because of the recent growth in US cell phone use, and the emergence of more complex cellular devices like smart phones, GSM's lack of an American presence puts limits on its growth as a worldwide network leader.
2) Phone Size: One disadvantage of GSM phones is that they must be large enough to incorporate a SIM card slot. This can place limits on the design of GSM-enabled phones, while cell phone manufacturers can work more freely on designs for phones that are intended for other cellular standards and require no SIM card.

1.3.4. GSM Success Factor

1) Comprehensive Services and System Features
   i) GSM offers telephony, short message, fax and data services and a very comprehensive range of supplementary services.
   ii) GSM offers network operators a choice of speech coding methods, full rate, dual rate (half and full rate) and enhanced full rate in order to meet their markets requirements in terms of capacity and quality.
   iii) GSM allows global roaming between more than 100 countries including U.S.A. and Canada. Interstandard roaming GSM/AMPS and GSM/PDC provides a bridge to/from US AMPS network resp. to/from Japan.

2) High Quality, Capacity, and Security
   i) GSM offers high quality telephony, data and fax services.
   ii) GSM offers high spectrum efficiency due to advanced TDMA technology with adaptive power control, slow frequency hopping and discontinuous transmission.
   iii) GSM's advanced security and anti-fraud measures secure the customers privacy and the operators' revenue.

3) Low Equipment Cost due to Vendor's Choice and Market Volume
   i) GSM is an open system standard, not dominated by the intellectual property rights of a single manufacturer.
   ii) GSM offers the widest choice of terminal, network system, and test system manufacturers.
   iii) GSM equipment has a very high market volume and therefore very attractive prices.

1.3.5. GSM Services

The speech- and data-services supported by GSM can be divided into the three categories, which are as follows (figure 1.8):

1) Teleservices,
2) Bearer services, and
3) Supplementary services.

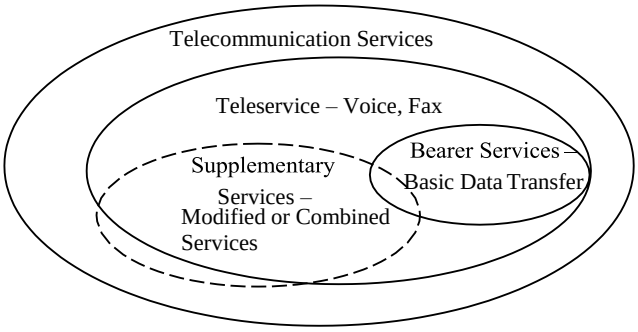


Figure 1.8: Types of Services

1.3.5.1. Teleservices

Teleservices are the services provided to user by mobile service network. It is a point-to-point service which means a service is from one terminal provided to another terminal. Teleservices provides many services some of them are SMS, MMS, Fax, Voice communication at full data rate 13.4 kbps, emergency numbers for emergency calls etc.

1) Voice Calls: The most basic teleservice supported by GSM is telephony. This includes full-rate speech at 13 Kbps and emergency calls, where the nearest emergency- service provider is notified by dialing three digits. A very basic example of emergency service is 100 service available for police in India.
2) Videotext and Facsmile: Another group of teleservices includes Videotext access, Teletex transmission, Facsimile alternate speech and facsimile Group 3, Automatic facsimile Group 3 etc.
3) Short Text Messages: SMS (Short Messaging Service) service is a text messaging which allow you to send and receive text messages on your GSM Mobile phone. Services available from many of the world's GSM networks today - in addition to simple user generated text message services - include news, sport, financial,

language and location based services, as well as many early examples of mobile commerce such as stocks and share prices, mobile banking facilities and leisure booking services.

A teleservice utilises the capabilities of a Bearer Service to transport data, defining which capabilities are required and how they should be set up.

1.3.5.2. Bearer Services

Bearer services are the services which are responsible for transmission of voice and data in terms of digital format over the network. Bearer services are transparent which uses only physical layer protocol for transmission or non-transparent which uses physical layer, data link layer and flow control layer protocol for transmission. These are telecommunication services providing the capability of transmission of signals between access points [the User Network Interfaces (UNIs) in ISDN]. For example, synchronous dedicated packet data access is a bearer service.

Categories of Bearer Services

1) Unrestricted Digital Information (UDI) is designed to offer a peer-to-peer digital link.
2) The 3.1kHz is external to the PLMN and provides a UDI service on the GSM network, interconnected with the ISDN or the PSTN by means of a modem.
3) PAD allows an asynchronous connection to a Packet Assembler/Disassembler (PAD). This enables the PLMN subscribers to access a Packet Switched Public Data Network (PSPDN).
4) Packet enables a synchronous connection to access a PSPDN network and alternate speech and data, providing the capability to switch between voice and data during a call.
5) Speech followed by data first provides a speech connection, and then allows switching during the call for a data connection. The user cannot switch back to speech after the data portion.

1.3.5.3. Supplementary Services

Supplementary services are provided on top of teleservices or bearer services, and include features such as caller identification, call forwarding, call waiting, multi-party conversations, and barring of outgoing (international) calls, among others.

In GSM, supplementary ISDN services are allowed and it is in digital form. It includes call diversion, closed user groups, identification facility and they are very popular. In analog mobile networks, these services are not possible. In addition, the short message services are also made possible in which a brief message/page is sent from between the subscriber and base station.

The page can be of any length within the limitation of 1607 bit ASCII characters. It is interesting to note that simultaneous voice transmission is made possible while a SMS is being sent. The SMS is sent for advisory and safety applications.

Subscriber Identity Module(SIM)

In GSM services another important feature is the memory device called as Subscriber Identity Module (SIM) that can store information like:

1) Subscriber's identification number,
2) User specific information,
3) Privacy keys, and
4) Network where the user is entitled.

This SIM is activated with a personal ID number (four digit number) and gets services from GSM phones. Only if the SIM card (or smart card) is inserted, the phone will be activated or the GSM phone will be in non-operational state.

Next to SMS, SIM services in GSM an on-the-air privacy is provided in the standard that guarantees high degree of privacy by encryption at transmitter end.

Supplementary services include several forms as follows:
1) Number identification services,
2) Call forwarding services,
3) Call completion services,
4) Multiparty services, and
5) Call restriction services.

1.3.6. Channel Structure / System Architecture

The structure of a GSM network relies on several functional entities, which have been specified in terms of functions and interfaces. It involves three main subsystems, each containing functional units and interconnected with the others through a series of standard interfaces as shown in figure 1.10.

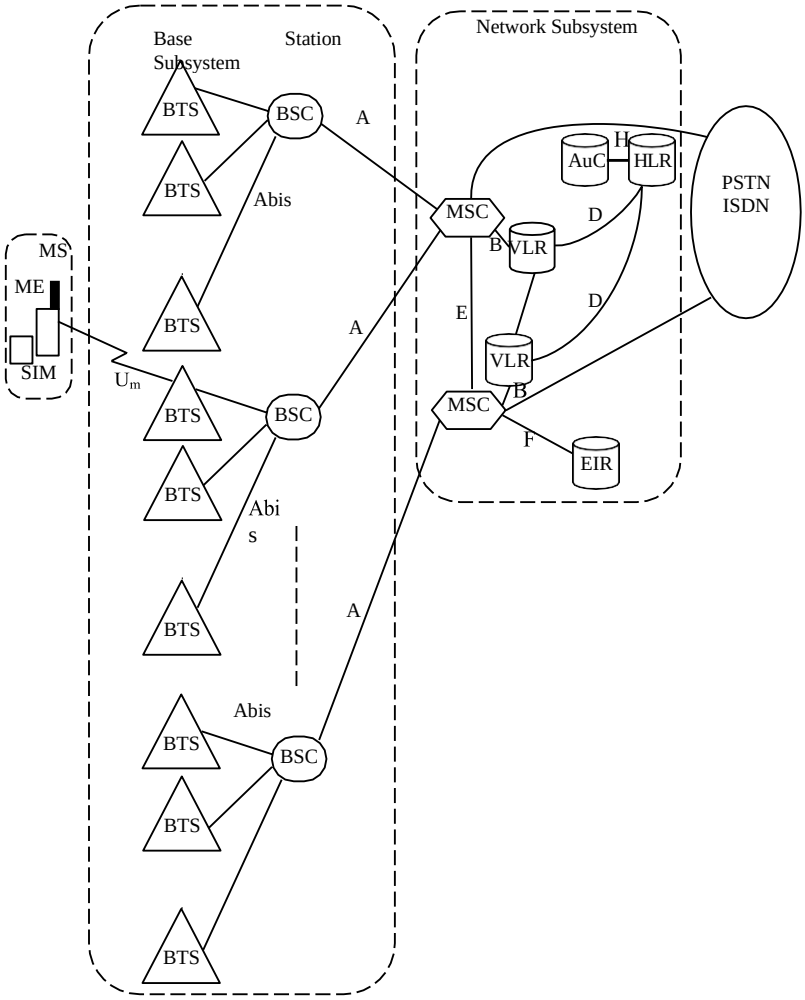


Figure 1.10: System Architecture of a GSM Network

The main components of a GSM network are:
1) MS (Mobile Station),
2) Base Station Subsystem (BSS), and
3) Network and Switching Subsystem (NSS).

1.3.6.1. MS (Mobile Station)

It is the handheld mobile terminal. The MS is made-up of the Mobile Equipment (ME), and a SIM. It performs the following functions:
1) Radio transmission and reception;

2) Source and channel coding and decoding, modulation and demodulation;
3) Audio functions (amplifiers, microphone, earphone);
4) Protocols to handle radio functions – power control, frequency hopping, rules for access to the radio medium;
5) Protocol to handle call control and mobility; and
6) Security algorithms (encryption techniques).

As the SIM enables the user to have access to subscribed services irrespective of a specific terminal. The insertion of the SIM card into any GSM terminal allows the user to receive calls on that terminal, to make calls from that terminal, and to use the other subscribed services. The ME is identified with an International Mobile Equipment Identity (IMEI).

The SIM card contains, among other information, the International Mobile Subscriber Identity (IMSI) used to identify the subscriber to the system, and a secret key for authentication. The IMEI and the IMSI are independent, thereby allowing personal mobility.

1.3.6.2. Base Station Subsystem (BSS)

BSS controls the radio link with the MS.

Components of BSS

Basically it is composed of 2 parts. These two elements communicate across the Abis interface. The BSS is composed of:

1) BSC (Base Station Controller): The BSC manages the radio resources for one or more BTSs. It handles the management of the radio resource, and as such it controls the following functions:
   i) Allocation and release of radio channels,
   ii) Frequency hopping, power control algorithms,
   iii) Handover management,
   iv) Choice of the encryption algorithm, and
   v) Monitoring of the radio link.

2) BTS (Base Transceiver Station): The BTS contains the radio transceivers, responsible for the radio transmission with the MS.

   Functions of BTS
   i) Modulation and demodulation;
   ii) Channel coding and decoding;
   iii) Encryption process;
   iv) RF transmits and receives circuits (power control, frequency hopping, management of antenna diversity, discontinuous transmission).

   Types of BTS
   i) Normal BTS: the functions of normal BTS are rate adaptation, including encoding/decoding for digital channels, signal quality measurements, power change management, control channel management(paging and access) etc.
   ii) Micro BTS: It is different from a normal BTS in two ways, which are as follows:
      a) The range requirements are reduced, and the close proximity requirements are more stringent.
      b) The micro BTS is required to be small and affordable in order to allow external street deployment in large numbers.
   iii) Pico BTS: It is an extension of the micro BTS concept to the indoor environments. The RF performances of these different BTSs are slightly different.

1.3.6.3. Network and Switching Subsystem (NSS)

The Network and Switching Subsystem (NSS) handles the switching of GSM calls between external networks and the BSCs in the radio subsystem and is also responsible for managing and providing external access to

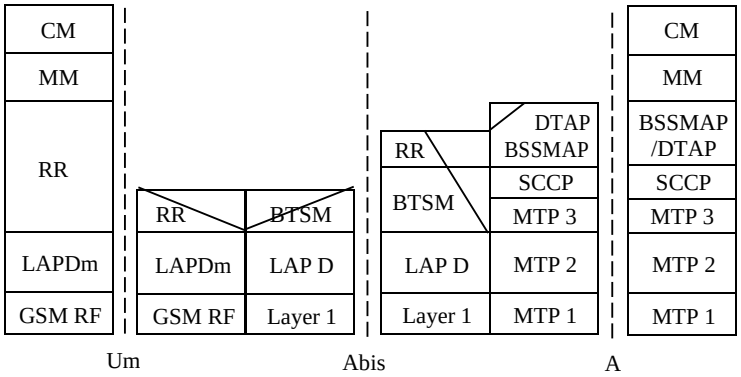
several customer databases. The MSC (Mobile Switching Center) is the central unit in the NSS and controls the traffic among all of the BSCs. The NSS has three different databases called:

1) Home Location Register (HLR): The HLR is a database which contains subscriber information and location information for each user who resides in the same city as the MSC. Each subscriber in a particular GSM market is assigned a unique International Mobile Subscriber Identity (IMSI), and this number is used to identify each home user.
2) Visitor Location Register (VLR): The VLR is a database which temporarily stores the IMSI and customer information for each roaming subscriber who is visiting the coverage area of a particular MSC. The VLR is lined between several adjoining MSCs in a particular market or geographic region and contains subscription information of every visiting user in the area. Once a roaming mobile is logged in the VLR, the MSC sends the necessary information to the visiting subscriber's HLR so that calls to the roaming mobile can be appropriately routed over the PSTN by the roaming user's HLR.
3) Authentication Center (AUC): The Authentication Center is a strongly protected database which handles the authentication and encryption keys for every single subscriber in the HLR and VLR. The Authentication Center contains a register called the Equipment Identity Register (EIR) which identifies stolen or fraudulently altered phones that transmit identity data that does not match with information contained in either the HLR or VLR.

1.3.7. GSM Communication Protocols

The layered model of the GSM architecture integrates and links the peer-to-peer communications between two different systems. The underlying layers satisfy the services of the upper-layer protocols. Notifications are passed from layer to layer to ensure that the information has been properly formatted, transmitted, and received.

The GMS protocol stacks diagram is shown below:



1) MS Protocols: The signalling protocol in GSM is structured into three general layers, depending on the interface:
   i) Layer 1: The physical layer, which uses the channel structures over the air interface.
   ii) Layer 2: The data-link layer. Across the Um interface, the data-link layer is a modified version of the Link Access Protocol for the D channel (LAP-D) protocol used in ISDN, called Link Access Protocol on the Dm channel (LAP-Dm). Across the A interface, the Message Transfer Part (MTP), Layer 2 of SS7 is used.
   iii) Layer 3: The third layer of the GSM signalling protocol is divided into three sublayers:
      a) Radio Resource Management (RR),
      b) Mobility Management (MM), and
      c) Connection Management (CM).

2) MS to BTS Protocols: The RR layer oversees the establishment of a link, both radio and fixed, between the MS and the MSC. The main functional components involved are the MS, the BSS, and the MSC. The RR layer is concerned with the management of an RR-session, which is the time that a mobile is in dedicated mode, as well as the configuration of radio channels, including the allocation of dedicated channels.

The MM layer is built on top of the RR layer and handles the functions that arise from the mobility of the subscriber, as well as the authentication and security aspects. Location management is concerned with the procedures that enable the system to know the current location of a powered-on MS so that incoming call routing can be completed.

The CM layer is responsible for CC. supplementary service management, and Short Message Service (SMS) management. Each of these may be considered as a separate sublayer within the CM layer. Other functions of the CC sublayer include call establishment, selection of the type of service (including alternating between services during a call), and call release.

3) BSC Protocols: After the information is passed from the BTS to the BSC, a different set of interfaces is used. The Abis interface is used between the BTS and BSC. At this level, the radio resources at the lower portion of Layer 3 are changed from the RR to the Base Transceiver Station Management (BTSM). The BTS management layer is a relay function at the BTS to the BSC.

   The RR protocols are responsible for the allocation and re-allocation of traffic channels between the MS and the BTS. These services include controlling the initial access to the system, paging for MT calls, the handover of calls between cell sites, power control, and call termination. The RR protocols provide the procedures for the use, allocation, re-allocation, and release of the GSM channels. The BSC still has some radio resource management in place for the frequency coordination, frequency allocation, and the management of the overall network layer for the Layer 2 interfaces.

   From the BSC, the relay is using SS7 protocols so the MTP 1-3 is used as the underlying architecture, and the BSS mobile application part or the direct application part is used to communicate from the BSC to the MSC.

4) MSC Protocols: At the MSC, the information is mapped across the A interface to the MTP Layers 1 through 3 from the BSC. Here, the equivalent set of radio resources is called the BSS MAP. The BSS MAP/DTAP and the MM and CM are at the upper layers of Layer 3 protocols. This completes the relay process. Through the control-signalling network, the MSCs interact to locate and connect to users throughout the network. Location registers are included in the MSC databases to assist in the role of determining how and whether connections are to be made to roaming users.

   Each user of a GSM MS is assigned a HLR that is used to contain the user's location and subscribed services. A separate register, the VLR, is used to track the location of a user. As the users roam out of the area covered by the HLR, the MS notifies a new VLR of its whereabouts. The VLR in turn uses the control network (which happens to be based on SS7) to signal the HLR of the MS's new location. Through this information, MT calls can be routed to the user by the location information contained in the user's HLR.

1.3.8. GSM Interfaces

One of the main purposes behind the GSM specifications is to define several open interfaces, which then limit certain parts of the GSM system. Because of this interface openness, the operator maintaining the network may obtain different parts of the network from different GSM network suppliers. When an interface is open, it also strictly defines what is happening through the interface, and this in turn strictly defines what kind of actions/procedures/functions must be implemented between the interfaces.

The three interfaces are:

1) Air/Radio Interface: The "air" or radio interface standard that is used for exchanges between a mobile (ME) and a base station (BTS / BSC). For signalling, a modified version of the ISDN LAPD, known as LAPDm is used.
2) Abis Interface: This is a BSS internal interface linking the BSC and a BTS, and it has not been totally standardised. The Abis interface allows control of the radio equipment and radio frequency allocation in the BTS.

3) A Interface: The A interface is used to provide communication between the BSS and the MSC. The interface carries information to enable the channels, timeslots and the like to be allocated to the mobile equipments being serviced by the BSSs. The messaging required within the network to enable handover etc to be undertaken is carried over the interface.

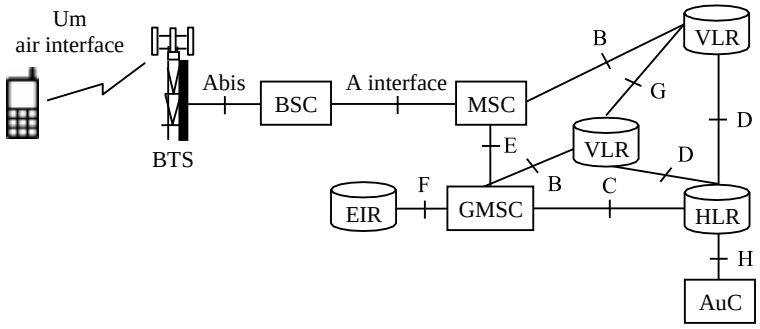


Figure 3.11: GSM Interfaces

1.3.8.1. Air/Radio Interface (MS to BTS)

The air interface between the BTS and MS is known as the Um interface. The manufacturers of network and MS might not be same, but these have to be complaint with each other, in order to work together in a GSM system.

The air interface is defined, so that MS and network manufacturers can design their equipment independently following the standards so that the outcomes will be compatible.

The air interface (RF Interface) uses the Time Division Multiple Access (TDMA) technique to transmit and receive traffic and signalling information between the GSM BTS and GSM Mobile Station.

The Um radio interface (between MS and Base Transceiver Stations [BTS]) is the most important in any mobile radio system, in that it addresses the demanding characteristics of the radio environment. The physical layer interfaces to the data-link layer and radio resource management sublayer in the MS and BS and to other functional units in the MS and network subsystem (which includes the BSS and MSC) for supporting traffic channels. The physical interface comprises a set of physical channels accessible through FDMA and TDMA.

Each physical channel supports a number of logical channels used for user traffic and signalling. The physical layer (or Layer 1) supports the functions required for the transmission of bit streams on the air interface. Layer 1 also provides access capabilities to upper layers. The physical layer is described in the GSM Recommendation 05 series (part of the ETSI documentation for GSM). At the physical level, most signalling messages carried on the radio path are in 23-octet blocks. The data-link layer functions are multiplexing, error detection and correction, flow control, and segmentation to allow for long messages on the upper layers. The radio interface uses the Link Access Protocol on Dm channel (LAPDm). This protocol is based on the principles of the ISDN Link Access Protocol on the D channel (LAPD) protocol. Layer 2 is described in GSM Recommendations 04.05 and 04.06. The following logical channel types are supported:

1) Speech Traffic Channels (TCH)
   i) Full-rate TCH (TCH/F)
   ii) Half-rate TCH (TCH/H)
2) Broadcast Channels (BCCH)
   i) Frequency Correction Channel (FCCH)
   ii) Synchronisation Channel (SCH)
   iii) Broadcast Control Channel (BCCH)
3) Common Control Channels (CCCH)
   i) Paging Channel (PCH)
   ii) Random Access Channel (RACH)
   iii) Access Grant Channel (AGCH)

4) Cell Broadcast Channel (CBCH)
   Cell Broadcast Channel (CBCH) (the CBCH uses the same physical channel as the DCCH)
5) Dedicated Control Channels (DCCH)
   i) Slow Associated Control Channel (SACCH)
   ii) Standalone Dedicated Control Channel (SDCCH)
   iii) Fast Associated Control Channel (FACCH)

The radio resource layer manages the dialog between the MS and BSS concerning the management of the radio connection, including connection establishment, control, release, and changes (e.g., during handover). The mobility management layer deals with supporting functions of location update, authentication, and encryption management in a mobile environment. In the connection management layer, the call control entity controls end-to-end call establishment and management, and the supplementary service entity supports the management of supplementary services. Both protocols are similar to those used in the fixed wireline network. The SMS protocol of this layer supports the high-level functions related to the transfer and management of short message services.

1.3.8.2. A_{bis} Interface (BTS to BSC)

The interface between BTS and BSC is known as A_{bis} standard interface. The primary functions carried-over this interface are traffic channel transmissions, radio channel management, and terrestrial channel management. This interface mainly supports two types of communication links:

1) Traffic channels at 64 kbps, which carry speech or user data for a full- or half-rate radio traffic channel, and
2) Signaling channels at 16 kbps, which carry information for BSC-BTS and BSC-MSC signaling. The BSC handles LAPD channel signaling for every BTS carrier. The lower three layer are based on the OSI/ITU-T recommendation

1.3.8.3. A Interface (BSC to MSC)

The A interface allows interconnection between the BSS radio base subsystem and the MSC. The physical layer of the A interface is a 2-Mbps standard Consultative Committee on Telephone and Telegraph (CCITT) digital connection. The signalling transport uses Message Transfer Part (MTP) and Signalling Connection Control Part (SCCP) of SS7. Error-free transport is handled by a subset of the MTP, and logical connection is handled by a subset of the SCCP. The application parts are divided between the BSS Application Part (BSSAP) and BSS Operation and Maintenance Application Part (BSSOMAP). The BSSAP is further divided into Direct Transfer Application Part (DTAP) and BSS Management Application Part (BSSMAP). The DTAP is used to transfer Layer 3 messages between the MS and the MSC without BSC involvement. The BSSMAP is responsible for all aspects of radio resource handling at the BSS. The BSSOMAP supports all the operation and maintenance communications of BSS.

1.3.9. Location Management

Location management deals with how to keep track of an active mobile station within the cellular network. A mobile station is active if it is powered on. Since the exact location of a mobile station must be known to the network during a call, location management usually means how to track an active mobile station between two consecutive phone calls.

There are two basic operations involved with location management:

1) Paging: It is performed by the cellular network. When an incoming call arrives for a mobile station, the cellular network will page the mobile station in all possible cells to find out the cell in which the mobile station is located so the incoming call can be routed to the corresponding base station. This process is called paging. The number of all possible cells to be paged is dependent on how the location update operation is performed. The location update operation is performed by an active mobile station.
2) Location Update: This scheme can be classified as either global or local. A location update scheme is global if all subscribers update their locations at the same set of cells, and a scheme is local if an individual subscriber is allowed to decide when and where to perform location update. A local scheme is also called individualised or per-user based. From another point of view, a location update scheme can be classified as either static or dynamic. A location update scheme is static if there is a predetermined set of cells at which

location updates must be generated by a mobile station regardless of its mobility. A scheme is dynamic if a location update can be generated by a mobile station in any cell depending on its mobility. A global scheme is based on aggregate statistics and traffic patterns, and it is usually static too. Location areas and reporting centres are two examples of global static schemes. A global scheme can be dynamic. For example, the time-varying location areas scheme is both global and dynamic. A per-user based scheme is based on the statistics and/or mobility patterns of an individual subscriber, and it is usually dynamic. The time-based, movement-based and distance-based schemes are three excellent examples of individualised dynamic schemes. An individualised scheme is not necessary dynamic. For example, the individualised location areas scheme is both individualised and static.

Location management involves signalling in both the wireline portion and the wireless portion of the cellular network. However, most researchers only consider signalling in the wireless portion due to the fact that the radio frequency bandwidth is limited while the bandwidth of the wireline network is always expandable. Location update involves reverse control channels while paging involves forward control channels. The total location management cost is the sum of the location update cost and the paging cost. There is a trade-off between the location update cost and the paging cost. If a mobile station updates its location more frequently (incurring higher location update cost), the network knows the location of the mobile station better. Then the paging cost will be lower when an incoming call arrives for the mobile station. Therefore both location update and paging costs cannot be minimised at the same time. However, the total cost cannot be minimised or one cost can be minimised by putting a bound on the other cost. For example, many researchers try to minimise the location update cost subject to a constraint on the paging cost.

The cost of paging a mobile station over a set of cells or location areas has been studied against the paging delay. There is a trade-off between the paging cost and the paging delay. If there is no delay constraint, the cells can paged sequentially in order of decreasing probability, which will result in the minimal paging cost. If all cells are paged simultaneously, the paging cost reaches the maximum while the paging delay is the minimum, many researchers try to minimise the paging cost under delay constraints.

1.3.9.1. HLR/VLR Scheme

The currently used two popular location management standards are:
1) GSM and
2) IS-41.

They make use of two types of Registers:
1) Home location register (HLR) and
2) Visitor location register (VLR),

to store the location information of the mobile terminals. Figure 1.17 shows the basic architecture under this two-level hierarchy.

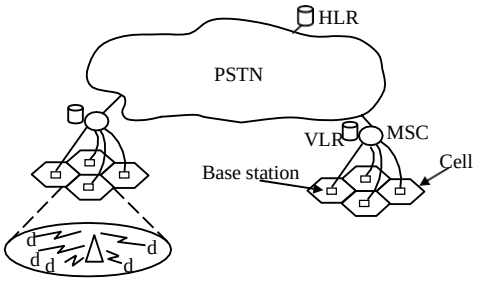


Figure 1.17: Standard PCN Architecture

The HLR stores the user profiles of its assigned subscribers. These user profiles contain information such as the type of services subscribed, the quality-of-service (QoS) requirements and the current location of the mobile terminals.

Each VLR stores replications of the user profiles of the subscribers currently residing in its associated LA (location Area). In order to effectively locate a mobile terminal when a call arrives, each mobile terminal is required to report its location whenever it enters a new LA. This reporting process is called location update.

On receiving a location update message, the MSC updates its associated VLR and transmits the new location information to the HLR. This register update process is called as location registration.

The HLR will acknowledge the MSC for the successful registration and it will also deregister the mobile terminal at the VLR of old LA. In order to locate a mobile terminal for call delivery, the HLR is queried to

determine the serving MSC of the target mobile terminal. The HLR then sends a message to this MSC which, in turn, will determine the serving base station of the mobile terminal by paging all cells within its associated LA.

This location tracking scheme requires the exchange of signaling messages between the HLR and the new and old MSC's whenever the mobile terminal crosses an LA boundary. This may result in significant traffic load to the network especially when the current location of the mobile terminal is far away from its HLR and the mobile terminal is making frequent movements among location area's (LA). Besides, the HLR may experience excessively high database access traffic as it is involved in every location registration and call delivery. This may result in increased connection set up delay during periods of high network utilization.

1.3.9.1.1. Location Registration Scheme

The major steps of the IS-41 location registration scheme are as follows (figure 1.18):

Step 1) The mobile terminal moves into a new LA and sends a location update message to the nearby base station.

Step 2) The base station forwards this message to the new serving MSC.

Step 3) The new MSC updates its associated VLR, indicating that the mobile terminal is now residing in its services area and sends a location registration message to the HLR.

Step 4) The HLR sends a registration acknowledgment message to the new MSC/VLR together with a copy of the subscriber's user profile.

Step 5) The HLR sends a registration cancellation message to the old MSC/VLR.

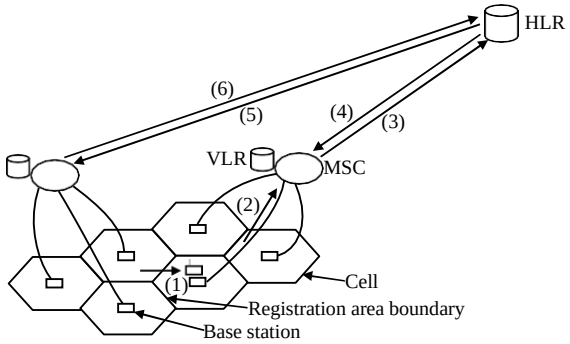


Figure 1.18: Location Registration

Step 6) The old MSC removes the record for the mobile terminal at its associated VLR and sends a cancellation acknowledgment message to the HLR.

1.3.9.1.2. Call Delivery Scheme

The IS-41 call delivery scheme is outlined as follows (figure 1.19):

Step 1) The calling mobile terminal sends a call initiation signal to its serving MSC through the nearby base station.

Step 2) The MSC of the calling mobile terminal sends a location request message to the HLR of the mobile terminal.

Step 3) The HLR determines the current serving MSC of the called mobile terminal (MT) and sends a route request message to this MSC.

Step 4) The MSC determines the cell location of the called mobile terminal and assigns a temporary location directory number (TLDN) to the called mobile terminal. The MSC then sends this TLDN to the HLR.

Step 5) The HLR sends the TLDN to the MSC of the calling mobile terminal. The calling MSC can now set up a connection to the called MSC through the PSTN.

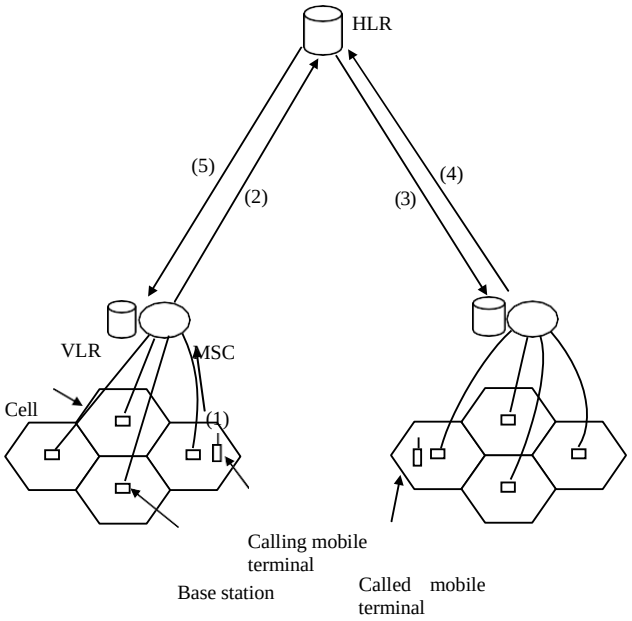


Figure 1.19: Call Delivery

1.3.9.2. Hierarchical Management of Location Information

The GSM network architecture is divided into hierarchical boundaries to facilitate efficient location management. At the lowest layer are the cells, which form the basic building blocks of a mobile network. A cell is uniquely and globally identified by using the Cell Global Identifier (CGI). A collection of cells forms a Location Area (LA), which is identified by a Location Area Identity(LAI).

At any point in time, the exact location of an MS is known by the CGI of the cell where the MS is present. The current location of the MS is maintained and tracked by the network elements (both CN and AN). Considering the fact that the network elements are themselves hierarchically organized, the location information can also be hierarchically maintained across the network elements in a distributed fashion. The information maintained in each network element is as shown in table 1.3.

Table 1.3: Maintenance of Location Information

Location Identifier	BSC	Network Element MSC/VLR	HLR
Cell Identity	Yes *	No	No
Location Area Identity	No	Yes	No
MSC/VLR Number	No	N/A	Yes

* Only in Connected State (refer to MM state model) N/A: Not Applicable

The HLR of a subscriber is the permanent store for all information about the subscriber. Apart from other subscription information, it also maintains the location information of the subscriber. This includes information about the existing MSC and the VLR in which the subscriber is currently registered. Thus, while the HLR stores information on the existing MSC/VLR of the subscriber, the MSC/ VLR itself stores information of the current LA of the subscriber. Further, for an MS in the connected state, the GSM RAN (BSQ to be precise) maintains information on the exact cell in which the subscriber is available. This is implicitly known to the BSC since in the connected state, me MS maintains a radio connection with the network. As a result of this radio connection, the BTS with which the MS is communicating is known, and from this, its location at the cell level On the other hand, for an MS in the idle mode, the GSM RAN does not maintain any location information, and hence, the exact cell level information is not available with the network in this state.

Figure 4 shows hierarchical location management structure

1.3.10. Handsoff / Handover

When a mobile terminal moves outside the coverage area of its base station, the network management is assumed to take appropriate measures. A "handover" or "handoff" to another base station is required to ensure sufficient quality of reception, including acceptable interference power levels.

In cellular telecommunication, the term handover or handoff refers to the process of transferring an ongoing call or data session from one channel connected to the core network to another. In satellite communications, it is the process of transferring satellite control responsibility from one earth station to another without loss or interruption of service.

Call handover is the process of transferring a call between base stations. Handover is typically called handoff in North America. Handover is necessary because mobile stations often move out of range of one base station and into the radio coverage area of another base station.

Figure 1.20 shows the basic call handover process:

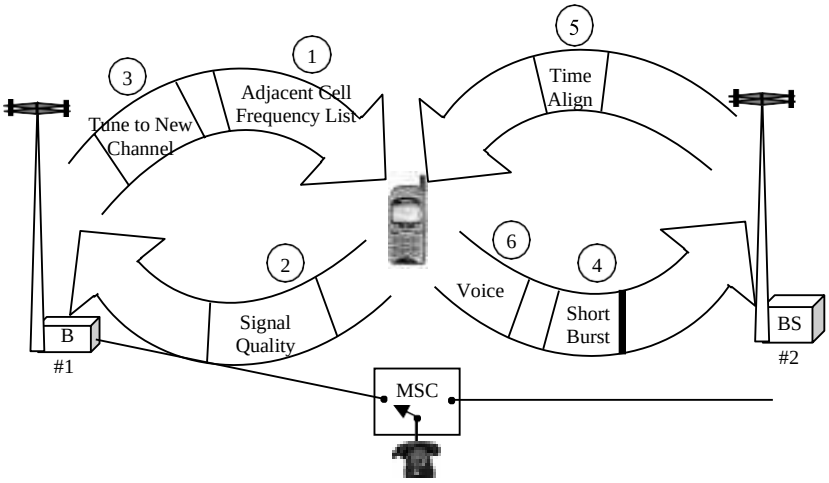


Figure 1.20: Call Handover

In this figure 1.20, an MS is communicating with base station #1:
Step 1: Adjacent Cell Frequency List: Base station #1 provides the MS with a list of radio carrier channels to measure of nearby base stations.
Step 2: Signal Quality Levels: After the MS measures the quality of the radio carrier channels, it returns this information to the serving base station.

Step 3: Tune to New Channel: Using this information and information from neighboring base stations, the serving base station sends a handover message, which instructs the MS to tune to a new radio carrier channel of the adjacent base station #2.

Step 4: Short Burst: The MS begins transmission on the new channel by sending a short burst.

Step 5: Time Align: The new base station uses this information to send a command to adjust the relative timing of the MS.

Step 6: Voice: After the MS has adjusted, the voice channel from the MSC is switched from base station #1 to base station #2 and voice conversation can continue.

Types of Handover in GSM
Figure 2.16 shows four possible handover scenarios in GSM, which are as follows:

1) Intra-Cell Handover: Within a cell, narrow-band interference could make transmission at a certain frequency impossible. The BSC could then decide to change the carrier frequency (scenario 1).

2) Inter-Cell, Intra-BSC Handover: This is a typical handover scenario. The mobile station moves from one cell to another, but stays within the control of the same BSC. The BSC then performs a handover, assigns a new radio channel in the new cell and releases the old one (scenario 2).

3) Inter-BSC, Intra-MSC Handover: As a BSC only controls a limited number of cells; GSM also has to perform handovers between cells controlled by different BSCs. This handover then has to be controlled by the MSC (scenario 3).

4) Inter-MSC Handover: A handover could be required between two cells belonging to different MSCs. Now both MSCs perform the handover together (scenario 4).

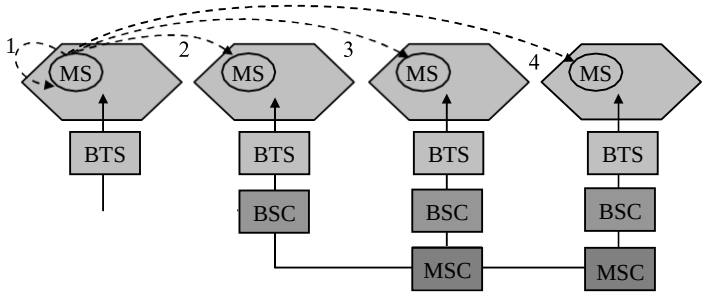


Figure 1.21: Types of Handover in GSM
1) Intra-cell: change in frequency
2) Inter-cell, Intra-BSC
3) Inter-BSC, Intra-MSC
4) Inter-MSC

1.4. CDMA

1.4.1. Introduction

Code Division Multiple Access (CDMA) systems, the narrowband message signal is multiplied by a very large bandwidth signal called the spreading signal. The spreading signal is a pseudo-noise code sequence that has a chip rate, which is order of magnitudes greater than the data rate of the message. All users in a CDMA system, as seen from figure 1.31, use the same carrier frequency and may transmit simultaneously. Each user has its own pseudorandom codeword which is approximately orthogonal to all other codewords. The receiver performs a time correlation operation to detect only the specific desired codeword. All other codewords appear as noise due to decorrelation. For detection of the message signal, the receiver needs to know the codeword used by the transmitter. Each user operates independently with no knowledge of the other users.

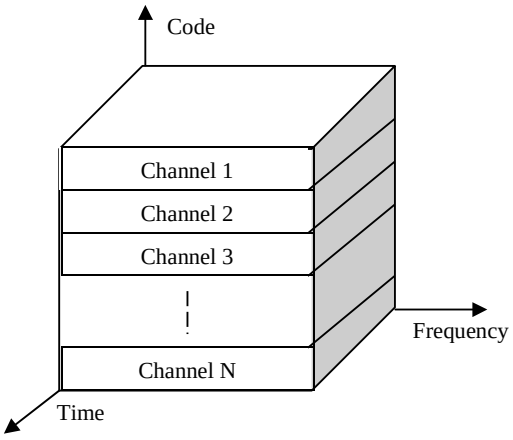


Figure 1.31: Spread Spectrum Multiple Access in which each channel is assigned a unique PN code which is orthogonal or approximately orthogonal to PN codes used by other users

Power control is used in most CDMA implementations. Power control is provided by each base station in a cellular system and assures that each mobile within the base station coverage area provides the same signal level to the base station receiver. This solves the problem of a nearby subscriber overpowering the base station receiver and drowning out the signals of faraway subscribers. Power control is implemented at the base station by rapidly sampling the Radio Signal Strength Indicator (RSSI) levels of each mobile and then sending a power change command over the forward radio link. Despite the use of power control within each cell, out-of-cell mobiles provide interference which is not under the control of the receiving base station.

Table: Different CDMA Standards

CDMA Standards	Description
2.5G+ Standards	
cdmaOne/IS-95 (interim standards 95)	Founded in 1991, developed by QUALCOM, U.S.A., operates at 824-849MHz and 869-894MHz, can transfer signals from multiple sources and users, uses M-sequence PN codes and Walsh codes.
3G Standards	
WCDMA(wide CDMA)/3GPP(3G partnership project)	It supports 2Mbps or higher rates for short distance and 384kbps for long distance transmission. It has a 3.84Mbps chipping rate. It uses both short and long scrambling codes for uplink only.
CDMA 2000	Started in 2001 and accepted as CDMA standards in 2004, used for voice, data and multimedia applications. Employs multiple n carriers, multi-chip rate, frequency division duplex for forward and reverse links. Uses M-sequence PN codes, variable length Walsh codes for data of variable rates.
UMTS (universal telecommunication system)	Communicates at data rate of 100kbps to 2Mbps, supports several technologies for transmission, supports a 3.84Mbps chipping rate, e.g., BlackBerry 7130e.

1.4.2. Features of CDMA

The features of CDMA including the following:

1) Many users of a CDMA system share the same frequency. Either Time Division Demultiplexing or Frequency Division Demultiplexing may be used.
2) Unlike TDMA or FDMA, CDMA has a soft capacity limit. Increasing the number of users in a CDMA system raises the noise floor in a linear manner. Thus, there is no absolute limit on the number of users in CDMA. Rather, the system performance gradually degrades for all users as the number of users is increased, and improves as the number of users is decreased.
3) Multipath fading may be substantially reduced because the signal is spread over a large spectrum. If the spread spectrum bandwidth is greater than the coherence bandwidth of the channel, the inherent frequency diversity will mitigate the effects of small-scale fading.
4) Channel data rates are very high in CDMA systems. Consequently, the symbol (chip) duration is very short and usually much less than the channel delay spread. Since Pseudo Noise sequences have low autocorrelation, multipath which is delayed by more than a chip will appear as noise.
5) Since CDMA uses co-channel cells, it can use macroscopic spatial diversity to provide soft handoff. Soft handoff is performed by the Mobile Switching Center, which can simultaneously monitor a particular user from two or more base stations. The MSC may choose the best version of the signal at any time without switching frequencies.
6) Self-jamming is a problem in CDMA system. Self-jamming arises from the fact that the spreading sequences of different users are not exactly orthogonal, hence in the despreading of a particular PN code, non-zero contributions to the receiver decision statistic for a desired user arise from the transmissions of other users in the system.
7) The near-far problem occurs at a CDMA receiver if an undesired user has a high detected power as compared to the desired user.

1.4.3. CDMA Spread Spectrum

CDMA is based around the use of direct sequence spread spectrum techniques. Essentially CDMA is a form of spread spectrum transmission which uses spreading codes to spread the signal out over a wider bandwidth then would normally be required. By using CDMA spread spectrum technology, many users are able to use the same channel and gain access to the system without causing undue interference to each other.

The key element of code division multiple access CDMA is its use of a form of transmission known as direct sequence spread spectrum(DSSS).

CDMA Spread Spectrum Encode/Decode Process

When transmitting a CDMA spread spectrum signal, the required data signal is multiplied with what is known as a spreading or chip code data stream. The resulting data stream has a higher data rate than the data itself. Often the data is multiplied using the XOR (exclusive OR) function.

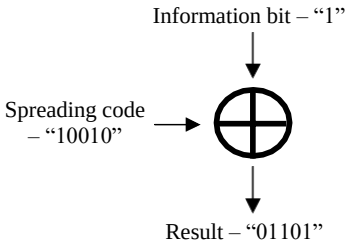


Figure 1.49: CDMA Spreading

Each bit in the spreading sequence is called a chip, and this is much shorter than each information bit. The spreading sequence or chip sequence has the same data rate as the final output from the spreading multiplier. The rate is called the chip rate, and this is often measured in terms of a number of M chips / sec.

The baseband data stream is then modulated onto a carrier and in this way the overall the overall signal is spread over a much wider bandwidth than if the data had been simply modulated onto the carrier. This is because; signals with high data rates occupy wider signal bandwidths than those with low data rates.

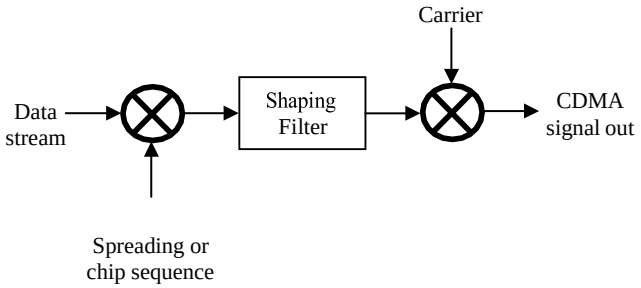


Figure 1.50: CDMA Spread Spectrum Generation

To decode the signal and receive the original data, the CDMA signal is first demodulated from the carrier to reconstitute the high speed data stream. This is multiplied with the spreading code to regenerate the original data. When this is done, then only the data with that was generated with the same spreading code is regenerated, all the other data that is generated from different spreading code streams is ignored.

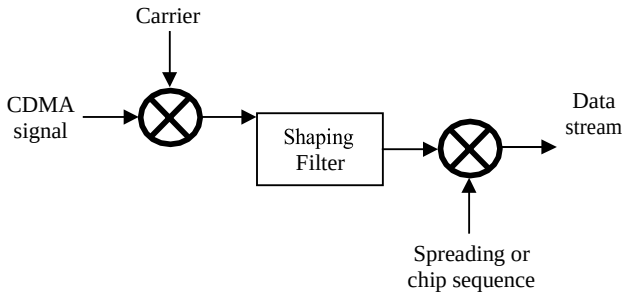


Figure 1.51: CDMA Spread Spectrum Decoding

1.4.4. CDMA Orthogonal Spreading Codes

CDMA orthogonal spreading codes are one of the major elements within the whole CDMA system. The CDMA orthogonal spreading codes are combined with the data stream to be transmitted in such a way that the bandwidth required is increased and the benefits of the spread spectrum system can be gained.

The CDMA codes are specific to each channel / user so that the different users can gain access to the system and communicate as required.

CDMA Codes and Correlation

The concept of CDMA is based around the fact that a data sequence is multiplied by a spreading code or sequence which increases the bandwidth of the signal. Then within the receiver the same spreading code or sequence is used to extract the required data. Only when the required code is used, does the required data appear from the signal.

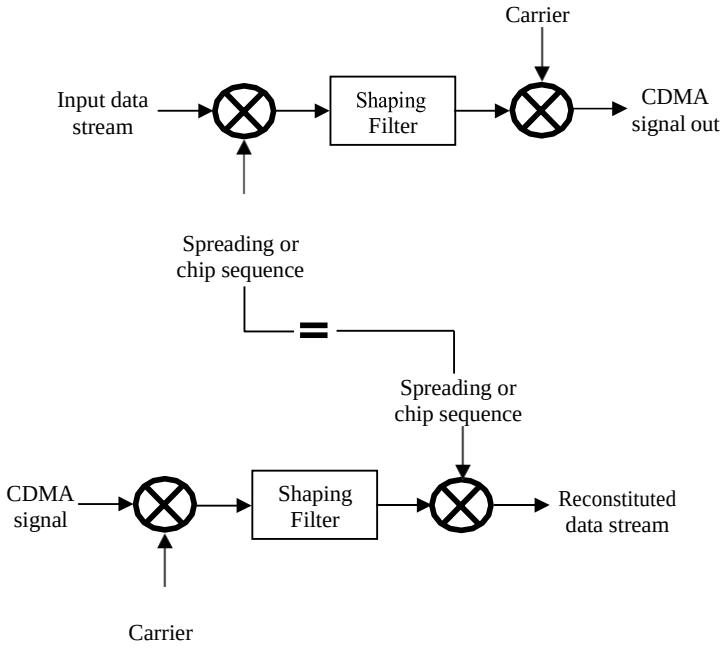


Figure 1.52: CDMA System showing Use of Spreading Codes

The process of extracting the data is called correlation. When a code exactly the same as that used in the transmitter is used, then it is said to have a correlation of one and data is extracted.

When a spreading code that does not correlate is used, then the data will not be extracted and a different set of data will appear. This means that it is necessary for the same spreading code to be used within the transmitter and receiver for the data to be extracted.

CDMA Code Types

There are several types of codes that can be used within a CDMA system for providing the spreading function:

1) PN Codes: Pseudo-random number codes (pseudo-noise or PN code) can be generated very easily. These codes will sum to zero over a period of time. Although the sequence is deterministic because of the limited length of the linear shift register used to generate the sequence, they provide a PN code that can be used within a CDMA system to provide the spreading code required. They are used within many systems as there is a very large number that can be used.

   A feature of PN codes is that if the same versions of the PN code are time shifted, then they become almost orthogonal, and can be used as virtually orthogonal codes within a CDMA system.

2) Truly Orthogonal Codes: Two codes are said to be orthogonal if when they are multiplied together the result is added over a period of time they sum to zero. For example, a codes 1 −1 −1 1 and 1 −1 1 −1 when multiplied together give 1 1 −1 −1 which gives the sum zero.

   An example of an orthogonal code set is the Walsh codes used within the IS95 / CDMA2000 system.

1.4.5. Advantages of CDMA

1) Increased cellular communications security.
2) Simultaneous conversations.
3) Increased efficiency, meaning that the carrier can serve more subscribers.
4) Smaller phones.
5) Low power requirements and little cell-to-cell coordination needed by operators.
6) Extended reach - beneficial to rural users situated far from cells.

1.4.6. Disadvantages of CDMA

1) Due to its proprietary nature, all of CDMA's flaws are not known to the engineering community.
2) CDMA is relatively new, and the network is not as mature as GSM.
3) CDMA cannot offer international roaming, a large GSM advantage.

1.5. GPRS

1.5.1. Introduction

GSM was the most successful second generation cellular technology, but the need for higher data rates spawned new developments to enable data to be transferred at much higher rates.

The first system to make an impact on the market was GPRS. The GPRS stand for General Packet Radio System. GPRS technology enabled much higher data rates to be conveyed over a cellular network when compared to GSM that was voice centric.

GPRS technology became the first stepping-stone on the path between the second-generation GSM cellular technology and the 3G W-CDMA / UMTS system. With GPRS technology offering data services with data rates up to a maximum of 172 kbps, facilities such as web browsing and other services requiring data transfer became possible. Although some data could be transferred using GSM, the rate was too slow for real data applications. It is a packet-based wireless communication service.

1.5.2. Packet Switching

The key element of GPRS technology is that it uses packet switched data rather than circuit switched data, and this technique makes much more efficient use of the available capacity. This is because most data transfer occurs in what is often termed a "bursty" fashion. The transfer occurs in short peaks, followed by breaks when there is little or no activity.

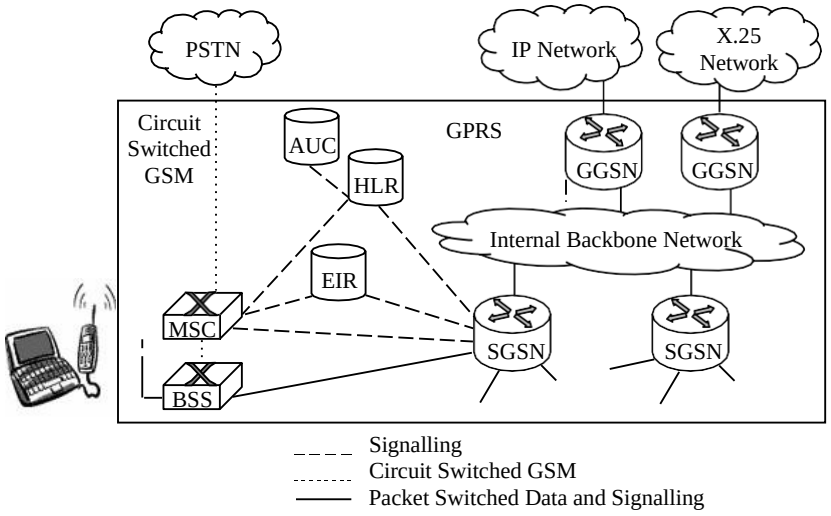
Using a traditional approach a circuit is switched permanently to a particular user. This is known as a circuit switched mode. In view of the bursty nature of data transfer it means that there are periods when it will not be carrying data.

To improve the situation the overall capacity can be shared between several users. To achieve this, the data is split into packets and tags inserted into the packet to provide the destination address. Packets from several sources can then be transmitted over the link. As it is unlikely that the data burst for different users will occur all at the same time, by sharing the overall resource in this fashion, the channel, or combined channels can be used far more efficiently. This approach is known as packet switching, and it is at the core of many cellular data systems, and in this case GPRS.

1.5.3. GPRS Architecture

GPRS architecture works on the same procedure like GSM network, but has additional entities that allow packet data transmission. This data network overlaps a second-generation GSM network providing packet data transport at the rates from 9.6 to 171kbps. Alongwith the packet data transport the GSM network accommodates multiple users to share the same air interface resources concurrently.

Following is the GPRS Architecture diagram:



GPRS attempts to re-use the existing GSM network elements as much as possible, but to effectively build a packet-based mobile cellular network, some new network elements, interfaces, and protocols for handling packet traffic are required.

Therefore, GPRS requires modifications to numerous GSM network elements as summarised below:

GSM Network Element	Modification or Upgrade Required for GPRS
Mobile Station (MS)	New mobile station is required to access GPRS services. These new terminals will be backward compatible with GSM for voice calls.
BTS	A software upgrade is required in the existing Base Transceiver Station (BTS).
BSC	The Base Station Controller (BSC) requires a software upgrade and the installation of new hardware called the Packet Control Unit (PCU). The PCU directs the data traffic to the GPRS network and can be a separate hardware element associated with the BSC.
GPRS Support Nodes (GSNs)	The deployment of GPRS requires the installation of new core network elements called the Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN).
Databases (HLR, VLR, etc.)	All the databases involved in the network will require software upgrades to handle the new call models and functions introduced by GPRS.

GPRS System architecture includes:

1) Serving GPRS Support Node (SGSN): A Serving GPRS Support Node (SGSN) is at the same hierarchical level as the MSC. Whatever functions MSC does for voice, SGSN does the same for packet data. SGSN's tasks include packet switching, routing and transfer, mobility management (attach/detach and location management), logical link management, and authentication and charging functions. SGSN processes registration of new mobile subscribers and keeps a record of their location inside a given service area. The location register of the SGSN stores location information (e.g., current cell, current VLR) and user profiles of all GPRS users registered with this SGSN. SGSN sends queries to Home Location Register (HLR) to obtain profile data of GPRS subscribers. The SGSN is connected to the base station system with Frame Relay.
2) Gateway GPRS Support Node (GGSN): A gateway GPRS Support Node (GGSN) acts as an interface between the GPRS backbone network and the external packet data networks. GGSN's function is similar to that of a router in a LAN. GGSN maintains routing information that is necessary to tunnel the Protocol Data Units (PDUs) to the SGSNs that service particular mobile stations. It converts the GPRS packets coming from the SGSN into the appropriate Packet Data Protocol (PDP) format for the data networks like Internet or X.25. PDP sends these packets out on the corresponding packet data network.

In the other direction, PDP receives incoming data packets from data networks and converts them to the GSM address of the destination user. The re-addressed packets are sent to the responsible SGSN. For this purpose, the GGSN stores the current SGSN address of the user and his or her profile in its location register. The GGSN also performs authentication and charging functions related to data transfer.

In addition to the new GPRS components (SGSN and GGSN), some existing GSM network elements must also be enhanced in order to support packet data. These are:

1) Base Station System (BSS): BSS system needs enhancement to recognise and send packet data. This includes BTS upgrade to allow transportation of user data to the SGSN. Also, the BTS needs to be upgraded to support packet data transportation between the BTS and the MS (Mobile Station) over the radio.

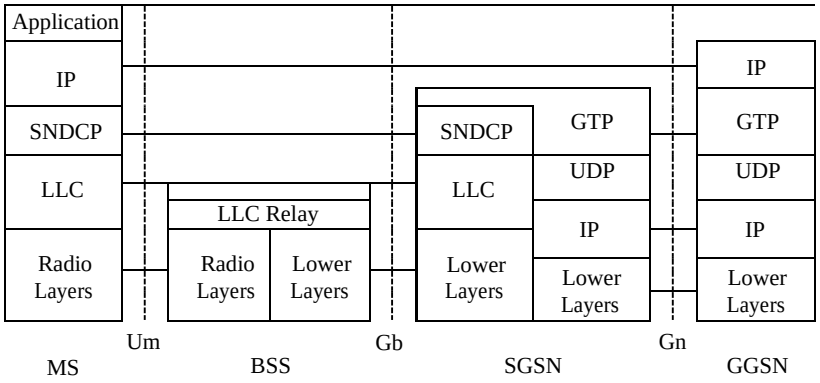
2) Home Location Register (HLR): HLR needs enhancement to register GPRS user profiles and respond to queries originating from GSNs regarding these profiles.

3) Mobile Station (MS): The mobile station or the mobile phone for GPRS is different from that of GSM.

4) SMS Nodes: SMS-GMSCs and SMS-IWMSCs are upgraded to support SMS transmission via the SGSN. Optionally, the MSC/VLR can be enhanced for more efficient coordination of GPRS and non-GPRS services and functionality.

1.5.4. GPRS Protocol Stack

The flow of GPRS protocol stack and end-to-end message from MS to the GGSN is displayed in the below diagram:



GTP is the protocol used between the SGSN and GGSN using the Gn interface. This is a Layer 3 tunnelling protocol.

The process that takes place in the application looks like a normal IP sub-network for the users both inside and outside the network. The vital thing that needs attention is, the application communicates via standard IP, that is carried through the GPRS network and out through the gateway GPRS. The packets that are mobile between the GGSN and the SGSN use the GPRS tunnelling protocol, this way the IP addresses located on the external side of the GPRS network do not have deal with the internal backbone. UDP and IP are run by GTP.

SubNetwork Dependent Convergence Protocol (SNDCP) and Logical Link Control (LLC) combination used in between the SGSN and the MS. The SNDCP flattens data to reduce the load on the radio channel. A safe logical link by encrypting packets is provided by LLC and the same LLC link is used as long as a mobile is under a single SGSN.

In case, the mobile moves to a new routing area that lies under a different SGSN; then, the old LLC link is removed and a new link is established with the new Serving GSN X.25. Services are provided by running X.25 on top of TCP/IP in the internal backbone.

1.5.5. Quality of Service

Quality of Service (QoS) requirements of conventional mobile packet data applications are in assorted forms. The QoS is a vital feature of GPRS services as there are different QoS support requirements for assorted GPRS applications like real-time multimedia, web browsing, and e-mail transfer.

GPRS allows defining QoS profiles using the following parameters:

1) Service Precedence,
2) Reliability,
3) Delay, and
4) Throughput.

These parameters are described below:

1) Service Precedence: The preference given to a service when compared to another service is known as Service Precedence. This level of priority is classified into three levels called:
   i) High,
   ii) Normal, and
   iii) Low.

   When there is network congestion, the packets of low priority are discarded as compared to high or normal priority packets.
2) Reliability: This parameter signifies the transmission characteristics required by an application. The reliability classes are defined which guarantee certain maximum values for the probability of loss, duplication, missequencing, and corruption of packets.
3) Delay: It is defined as the end-to-end transfer time between two communicating mobile stations or between a mobile station and the GI interface to an external packet data network.

   This includes all delays within the GPRS network, e.g., the delay for request and assignment of radio resources and the transit delay in the GPRS backbone network. Transfer delays outside the GPRS network, e.g., in external transit networks, are not taken into account.
4) Throughput: It specifies the maximum/peak bit rate and the mean bit rate.

Using these QoS classes, QoS profiles can be negotiated between the mobile user and the network for each session, depending on the QoS demand and the available resources.

The billing of the service is then based on the transmitted data volume, the type of service, and the chosen QoS profile.

1.5.6. Advantages of GPRS

1) Speed: GPRS is packet switched. Higher connection speeds are attainable at around 56-118kbps, a vast improvement on circuit switched networks of 9.6kbps. By combining standard GSM time slots theoretical speeds of 171.2kbps are attainable. However in the very short-term, speeds of 20-50kbps are more realistic.
2) Always-On Connectivity: GPRS is an always-on service. There is no need to dial up like you have to on a home PC for instance. This feature is not unique to GPRS but is an important standard that will no doubt be a key feature for migration to 3G. It makes services instantaneously available to a device.
3) New and Better Applications: Due to its high-speed connection and always-on connectivity GPRS enables full internet applications and services such as videoconferencing straight to your desktop or mobile device. Users are able to explore the internet or their own corporate networks more efficiently than they could when using GSM. There is often no need to re-develop existing applications.
4) GSM Operator Costs: GSM network providers do not have to start from scratch to deploy GPRS. GPRS is an upgrade to the existing network that sits along side the GSM network. This makes it easier to deploy, there is little or no downtime of the existing GSM network whilst implementation takes place, most updates are software so they can be administered remotely and it allows GSM providers to add value to their business at relatively small costs.

The GSM network still provides voice and the GPRS network handles data, because of this voice and data can be sent and received at the same time.

1.5.7. Disadvantages of GPRS

1) Overhead of Standby: Since GPRS uses the cellular network's GSM band to transmit data, more often than not, when a connection is active, calls and other network-related functions cannot be used. The data session will go on standby. This is a characteristic typical of the Class B GPRS device. There are Class A devices as well, where there are two radios incorporated into the device, allowing both features to run simultaneously. However, Class A devices tend to be more expensive, and by extension, less popular. Most mobile phones fall in the Class B category.

2) Heavy Bills: GPRS is usually billed per megabyte or kilobyte, depending on the individual service provider. However, this has changed in many places, where GPRS downloads are no longer charged as per usage, but are unlimited, and there is merely a flat fee to be paid every month.

LECTURE NOTESON

MOBILE COMPUTING

UNIT – II

Wireless Networking, Wireless LAN Overview: MAC issues, IEEE 802.11, Blue Tooth, Wireless multiple access protocols, TCP over wireless, Wireless applications, data broadcasting, Mobile IP, WAP : Architecture, Traditional TCP, Classical TCP, improvements in WAP, WAP applications.

Wireless Network

Wireless networks are computer networks that are not connected by cables of any kind. The use of a wireless network enables enterprises to avoid the costly process of introducing cables into buildings or as a connection between different equipment locations. The basis of wireless systems are radio waves, an implementation that takes place at the physical level of network structure.

Wireless networks use radio waves to connect devices such as laptops to the Internet, the business network and applications. When laptops are connected to Wi-Fi hot spots in public places, the connection is established to that business's wireless network.

There are four main types of wireless networks:

- Wireless Local Area Network (LAN): Links two or more devices using a wireless distribution method, providing a connection through access points to the wider Internet.
- Wireless Metropolitan Area Networks (MAN): Connects several wireless LANs.
- Wireless Wide Area Network (WAN): Covers large areas such as neighboring towns and cities.
- Wireless Personal Area Network (PAN): Interconnects devices in a short span, generally within a person's reach.

WIRELESS MAC ISSUES

The three important issues are:

1. Half Duplex operation –> either send or receive but not both at a given time
2. Time varying channel
3. Burst channel errors

1. Half Duplex Operation

In wireless, it's difficult to receive data when the transmitter is sending the data, because: When node is transmitting, a large fraction of the signal energy leaks into the receiver path. The transmitted and received power levels can differ by orders of magnitude. The leakage signal typically has much higher power than the received signal -—Impossible to detect a received signal, while transmitting data‖. Collision detection is not possible, while sending data. As collision cannot be detected by the sender, all proposed protocols attempt to minimize the probability of collision - Focus on collision avoidance.

2. Time Varying Channel

Three mechanisms for radio signal propagation

· **Reflection** – occurs when a propagating wave impinges upon an object that has very large dimensions than the wavelength of the radio wave e.g. reflection occurs from the surface of the earth and from buildings and walls

· **Diffraction** – occurs when the radio path between the transmitter and the receiver is obstructed by a surface with sharp edges

· **Scattering** – occurs when the medium through which the wave travels consists of objects with

The received signal by a node is a superposition of time-shifted and attenuated versions of the transmitted signals the received signal varies with time .The time varying signals (time varying channel) phenomenon also known as multipath propagation. The rate of variation of channel is determined by the coherence time of the channel Coherence time is defined as time within which When a node's received signal strength

drops below a certain threshold the node is said to be in fade .Handshaking is widely used strategy to ensure the link quality is good enough for data communication. A successful handshake between a sender and a receiver (small message) indicates a good communication link.

3. Burst Channel Errors

As a consequence of time varying channel and varying signals strengths errors are introduced in the transmission (Very likely) for wire line networks the bit error rate (BER) is the probability of packet error is small .For wire line networks the errors are due to random For wireless networks the BER is as high. For wireless networks the errors are due to node being in fade as a result errors occur in a long burst. Packet loss due to burst errors - mitigation techniques

· Smaller packets

· Forward Error Correcting Codes

· Retransmissions (Acks)

Location Dependent Carrier Sensing

Location Dependent Carrier Sensing results in three types of nodes that protocols need to deal with

Hidden Nodes: Even if the medium is free near the transmitter, it may not be free near the

intended receiver

Exposed Nodes: Even if the medium is busy near the transmitter, it may be free near the intended receiver

Capture: Capture occurs when a receiver can cleanly receive a transmission from one of two simultaneous transmissions

Hidden Node/Terminal Problem

A hidden node is one that is within the range of the intended destination but out of range of sender Node B can communicate with A and C both A and C cannot hear each other When A transmits to B, C cannot detect the transmission using the carrier sense mechanism C falsely thinks that the channel is idle

Exposed Nodes
An exposed node is one that is within the range of the sender but out of range of destination .when a node‘s received signal strength drops below a certain threshold the node is said to be in fade .Handshaking is widely used strategy to ensure the link quality is good enough for data communication. A successful handshake between a sender and a receiver (small message) indicates a good communication link.

In theory C can therefore have a parallel transmission with any node that cannot hear the transmission from B, i.e. out of range of B. But C will not transmit to any node because its an exposed node. Exposed nodes waste bandwidth.

Capture
Capture is said to occur when a receiver can cleanly receive a transmission from one of two simultaneous transmissions both within its range Assume node A and D transmit simultaneously to B. The signal strength received from D is much higher than that from A, and

D‘s transmission can be decoded without errors in presence of transmissions from A.D has captured A. Capture is unfair because it gives preference to nodes that are closer to the receiver. It may improve protocol performance.

IEEE 802.11

Wireless LANs are those Local Area Networks that use high frequency radio waves instead of cables for connecting the devices in LAN. Users connected by WLANs can move around within the area of network coverage. Most WLANs are based upon the standard IEEE 802.11 or WiFi.

IEEE 802.11 Architecture

The components of an IEEE 802.11 architecture are as follows

1) Stations (STA) − Stations comprise all devices and equipments that are connected to the wireless LAN. A station can be of two types:

- **Wireless Access Pointz (WAP) –** WAPs or simply access points (AP) are generally wireless routers that form the base stations or access.
- **Client. –** Clients are workstations, computers, laptops, printers, smartphones, etc.

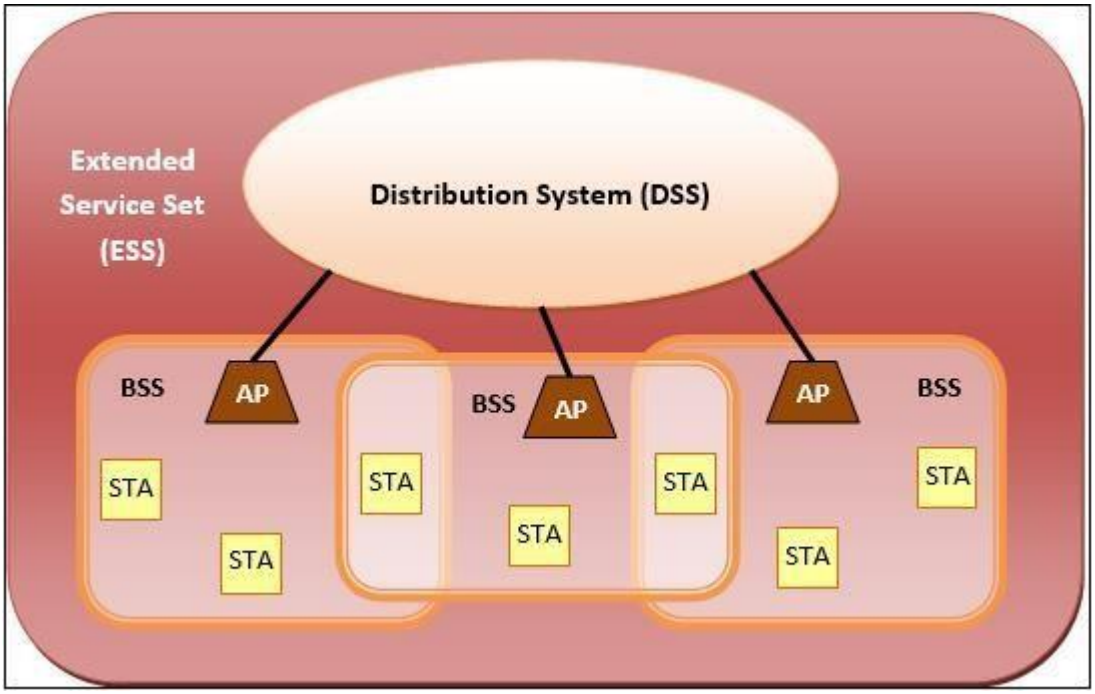
Each station has a wireless network interface controller.

2) Basic Service Set (BSS) −A basic service set is a group of stations communicating at physical layer level. BSS can be of two categories depending upon mode of operation:

- **Infrastructure BSS –** Here, the devices communicate with other devices through access points.
- **Independent BSS –** Here, the devices communicate in peer-to-peer basis in an ad hoc manner.

3) Extended Service Set (ESS) − It is a set of all connected BSS.

4) Distribution System (DS) − It connects access points in ESS.



Advantages of WLANs

- They provide clutter free homes, offices and other networked places.
- The LANs are scalable in nature, i.e. devices may be added or removed from the network at a greater ease than wired LANs.
- The system is portable within the network coverage and access to the network is not bounded by the length of the cables.

- Installation and setup is much easier than wired counterparts.
- The equipment and setup costs are reduced.

Disadvantages of WLANs

- Since radio waves are used for communications, the signals are noisier with more interference from nearby systems.
- Greater care is needed for encrypting information. Also, they are more prone to errors. So, they require greater bandwidth than the wired LANs.
- WLANs are slower than wired LANs.

IEEE 802.11

IEEE 802.11 is a set of media access control (MAC) and physical layer (PHY) specifications for implementing wireless local area network(WLAN) computer communication in the 900 MHz and 2.4, 3.6, 5, and 60 GHz frequency bands

The IEEE developed an international standard for WLANs. The 802.11 standard focuses on the bottom two layers of the OSI model, the physical layer (PHY) and data link layer (DLL).

The objective of the IEEE 802.11 standard was to define a medium access control (MAC) sublayer, MAC management protocols and services, and three PHYs for wireless connectivity of fixed, portable, and moving devices within a local area.

The three physical layers are an IR base band PHY, an FHSS radio in the 2.4 GHz band, and a DSSS radio in the GHz.

IEEE 802.11 Architecture:

The architecture of the IEEE 802.11 WLAN is designed to support a network where most decision making is distributed to mobile stations. This type of architecture has several advantages. It is tolerant of faults in all of the WLAN equipment and eliminates possible bottlenecks a centralized architecture would introduce. The architecture is flexible and can easily support both small, transient networks and large, semipermanent or permanent networks. In addition, the architecture and protocols offer significant power saving and prolong the battery life of mobile equipment without losing network connectivity

Two network architectures are defined in the IEEE 802.11 standard:

- **Infrastructure network:** An infrastructure network is the network architecture for providing communication between wireless clients and wired network resources. The transition of data from the wireless to wired medium occurs via an AP. An AP and its associated wireless clients define the coverage area. Together all the devices form a basic service set (refer figure 1).
- **Point-to-point (ad-hoc) network:** An ad-hoc network is the architecture that is used to support mutual communication between wireless clients. Typically, an ad-hoc network is created spontaneously and does not support access to wired networks. An ad-hoc network does not require an AP.

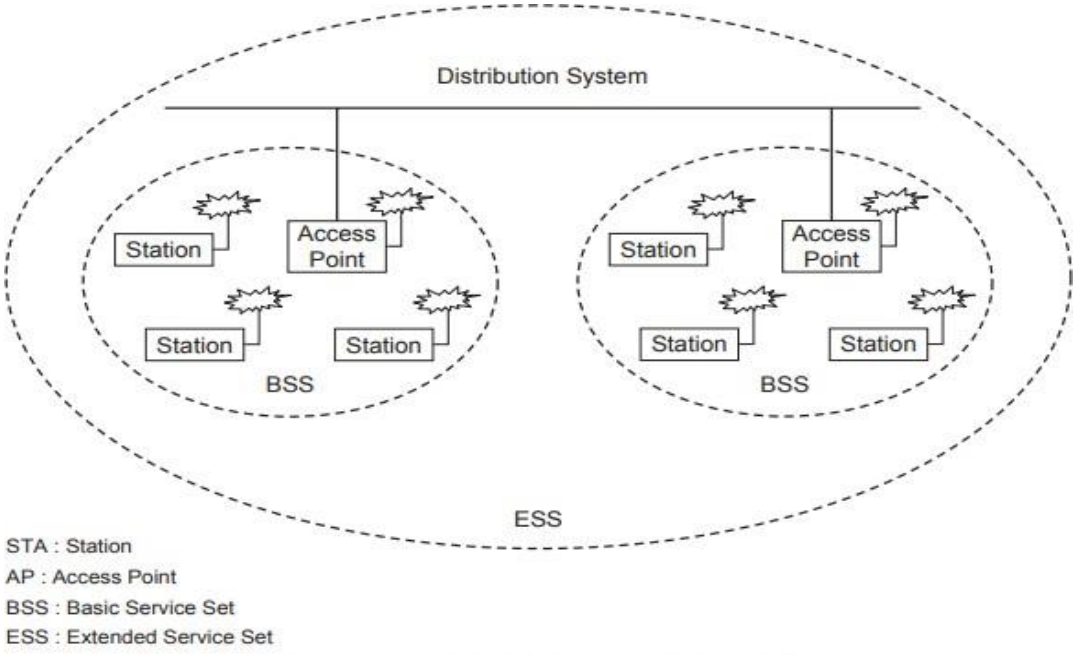


Fig1: BSS and ESS configuration of IEEE 802.11 WLAN

IEEE 802.11 supports three basic topologies for WLANs, the independent basic service set (IBSS), the basic service set, and the extended service set (ESS). The MAC layer supports implementations of IBSS, basic service set, and ESS configurations.

Independent basic service set: The IBSS configuration is referred to as an independent configuration or an ad-hoc network. An IBSS configuration is analogous to a peer-to-peer office network in which no single node is required to act as a server. IBSS WLANs include a number of nodes or wireless stations that communicate directly with one another on an ad-hoc, peer-to-peer basis. Generally, IBSS implementations cover a limited area and are not connected to any large network. An IBSS is typically a short-lived network, with a small number of stations, that is created for a particular purpose.

Basic service set: The basic service set configuration relies on an AP that acts as the logical server for a single WLAN cell or channel. Communications between station 1 and station 4 actually flow from station 1 to AP1 and then from AP1 to AP2 and then from AP2 to AP4 and finally AP4 to station 4 (refer to Figure 2). An AP performs a bridging function and connects multiple WLAN cells or channels, and connects WLAN cells to a wired enterprise LAN.

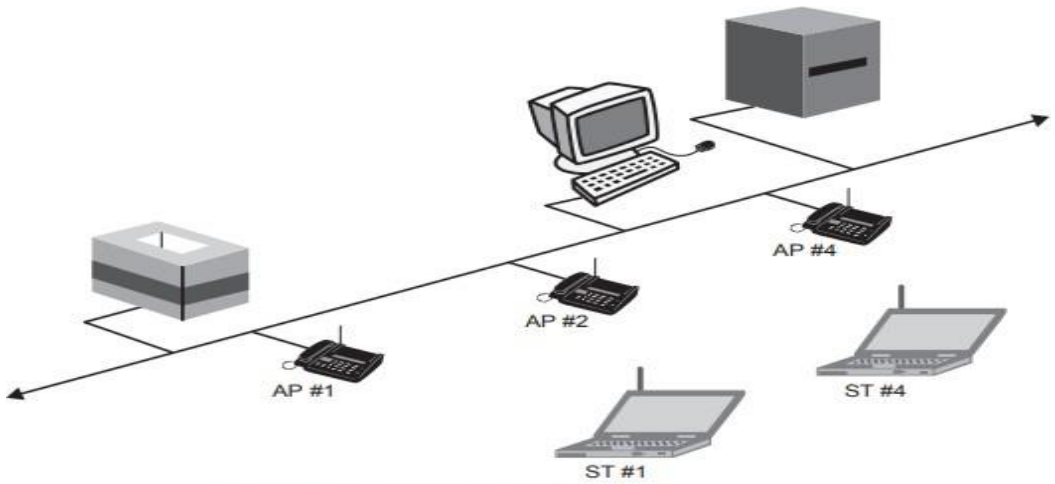


Fig.2 Access point-based topology

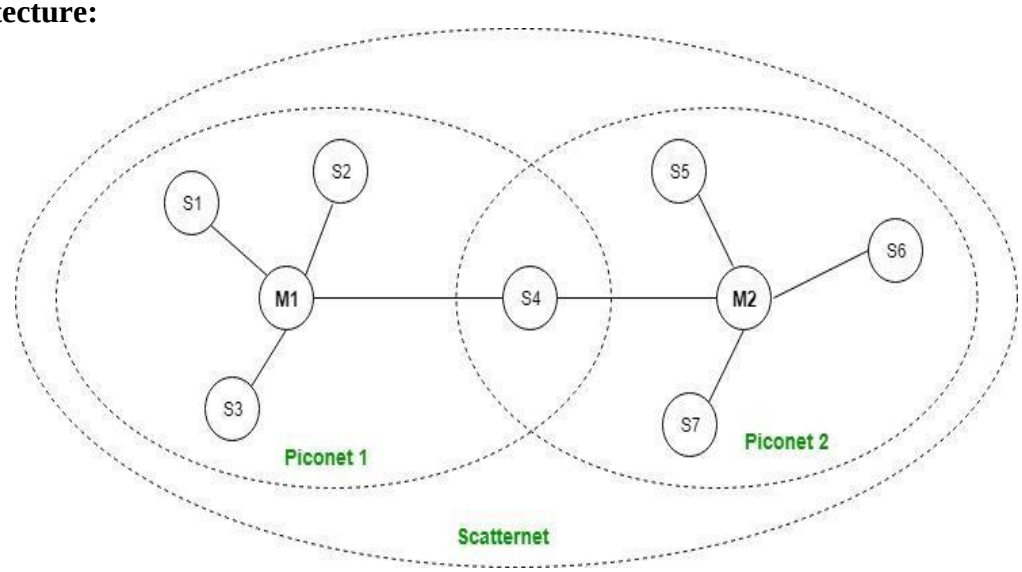
Extended service set: The ESS configuration consists of multiple basic service set cells that can be linked by either wired or wireless backbones called a distributed system. IEEE 802.11 supports ESS configurations in which multiple cells use the same channel, and configurations in which multiple cells use different channels to boost aggregate throughput. To network the equipment outside of the ESS, the ESS and all of its mobile stations appear

to be a single MAC layer network where all stations are physically stationary. Thus, the ESS hides the mobility of the mobile stations from everything outside the ESS (refer figure 1).

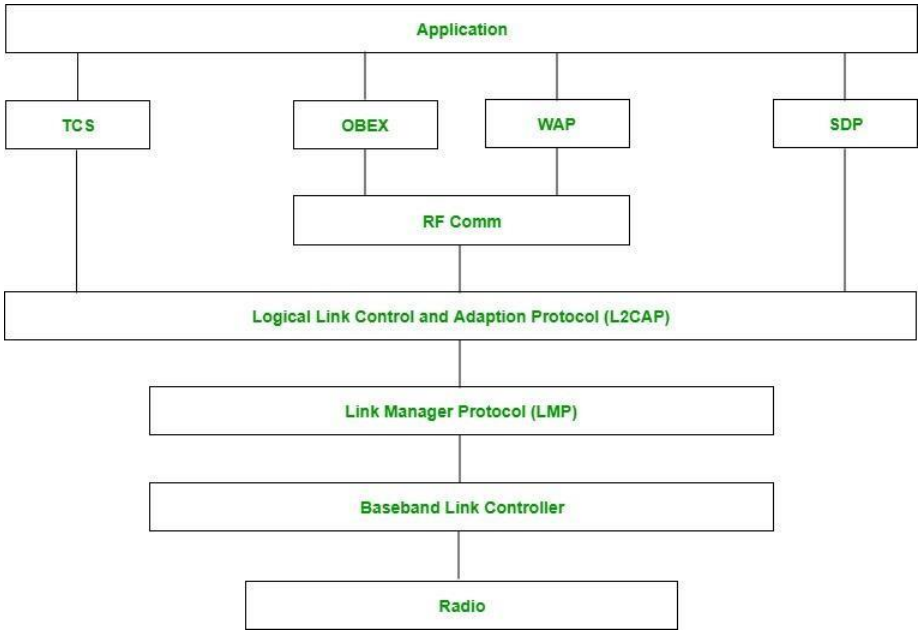
Bluetooth

It is a Wireless Personal Area Network (WPAN) technology and is used for exchanging data over smaller distances. This technology was invented by Ericson in 1994. It operates in the unlicensed, industrial, scientific and medical (ISM) band at 2.4 GHz to 2.485 GHz. Maximum devices that can be connected at the same time are 7. Bluetooth ranges upto 10 meters. It provides data rates upto 1 Mbps or 3 Mbps depending upon the version. The spreading technique which it uses is FHSS (Frequency hopping spread spectrum). A bluetooth network is called **piconet** and a collection of interconnected piconets is call **scatternet**.

Bluetooth Architecture:



Bluetooth protocol stack:



1. **Radio (RF) layer:**
   It performs modulation/demodulation of the data into RF signals. It defines the physical characteristics of bluetooth transceiver. It defines two types of physical link: connection-less and connection-oriented.

2. **Baseband Link layer:**
   It performs the connection establishment within a piconet.
3. **Link Manager protocol layer:**
   It performs the management of the already established links. It also includes authentication and encryption processes.
4. **Logical Link Control and Adaption protocol layer:**
   It is also known as the heart of the bluetooth protocol stack. It allows the communication between upper and lower layers of the bluetooth protocol stack. It packages the data packets received from upper layers into the form expected by lower layers. It also performs the segmentation and multiplexing.
5. **SDP layer:**
   It is short for Service Discovery Protocol. It allows to discover the services available on another bluetooth enabled device.
6. **RF comm layer:**
   It is short for Radio Frontend Component. It provides serial interface with WAP and OBEX.
7. **OBEX:**
   It is short for Object Exchange. It is a communication protocol to exchange objects between 2 devices.
8. **WAP:**
   It is short for Wireless Access Protocol. It is used for internet access.
9. **TCS:**
   It is short for Telephony Control Protocol. It provides telephony service.
10. **Application layer:**
    It enables the user to interact with the application.

Advantages:

- Low cost.
- Easy to use.
- It can also penetrate through walls.
- It creates an adhoc connection immediately without any wires.
- It is used for voice and data transfer.

Disadvantages:

- It can be hacked and hence, less secure.
- It has slow data transfer rate: 3 Mbps.
- It has small range: 10 meters.

Multiple Access Techniques

Multiple access schemes are used to allow many mobile users to share simultaneously a finite amount of radio spectrum.

In wireless communication systems, it is often desirable to allow the subscriber to send information simultaneously from the mobile station to the base station while receiving information from the base station to the mobile station.

A cellular system divides any given area into cells where a mobile unit in each cell communicates with a base station. The main aim in the cellular system design is to be able to **increase the capacity of the channel**, i.e., to handle as many calls as possible in a given bandwidth with a sufficient level of quality of service.

There are several different ways to allow access to the channel.

These includes mainly the following −

- Frequency division multiple-access (FDMA)
- Time division multiple-access (TDMA)
- Code division multiple-access (CDMA)
- Space division multiple access (SDMA)

Depending on how the available bandwidth is allocated to the users, these techniques can be classified as **narrowband** and **wideband** systems.

Narrowband Systems

Systems operating with channels substantially narrower than the coherence bandwidth are called as Narrow band systems. Narrow band TDMA allows users to use the same channel but allocates a unique time slot to each user on the channel, thus separating a small number of users in time on a single channel.

Wideband Systems

In wideband systems, the transmission bandwidth of a single channel is much larger than the coherence bandwidth of the channel. Thus, multipath fading doesn't greatly affect the received signal within a wideband channel, and frequency selective fades occur only in a small fraction of the signal bandwidth.

Frequency Division Multiple Access (FDMA)

FDMA is the basic technology for advanced mobile phone services. The features of FDMA are as follows.

- FDMA allots a different sub-band of frequency to each different user to access the network.
- If FDMA is not in use, the channel is left idle instead of allotting to the other users.
- FDMA is implemented in Narrowband systems and it is less complex than TDMA.
- Tight filtering is done here to reduce adjacent channel interference.
- The base station BS and mobile station MS, transmit and receive simultaneously and continuously in FDMA.

Time Division Multiple Access (TDMA)

In the cases where continuous transmission is not required, there TDMA is used instead of FDMA. The features of TDMA include the following.

- TDMA shares a single carrier frequency with several users where each users makes use of non-overlapping time slots.
- Data transmission in TDMA is not continuous, but occurs in bursts. Hence handsoff process is simpler.
- TDMA uses different time slots for transmission and reception thus duplexers are not required.
- TDMA has an advantage that is possible to allocate different numbers of time slots per frame to different users.
- Bandwidth can be supplied on demand to different users by concatenating or reassigning time slot based on priority.

Code Division Multiple Access (CDMA)

Code division multiple access technique is an example of multiple access where several transmitters use a single channel to send information simultaneously. Its features are as follows.

- In CDMA every user uses the full available spectrum instead of getting allotted by separate frequency.
- CDMA is much recommended for voice and data communications.
- While multiple codes occupy the same channel in CDMA, the users having same code can communicate with each other.
- CDMA offers more air-space capacity than TDMA.
- The hands-off between base stations is very well handled by CDMA.

Space Division Multiple Access (SDMA)

Space division multiple access or spatial division multiple access is a technique which is MIMO (multiple-input multiple-output) architecture and used mostly in wireless and satellite communication. It has the following features.

- All users can communicate at the same time using the same channel.

- SDMA is completely free from interference.
- A single satellite can communicate with more satellites receivers of the same frequency.
- The directional spot-beam antennas are used and hence the base station in SDMA, can track a moving user.
- Controls the radiated energy for each user in space.

Spread Spectrum Multiple Access

Spread spectrum multiple access (SSMA) uses signals which have a transmission bandwidth whose magnitude is greater than the minimum required RF bandwidth.

There are two main types of spread spectrum multiple access techniques −

- Frequency hopped spread spectrum (FHSS)
- Direct sequence spread spectrum (DSSS)

Frequency Hopped Spread Spectrum (FHSS)

This is a digital multiple access system in which the carrier frequencies of the individual users are varied in a pseudo random fashion within a wideband channel. The digital data is broken into uniform sized bursts which is then transmitted on different carrier frequencies.

Direct Sequence Spread Spectrum (DSSS)

This is the most commonly used technology for CDMA. In DS-SS, the message signal is multiplied by a Pseudo Random Noise Code. Each user is given his own code word which is orthogonal to the codes of other users and in order to detect the user, the receiver must know the code word used by the transmitter.

The combinational sequences called as **hybrid** are also used as another type of spread spectrum. **Time hopping** is also another type which is rarely mentioned.

Since many users can share the same spread spectrum bandwidth without interfering with one another, spread spectrum systems become **bandwidth efficient** in a multiple user environment.

The wireless channel is susceptible to a variety of transmission impediments such as **path loss, interference** and **blockage**. These factors restrict the range, data rate, and the reliability of the wireless transmission.

Types of Paths

The extent to which these factors affect the transmission depends upon the environmental conditions and the mobility of the transmitter and receiver. The path followed by the signals to get to the receiver, are two types, such as −

Direct-path

The transmitted signal, when reaches the receiver directly, can be termed as a **directpath** and the components presents that are present in the signal are called as **directpath components**.

Multi-path

The transmitted signal when reaches the receiver, through different directions undergoing different phenomenon, such a path is termed as **multi-path** and the components of the transmitted signal are called as **multi-path components**.

They are reflected, diffracted and scattered by the environment, and arrive at the receiver shifted in amplitude, frequency and phase with respect to the direct path component.

Characteristics of Wireless Channel

The most important characteristics of wireless channel are −

- Path loss
- Fading
- Interference
- Doppler shift

In the following sections, we will discuss these channel characteristics one by one.

Path Loss

Path loss can be expressed as the ratio of the power of the transmitted signal to the power of the same signal received by the receiver, on a given path. It is a function of the propagation distance.

- Estimation of path loss is very important for designing and deploying wireless communication networks
- Path loss is dependent on a number of factors such as the radio frequency used and the nature of the terrain.
- The free space propagation model is the simplest path loss model in which there is a direct-path signal between the transmitter and the receiver, with no atmosphere attenuation or multipath components.

In this model, the relationship between the transmitted power P_t and the received power P_r is given by

$$P_{r} = P_{t}G_{t}G_{r}(\frac{\lambda}{4\Pi d})^2$$

Where

- G_t is the transmitter antenna gain
- G_r is the receiver antenna gain
- **d** is the distance between the transmitter and receiver
- **λ** is the wavelength of the signal

Two-way model also called as two path models is widely used path loss model. The free space model described above assumes that there is only one single path from the transmitter to the receiver.

In reality, the signal reaches the receiver through multiple paths. The two path model tries to capture this phenomenon. The model assumes that the signal reaches the receiver through two paths, one a line-of-sight and the other the path through which the reflected wave is received.

According to the two-path model, the received power is given by

$$P_{r} = P_{t}G_{t}G_{r}(\frac{h_{t}h_{r}}{d^2})^2$$

Where

- p_t is the transmitted power
- G_t represent the antenna gain at the transmitter
- G_r represent the antenna gain at the receiver
- **d** is the distance between the transmitter and receiver
- h_t is the height of the transmitter
- h_r are the height of the receiver

Fading

Fading refers to the fluctuations in signal strength when received at the receiver. Fading can be classified in to two types −

- Fast fading/small scale fading and
- Slow fading/large scale fading

Fast fading refers to the rapid fluctuations in the amplitude, phase or multipath delays of the received signal, due to the interference between multiple versions of the same transmitted signal arriving at the receiver at slightly different times.

The time between the reception of the first version of the signal and the last echoed signal is called **delay spread**. The multipath propagation of the transmitted signal, which causes fast fading, is because of the three propagation mechanisms, namely −

- Reflection
- Diffraction
- Scattering

The multiple signal paths may sometimes add constructively or sometimes destructively at the receiver causing a variation in the power level of the received signal. The received single envelope of a fast fading signal is said to follow a **Rayleigh distribution** to see if there is no line-of-sight path between the transmitter and the receiver.

Slow Fading

The name Slow Fading itself implies that the signal fades away slowly. The features of slow fading are as given below.

- Slow fading occurs when objects that partially absorb the transmission lie between the transmitter and receiver.
- Slow fading is so called because the duration of the fade may last for multiple seconds or minutes.
- Slow fading may occur when the receiver is inside a building and the radio wave must pass through the walls of a building, or when the receiver is temporarily shielded from the transmitter by a building. The obstructing objects cause a random variation in the received signal power.
- Slow fading may cause the received signal power to vary, though the distance between the transmitter and receiver remains the same.
- Slow fading is also referred to as **shadow fading** since the objects that cause the fade, which may be large buildings or other structures, block the direct transmission path from the transmitter to the receiver.

Interference

Wireless transmissions have to counter interference from a wide variety of sources. Two main forms of interference are −

- Adjacent channel interference and
- Co-channel interference.

In Adjacent channel interference case, signals in nearby frequencies have components outside their allocated ranges, and these components may interfere with on-going transmission in the adjacent frequencies. It can be avoided by carefully introducing guard bands between the allocated frequency ranges.

Co-channel interference, sometimes also referred to as **narrow band interference**, is due to other nearby systems using the same transmission frequency.

Inter-symbol interference is another type of interference, where distortion in the received signal is caused by the temporal spreading and the consequent overlapping of individual pulses in the signal.

Adaptive equalization is a commonly used technique for combating inter symbol interference. It involves gathering the dispersed symbol energy into its original time interval. Complex digital processing algorithms are used in the equalization process.

The original TCP/IP protocol was defined as four software layers built upon the hardware. Today, however, TCP/IP is thought of as a five-layer model with the layers named similar to the ones in the OSI model.

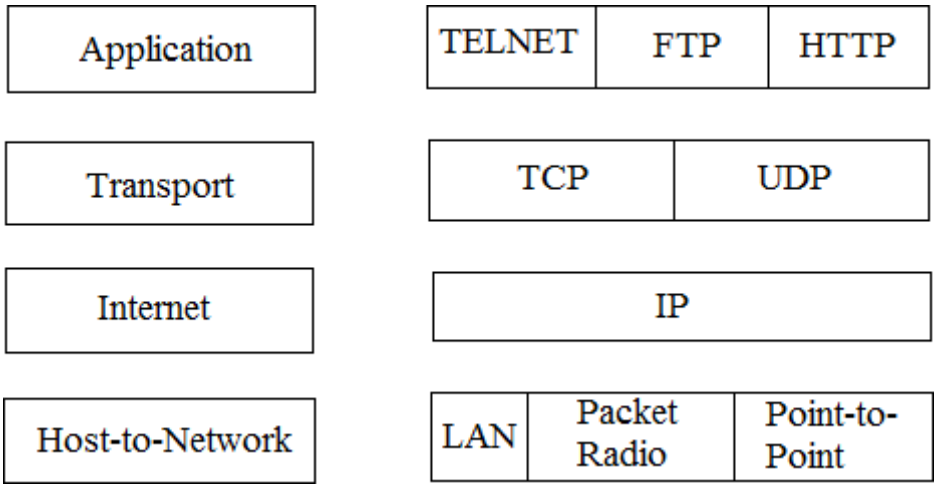
Comparison between OSI and TCP/IP Suite

When we compare the two models, we find that two layers, session and presentation, are missing from the TCP/IP protocol. The application layer in the suite is usually considered to be the combination of three layers in the OSI model.

The OSI model specifies which functions belong to each of its layers but the layers of the TCP/IP protocol suite contain relatively independent protocols that can be mixed and matched, depending on the needs of the system. The term hierarchical means that each upper level protocol is supported by one or more lower level protocols.

Layers in the TCP/IP Suite

The four layers of the TCP/IP model are the host-to-network layer, internet/network layer, transport layer and the application layer. The purpose of each layer in the TCP/IP protocol suite is detailed below.



The above image represents the layers of TCP/IP protocol suite.

Physical Layer

TCP/IP does not define any specific protocol for the physical layer. It supports all of the standard and proprietary protocols.

- At this level, the communication is between two hops or nodes, either a computer or router. The unit of communication is a **single bit**.
- When the connection is established between the two nodes, a stream of bits is flowing between them. The physical layer, however, treats each bit individually.

The responsibility of the physical layer, in addition to delivery of bits, matches with what mentioned for the physical layer of the OSI model, but it mostly depends on the underlying technologies that provide links.

Data Link Layer

TCP/IP does not define any specific protocol for the data link layer either. It supports all of the standard and proprietary protocols.

- At this level also, the communication is between two hops or nodes. The unit of communication however, is a packet called a **frame**.
- A **frame** is a packet that encapsulates the data received from the network layer with an added header and sometimes a trailer.
- The head, among other communication information, includes the source and destination of frame.
- The **destination address** is needed to define the right recipient of the frame because many nodes may have been connected to the link.
- The **source address** is needed for possible response or acknowledgment as may be required by some protocols.

LAN, Packet Radio and Point-to-Point protocols are supported in this layer

Network Layer

At the network layer, TCP/IP supports the Internet Protocol (IP). The Internet Protocol (IP) is the transmission mechanism used by the TCP/IP protocols.

- IP transports data in packets called **datagrams**, each of which is transported separately.
- Datagrams can travel along different routes and can arrive out of sequence or be duplicated.

IP does not keep track of the routes and has no facility for reordering datagrams once they arrive at their destination.

Transport Layer

There is a main difference between the transport layer and the network layer. Although all nodes in a network need to have the network layer, only the two end computers need to have the transport layer.

- The network layer is responsible for sending individual datagrams from computer A to computer B; the transport layer is responsible for delivering the whole message, which is called a **segment**, from A to B.
- A segment may consist of a few or tens of **datagrams**. The segments need to be broken into datagrams and each datagram has to be delivered to the network layer for transmission.
- Since the Internet defines a different route for each datagram, the datagrams may arrive out of order and may be lost.
- The transport layer at computer B needs to wait until all of these datagrams to arrive, assemble them and make a segment out of them.

Traditionally, the transport layer was represented in the TCP/IP suite by two protocols: **User Datagram Protocol (UDP)** and **Transmission Control Protocol (TCP)**.

A new protocol called **Stream Control Transmission Protocol (SCTP)** has been introduced in the last few years.

Application Layer

The application layer in TCP/IP is equivalent to the combined session, presentation, and application layers in the OSI model.

- The application layer allows a user to access the services of our private internet or the global Internet.
- Many protocols are defined at this layer to provide services such as electronic mail file transfer, accessing the World Wide Web, and so on.
- The protocols supported in this layer are **TELNET, FTP** and **HTTP**.
-

Applications of Wireless Communication

Following is a list of applications in wireless communication:

Vehicles

Many wireless communication systems and mobility aware applications are used for following purpose:

- Transmission of music, news, road conditions, weather reports, and other **broadcast information** are received via digital audio broadcasting (DAB) with 1.5Mbit/s.
- For **personal communication,** a universal mobile telecommunications system (UMTS) phone might be available offering voice and data connectivity with 384kbit/s.
- For **remote areas,** satellite communication can be used, while the current position of the car is determined via the GPS (Global Positioning System).
- A local ad-hoc network for the fast **exchange of information** (information such as distance between two vehicles, traffic information, road conditions) in emergency situations or to help each other keep a safe distance. Local ad-hoc network with vehicles close by to prevent guidance system, accidents, redundancy.
- Vehicle data from buses, trucks, trains and high speed train can be transmitted in advance for **maintenance.**
- In ad-hoc network, car can comprise personal digital assistants (PDA), laptops, or mobile phones connected with each other using the Bluetooth technology.

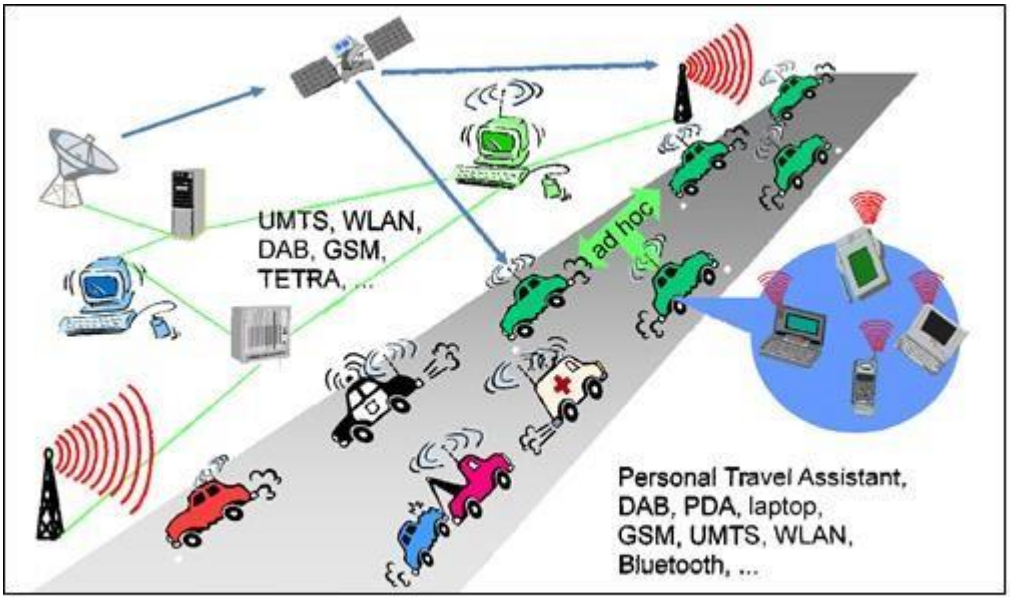


Fig: A Typical Application of Mobile Communication in Road Traffic

Emergency

Following services can be provided during emergencies:

- **Video communication:** Responders often need to share vital information. The transmission of real time situations of video could be necessary. A typical scenario includes the transmission of live video footage from a disaster area to the nearest fire department, to the police station or to the near NGOs etc.
- **Push To Talk (PTT):** PTT is a technology which allows half duplex communication between two users where switching from voice reception mode to the transmit mode takes place with the use of a dedicated momentary button. It is similar to walkie-talkie.
- **Audio/Voice Communication:** This communication service provides full duplex audio channels unlike PTT. Public safety communication requires novel full duplex speech transmission services for emergency response.
- **Real Time Text Messaging (RTT):** Text messaging (RTT) is an effective and quick solution for sending alerts in case of emergencies. Types of text messaging can be email, SMS and instant message.

Business

Travelling Salesman

- Directly access to customer files stored in a central location.
- Consistent databases for all agents
- Mobile office
- To enable the company to keep track of all the activities of their travelling employees.

In Office

- **Wi-Fi** wireless technology saves businesses or companies a considerable amount of money on installations costs.

- There is no need to physically setup wires throughout an office building, warehouse or store.
- **Bluetooth** is also a wireless technology especially used for short range that acts as a complement to Wi-Fi. It is used to transfer data between computers or cellphones.

Transportation Industries

- In transportation industries, GPS technology is used to find efficient routes and tracking vehicles.

Replacement of Wired Network

- Wireless network can also be used to replace wired network. Due to economic reasons it is often impossible to wire remote sensors for weather forecasts, earthquake detection, or to provide environmental information, wireless connections via satellite, can help in this situation.
- Tradeshows need a highly dynamic infrastructure, since cabling takes a long time and frequently proves to be too inflexible.
- Many computers fairs use WLANs as a replacement for cabling.
- Other cases for wireless networks are computers, sensors, or information displays in historical buildings, where excess cabling may destroy valuable walls or floors.

Location dependent service

It is important for an application to know something about the location because the user might need location information for further activities. Several services that might depend on the actual location can be described below:

- **Follow-on Services:**
- **Location aware services:** To know about what services (e.g. fax, printer, server, phone, printer etc.) exist in the local environment.
- **Privacy:** We can set the privacy like who should get knowledge about the location.
- **Information Services:** We can know about the special offers in the supermarket. Nearest hotel, rooms, cabs etc.

Infotainment: (Entertainment and Education)

- Wireless networks can provide information at any appropriate location.
- Outdoor internet access.
- You may choose a seat for movie, pay via electronic cash, and send this information to a service provider.
- Ad-hoc network is used for multiuser games and entertainment.

Mobile and Wireless devices

Even though many mobile and wireless devices are available, there will be many more devices in the future. There is no precise classification of such devices, by sizes, shape, weight, or computing power. The following list of given examples of mobile and wireless devices graded by increasing performance (CPU, memory, display, input devices, etc.)

Sensor: Wireless device is represented by a sensor transmitting state information. 1 example could be a switch, sensing the office door. If the door is closed, the switch transmits this information to the mobile

phone inside the office which will not accept incoming calls without user interaction; the semantics of a closed door is applied to phone calls.

Embedded Controller: Many applications already contain a simple or sometimes more complex controller. Keyboards, mouse, headsets, washing machines, coffee machines, hair dryers and TV sets are just some examples.

Pager: As a very simple receiver, a pager can only display short text messages, has a tiny display, and cannot send any messages.

Personal Digital Assistant: PDAs typically accompany a user and offer simple versions of office software (calendar, notepad, mail). The typically input device is a pen, with built-in character recognition translating handwriting into characters. Web browsers and many other packages are available for these devices.

Pocket computer: The next steps towards full computers are pocket computers offering tiny keyboards, color displays, and simple versions of programs found on desktop computers (text processing, spreadsheets etc.)

Notebook/laptop: Laptops offer more or less the same performance as standard desktop computers; they use the same software - the only technical difference being size, weight, and the ability to run on a battery. If operated mainly via a sensitive display (touch sensitive or electromagnetic), the device are also known as notepads or tablet PCs.

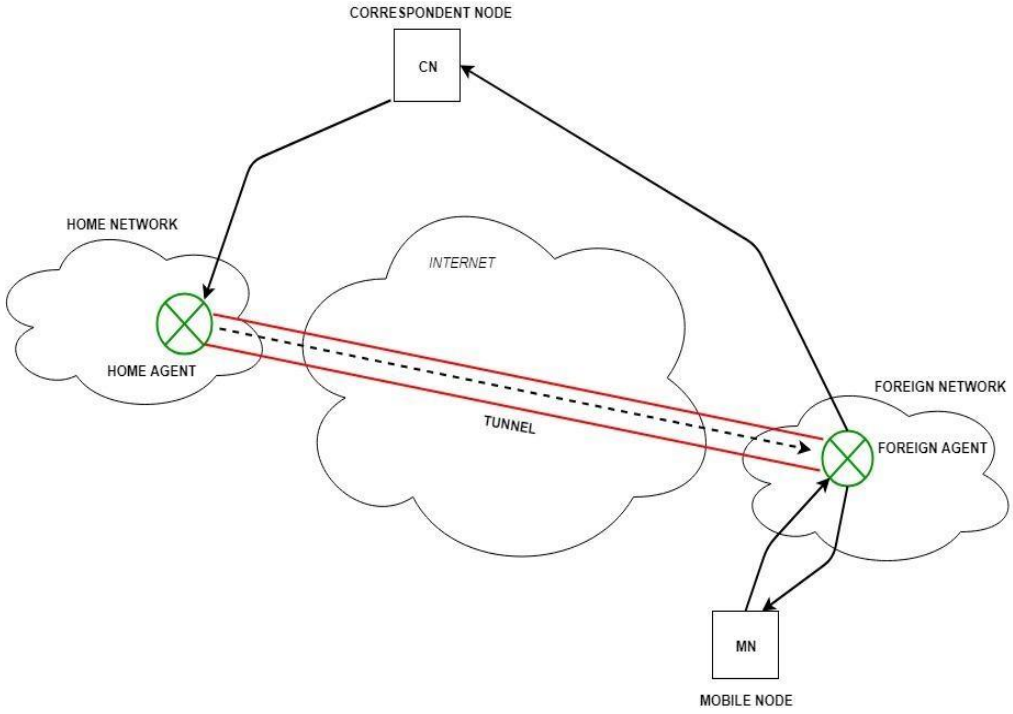
Datacasting (data broadcasting) is the broadcasting of data over a wide area via radio waves. It most often refers to supplemental information sent by television stations along with digital terrestrial television, but may also be applied to digital signals on analog TV or radio. It generally does not apply to data which is inherent to the medium, such as PSIP data which defines virtual channels for DTT or direct broadcast satellite systems; or to things like cable modem or satellite modem, which use a completely separate channel for data.

Mobile Internet Protocol (or Mobile IP)

Mobile IP is a communication protocol (created by extending Internet Protocol, IP) that allows the users to move from one network to another with the same IP address. It ensures that the communication will continue without user's sessions or connections being dropped.

Terminologies:

- **Mobile Node (MN):**
  It is the hand-held communication device that the user caries e.g. Cell phone.
- **Home Network:**
  It is a network to which the mobile node originally belongs to as per its assigned IP address (home address).
- **Home Agent (HA):**
  It is a router in home network to which the mobile node was originally connected
- **Home Address:**
  It is the permanent IP address assigned to the mobile node (within its home network).
- **Foreign Network:**
  It is the current network to which the mobile node is visiting (away from its home network).
- **Foreign Agent (FA):**
  It is a router in foreign network to which mobile node is currently connected. The packets from the home agent are sent to the foreign agent which delivers it to the mobile node.
- **Correspondent Node (CN):**
  It is a device on the internet communicating to the mobile node.
- **Care of Address (COA):**
  It is the temporary address used by a mobile node while it is moving away from its home network.



Working:

Correspondent node sends the data to the mobile node. Data packets contains correspondent node's address (Source) and home address (Destination). Packets reaches to the home agent. But now mobile node is not in the home network, it has moved into the foreign network. Foreign agent sends the care-of-address to the home agent to which all the packets should be sent. Now, a tunnel will be established between the home agent and the foreign agent by the process of tunneling.

Tunneling establishes a virtual pipe for the packets available between a tunnel entry and an endpoint. It is the process of sending a packet via a tunnel and it is achieved by a mechanism called encapsulation.

Now, home agent encapsulates the data packets into new packets in which the source address is the home address and destination is the care-of-address and sends it through the tunnel to the foreign agent. Foreign agent, on other side of the tunnel receives the data packets, decapsulates them and sends them to the mobile node. Mobile node in response to the data packets received, sends a reply in response to foreign agent. Foreign agent directly sends the reply to the correspondent node.

Key Mechanisms in Mobile IP:

1. **Agent Discovery:**
   Agents advertise their presence by periodically broadcasting their agent advertisement messages. The mobile node receiving the agent advertisement messages observes whether the message is from its own home agent and determines whether it is in the home network or foreign network.
2. **Agent Registration:**
   Mobile node after discovering the foreign agent, sends registration request (RREQ) to the foreign agent. Foreign agent in turn, sends the registration request to the home agent with the care-of-address. Home agent sends registration reply (RREP) to the foreign agent. Then it forwards the registration reply to the mobile node and completes the process of registration.
3. **Tunneling:**
   It establishes a virtual pipe for the packets available between a tunnel entry and an endpoint. It is the process of sending a packet via a tunnel and it is achieved by a mechanism called encapsulation. It takes place to forward an IP datagram from the home agent to the care-of-address. Whenever home agent receives a packet from correspondent node, it encapsulates the packet with source address as home address and destination as care-of-address.

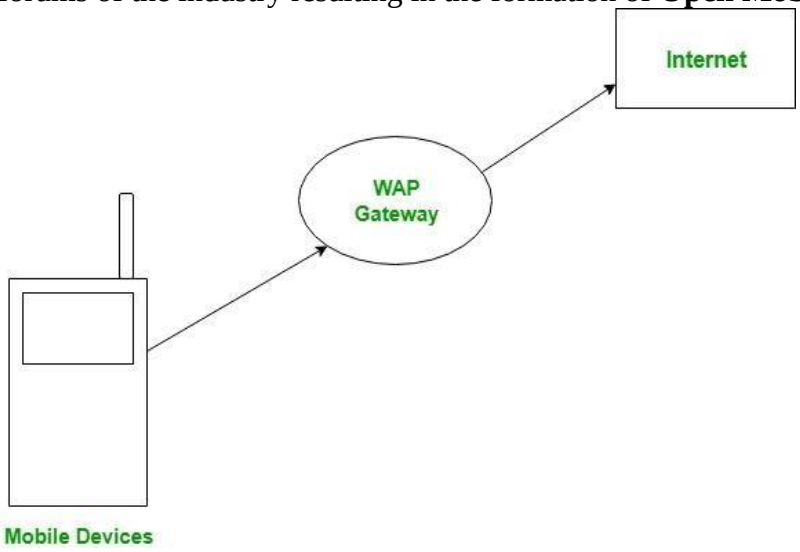
Route Optimization in Mobile IP:

The route optimization adds a conceptual data structure, the binding cache, to the correspondent node. The binding

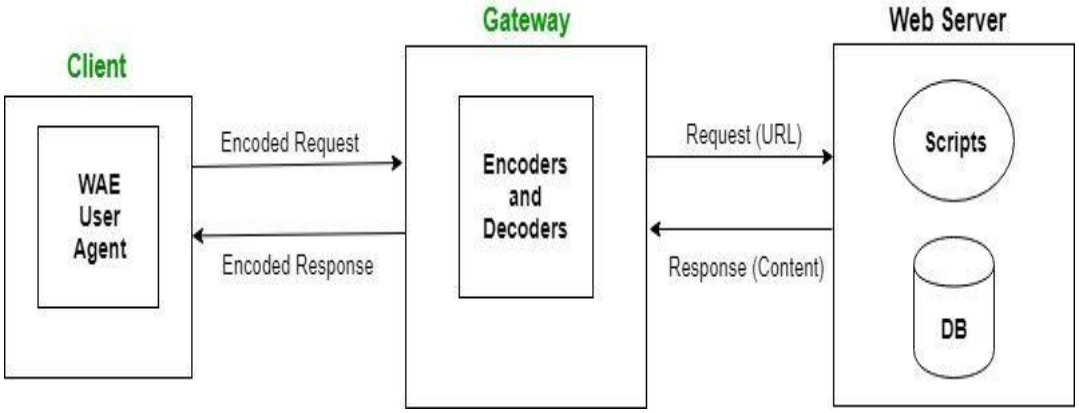
cache contains bindings for mobile node's home address and its current care-of-address. Every time the home agent receives a IP datagram that is destined to a mobile node currently away from the home network, it sends a binding update to the correspondent node to update the information in the correspondent node's binding cache. After this the correspondent node can directly tunnel packets to the mobile node.

Wireless Application Protocol

WAP stands for **Wireless Application Protocol**. It is a protocol designed for micro-browsers and it enables the access of internet in the mobile devices. It uses the mark-up language WML (Wireless Markup Language and not HTML), WML is defined as XML 1.0 application. It enables creating web applications for mobile devices. In 1998, *WAP Forum* was founded by Ericson, Motorola, Nokia and Unwired Planet whose aim was to standardize the various wireless technologies via protocols.
WAP protocol was resulted by the joint efforts of the various members of WAP Forum. In 2002, WAP forum was merged with various other forums of the industry resulting in the formation of **Open Mobile Alliance (OMA)**.

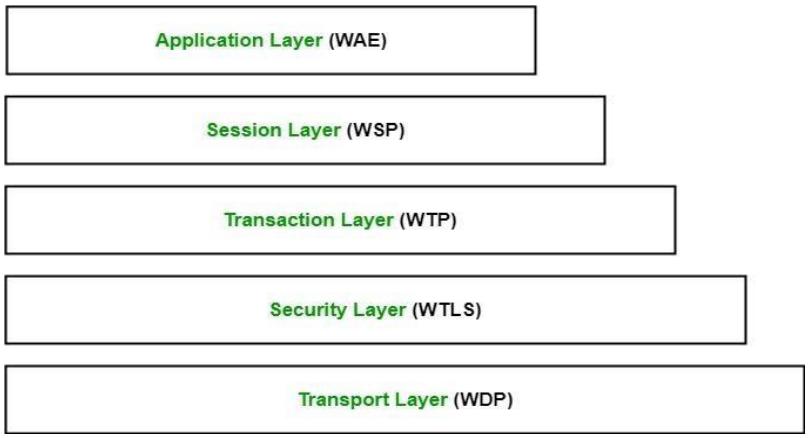


WAP Model:
The user opens the mini-browser in a mobile device. He selects a website that he wants to view. The mobile device sends the URL encoded request via network to a WAP gateway using WAP protocol.



The WAP gateway translates this WAP request into a conventional HTTP URL request and sends it over the internet. The request reaches to a specified Web server and it processes the request just as it would have processed any other request and sends the response back to the mobile device through WAP gateway in WML file which can be seen in the micro-browser.

WAP Protocol stack:



1. **Application Layer:**
   This layer contains the *Wireless Application Environment (WAE)*. It contains mobile device specifications and content development programming languages like WML.
2. **Session Layer:**
   This layer contains *Wireless Session Protocol (WSP)*. It provides fast connection suspension and reconnection.
3. **Transaction Layer:**
   This layer contains *Wireless Transaction Protocol (WTP)*. It runs on top of UDP (User Datagram Protocol) and is a part of TCP/IP and offers transaction support.
4. **Security Layer:**
   This layer contains *Wireless Transaction Layer Security (WTLS)*. It offers data integrity, privacy and authentication.
5. **Transport Layer:**
   This layer contains *Wireless Datagram Protocol*. It presents consistent data format to higher layers of WAP protocol stack.

Traditional TCP

Transmission Control Protocol (TCP) is the transport layer protocol that serves as an interface between client and server. The TCP/IP protocol is used to transfer the data packets between transport layer and network layer. Transport protocol is mainly designed for fixed end systems and fixed, wired networks. In simple terms, the traditional TCP is defined as a wired network while classical TCP uses wireless approach. Mainly TCP is designed for fixed networks and fixed, wired networks.

The main research activities in TCP are as listed below.

1. Congestion control:

During data transmission from sender to receiver, sometimes the data packet may be lost. It is not because of hardware or software problem. Whenever the packet loss is confirmed, the probable reason might be the temporary overload at some point in the transmission path. This temporary overload is otherwise called as Congestion.

Congestion is caused often even when the network is designed perfectly. The transmission speed of receiver may not be equal to the transmission speed of the sender. if the capacity of the sender is more than the capacity of output link, then the packet buffer of a router is filled and the router cannot forward the packets fast enough. The only thing the router can do in this situation is to drop some packets.

The receiver sense the packet loss but does not send message regarding packet loss to the sender. Instead, the receiver starts to send acknowledgement for all the received packets and the sender soon identifies the missing acknowledgement. The sender now notices that a packet is lost and slows down the transmission process. By this, the congestion is reduced. This feature of TCP is one of the reason for its demand even today.

2. Slow start:

The behavior TCP shows after the detection of congestion is called as slow start. The sender always calculates a congestion window for a receiver. At first the sender sends a packet and waits for the acknowledgement. Once the acknowledgement is back it doubles the packet size and sends two packets. After receiving two

acknowledgements, one for each packet, the sender again doubles the packet size and this process continues. This is called Exponential growth.
It is dangerous to double the congestion window each time because the steps might become too large. The exponential growth stops at congestion threshold. As it reaches congestion threshold, the increase in transmission rate becomes linear (i.e., the increase is only by 1). Linear increase continues until the sender notices gap between the acknowledgments. In this case, the sender sets the size of congestion window to half of its congestion threshold and the process continues.

3. Fast re-transmission:
In TCP, two things lead to a reduction of the congestion threshold. One of those is sender receiving continuous acknowledgements for the single packet. By this it can convey either of two things. One such thing is that the receiver received all the packets up to the acknowledged one and the other thing is the gap is due to packet loss. Now the sender immediately re-transmit the missing packet before the given time expires. This is called as Fast re-transmission.

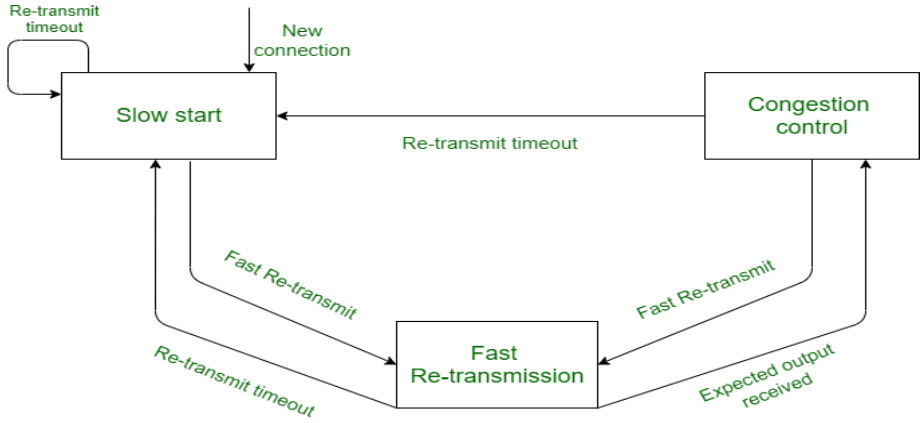


Figure: Traditional TCP

Example:
Assume that few packets of data are being transferred from sender to receiver, and the speed of sender is 2 Mbps and the speed of receiver is 1 Mbps respectively. Now the packets that are being transferred from sender sender to receiver makes a traffic jam inside the network. Due to this the network may drop some of the packets. When these packets are lost, the receiver sends the acknowledgement to the sender and the sender identifies the missing acknowledgement. This process is called as congestion control.
Now the slowstart mechanism takes up the plan. The sender slows down the packet transfer and then the traffic is slightly reduces. After sometime it puts a request to fast re-transmission through which the missing packets can be sent again as fast as possible. After all these mechanisms, the process of next packet begins.

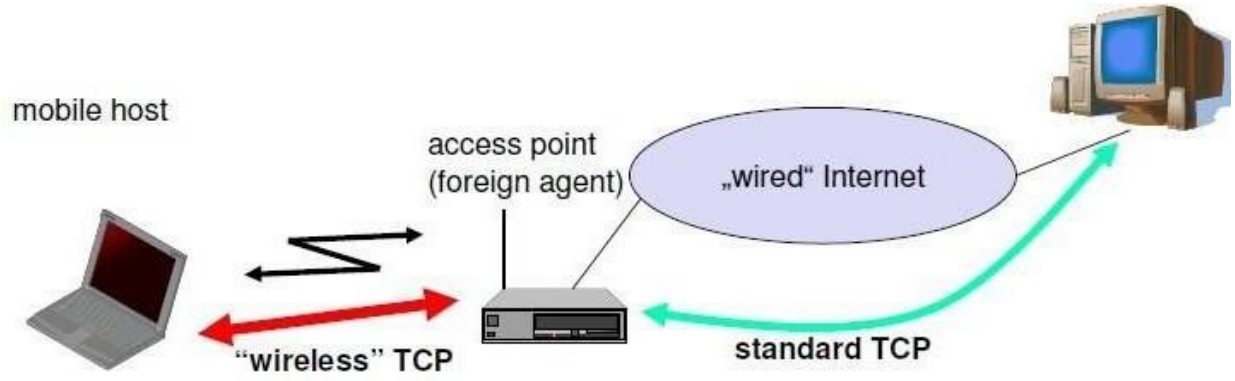
Problems with Traditional TCP in wireless environments

- Slow Start mechanism in fixed networks decreases the efficiency of TCP if used with mobile receivers orsenders.
- Error rates on wireless links are orders of magnitude higher compared to fixed fiber or copper links. Thismakescompensationforpacketlossby TCPquitedifficult.
- Mobility itself can cause packet loss. There are many situations where a soft handover from one access point to another is not possible for a mobileend-system.
- Standard TCP reacts with slow start if acknowledgements are missing, which does not help in the case of transmission errors over wireless links and which does not really help during handover. This behavior results in a severe performance degradation of an unchanged TCP if used together with wireless links or mobilenodes

Classical TCP Improvements

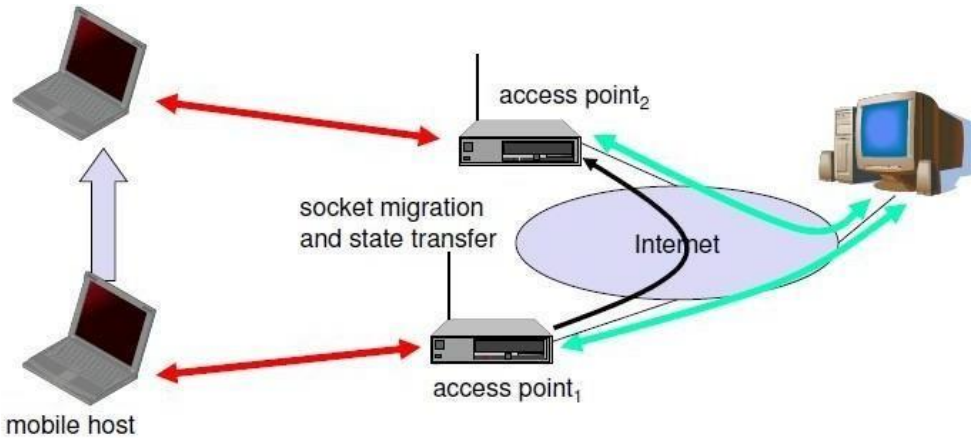
Indirect TCP (I-TCP)

Indirect TCP segments a TCP connection into a fixed part and a wireless part. The following figure shows an example with a mobile host connected via a wireless link and an access point to the 'wired' internet where the correspondent hostresides.



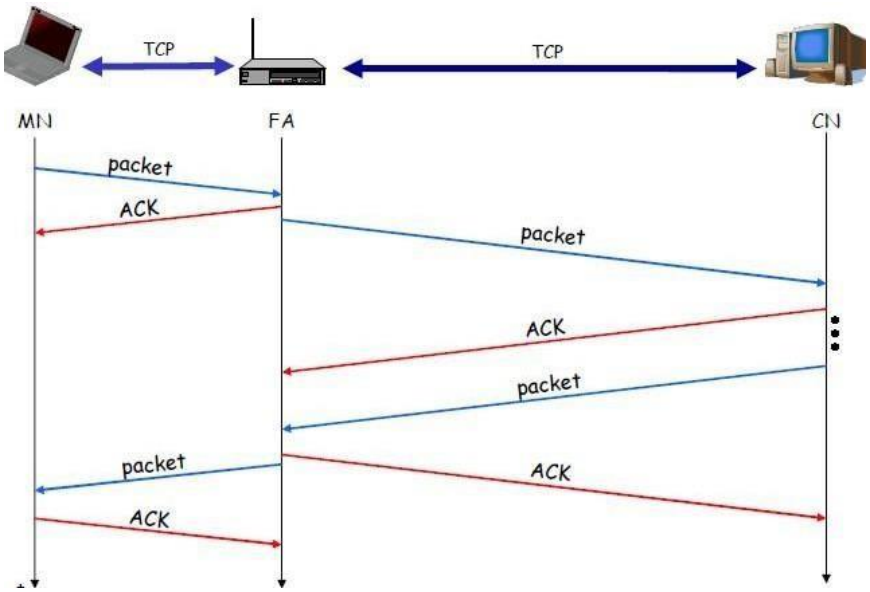
Standard TCP is used between the fixed computer and the access point. No computer in the internet recognizes any changes to TCP. Instead of the mobile host, the access point now terminates the standard TCP connection, acting as a proxy. This means that the access point is now seen as the mobile host for the fixed host and as the fixed host for the mobile host. Between the access point and the mobile host, a special TCP, adapted to wireless links, is used. However, changing TCP for the wireless link is not a requirement. A suitable place for segmenting the connection is at the foreign agent as it not only controls the mobility of the mobile host anyway and can also handover the connection to the next foreign agent when the mobile host moves on.

The foreign agent acts as a proxy and relays all data in both directions. If CH (correspondent host) sends a packet to the MH, the FA acknowledges it and forwards it to the MH. MH acknowledges on successful reception, but this is only used by the FA. If a packet is lost on the wireless link, CH doesn't observe it and FA tries to retransmit it locally to maintain reliable data transport. Ifthe MHsends apacket, the FA acknowledges itand forwards it to CH. If the packet is lost on the wireless link, the mobile hosts notice this much faster due to the lower round trip time and can directly retransmit the packet. Packet loss in the wired network is now handled by the foreign agent.



Socket and state migration after handover of a mobile host

During handover, the buffered packets, as well as the system state (packet sequence number, acknowledgements, ports, etc), must migrate to the new agent. No new connection may be established for the mobile host, and the correspondent host must not see any changes in connection state. Packet delivery in I-TCP is shown below:



~~Advantages of I-TCP~~

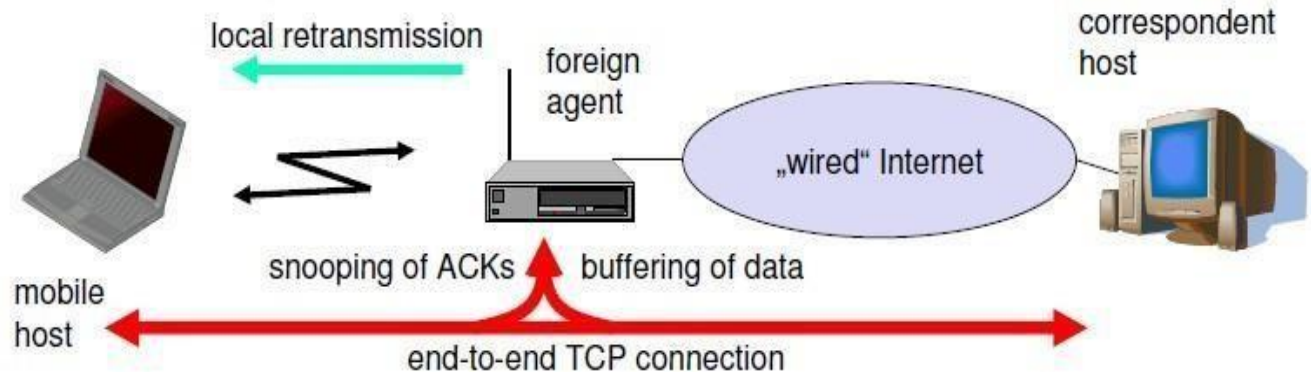
- Nochanges inthe fixed networknecessary, nochanges for thehosts (TCP protocol) necessary, all current optimizations to TCP stillwork
- Simpletocontrol,mobile TCP isusedonlyforonehop between,e.g.,aforeignagentand mobile host
  1. transmission errors onthe wireless link do not propagate into the fixednetwork
  2. therefore, avery fast retransmission ofpackets is possible, the short delay on the mobile hop sknown
- It isalwaysdangerous tointroducenewmechanismsinahugenetworkwithout knowing exactly how theybehave.
  - Newoptimizationscanbetestedatthelasthop,withoutjeopardizingthestabilityofthe Internet.
- It is easy to use different protocols for wired and wireless networks.

~~Disadvantages of I-TCP~~

- Loss ofend-to-end semantics:- an acknowledgement toa sender nolongermeansthata receiverreally has received a packet, foreign agents mightcrash.
- Higherlatencypossible:-duetobufferingofdatawithintheforeignagentandforwardingto a new foreignagent
- Security issue:- The foreign agent must be a trusted entity

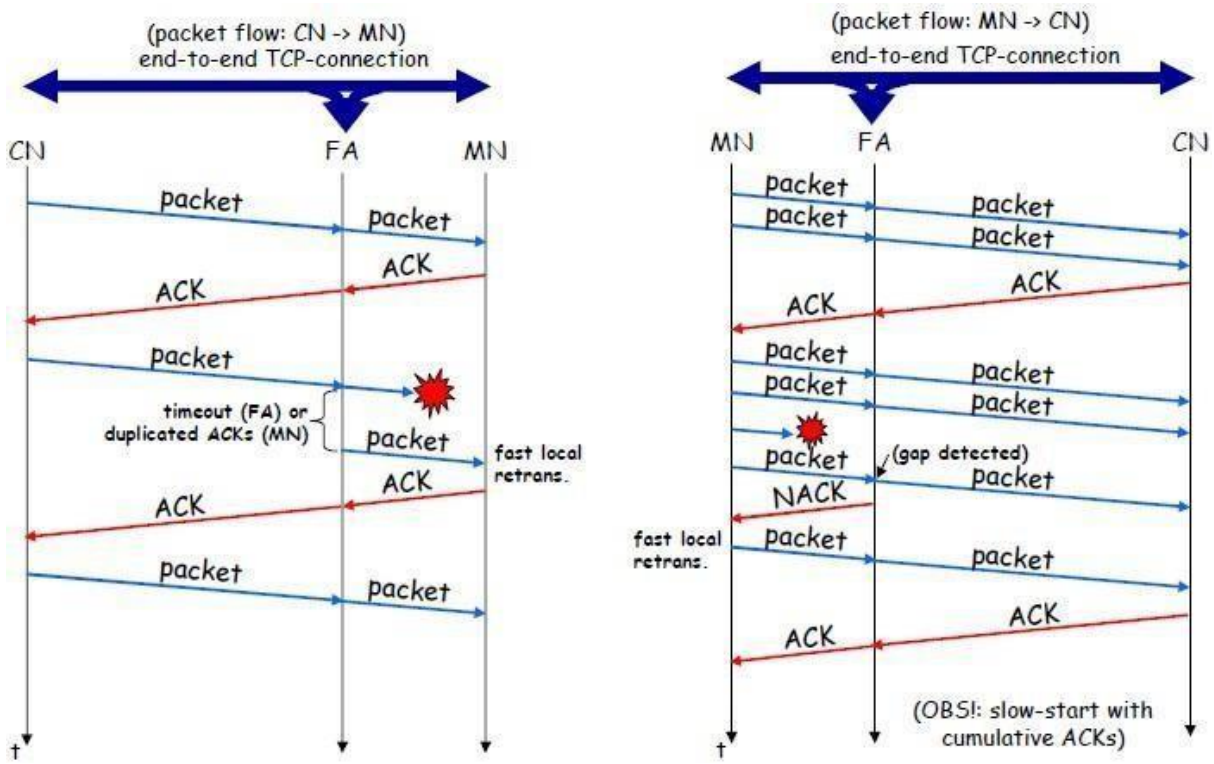
Snooping TCP

The main drawback of I-TCP is the segmentation of the single TCP connection into two TCP connections, which loses the original end-to-end TCP semantic. A new enhancement, which leaves the TCP connection intact and is completely transparent, is Snooping TCP. The main function is tobuffer data close to the mobile host to perform fast local retransmission in caseof packet loss.



Snooping TCP as a transparent TCP extension

Here, the foreign agent buffers all packets with **destination mobile host** and additionally 'snoops' the packet flow in both directions to recognize acknowledgements. The foreign agentbuffers every packet until it receives anacknowledgement from themobile host. If the FA does not receive an acknowledgement from the mobile host within a certain amount of time, either the packet or the acknowledgement has been lost. Alternatively, the foreign



agent could receive a duplicate ACK which also shows the loss of a packet. Now, the FA retransmits the packet directly from the buffer thus performing a faster retransmission compared to the CH. For transparency, the FA does not acknowledge data to the CH, which would violate end-to-end semantic in case of a FA failure. The foreign agent can filter the duplicate acknowledgements to avoid unnecessary retransmissions of data from the correspondent host. If the foreign agent now crashes, the time-out of the correspondent host still works and triggers a retransmission. The foreign agent may discard duplicates of packets already retransmitted locally and acknowledged by the mobile host. This avoids unnecessary traffic on the wireless link.For data transfer from the mobile host with **destination correspondent host**, the FA snoops into the packet stream to detect gaps in the sequence numbers of TCP. As soon as the foreign agent detects a missing packet, it returns a negative acknowledgement (NACK) to the mobile host. The mobile host can now retransmit the missing packet immediately. Reordering of packets is done automatically at the correspondent host byTCP.

Snooping TCP: Packet delivery

~~*Advantages of snooping TCP:*~~

- The end-to-end TCP semantic is preserved.
- Most oftheenhancements are done in theforeign agent itself whichkeeps correspondent host unchanged.
- Handover of state is not required as soon as the mobile host moves to another foreign agent. Even though packets are present in the buffer, time out at the CH occurs and the packets are transmitted to the new COA.
- No problem arises if the new foreign agent uses the enhancement or not. If not, the approach automatically falls back to the standard solution.

Disadvantages of snooping TCP

- Snooping TCP does not isolate the behavior of the wireless link as well as I-TCP. Transmission errors may propagate till CH.
- Using negative acknowledgements between the foreign agent and the mobile host assumes additional mechanisms on the mobile host. This approach is no longer transparent for arbitrary mobile hosts.
- Snooping and buffering data may be useless if certain encryption schemes are applied end- to-end between the correspondent host and mobile host. If encryption is used above the transport layer, (eg. SSL/TLS), snooping TCP can be used.

Mobile TCP

Both I-TCP and Snooping TCP does not help much, if a mobile host gets disconnected. The **M-TCP (mobile TCP)** approach has the same goals as I-TCP and snooping TCP: to prevent the sender window from shrinking if bit errors or disconnection but not congestion cause current problems. M-TCP wants to improve overall throughput, to lower the delay, to maintain end-to-end semantics of TCP, and to provide a more efficient handover. Additionally, M-TCP is especially adapted to the problems arising from lengthy or frequent disconnections. M-TCP splits the TCP connection into two parts as I-TCP does. An unmodified TCP is used on the standard host-**supervisory host (SH)** connection, while an optimized TCP is used on the SH-MH connection.

The SH monitors all packets sent to the MH and ACKs returned from the MH. If the SH doesnot receive an ACK forsome time, it assumes thatthe MH is disconnected. It then chokes the sender by setting the sender's window size to 0. Setting the window size to 0 forces the sender to go into **persistent mode**, i.e., the state of the sender will not change no matter how long the receiver is disconnected. This meansthatthe sender will not try to retransmit data. As soon as the SH (either the old SH or a new SH) detects connectivity again, it reopens the window of the sender to the old value. The sender can continue sending at full speed. This mechanism does not require changes to the sender's TCP. The wireless side uses an adapted

TCP that can recover from packet loss much faster. This modified TCP does not use slow start, thus, M-TCP needsa **bandwidth manager** to implement fair sharing over the wireless link.

Advantages of M-TCP:

- ItmaintainstheTCPend-to-endsemantics. TheSHdoesnotsendanyACKitself but forwards the ACKs from the MH.
- If the MHisdisconnected, it avoids useless retransmissions, slow starts or breaking connections by simply shrinking the sender's window to 0.
- Asnobufferingisdone as in I-TCP, thereis no need to forward bufferstoanew SH. Lost packets will be automatically retransmitted to the SH.

Disadvantages of M-TCP:

- As the SH does not act as proxy as in I-TCP, packet loss on the wireless link due to bit errors is propagated to the sender. M-TCP assumes low bit error rates, which is not always a valid assumption.
- A modified TCP on the wireless link not only requires modifications to the MH protocol software but also new network elements like the bandwidth manager.

Transmission/time-out freezing

Often, MAC layer notices connection problems even before the connection is actually interrupted from a TCP point of view and also knows the real reason for the interruption. The MAC layer can inform the TCP layer of an upcoming loss of connection or that the current interruption is not caused by congestion. TCP can now stop sending and 'freezes' the current state of its congestion window and further timers. If the MAC layer notices the upcoming interruption early enough, both the mobile and correspondent host can be informed. With a fast interruption of the wireless link, additional mechanisms in the access point are needed to inform the correspondent host of the reason for interruption. Otherwise, the correspondent host goes into slow start assuming congestion and finally breaks the connection.

As soon asthe MAC layer detects connectivity again, it signals TCP that it can resume operationat exactly thesame point where it hadbeen forcedto stop. For TCP time simply does not advance, so no timers expire.

Advantages:

- It offers away to resume TCPconnections evenafterlonginterruptionsoftheconnection.
- ItcanbeusedtogetherwithencrypteddataasitisindependentofotherTCPmechanisms such as sequence no or acknowledgements

Disadvantages:

- Lots of changes have to be made in software of MH, CH andFA.

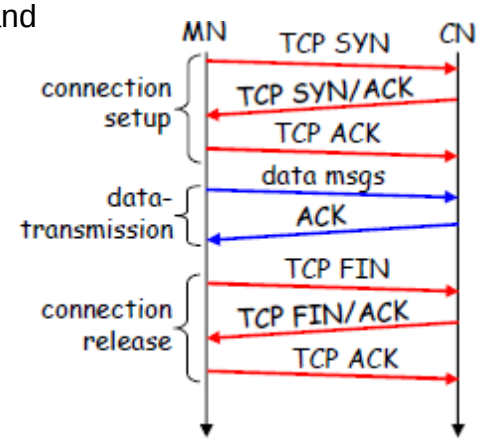
Selective retransmission

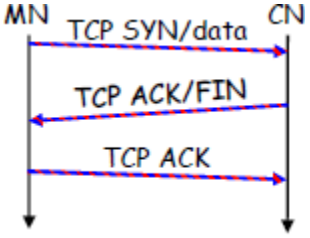
A very useful extension of TCP is the use of selective retransmission. TCP acknowledgements are cumulative, i.e., they acknowledge in-order receipt of packets up to a certain packet. A single acknowledgement confirms reception of all packets upto a certain packet. If a single packet is lost, thesenderhasto retransmit everything starting from the lost packet (go-back-n retransmission). This obviously wastes bandwidth, not just in the case of a mobile network, but for anynetwork.

Using selective retransmission, TCP can indirectly request a selective retransmission of packets. The receiver can acknowledge single packets, not only trains of in-sequence packets. The sender can now determine precisely which packet is needed and can retransmit it. The **advantage** of this approach is obvious: a sender retransmits only the lost packets. This lowers bandwidth requirements and is extremely helpful in slow wireless links. The disadvantage is that a more complex software on the receiver side is needed. Also more buffer space is needed to resequence data and to wait for gaps to be filled.

Transaction-oriented TCP

Assume an application running on the mobile host that sends a short request toa server from time to time, which responds with a short message and it requires reliable TCP transport of the packets. For it touse normal TCP, it is inefficient because of the overhead involved. Standard TCP is made up of three phases: setup, data transfer and release. First, TCP uses a three-way handshake to establish the connection. At least oneadditional packet is usually needed for transmission of the request, and requires three more packets to close the connection via a three-way handshake. So, for sending one data packet, TCP may need seven packets altogether. This kind of overhead is acceptable for long sessions in fixed networks, but is quite inefficient for short messages or sessions in wireless networks. This led to the development of transaction-oriented TCP(T/TCP).





T/TCP can combine packets for connection establishment and connection release with user data packets. This can reduce the number of packets down to two instead of seven. The obvious **advantage** for certain applications is the reduction in the overhead which standard TCP has for connection setup and connection release. Disadvantage is that it requires changes in the software in mobile host

and all correspondent hosts. This solution does not hide mobility anymore. Also, T/TCP exhibits several securityproblems.

Classical Enhancements to TCP for mobility: A comparison

Approach	Mechanism	Advantages	Disadvantages
Indirect TCP	splits TCP connection into two connections	isolation of wireless link, simple	loss of TCP semantics, higher latency at handover
Snooping TCP	"snoops" data and acknowledgements, local retransmission	transparent for end-to-end connection, MAC integration possible	problematic with encryption, bad isolation of wireless link
M-TCP	splits TCP connection, chokes sender via window size	Maintains end-to-end semantics, handles long term and frequent disconnections	Bad isolation of wireless link, processing overhead due to bandwidth management
Fast retransmit/ fast recovery	avoids slow-start after roaming	simple and efficient	mixed layers, not transparent
Transmission/ time-out freezing	freezes TCP state at disconnect, resumes after reconnection	independent of content or encryption, works for longer interrupts	changes in TCP required, MAC dependant
Selective retransmission	retransmit only lost data	very efficient	slightly more complex receiver software, more buffer needed
Transaction oriented TCP	combine connection setup/release and data transmission	Efficient for certain applications	changes in TCP required, not transparent

UNIT 5

Ad hoc Networks

5. AD HOC NETWORKS

4.1 Introduction

- An ad hoc network is a temporary type of Local Area Network (LAN). If you set up an ad hoc network permanently, it becomes a LAN. Multiple devices can use an ad hoc network at the same time, but this might cause a lull in performance.

- A typical example of an ad-hoc network is connecting two or more laptops (or other supported devices) to each other directly without any central access point, either wirelessly or using a cable. When to use an ad-hoc network: If you want to quickly set up a peer-to-peer (P2P) network between two devices.

- Ad hoc networks are created between two or more wireless PCs together, without the use of a wireless router or an access point. The computers communicate directly with each other. Ad hoc networks can be very helpful during meetings or in any location where a network doesn't exist and where people need to share files.

- Ad Hoc Network Characteristics: These networks are characterized by need for low power consumption and low levels of physical security and broadcast physical medium. Asymmetric techniques like RSA are not to be used as are inefficient and consume too much power.

5.1.1 Difference between Ad hoc and Cellular Network

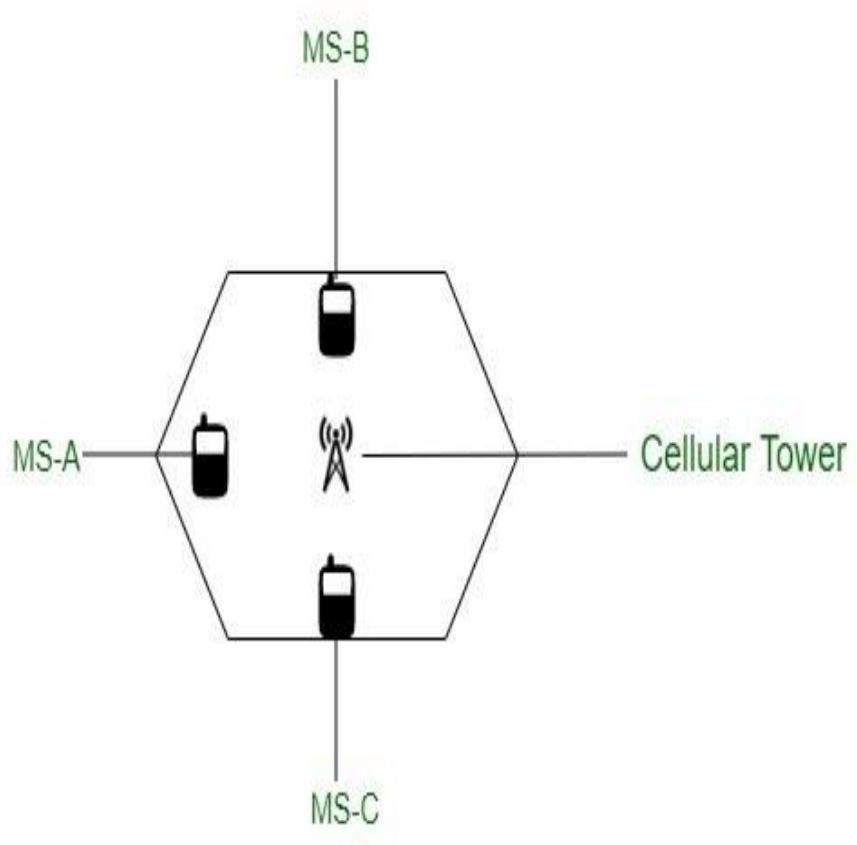
- The network routing of cellular network are centralized, all the traffic goes through the base station. The network routing of adhoc network are distributed, no centralized system such as base station needed. Database servers.

Difference between Ad hoc and Cellular Network

1. Ad hoc Network **:**

An ad hoc network is a network that is composed of individual devices communicating with each other directly. The term implies spontaneous or impromptu construction because these networks often bypass the gatekeeping hardware or central access point such as a router. Many ad hoc networks are local area networks where computers or other devices are enabled to send data directly to one another rather than going through a centralized access point.

2. Cellular Network **:**
A Cellular network or Mobile network is a radio network distributed over land areas called cells, each served by at least one fixed-location transceiver, known as a cell site or base station. At present sent cell remote systems, for example, GSM/CDMA/HSPA/LTE are foundation type. Cell organize comprises of focal element known as base station and cell phones as Mobile Subscribers(MS). In the event that MS-A needs to speak with MS-B, correspondence happens by means of base station(BTS) as appeared in the figure. It follows hexagonal pattern.



Let's see that the difference between Cellular Network and Ad hoc Network :

S.NO	Cellular Network	Ad hoc Network
1.	The network routing of cellular network are centralized, all the traffic goes through the base station.	The network routing of adhoc network are distributed, no centralized system such as base station needed.
2.	Circuit Switching are used.	Packet Switching are used.
3.	It has single hop type.	It has multiple hopes
4.	Star topology are used.	Mesh topology are used.
5.		

5.1. DATA REPLICATION FOR MOBILECOMPUTERS

5.1.1 Introduction

- The word replicate means "to produce a copy of itself" and originates from the Latinword'replicates'.
- Datareplication, initsbroadestsense, simplyreferstocopyingdatafromoneormoredatastoragelocations to one or more other data storage locations.
- These locations are virtual locations and not physical locations–it is not required for the virtual locations to be at different physical locations.
- Datareplicationisatechniquethatwasinitiallyusedintraditionaldistributedenvironmentstoincreasedata availability and improve system performance. These environments are characterized by a fixed infrastructure where the user uses fixed machines that have sufficient resources and are permanently connected to the network. These characteristics are not verified in mobile environments. In this case, user devices such as PDAs and mobile phones have limited capacity in terms of memory space, disk space and processor capacity. These limitations may prevent the replication system from creating and placing the replicas on the userdevice.
- Inthemobileenvironment,theusermayalsochangehisdeviceto accessaservice.Thediversityofdevices andconsequentlythecontextofuseofaserviceoranapplication mustbetakenintoaccount.Indeed,there is a change in device capacities such as screen size, storage capacity and battery power, and its network parameterssuchasthetypeandbandwidth.Thusthemodeofapplicationmustbeadaptedtoeachcontext.
- In short, mobile environments are characterized by a frequent change in their resources which comes from various sources such as the nature of the wireless network itself, the mobility of users and multi-terminal accesses.
- This change may influence data replication because the creation of and access to these data may need a set ofresources.Forexample,confidentialdatalikeacreditcardnumbermaynotbereplicatedandexchanged across non-secure nodes and links. Thus, variation in the level of security may prevent the user from accessing this data. So, a traditional system is not able to satisfy the client's request. Consequently, to ensure service continuity, the replication system functionalities like creation, placement, read, write and consistencyoperationsmustbeadaptedtoallvariationsinresourcesthatdatamayneed.

Data Replication

Data Replication in mobile computing means the sharing of information to ensure data consistency between software and hardware resources connected via the internet, to improve reliability, availability, fault-tolerance, and accessibility of data. In simpler terms, data replication is the process of storing different copies of the database at two or more sites in order to improve data availability in less time and at a cheaper cost.
Data replication in mobile computing is a popular fault tolerance technique for distributed databases.

Advantages of Data Replication

In modern mobile computing, scenario data replication has been adopted as an efficient way to ensure data availability, integrity, and an effective means to achieve fault tolerance. Data replication not only ensures the availability of the data but also minimizes the communication cost, increase data sharing, and enhance the security of sensitive data. Data replication in mobile computing also determines when and which location to store the replica of data, controlling different data replicas over a network for efficient utilization of the network resources.

Data Replication Benefits

Important benefits of data replication are as below-

- **Reliability** − Data replication provides the reliability of data. In case of failure of any site, the database system continues to work since a copy is available at another site(s).
- **Reduction in Network Load** − Since local copies of data are available through data replication. Therefore, query processing can be done with reduced network usage, particularly during prime hours.
- **Data updating can be done at non-prime hours** – Due to data replication data can be updated easily.
- **Quicker Response** − Availability of local copies of data ensures quick query processing and consequently quick response time.
- **Simpler Transactions** − Transactions require less number of joins of tables located at different sites and minimal coordination across the network. Thus, they become simpler in nature.

Disadvantages of Data Replication

- **Increased Storage Requirements** − Maintaining multiple copies of data is associated with increased storage costs. The storage space required is in multiples of the storage required for a centralized system.
- **Increased Cost and Complexity of Data Updating** − Each time a data item is updated, the update needs to be reflected in all the copies of the data at the different sites.
- **Undesirable Application – Database coupling** − If complex update mechanisms are not used, removing data inconsistency requires complex co-ordination at the application level.

Data Replication in Mobile Computing

Data replication is the process of making copies of data stored on various sites in order to improve reliability, efficiency, robustness, simpler transaction, fault tolerance, and reduce network load.

Goals of data replication

Data replication is performed with the purpose of

- Increasing the availability of data.
- Speeding up the query evaluation.

Types of data replication

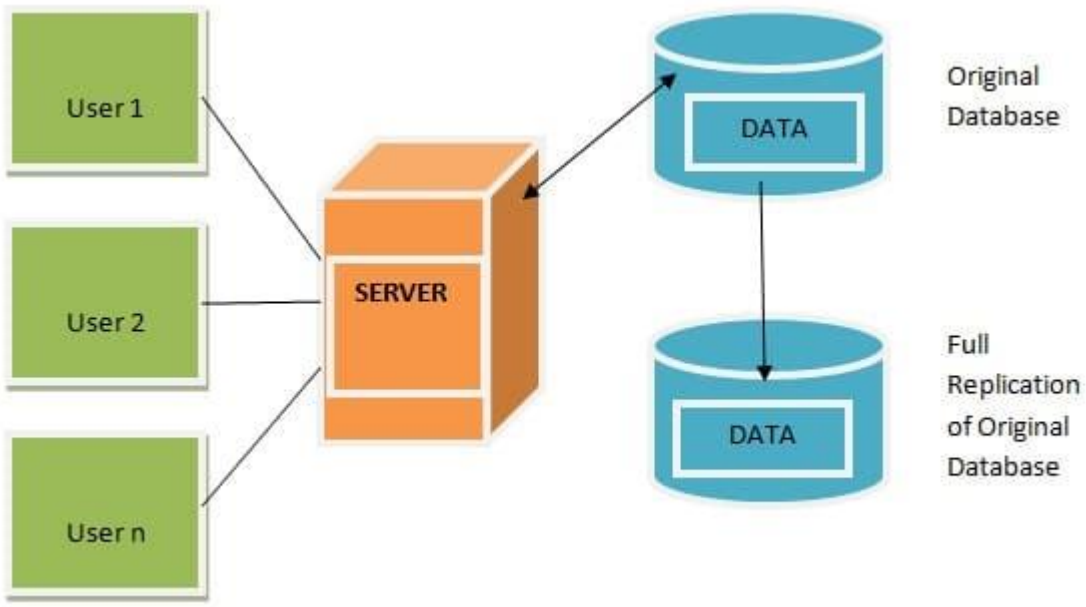
There are two types of data replication

1. **Synchronous Replication**
   In synchronous replication, the replica of the database is modified immediately after changes are made in the relation table. So there is no difference between the original data and the replicated data table.
2. **Asynchronous replication**
   In asynchronous replication, the replica will be modified after commit action is fired on to the database.

Replication Schemes

The three replication schemes are as follows:

1. **Full Replication scheme**
   In full replication scheme, the database is available at all the locations to ease the user in the communication network



Advantages of full replication

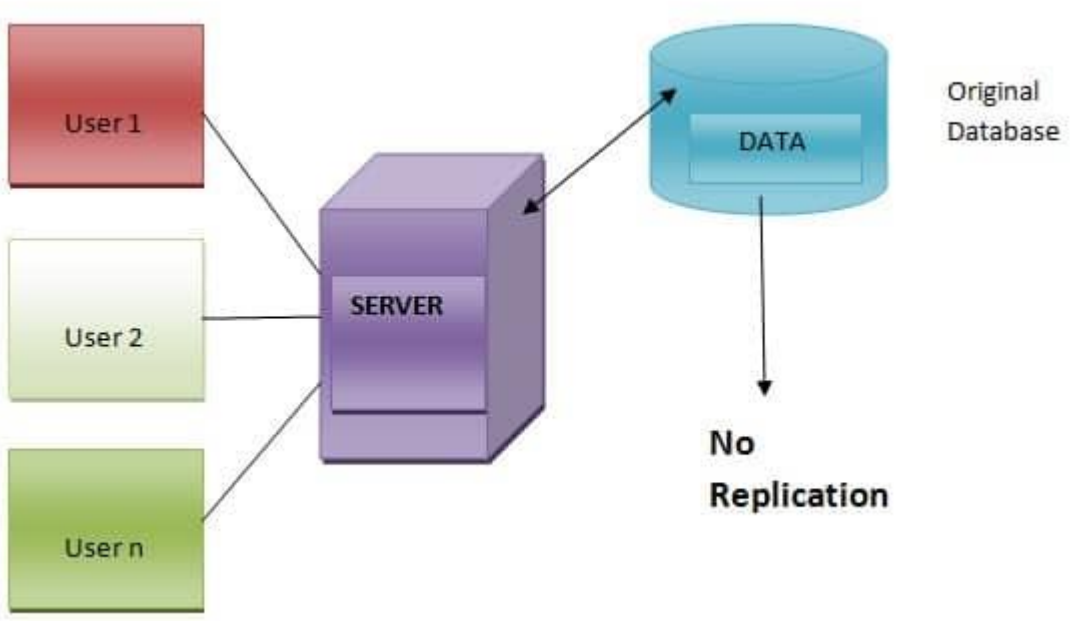
- It gives high availability of data. In this scheme, the database is available at each location.
- It supports faster execution of queries.

Disadvantages of full replication

- In a full replication scheme, concurrency control is difficult to achieve in full replication.
- During updating each and every side need to be updated therefore update operation is slower.

2. **No Replication**

No replication means each fragment is stored exactly at one location only.



Advantages of no replication

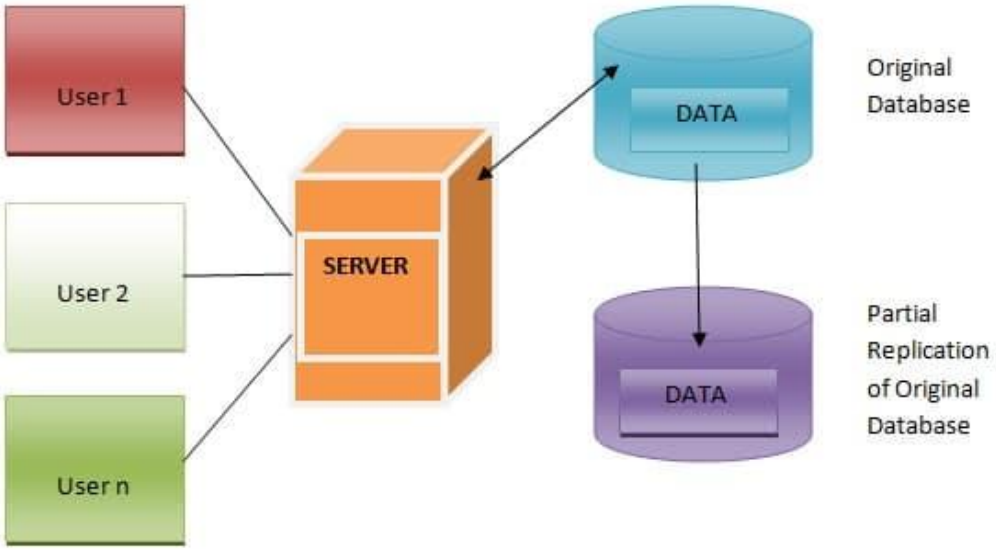
- Concurrency can be easily minimized.
- Easy recovery of data becomes easy.

Disadvantages of no replication

- Poor availability of data.
- Slows down the query execution process, because multiple clients are accessing the same data at the same server.

3. **Partial replication**

A partial replication scheme means only part of the or data fragments are replicated.



3.1.1. eplicationTerminology

1) Asynchronous: In the context of replication, this is the ability to log SQL operations targeted for another siteforlaterprocessingbythatsite.NotethatthisdoesNOTnecessarilyimplythattheactualprocessingof thetransactionswillbeasynchronouswhenthesubscriberfinallyconnectsto receivethechanges.
2) Bidirectional: In this form of replication, both the publishing (originator) and subscribing (receiving) site can update a particular data object. This is also referred to as“symmetric.”
3) Heterogeneous: This term implies replicating data among differentDBMS.

4) Heteromorphic: This term describes the ability to address and identify the same data element by different names(mapping)orinadifferentshape(differentunits,scale,displayformat,andsoon).
5) Horizontal partitioning: This is a form of data distribution in which a subset of the replication data for a table is selected through the use of a replication query that restricts the set of rows that are replicated to a specificsite.
6) Publisher: This is the originator of a replicated database change. Note that a publisher may also be a subscriber in bidirectionalschemes.
7) Push replication: A publisher site controls when replication is to occur and "pushes" the changes out to the subscribers.
8) PullReplication:Thesubscribersitesdeterminewhentheywishto receivereplication transactions.
9) Refresh:Thisisaprocesswherebyoneormoretablesatoneofthesubscribersitesarecompletelyrestored orupdated.
10) Snapshot:Thisisasingle-updaterformofreplicationwhereonlythepublishingsitecanupdatethedata.
11) Subscriber: This is a receiver of a replicated database change. Note that a subscriber may also be a publisher in bidirectionalschemes.
12) Symmetric: Asbidirectional.
13) Vertical partitioning: This is a form of data distribution in which a subset of the replication data for a table is selectedthroughtheuseofareplicationquerythatwillselectonlythecolumnsofinterestforaspecificsite.

3.1.2. Forms of Replication

Replication can be provided in numerous forms and combinations. There are two simple forms of replication are as below:

1) Single-Updater Systems: If one has a system where only one user has rights to update the data or one group whose updates can be serialized through one data source, replication problems will be fairly easy to solve. One shall has to deal with the remote configuration management issues involved with distributed systems, but several vendors are beginning to provide tools to that end, including Microsoft SMS, Oracle Battle Star, Stream, and Remote Ware Excellent.

   In addition, the replication is performed through read-only copies or "snapshots," and one does not have to worry about data collisions (the replication version of concurrency management) or sequence generator conflicts. One also does not have to worry if all of the remote nodes have synchronized in order to reorganize the server.

2) Multiple-Updater Systems: If any of the replicated tables must be updated at more than one site, things get much more complicated. The most obvious issue is data collision. This is a variation of the concurrency problem typical in OLTP (Online Transaction Processing) database systems. There is no locking mechanism available in asynchronous replicated systems (users at two or more autonomous sites can be updating the same data at about the same time). At replication time, the replication server must provide mechanisms to detect and resolve such datacollisions.

Thelatterarealsoreferredtoasbidirectionalorsymmetricsystems.Datacanbereplicatedatthediscretetable- row level or the transaction level. Replication can also be provided through a single distribution hub, or by forwarding through several intermediate distribution points to the final destination sites. Realistically, it is doubtfulthatthepropagationdelayandriskwouldbetolerablethroughmorethanaverysmallnumberofsites.

3.1.3. Steps in Data Replication

First, one needs data replication and synchronization for mobile devices because he/she assumes that most mobile devices are sometimes disconnected from a network and operating in isolation. This means that they need to access local data so that the device operator can continue to use the applications on the device. The amount of data to be stored varies based on the device capabilities, the application types, and the user preferences. Typically, the sequence of events taking place is as follows:

1) Initial Replication: Some data are replicated from the master to the replica. Normally, the master is a PC, server,orsomeotherdevicethathostssomedataandthereplicaisthemobiledevice,butthisdoesnothave tobethecase.Themobiledevicecanactasamasterandcreatereplicasonotherdevices.
2) Local Data Modification: This stage entails all of the user or machine interactions that modify the data for one node. Usage of the data while the device is weakly connected or disconnected may result in the data in other nodes becoming partially or completely obsolete (or in conflict if the same data are modified at some

other node with different information).

3) Synchronization: This stage involves exchange of synchronization messages that update the obsolete data and either resolve the conflicts in an automated fashion or present the user with the necessary information to resolve the conflicts manually.

After the initial replication, steps 2 and 3 can be repeated as many times as desired.

3.1.4. Data Replication Strategies

There are three data replication strategies:

1) Synchronous Data Replication: Under Synchronous Data Replication (SDR) strategy, updates are applied to all the database replicas of an object as part of the original transaction. The database replicas are then kept in a state of synchronization at all the network nodes by updating all the replicas as a part of one atomic transaction.

   In synchronous data replication, data replication takes place as soon as the source data is changed, provided thatallhardwarecomponentsandnetworksinthereplicationsystemareavailable.Atransactionisapplied only if all the interconnected sites agree. Synchronous data replication is appropriate for applications that require immediate data synchronization (figure3.2).

   It prevents two users from editing the same record at the same time and is also concerned with serializing transactions for backup and recovery; SDR eliminates integrity anomalies during replication.

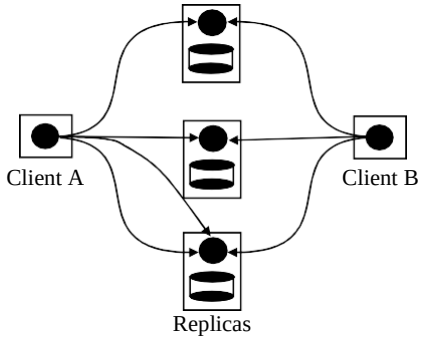


Figure 3.2: Synchronous Data Replication

   Basic scheme of this strategy is to connect each client (or front-end) with every replica, writes go to all replicas, but client can read from any replica (read-one-write-all replication).

   Synchronous Replication Issues

   i) Continuous Network Availability: Because the synchronous approach requires access to all slave databases at the time of update, 100 percent network availability is required to ensure that all transactions are completed successfully. The implication of 100 percent availability is obvious -- without the network connections a synchronous replication will not process, causing all update processing to be suspended and potentially locking theapplication.

   ii) High Cost: Network managers working with synchronous replication need to plan for a high network availability with some sort of circuit redundancy, either high-speed dial-back (Switched 56 or higher speed) or self-healing networks. The application's users are the people to decide what type of redundancy will be required (with the network groups assistance) so that an appropriate cost justification can beprepared.

   iii) High Network Requirements: Transaction volume and the speed at which transactions occur and are processed will dictate network requirements. High-speed connections do not guarantee successful implementation of an application. If the slave database's processor is not capable of processing the transactions in a timely manner, the network capacity will only increase the transaction processing backlogoftheslavesystemandslowdowntheprocessingofthemastersystem

2) Asynchronous Data Replication: In the asynchronous data replication (ASDR) strategy, the database transactions commit the source database, producing an identical copy of the data at both ends of the transaction resulting in an efficient technique. However, ASDR does not enforce consistency between the replicated databases (figure3.3).

The basic idea is to build available/scalable information services with read-any-write-any replication and a weak consistency model:

i) Nodenialofserviceduringtransientnetworkpartitions.
ii) Supportsmassivereplicationwithoutmassiveoverhead.
iii) Ideal for the Internet and mobilecomputing.

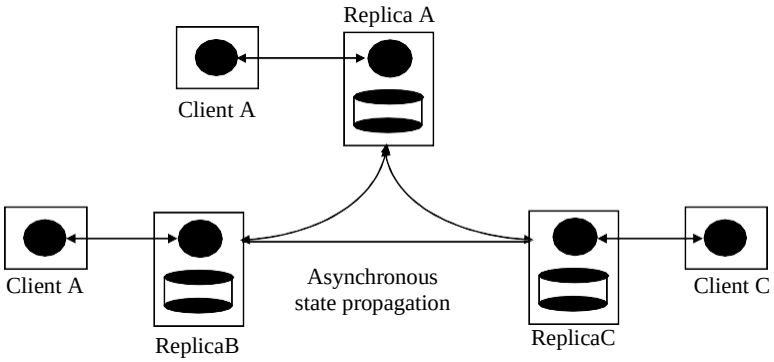


Figure 3.3: Asynchronous DataReplication

ThemainproblemwiththisstrategyisReplicasmaybeoutofdate,mayacceptconflictingwrites,andmay receive updates in differentorders.

Asynchronous ReplicationIssues

i) Integrity Problems: Asynchronous replication is what most vendors recommend and is the type o f replication currently in vogue. Asynchronous replication typically involves some form of replication engine or server that tracks all updates and then ensures that these updates are shared with all other systems. If a system is unavailable, the server continues to track the unapplied updates and will update the system when it is available. If the replication process cannot be completed with all systems, all databases may not reflect currentinformation.

ii) High Demands on Network: Asynchronous replication can be highly demanding on a network, because of the number of variables that need to be considered. Besides transaction volumes, other variables include line speeds, type of connections, speed and number of processors involved, data timeliness and number of replicationservers.

iii) Heavy Replication Queues: While the implementation of asynchronous replication eases the real-time, immediate demands on a network, the overall demand on the network is either shifted or spread out. Shifting of demand could involve the processing of replication overnight, however by shifting processing, the network capacity needs to be sufficient to handle high-volume updates to the remote databases. In spreading out demand, the network manager needs to provide reliable connections to ensurethatreplicationqueuesdonotbecomeoverlylarge,thuscausingahigh-volumeupdatesituation.

3.1.5. Internal View of ReplicationSystem

- The principal functionalities of a replication systemare:
  i) Replicacreation
  ii) Replicaplacement
  iii) Read/Writeoperation
  iv) Replicaconsistency.

- Consequently, system contains three principal modules(figure1):
  i) Replica Planner: The replica planner is responsible for the creation and placement of replicas on nodes.
  ii) Localization Manager: Next, the localization manager locates replicas for read/write operations and then performs theseoperations.
  iii) Consistency Manager: Finally, the consistency manager ensures replica consistency by exchanging update messages after each write operation and resolving updateconflicts.

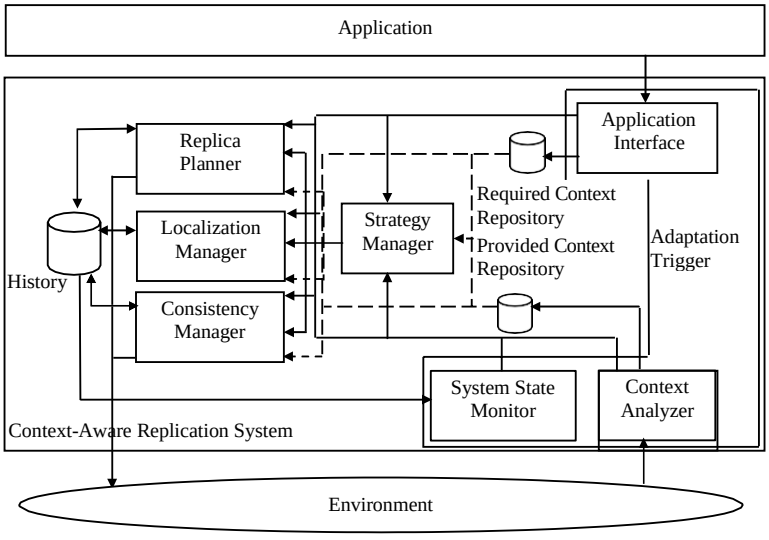


Figure 3.4: Internal View of Replication System

- Replicationsystemtakesintoaccounttheexecutioncontextandadaptsitsfunctionalitiestochangesinthis context. One consider three dimensions of adaptation accordingto:
  i) Adaptationtrigger(adaptationto contextvariations,adaptationtosystemstatevariations),

ii) Adaptationmoment(adaptationattheapplicationdeploymenttime,i.e.,configuration,adaptationatthe application run time i.e., reconfiguration),and
iii) Adaptation objects (replication scheme adaptation, replication strategyadaptation).

- Thus,onehastodealwiththereplicationschemeandreplicationstrategyconfigurationandreconfiguration as shown in tablebelow.

Table: Adaptation Classes

	Configuration	Reconfiguration
Replication Scheme	Replication scheme configuration	Replication scheme reconfiguration
Replication Strategy	Replication strategy configuration	Replication strategy reconfiguration

3.2.6.1. Replication SchemeConfiguration

- The replication scheme is represented by threeplans:
  i) Placement Plan: The placement plan indicates the placement of each replica onnodes.
  ii) Localization Plan: Next, the localization plan indicates for each user a set of nodes and links where he can reach eachreplica.
  iii) Consistency Plan: Finally, the consistency plan indicates for each replica a set of nodes and links where an update can be propagated from this replica to otherreplicas.
- These plans are provided by the replica planner, the localization manager and the consistency manager respectively.

- Thereplicationschemeisprovidedattheapplicationdeploymenttimeaccordingtothecurrentcontextand the current system state. This replication scheme is then stored in the history. The placement plan is immediately projected onto the environment. The localization and consistency plans are used at the read/writetime.

3.2.6.2. Replication SchemeReconfiguration

- Inordertomonitortheexecutioncontextandthesystemstate,oneproposesadaptationtriggermodules:
  i) ApplicationInterface,
  ii) ContextAnalyser,
  iii) System StateMonitor

- The application interface and context analyser detect the pertinent change in required context information and provided context information respectively. Next, they store this change in the required context repository and the provided context repository. Finally, they notify different modules (replica planner, localization manager and consistency manager) of thischange.

- These latter provide a replication scheme that is most adapted to the change in context by modifying their corresponding plans that are stored in the history. For example, if the bandwidth is decreased, the system changes some replica locations in order to avoid the use of weak bandwidthlinks.

- The basic objective of the replication technique is to improve the performance and to increase the data availability andconsistency.

- To reach these objectives, replication system monitors its own state. This state is represented by system performance parameters, data availability parameters, and data consistency parameters. For example, the response time is a parameter characterizing the system performance. Each parameter is measured by a metric.

- Differentparametersareevaluatedbythesystemandstoredinthehistory.Thesystemstatemonitordetects the pertinent change in these parameters and notifies the replica planner, localization manager and consistency manager of this change. These modules modify the replication scheme in order to redress the system state. For example, if the response time increases, the replica planner modifies some replica locations in order to reduce thisparameter.

3.2.6.3. Replication StrategyConfiguration

- Replication system adapts the replication strategy to context or system state variations. For example, the system changes its strategy from a pessimistic strategy to an optimistic one in order to improve the data

availability. This adaptation is based on a set of rules that specify the strategy application conditions. To managethisadaptation,oneproposesastrategymanagerthatensuressystemconsistencyandimplementsa newstrategy.

- The strategy configuration is carried out at the application deployment time. Based on the current context and the current system state, the strategy manager chooses the most adapted replication strategy and implements it in the replica planner, localization manager, and consistencymanager.

3.2.6.4. Replication StrategyReconfiguration

- The replication strategy reconfiguration is carried out at application runtime.

- After receiving the notification from the adaptation trigger modules, the strategy manager chooses the adapted strategy, ensures the system consistency and implements this strategy in some or all modules (replica planner, localization manager, and consistencymanager).

3.2.6.5. AdaptationProcess

- To summarize, at the application deployment time, the current context and system state are analyzed by adaptation trigger modules. Based on this analysis, the strategy manager chooses the most adapted replication strategy and implements it in the replica planner, localization manager, and consistency manager. Then these modules provide a replication scheme that is adapted to this context as shown in figure3.5.

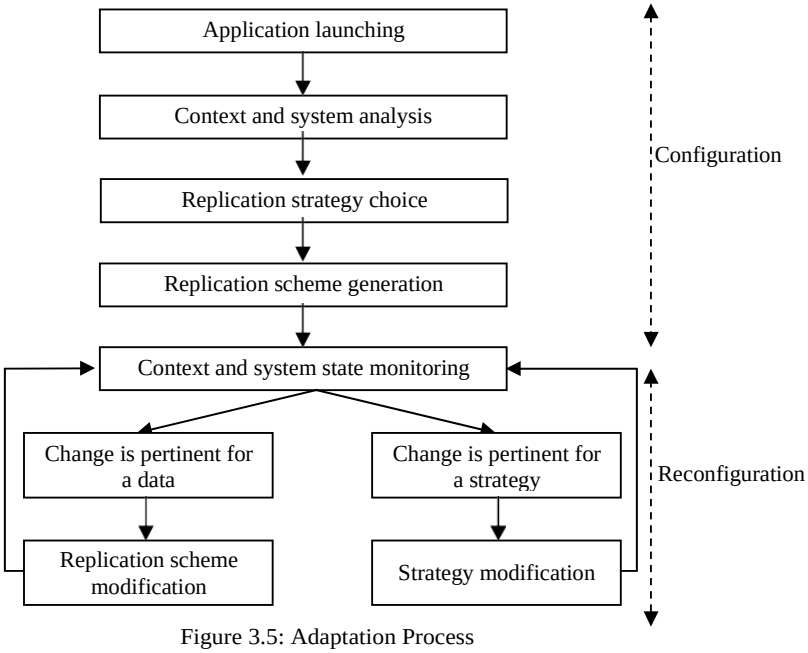


Figure 3.5: Adaptation Process

- At the application run time, adaptation trigger modules monitor context information and system state parameters and detect the pertinent change in thisinformation:
  i) Ifthischangeispertinentforareplicationstrategythenthesemodulesnotifythestrategymanager.
  ii) Otherwise, if it is pertinent for data then trigger modules notify the replicas planner, localization manager and consistencymanager.

- Intheformercase,thestrategymanagerchoosesthereplicationstrategythatisadaptedtonewinformation and ensures the system consistency. In the latter case, the replica planner, localization manager and consistency manager modify their plans that are stored in thehistory.

3.1.6. External View of ReplicationSystem

- Replicationsysteminteractswiththeapplication inordertocommunicatedatatobereplicatedandtheirre- source constraints as shown in figure3.6.
- It also interacts with the execution environment in order to collect its resources and to project the replication scheme onto thisenvironment.

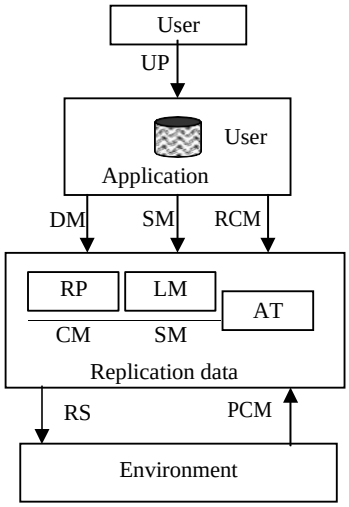


Figure 3.6: External View of the Replication System

- Theseinteractionsarecarriedoutaccordingtoasetofformats(meta-models)thataredefinedbythesystem designer.
- First,theapplicationdesignermustdefineanddescribethedatatobereplicated.Thisdescriptionallowsthe application to treat the data and to communicate them to the replication system. This latter can be based on astanderddataformatlikeobjectandcomponentdata,oraspecificformatthatdescribesthedatastructure.
- The application designer must also describe the data constraints before communicating them to the replicationsystem.Thisdescriptioniscarriedoutaccordingtoawelldefinedformatthatspecifiesthe mannerinwhichdeferentreplicationsystemmodulescanunderstanddataconstraintsemantics.Thisformat isdefinedbythesystemdesignerandrepresentsameta-modelfortherequiredcontextdescription.
- Provided context information must also be described in order to exchange and analyse them. This description is based on a meta-model for the provided context description. Replication strategies and their corresponding rules have to be described in order to adapt them to execution context and system state variations.
- Finally, the replication scheme also has to be described in order to apply the replication strategy that is based on thisscheme.

3.3. ADAPTIVE CLUSTERING FORMOBILE WIRELESSNETWORK

3.3.1. Introduction

- Wireless communication and the lack of centralized administration pose numerous challenges in mobile wireless ad-hoc networks (MANETs). Node mobility results in frequent failure and activation of links, causing a routing algorithm reaction to topology changes and hence increasing network control traffic. Ensuring effective routing and QoS (Quality of Service) support while considering the relevant bandwidth and power constraints remains a great challenge. Given that MANETs may comprise a large number of MNs, a hierarchical structure will scalebetter.
- Hence, one promising approach to address routing problems in MANET environments is to build hierarchies among the nodes, such that the network topology can be abstracted. This process is commonly referred to as clustering and the substructures that are collapsed in higher levels are calledclusters.

- The concept of clustering in MANETs is not new; many algorithms that consider different metrics and focusondiverseobjectiveshavebeenproposed.However,mostexistingalgorithmsfailtoguaranteestable clusterformations.Moreimportantly,theyarebasedonperiodicbroadcastingofcontrolmessagesresulting in increased consumption of network traffic and mobile hosts (MH)energy.

- In the network architecture of interest, nodes are organised into non-overlapping clusters. The clusters are independently controlled and are dynamically reconfigured as nodes move. This network architecture has three mainadvantages:
  i) It provides spatial reuse of the bandwidth due to nodeclustering.
  ii) Bandwidth can be shared or reserved in a controlled fashion in eachcluster.
  iii) The bandwidth allocation is robust in the face of topological changes caused by node motion, node failure and node insertion/removal because the cluster algorithm itself can efficiently adapt to such changes.

3.3.2. Single-Hopping *vs.*Multi-Hopping

- An important wireless network feature addressed is multi-hopping, i.e., the ability of the radios to relay packets from one to another without the use of base stations. Most of the nomadic computing applications today are based on a single hop radio connection to the wirednetwork.

- Figure 3.7 shows the cellular model commonly used in the wireless networks. A, B, C, and D are fixed base stations connected by a wired backbone. Nodes 1 through 8 are mobile nodes. A mobile node is only one hop away from a base station. Communications between two mobile nodes must be through fixed base stations and the wiredbackbone.

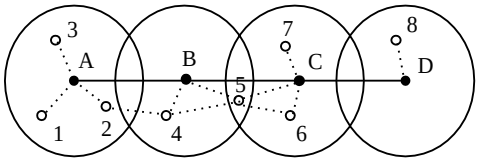


Figure 3.7: Conventional Cellular Networks (Single-hop)

- In parallel with (and separately from) the single hop cellular model, another type of model, based on radio toradiopacketmulti-hopping,hasbeenemergingtoserveagrowingnumberofapplicationswhichrelyon a fast deployable, wireless infrastructure. The classic examples are battlefield communications and (in the civiliansector)disasterrecovery(fire,earthquake)andsearchandrescue.Arecentadditiontothissetisthe "ad-hoc"personalcommunicationsnetwork,whichcouldberapidlydeployedonacampus,forexample,to support collaborative computing and access to the Internet during special events (concerts, festivals etc.). Multi-hopping through wireless repeaters strategically located on campus permits to reduce battery power and to increase network capacity. More precisely, by carefully limiting the power of radios, we conserve battery power. Furthermore, we also cause less interference to other transmissions further away; this gives theadditionalbenefitof"spatialreuse"ofchannelspectrum,thusincreasingthecapacityofthesystem.

- Interestingly,themulti-hoprequirementmayalsoariseincellularnetworks.Ifabasestationfails,amobile nodemaynotbeableto access thewirednetworkinasinglehop. In figure3.8, ifbasestationBfails,node 4mustaccessbasestationsAorCthroughnode2ornode5whichactaswirelessmulti-hoprepeaters.

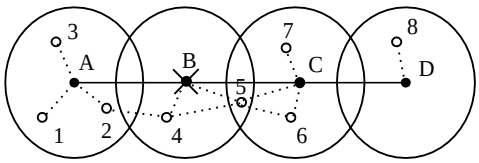


Figure 3.8: Multi-hop Situation Occurs when Base Station B Fails

3.3.3. Multi-ClusterArchitecture

- A major challenge in multi-hop, multimedia networks is the ability to account for resources so that bandwidth reservations (in a deterministic or statistical sense) can be placed on them. In cellular (single hop) networks such accountability is made easy by the fact that all stations learn of each other's requirements,eitherdirectly,orthroughacontrolstation(forexample,basestationincellularsystems).

- This solution can be extended to multi-hop networks by creating clusters of radios, in such a way that accesscanbecontrolledandbandwidthcanbeallocatedineachcluster.Thenotionofclusterhasbeenused alsoinearlierPacketRadionets,butmainlyforhierarchicalroutingratherthanforresourceallocation.

- Most hierarchical clustering architectures for mobile radio networks are based on the concept of clusterhead. The clusterhead acts as a local coordinator of transmissions within the cluster. It differs from the base station concept in current cellular systems, in that it does not have special hardware and in fact is dynamically selected among the set of stations. However, it does extra work with respect to ordinary stations, and therefore it may become the bottleneck of the cluster. To overcome these difficulties, one eliminate the requirement for a clusterhead altogether and adopt a fully distributed approach for cluster formation and intra-clustercommunications.

- The objective of the proposed clustering algorithm is to find an interconnected set of clusters covering the entire node population. Namely, the system topology is divided into small partitions (clusters) with independent control. A good clustering scheme will tend to preserve its structure when a few nodes are moving and the topology is slowly changing. Otherwise, high processing and communications overheads will be paid to reconstructclusters.

- Within a cluster, it should be easy to schedule packet transmissions and to allocate the bandwidth to real timetraffic.Acrossclusters,thespatialreuseofcodesmustbeexploited.Sincethereisnonotionofcluster-head,eachnodewithinaclusteristreatedequally.Thispermitsustoavoidvulnerablecentersandhotspots of packet trafficflow.

3.3.4. ClusteringAlgorithm

- In order to support multimedia traffic, the wireless network layer must guarantee QoS (bandwidth and delay) to real time traffic components. The approach to provide QoS(Quality of Service) to multimedia consists of the following twosteps:
  i) Partitioning of the multi-hop network into clusters, so that controlled, accountable bandwidth sharing can be accomplished in eachcluster;
  ii) Establishment of Virtual Circuits with QoSguarantee.

- The objective of the clustering algorithm is to partition the network into several clusters. Optimal cluster size is dictated by the tradeoff between spatial reuse of the channel (which drives toward small sizes), and delay minimization (which drives towards large sizes). Other constraints also apply, such as power consumption and geographical layout. Cluster size is controlled through the radio transmission power. For the cluster algorithm, one has so far assumed that transmission power is fixed and is uniform across the network.

- Within each cluster, nodes can communicate with each other in at most two hops. The clusters can be constructed based on node ID. The following algorithm partitions the multi-hop network into some non-overlapping clusters. One makes the following operational assumptions underlying the construction of the algorithminaradionetwork.Theseassumptionsarecommontomostradiodatalinkprotocols:
  i) A1: Every node has a unique ID and knows the IDs of its 1-hop neighbors. This can be provided by a physical layer for mutual location and identification of radionodes.
  ii) A2:Amessagesentbyanodeisreceivedcorrectlywithinafinitetimebyallits1-hopneighbors.
  iii) A3: Network topology does not change during the algorithmexecution.

```
Distributed Clustering Algorithm(Γ)
Γ: the set of lD's of my one-hop neighbors and myself
{
    if (my_id == min(Γ))
    {
        my_cid = my_id;
        broadcast cluster(my_id,my_cid);
        Γ=Γ-{my_id};
    }
    for(;;)
    {
        on receiving clustered, cid)
        {
            set the cluster ID of node id to cid;
            if (id==cid and (my_cid==UNKNOWN or my_cid>cid))
                my_cid = cid;
            Γ=Γ-{id};
            if (my_id == min(Γ))
            {
                if (my_cid==UNKNOWN) my_cid = my_id;
                broadcast cluster(my_id,my_cid);
                Γ=Γ-{my_id};
            }
        }
        if (Γ==∅) stop;
    }
}
```

Figure 3.9: Distributed Clustering Algorithm

- Onecanfindfromthisalgorithm(figure3.9)thateachnodeonlybroadcastsonedustermessagebeforethe algorithm stops, and the time complexity is O (|V|) where V is the set of nodes. The clustering algorithm convergesveryrapidly.Intheworstcase,theconvergenceislinear inthetotalnumberofnodes.

- For example, let consider the topology in figure3.10.

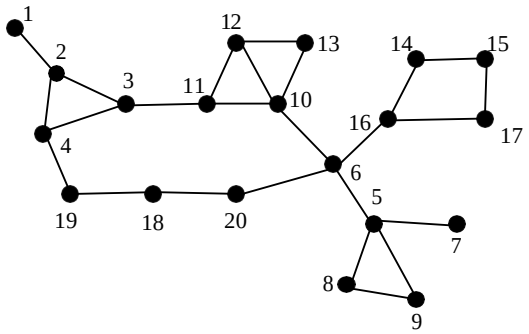


Figure 3.10: System Topology

- After clustering, in figure 3.11, one can find six clusters in the system, which are {1,2},{3,4,11}, {5,6,7,8,9}, {10,12,13}, {14,15,16,17}, {18,19,20}. To prove the correctness of the algorithm one has to show that:
  i) Every node eventually determines itscluster;
  ii) In a cluster, any two nodes arc at most two hopsaway;
  iii) The algorithmterminates.

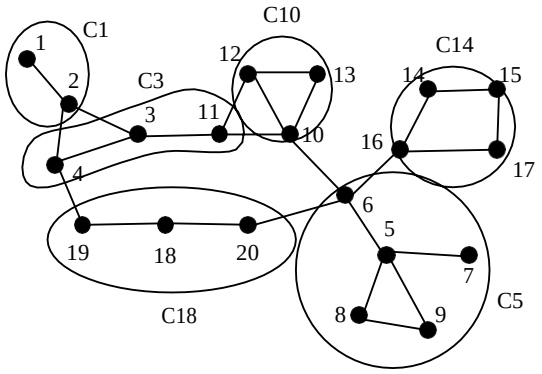


Figure 3.11: Clustering

Lemma 1: Every node can determine its cluster and only one cluster.

Proof:TheclusterIDofeachnodeiseitherequaltoitsnodeIDorthelowestclusterIDofitsneighbors.Every node has to decide its cluster ID once it becomes the lowest ID node in its locality. Hence, every node can determine its cluster and only onecluster.

Lemma 2: In a cluster, any two nodes are two hops away at most.

Proof: Consider nodes in the same cluster. Every node can reach the node which node ID is equal to cluster ID in one hop. Thus, any two nodes are two hops away at most.

Theorem 1: Eventually the algorithm terminates.

Proof: Since every node can determine its cluster (Lemma 1), the set Γ will eventually become empty. Thus, the algorithm will terminate.

Theorem 2: Each node transmits only one message during the algorithm.

Proof:AnodebroadcastsmessagesonlyatthetimeitdecidesitsclusterID.Thus,onlyonemessageissentout before the algorithmstops.

Theorem 3: The time complexity of the algorithm is O(|V|).

Proof: From the distributed clustering algorithm, each message is processed by a fixed number of computation steps. From Theorem 2, there are only |V| messages in the system. Thus, the time complexity is O(|V|).

3.3.5. Cluster Maintenance in the Presence ofMobility

- In the dynamic radionetwork:
  i) Nodes can change location,
  ii) Nodes can beremoved,
  iii) Nodes can beadded.

- A topological change occurs when a node disconnects and connects from/to all or part of its neighbors, thus,alteringtheclusterstructure.Systemperformanceisaffectedbyfrequentclusterchanges.Therefore,it is important to design a cluster maintenance scheme to keep the cluster infrastructure as stable as possible. Inthisrespect,theproposedclusteralgorithmismorerobust,sincetherearefewerrestrictionsonclusters.

- Theclustermaintenanceschemewasdesignedtominimizethenumberofnodetransitionsfromonecluster toanother.

- For example, let consider figure 3.12(a). There are 5 nodes in the cluster and the hop distance is no more than2.Becauseofmobility,thetopologychangestotheconfigurationshowninfigure3.12(b).Atthistime, d(1,5) = d(2,5) = 3 > 2, where d(i, j) is the hop distance between node i and j. So the cluster needs to be reconfigured.Namely,oneshoulddecidewhichnode(s)shouldberemovedfromthecurrentcluster.Onelet the highest connectivity node and its neighbors to stay in the original cluster, and remove the other nodes. Recallthateachnodeonlykeepstheinformationofits"locality",thatis,oneandtwohopneighbors.Upon discoveringthatamember,sayx,ofitscluster isnolongerinitslocality;nodeyshouldcheckifthehighest connectivitynodeisaonehopneighbor.Ifso,yremovesxfromitscluster,otherwise,ychangescluster.

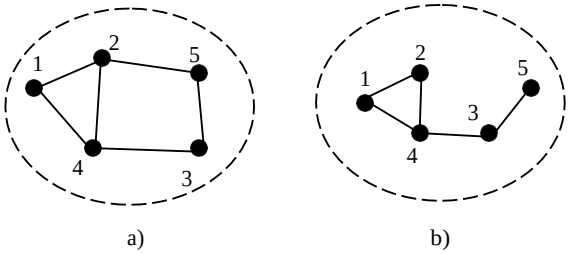


Figure 3.12: Re-Clustering

- Two steps are required to maintain the clusterarchitecture:
  Step 1: Check if there is any member of my cluster has moved out of my locality.
  Step 2: If Step 1 is successful, decide whether I should change cluster or remove the nodes not in my locality from my cluster.

3.3.6. CodeAssignment

- Each node has a transceiver which can either transmit or receive at any given time. In the spread-spectrum code-division system, the receiver should be set to the same code as the designated transmitter. For simplicity (and to be conservative) one assumes no capture. That is, if two or more transmissions interfere at the same receiver, none is received, regardless of thecode.
- One assumes that there is a small set of "good" spread-spectrum codes which are low cross-correlation. Since the number of codes we can use is very limited, spatial reuse of codes will be important. Thus each cluster is assigned a single code which is different from the codes using in the neighbor cluster. The problem of the code selection can be formulated as a graph coloringproblem.
- There are three options for using the dedicated code within acluster:
  i) Receiver-Based Code Assignment: Every node within a cluster is assigned a common receiving code. All neighbor nodes send packets to a node using its code. In this scheme, a receiver only listens to one code, but both inter-cluster and intra-cluster collisions canoccur.
  ii) Transmitter-Based Code Assignment: Within a cluster, every node uses a common transmitting code so that there is no inter-cluster collision. If no two nodes in a cluster are transmitting simultaneously, there will be no intra-clustercollision.
  iii) Transmitter-Receiver Pairs within a Cluster: Another approach is to assign a common code to all transmitter-receiver pairs within a cluster. This code assignment requires that some other codes be assigned for inter-clustercommunications.

3.3.7. NetworkInitialization

- Anodewhichdoesnotyetbelongtoaclusterlistenstothecontrolcodeuntiltimeout.Then,ittransmitsits ownID(usingthecontrolcode)andrepeatstheprocedureuntilithearsfromoneoftheneighbors.Channel access in this phase is CSMA, This basic communications facility allows nodes to organize themselves in clusters following the algorithm describedabove.

- Once a cluster is formed, the cluster leader communicates with the neighbors (using the control code) to selectthecodes.Onlywhenthecodeassignmentiscompleted(i.e.,eachclusterhasbeenassigneditscode) can user data be accepted by the nodes and transmitted in thenetwork.

3.4. FILESYSTEMS

3.4.1. Introduction

- File systems are fundamental middleware, joining applications to the operating system. A file system that has the same appearance and behavior, whether at the office, home, or in transit, can support applications without requiring that they be modified to accommodate mobility. This important aspect of distributed systems location transparency is lost if files have different names, are found in different places on different computers, or have to be moved by hand before they can beused.
- The goal of file system is to extend the location transparency of distributed file systems to mobile computers and also efficient and transparent access to shared files within a mobile environment while maintaining dataconsistency.
- Network communication is necessary to convey updates from the client to the server, to keep cached data consistent with other clients and servers, to fetch data missing from the client cache, and to service informational or maintenance requests. A mobile client that provides all of these operates in connected mode. Eliminating these services places a client in disconnectedmode.

3.4.2. Mobile File System Modes ofOperation

- This figure 5.1 shows the modes of operation as cache management operations are modified to avoid network communications. In connected mode, the file system offers the usual semantics, but places the greatest demands on thenetwork.

- Indisconnectedmode,manyofthesemanticguaranteesofferedbythefilesystemareeliminated,butsoare all network requirements. In between are fine-grained modes of operation that trade communications for cacheconsistency.

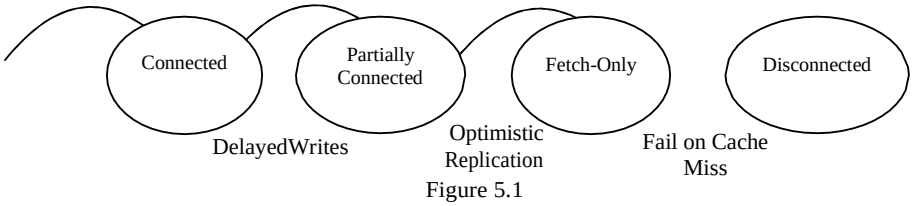


Figure 5.1

3.4.3. Goals of FileSystem

The goals of a file system is to support efficient, transparent, and consistent access to files, no matter where the client requesting files or the servers offering files arelocated.

1) Efficiency:Efficiencyisofspecialimportanceforwirelesssystemsasthebandwidthislowsotheprotocol overhead and updating operations etc. should be kept at aminimum.

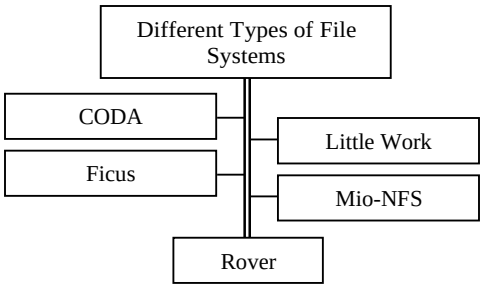
2) Transparency: Transparency addresses the problems of location-dependent views on a file system. To supportmobility,thefilesystemshouldprovideidenticalviewsondirectories,filenames,accessrightsetc., independent of the currentlocation.

3) Consistency: The basic problem for distributed file system that allow replication of data for performance reasonsistheconsistencyofreplicatedobjectslikefiles,partsoffiles,partsofdatastructureetc.

   To avoid inconsistencies many traditional systems apply mechanisms to maintain a permanent consistent view for all users of a file system. This consistency is achieved by atomic updates similar to database systems.Mobilesystemshavetouseaweakconsistencymodelforfilesystems.Weakconsistencyimplies certain periods of inconsistency that have to be tolerated for performance reason. However, the overall file system should remain consistent so conflict resolution strategies are needed for reintegration. Reintegrationistheprocessofmergingobjectsfromdifferentusersresultinginoneconsistentfilesystem.

3.4.4. Different Types of FileSystems

Figure below shows the different types of distributed file systems:



3.4.4.1. Constant Data Availability(CODA)

- The predecessor of many distributed file systems that can be used for mobile operation is the Andrew File System(AFS).
- AFS was designed to support the entire CMU community, which implied that approximately 10,000 workstationswouldneedtohaveaccesstothesystem.Tomeetthisrequirement,AFSnodesarepartitioned into two groups. One group consists of a relatively small number of dedicated vice file servers, which are centrallyadministered.Theothergroupconsistsofaverymuchlargercollectionofvirtueworkstationsthat give users and processes access to the file system, as shown in figure5.2.

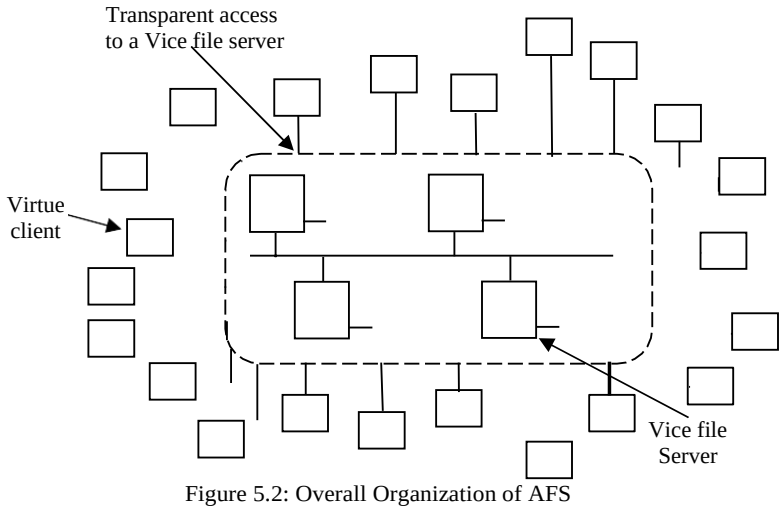


Figure 5.2: Overall Organization of AFS

- CODAisthesuccessorofAFSandofferstwodifferenttypesofreplication–serverreplicationandcaching on clients. Disconnected clients work only on the cache, i.e., applications use only cached replicated files. Figure 5.3 shows the cache between an application and theserver:

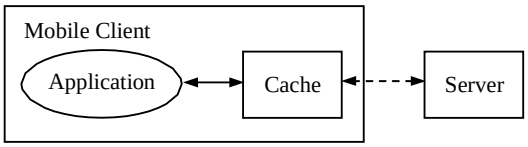


Figure 5.3: Application, Cache, and Server in Coda

- CODA is a transparent extension of the client's cache manager. This very general architecture is valid for most of today's mobile systems that utilise acache.
- To provide all the necessary files for disconnected work, CODA offers extensive mechanisms for pre-fetching of files while still connected, called hoarding. If the client is connected to the server with a strong connection (figure 5.4), hoarding transparently pre-fetches files currentlyused.

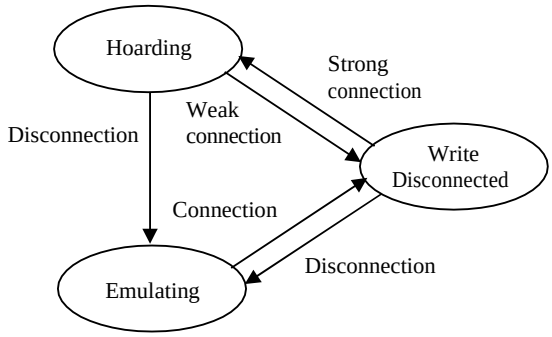


Figure 5.4: States of a Client in Coda

- This automatic data collection is necessary for it is impossible for a standard user to know all the files currently used. While standard programs and application data may be familiar to a user, he/she typically does not know anything about the numerous small system files needed in addition (e.g., profiles, shared libraries, drivers,fonts).
- A user can predetermine a list of files, which CODA should explicitly pre-fetch. Additionally, a user can assign priorities to certain programs. CODA now decides on the current cache content using the list and a Least-Recently-Used (LRU)strategy.
- As soon as the client is disconnected, applications work on the replicates (figure 5.4, emulating). CODA follows an optimistic approach and allows read and write access to all files. The system keeps a record of changed files, but does not maintain a history of changes for each file. The cache always has only one replicate (possibly changed). After re-connection, CODA compares the replicates with the files on the server. If CODA notices that two different users have changed a file, re-integration of this file fails and CODAsavesthechangedfileasacopyontheservertoallowformanualre-integration.

3.4.4.2. LittleWork

ThedistributedfilesystemLittleWorkis,likeCODA,anextensionofAFS.LittleWorkonlyrequireschanges to the cache manager of the client and detects write conflicts during re-integration. Little Work has no specific tools for re-integration and offers no transactionservice.

However, Little Work uses more client states to maintain consistency:

1) Connected: The operation of the client is normal, i.e., no special mechanisms from Little Work are required. This mode needs a continuous high bandwidth as available in typical office environments using, e.g., aWLAN.
2) Partially Connected: If a client has only a lower bandwidth connection, but still has the possibility to communicate continuously, it is referred to as partiallyconnected.
3) Fetch Only: If the only network available offers connections on demand, the client goes into the fetch only state.NetworksofthistypearecellularnetworkssuchasGSMwithcostspercall.
4) Disconnected: Without any network, the client is disconnected. Little Work now aborts if a cache misoccurs, otherwise replicates areused.

3.4.4.3. Ficus

- Ficusisadistributedfilesystem,whichisnotbasedonaclient/serverapproach.Ficusallowstheoptimistic use of replicates, detects write conflicts, and solves conflicts on directories. Ficus uses so-called gossip protocols.
- Amobilecomputerdoesnotnecessarilyneedtohaveadirectconnectiontoaserver.Withthehelpofother mobile computers, it can propagate updates through the network until it reaches a fixed network and the server. Thus, changes on files propagate through the networkstep-by-step.
- Ficus tries to minimize the exchange of files that are valid only for a short time, e.g., temporary files. A critical issue for gossip protocols is how fast they propagate to the client that needs this information and howmuchunnecessarytrafficitcausestopropagateinformationtoclientsthatarenotinterested.

3.4.4.4. MIo-NFS

- The system mobile integration of NFS (MIo-NFS) is an extension of the Network File System (NFS). In contrast to many other systems, MIo-NFS uses a pessimistic approach with tokens controlling access to files. Only the token-holder for a specific file may change this file, so MIo-NFS avoids write conflicts. Read/write conflicts as cannot beavoided.

- MIo-NFS supports three differentmodes:
  i) Connected: The server handles all access to files asusual.
  ii) Loosely Connected: Clients use local replicates, exchange tokens over the network, and update files via thenetwork.
  iii) Disconnected: The client uses only local replicates. Writing is only allowed if the client is token-holder.

3.4.4.5. Rover

Compared to CODA, the Rover platform uses another approach to support mobility. Instead of adapting existing applications for mobile devices, Rover provides a platform for developing new, mobility aware applications. Two new components have been introduced in Rover:

1) Relocatable Dynamic Objects: These are objects that can be dynamically loaded into a client computer from a server (or *vice versa*) to reduce client-server communication. A trade-off between transferring objects and transferring only data for objects has to be found. If a client needs an object quite often, it makessensetomigratetheobject.Objectmigrationforasingleaccess,ontheotherhand,createstoomuch overhead.
2) Queued Remote Procedure Calls: These calls allow for non-blocking RPCs even when a host is disconnected. Requests and responses are exchanged as soon as a connection is available again. Conflict resolution is done in the server and is applicationspecific.

3.5. DISCONNECTED OPERATIONS

3.5.1. Introduction

- Disconnected operation is a mode of computer operation that enables a network client system to continue accessingcriticaldataduringtemporarynetworkfailures.Theabilitytooperatedisconnectedcanbeuseful even when connectivity isavailable.

- Disconnected operation for mobile computing refers to situations such as a person taking his laptop home from work at night or on vacation for a week, or a person suspending wireless communications for some period of time to conserve battery power or save money on an expensive connection, but still being able to use his computer as though connected. The goal is that the disconnection be transparent to the user by keepingtheenvironmentthesameandtheresourceshighlyavailable.Keepingsystemandnetworkfailures transparenttoauserbyusingreplicationtokeephisfilesavailableisessentiallyfulfillingthesamegoal.

- Although not originally designed with this in mind, both Coda and Ficus were capable of being used for primarily disconnectedoperation.

- Cache misses during periods of disconnection are very serious since they cannot be serviced, and thus could possibly prevent the user from being able to work at all. Research on both projects determined that simple caching techniques, like least-recently-used algorithms, were not good enough alone to provide the availabilityoffilesnecessaryfordisconnectedoperation.RelatedtotheFicusproject,UCLAisdeveloping the Seer Predictive Caching System to attempt to reduce the frequency of cache misses. It attempts to measure the semantic relationships between objects to determine which ones are necessary to the user's tasks and put those in the cache, transparently, without enlisting the user'sinput.

- Coda, on the other hand, has the user actively participate in cache management, through a means called hoard profiles, that the user can create to tell the caching software what files he is interested in having available,andtheirrelativepriorities.Itkeepsthecacheup-to-dateusingperiodichoardwalkswhich

determine the validity and suitability of objects in the cache and replaces them if necessary. The user can explicitlyaskforonetobedoneatanytime.Italsohasafacilitycalledreferencespyingtoassisttheuserin creating a hoard profile. This facility records all file references between a given start and stop event. It is usefulfordiscoveringcertainfilesthattheusermaynotevenknowarebeingusedbysomeapplication.

- Disconnected operation makes the assumption that the mobile computing platform has some autonomy; it has its own disk, system and application software, etc., and a reasonable amount of processing and user interface capability. But the reality that resources will be limited on a mobile platform must be taken into account.Codahasoptimized themechanismthatlogsupdatesonthedisconnected clientinordertoreduce log sizes. One such optimization is the automatic removal of log records of previous actions whose effect has been cancelled out by a currentaction.

  For example, there needs to be no record of a file being created, updated, and then deleted during the same period of disconnection. Also, because Coda does whole-file caching, there is no reason to log the set of intermediate writes that occur between the open and close of a file. Reducing the log size conserves disk space on the mobile client and improves performance during reintegration when connection is restored.

- Thelongeraperiodofdisconnection,themorelikelyitisthattherewillbeupdateconflictstoresolveafter connectivity is restored. Coda's reintegration process resolves what conflicts it can automatically, and then it consults the user for instructions on how to resolve any remaining conflicts. Coda also has the notion of application-specificresolversthatcanbeinstalledintothesystem,thenbeinvokedautomaticallywhenever a conflict occurs involving an application-specific file. Reconciliation in Ficus and reintegration in Coda handle the detection and resolution of write-write conflicts only; there is no support for identifying read-writeconflicts.

  For example, a read-write conflict is a disconnected programmer compiling a program using a library that has meanwhile been updated by someone else. Once the programmer regains connection, no write-write conflict will have occurred, but executable will inconsistent with the library and there will be no warning. Toaddressthisissue,CMUisdevelopingadesignforisolation-onlytransactions,tobeavailableasanAPI that extends the normal Unix file facilities to give applications access to improved consistency for mobile computing,therebyallowingimportantread-writeinconsistencies tobedetectedandhandled.

3.5.2. Requirements for Connected State

1) Cache Management: Monitor requests for file input output (I/O) to the network server. Intercept monitored I/O requests and respond to them locally, satisfying client requests for network server files from thelocalcachewheneverpossible.Otherwise,contacttheservertosatisfytherequest.

   Many factors complicate the implementation of hoarding: File reference behavior, especially in the distant future, cannot be predicted with certainty. Disconnections and reconnection are often unpredictable. The true cost of a cache miss while disconnected is highly variable and hard to quantify. Activity at other clients must be accounted for, so that the latest version of an object is in the cache at disconnection. Since cache space is finite, the availability of less critical objects may have to be sacrificed in favor of more critical objects.

2) Response Time: Respond to network I/O requests at least as fast as, or faster than, the network server, on average.

3) Preparation for Disconnected Operation: Pre-fetch data from the server for possible future use in disconnectedmode.Movetothedisconnectedstateifthenetworkisnotavailableorexcessivelyexpensive.

3.5.3. Requirements for Disconnected State

1) Cache Management: Monitor requests for file I/O to the network server. Intercept monitored I/O requests and satisfy them from the previously stored cache whenever possible. Record when cached files have been modified or use during reintegration. Handle cache misses gracefully by reporting an error to the application.

2) Preparation for Reintegration: Monitor for network availability. Move to the connected state if the network is available upon user'srequest.

3.5.4. Requirements for Reintegrating State

- Synchronizeapreviouslystoredcachewiththenetworkserverbycopyingmodifiedfilestotheserver.

- Notifytheuserofupdateconflicts,andprovideforanapplicationinterfacetoresolvetheseconflicts.

3.5.4.1. First *vs*. Second ClassReplication

- It is appropriate to distinguish between first-class replicas on servers, and second-class replicas (i.e. cache copies)onclients.First-classreplicasareofhigherquality.Theyaremorepersistent,widelyknown,secure, available,completeandaccurate.Second-classreplicasareinferioralongallthesedimensions.

- Only by periodic revalidation with respect to a first-class replica can a second-class replica be useful. The function of a cache coherence protocol is to combine the performance and scalability advantages of a second-class replica with the quality of a first-class replica. When disconnected, the quality of the second-classreplicamaybedegradedbecausethefirst-classreplicauponwhichitiscontingentisinaccessible.The longerthedurationofdisconnection,thegreateristhepotentialfordegradation.Whereasserverreplication preservesthequalityofdatainthefaceoffailures,disconnectedoperationforsakesqualityforavailability.

3.5.4.2. Optimistic *vs.* Pessimistic ReplicaControl

- The choice between two families of replica control strategies, pessimistic and optimistic, is central to the design of disconnected operation. A pessimistic strategy avoids conflicting operations by disallowing all partitioned writes or by restricting reads and writes to a single partition. An optimistic strategy provides muchhigheravailabilitybypermittingreadsandwriteseverywhere,anddealswiththeattendantdangerof conflicts by detecting and resolving them after theiroccurrence.
- A pessimistic approach towards disconnected operation would require a client to acquire shared or exclusive control of a cached object prior to disconnection, and to retain such control until reconnection. Possession of exclusive control by a disconnected client would preclude reading or writing at all other replicas.

- Possession of shared control would allow reading at other replicas, but writes would still be forbidden everywhere. Acquiring control prior to voluntary disconnection is relatively simple. It is more difficult when disconnection is involuntary, because the system may have to arbitrate among multiple requesters. Unfortunately,theinformationneededtomakeawisedecisionisnotreadilyavailable.

- For example, the system cannot predict which requesters would actually use the object, when they would release control, or what the relative costs of denying them access would be. Retaining control until reconnectionisacceptableinthecaseofbriefdisconnections.Butitisunacceptableinthecaseofextended disconnections.Adisconnectedclientwithsharedcontrolofanobjectwouldforcetherestofthesystemto defer all updates until it reconnected. With exclusive control, it would even prevent other users from making a copy of the object. Coercing the client to reconnect may not be feasible, since its whereabouts may not be known. Thus, an entire user community could be at the mercy of a single errant client for an unbounded amount oftime.
- Placingatimeboundonexclusiveorsharedcontrolavoidsthisproblembutintroducesothers.Oncealease expires,adisconnectedclientlosestheabilitytoaccessacachedobject,evenifnooneelseinthesystemis interested in it. This, in turn, defeats the purpose of disconnected operation which is to provide high availability. Worse, updates already made while disconnected have to bediscarded.
- Anoptimisticapproachhasitsowndisadvantages.Anupdatemadeatonedisconnectedclientmayconflict with an update at another disconnected or connected client. For optimistic replication to be viable, the system has to be more sophisticated. There needs to be machinery in the system for detecting conflicts, for automatingresolutionwhenpossible,andforconfiningdamageandpreservingevidenceformanualrepair. Having to repair conflicts manually violates transparency, is an annoyance to users, and reduces the usability of thesystem.
- Usinganoptimisticstrategythroughoutpresentsauniformmodelofthesystemfromtheusersperspective. At any time, he is able to read the latest data in his accessible universe and his updates are immediately visible to everyone else in that universe. His accessible universe is usually the entire set of servers and clients. When failures occur, his accessible universe shrinks to the set of servers he can contact, and the set of clients that they, in turn, can contact. In the limit, when he is operating disconnected, his accessible universe consists of just his machine. Upon reconnection, his updates become visible throughout his now- enlarged accessibleuniverse.

3.5.5. Weakly-Connected Operation

- Weakly-connected operation refers to the condition arising from wireless communications links that are slow and have unreliable or intermittentconnectivity.
- The goal is the same as for disconnected operation, that connectivity problems remain transparent and resources available; however, as designed, neither Coda nor Ficus were originally able to address weakly-connected operation because many of their algorithms were designed to take advantage of high-bandwidth LAN connections during all times ofconnection.
- The Coda team has begun researching and developing a number of enhancements to improve Coda's performance in the case of weakly-connected operation. One such enhancement is speeding up cache validationbyincreasingthegranularityofcoherencecheckingtotwolevels.Theirdatahasshownthatmost cacheitemsarestillvalidatreconnectiontime.Sotheyhaveadded versionstampsatthevolume(groupof file objects) level as well as the file level. Volume version stamps are checked first, and if no object in a volume has become invalid, the volume stamp will still be valid, so the individual files belonging to that volumedonothavetobevalidatedseparately.Thusonlyonecheckisnecessarytovalidateanentiregroup of files. If the volume stamp has become invalid, however, then all the files in the volume will have to be checked.
- Another such enhancement is called trickle reintegration. Instead of forcing full reintegration immediately upon reconnection, normal user operation is allowed to continue in the foreground, and reintegration takes place in the background, proceeding as the strength of the connectivity allows. The user can request a full foreground reintegration if necessary. Also, if a cache miss occurs during trickle reintegration, then system calculatesthetimeitwouldtaketoserviceitgiventhecurrentqualityoftheconnection.Basedonthat,and the file's priority as given in the hoard profile, and a patience threshold, it determines whether or not to service the miss or return an error. The user has the opportunity view cache miss statistics at anytime and force certain files to be fetched, if sodesired.

3.5.6. Advantages of Disconnected Operation

1) Disconnectedoperationcanextendbatterylifebyavoidingwirelesstransmissionandreception.
2) It can reduce network charges, an important feature when rates arehigh.
3) Itallowsradiosilencetobemaintained,avitalcapabilityinmilitaryapplications.
4) Itisaviablefallbackpositionwhennetworkcharacteristicsdegradebeyondusability.
5) It can be viewed as the extreme case of weakly-connected operation: The mobile client is effectively using a network of zero bandwidth and infinitelatency.

3.5.7. Disadvantages of DisconnectedOperation

1) Updates are not visible to otherclients
2) Cache misses may impedeprogress
3) Updates are at risk due totheft
4) Loss or damage, Update conflicts become morelikely
5) Exhaustion of cache space is aconcern.

3.6. EXERCISE

UNIT 4

Mobile Agents

4. MOBILE AGENTS

4.1 Introduction

Mobile agents are autonomous programs that can travel from computer to computer in a network, at times and to places of their own choosing. The state of the running program is saved, by being transmitted to the destination. The program is resumed at the destination continuing its processing with the saved state. They can provide a convenient, efficient, and robust framework for implementing distributed applications and smart environments for several reasons, including improvements to the latency and bandwidth of client-server applications and reducing vulnerability to network disconnection. In fact, mobile agents have several advantages in the development of various services in smart environments in addition to distributed applications.

• **Reduced communication costs:** Distributed computing needs interactions between different computers through a network. The latency and network traffic of interactions often seriously affect the quality and coordination of two programs running on different computers. As we can see from Figure 1, if one of the programs is a mobile agent, it can migrate to the computer the other is running on communicate with it locally. That is, mobile agent technology enables remote communications to operate as local communications.

• **Asynchronous execution:** After migrating to the destination-side computer, a mobile agent does not have to interact with its source-side computer. Therefore, even when the source can be shut down or the network between the destination and source can be disconnected; the agent can continue processing at the destination. This is useful within unstable communications, including wireless communication, in smart environments.

• Direct manipulation A mobile agent is locally executed on the computer it is visiting. It can directly access and control the equipment for the computer as long as the computer allows it to do so. This is helpful in network management, in particular in detecting and removing device failures. Installing a mobile agent close to a real-time system may prevent delays caused by network congestion.

• Dynamic-deployment of software Mobile agents are useful as a mechanism for the deployment of software, because they can decide their destinations and their code and data can be dynamically deployed there, only while they are needed. This is useful in smart environments, because they consist of computers whose computational resources are limited.

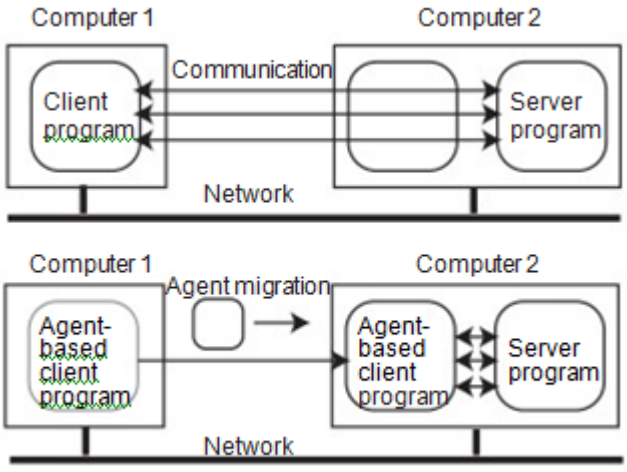


Figure 1: Reduced communication

. Easy-development of distributed applications Most distributed applications consist of at least two programs, i.e., a client-side program and a server side program and often spare codes for communications, including exceptional handling. However, since a mobile agent itself can carry its information to another computer, we can only write a single program to distributed computing. A mobile agent program does not have to communications with other computers. Therefore, we can easily modify standalone programs as mobile agent programs.

As we can see from Figure 2, mobile agents can save themselves through persistent storage, duplicate themselves, and migrate themselves to other computers under their own control so that they can support various types of processing in distributed systems.

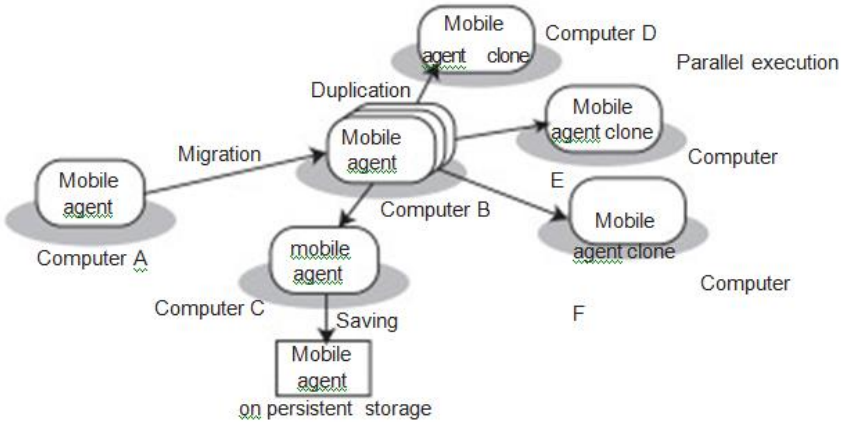


Figure 2: Functions of mobile agents in distributed system

4.2 Mobility and Distribution

It provided a description of mobile software paradigms for distributed applications. These are classified as client/server (CS), remote evaluation (REV), code on demand (COD), and mobile agent (MA) approaches. By decompiling distributed applications into code, data, and execution, most distributed executions can be modeled as primitives of these approaches as we can see from Figure 3.

• The client-server approach is widely used in traditional and modern distributed systems (Figure 3 a)). The code, data, and execution remain at computer A. Computer B requests a service from the server with some data arguments of the request. The code and remaining data to provide the service are resident within computer B. As a response, computer B provide the service requested by accessing computational resources provided in it. Computer B returns the results of the execution to computer A.

• The remote evaluation approach assumes that the code to perform the execution is stored at computer A (Figure 3 b)). Both the code and data are sent to computer B. As a response, computer B executes the code and data by accessing computational resources, including data, provided in them. An additional interaction returns the results from computer B to computer A.

• The code-on-demand approach is an inversion of the remote evaluation approach (3 c)). The code and data are stored at computer A and execution is done at computer B. Computer A fetches code and data from computer B and then executes the code with its local data as well as the imported data. An example of this is Java applets, which are Java codes that web-browsers download from remote HTTP servers to execute locally.

• The mobile agent approach assume that the code and data are initially hosted by computer A (Figure 3 d)). Computer A migrates the data and code it need to computer B. After it has moved to computer B, the code is executed with the data and the resources available on computer B.

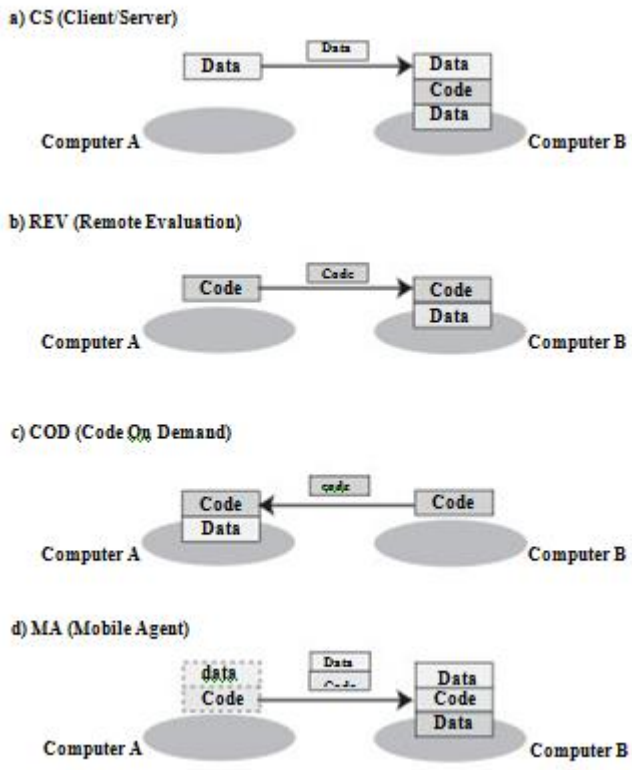


Figure 3: Client/server, remote evaluation, code on demand, and mobile agent

4.3 Mobile Agent Platform

Mobile agent platforms consist of two parts: mobile agents and runtime systems. The latter are called agent platforms, agent systems, and agent servers, and support their execution and migration. The same architecture exists on all computers at which agents are reachable. That is, each mobile agent runs within a runtime systems on its current computer. When an agent requests the current runtime system to migrate itself, the runtime system can migrate the agent to a runtime system on the destination computer, carrying its state and code with it. Each runtime system itself runs on top of the operating system as a middleware. It provides interpreters or virtual machines for executing agent programs, or the system themselves are provided on top of virtual machines, e.g., the Java virtual machine (JVM).

4.3.1 Remote procedure call

Agent migration is similar to RPC (Remote Procedure Calling) or RMI (Remote Method Invocation). RPC enables a client program to call a procedure for server programs running in separate processes, generally in different computers from the client [2]. RMI is an extension of local method invocation that allows an object to invoke the methods of the object on a remote computer. RPC or RMI can pass arguments to a procedure or method of a program on the server and receives a return value from the server. The mechanism for passing arguments and results between two computers through RPC or RMI correspond to that for agent migration between two computers. Figure 4 shows row for the basic mechanism of RPC between two computers.

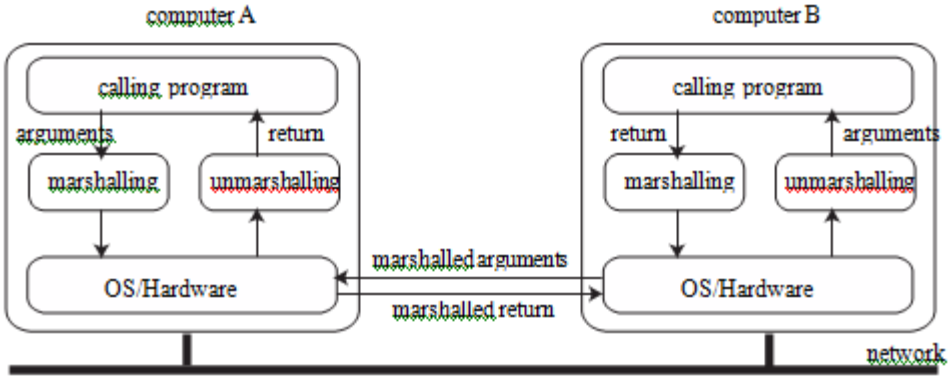


Figure 4: Remote procedure call between two computers

4.3.1.1 Agent marshalling

Data items, e.g., objects and values, in a running program cannot be directly transmitted over a network. They must be transformed into external data representation, e.g., a binary form or text form, before migrating them (Figure 5). Marshalling is the process of collecting data items and assembling them into a form suitable for transmission in a message. Unmarshalling is the process of disassembling them on arrival to produce an equivalent collection of data items at the destination.[1] The marshalling and unmarshalling processes are carried out by runtime systems in mobile agent systems. The runtime system at the left (at sender-side computer) of Figure 6 marshals an agent to transmit it to a destination through a communication channel or message and then the runtime system at the right (at receiver-side computer) of Figure 6 receives the data and unmarshals the agent.

4.3.1.2 Agent migration

Figure 6 shows the basic mechanism for agent migration between two computers.

Step.1 The runtime system on the sender-side computer suspends the execution of the agent.

Step.2 It marshals the agent into a bit-chunk that can be transmitted over a network.

Step.3 It transmits the chunk to the destination computer through the underlying network protocol.

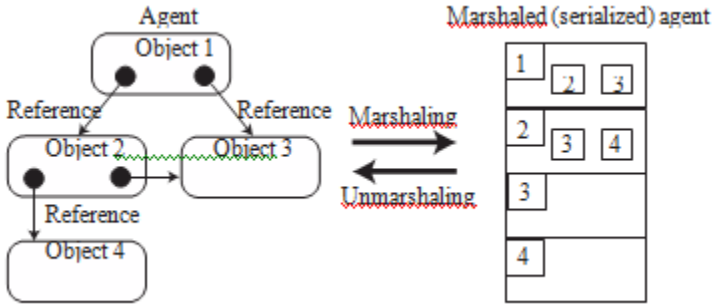


Figure 5: Marshaling agent

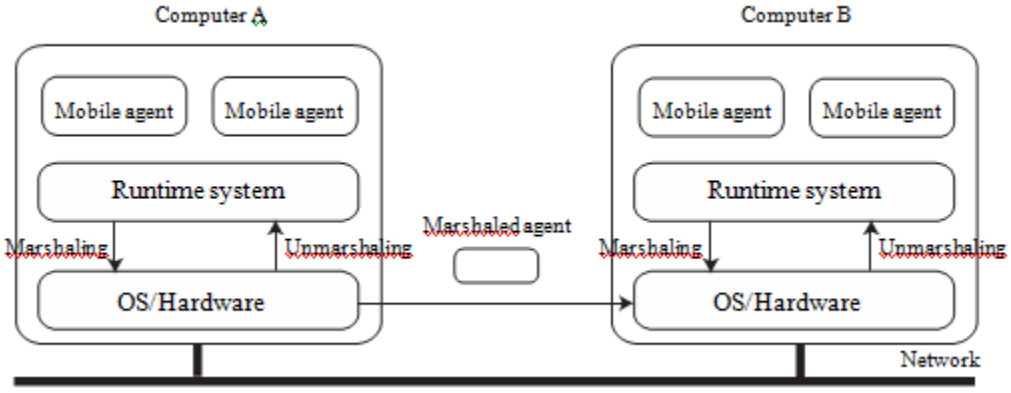


Figure 6: Agent migration between two computers

Step.4 The runtime system on the receiver-side computer receives the chunk.
Step.5 It unmarshals the chunk into the agent and resumes the agent.
Most existing mobile agent systems use TCP channels, SMTP, or HTTP as their underlying communication protocols. Mobile agents themselves are separated from the underlying communication protocols.

4.3.1..3 Strong migration vs. weak migration
The state of execution is migrated with the code so that computation can be resumed at the destination. According to the amount of detail captured in the state, we can classify agent migration into two types: strong and weak.
• Strong migration: is the ability of an agent to migrate over a network, carrying the code and execution state, where the state includes the program counter, saved processor registers, and local variables, which correspond to variables allocated in the stack frame of the agent's memory space, global variables. These correspond to variables allocated in the heap frame. The agent is suspended, marshaled, transmitted, unmarshaled and then restarted at the exact position where it was previously suspended on the destination node without loss of data or execution state.
• Weak migration: is the ability of an agent to migrate over a network, carrying the code and partial execution state, where the state is variables in the heap frame, e.g., instance variables in object oriented programs, instead of its program counter and local variables declared in methods or functions. The agent is moved to and restarted on the destination with its global variables. The runtime system may explicitly invoke specified agent methods.
Strong migration can cover weak migration, but it is a minority. This is because the execution state of an agent tends to be large and the marshaling and transmitting of the state over a network need heavy processing. Moreover, like the latter, the former cannot migrate agents that access the computational resources only available in current computers, e.g., input-and-output equipment and networks.
The program code for an agent needs to be available at the destination where the agent is running. The code must to be deployed at the source at the time of creation and at the destination to which it moves. Therefore, existing runtime systems offer a facility for statically deploying program code that is needed to execute the agent, for loading the program code on demand, or for transferring the program code along with the agent.

4..4 Mobile agent languages
Since mobile agents are programming entities, programming languages for destining mobile agents are needed. There has been a huge number of programming languages, but all of these are not available for mobile agents. Programming languages for mobile agents must support the following functions. They should enable programs to be marshaled into data and vice versa. They should also download code from remote computers and link it at run-time.

Table 1: Functions available in agents

command	parameters	function
go	destination address, agent-identi□er	agent migration
terminate	agent-identi□er	agent termination
duplicate	agent-identi□er	agent duplication
identify	agent-type	identi□cation
lookup	agent-type, runtime system address	discovery of available agents
communicate	agent-identi□er	inter-agent communication

A few researchers have provided newly designed languages for defining mobile agents, e.g., Telescript [24], and most current mobile agent systems use existing general-purpose programming languages that can satisfy the above requirements, e.g., Java [1]. Telescript provides primitives for de□ning mobile agents, e.g, go operation, and enables a thread running on an interpreter to migrate to another computer. The Java language itself offers no support for the migration of executing code, but offers dynamic class loading, a programmable class loader, and a language-level marshaling mechanism, where these can be directly exploited to enable code mobility. Creating distributed systems based on mobile agents is a relatively easy paradigm because most existing mobile agents are object oriented programs, e.g., Java, and can be developed by using rapid application development (RAD) environments.
Distributed systems are characterized by heterogeneity in hardware architectures and operating systems. To achieve heterogeneity, the state and code of an agent need to be saved in a platform-independent representation. Hidden differences between platforms is provided at the language level, by using intermediate byte code representation in Java or by relying on scripting languages, such as Python and Ruby. Therefore, Java-based mobile agents are executed on Java virtual machines. The costs of running agents in a Java virtual machine on a device are decreasing by using just-in-compiler technologies.

4.5 Agent execution management

The runtime system manages execution and monitoring of all agents on a computer. It allows several hundred agents to be present at any one time on a computer. It also provide these agents with an execution environment and executes them independently of one another. It manages the life-cycle of its agents, e.g., creation, termination, and migration.

Each agent program can access basic functions provided by its runtime system by invoking APIs (Table 1). The agent uses the go command to migrate from one computer to another with the destination system address (and its target agents identifier) and does not need to concern itself with any other details of migration. Instead, the runtime system supports the migration of the agent. It stops the agent's execution and then marshals the agent's data items to the destination via the underlying communication protocol, e.g., TCP channel, HTTP (hyper text transfer protocol), and SMTP (simple mail transfer protocol). The agent is unpacked and reconstituted on the destination.

4.6 Inter-agent communication

Mobile agents can interact with other agents residing within the same computer or with agents on remote computers as other multi-agents. Existing mobile agent systems provide various inter-agent communication mechanisms, e.g., method invocation, publish/subscribe-based event passing, and stream-based communications.

4.7 Locating mobile agents

Since mobile agents can autonomously travel from computer to computer, a mechanism for tracking the location of agents is needed by the users to control their agents and for agents to communicate with other agents. Several mobile agent systems provide such mechanisms, which can be classified into three schemes:

• A name server multicasts query messages about the location of an agent the to computers and receives a reply message from a computer hosting the agent (Figure 7 (a)).

• An agent registers its current location at a predefined name server whenever it arrives at another computer (Figure 7 (b)).

• An agent leaves a footprint specifying its destination at its current computer whenever it migrates to another computer to track the trails of the agent (Figure 7 (c)).

In many cases, locating agents is application specific. For example, the first scheme is suitable for an agent moving within a local region. It is not suitable for agents visiting distant nodes. The second scheme is suitable for an agent migrating within a far away region; in the case of a large number of nodes, registering nodes are organized hierarchically. However, it is not suitable for a large number of migrations. The third scheme is suitable for a small number of migrations; it is not appropriate for long chains.

4.4 Transaction Processing in Mobile Computing Environment

This topic focuses on the main topic of this thesis: mobile transaction processing systems. The main objective of this chapter is to identify a set of requirements that must be fulfilled by a mobile transaction processing system in order to efficiently support transaction processing in mobile environments.

4.4.1 Introduction

Unlike distributed environments, transaction processing in mobile environments must take into account three new challenging characteristics of mobile environment - that are: the mobility of mobile computing hosts, the limitation of wireless communications and the resource constraints of mobile computing devices [PS98]. These three challenging characteristics have a strong impact on the processing of transactions in terms of concurrency control, data availability, and recovery strategies [Mad+02]. Because of these unique characteristics of the mobile environments, the standard transaction ACID properties can be too strict to be applied in mobile environments. In other words, we need to define a set of requirements that broadens these properties in the context of the mobile environments.

4.4.2 Characteristics of mobile environments

In this section, we discuss the characteristics of the mobile environments that could have strong impact on mobile transactions in terms of transaction specification and transaction processing. There are other important issues like authentication and security; however, they are not in the scope of this thesis. The main characteristics of the mobile environments that are addressed in this section include: the mobility of mobile computing hosts, the limitation of wireless communications and the resource constraints of mobile computing devices. In this chapter, we will use the *mobile transaction* terminology for specifying transactions in mobile environments.

4.4.2.1 Mobile hosts

Mobility is the main characteristic that distinguishes the mobile environments from the traditional distributed environments. In traditional distributed environments, computers are stationary hosts. In mobile environments, mobile computers are continuously moving from one geographical location to another.

The features of the mobility characteristic are discussed as follows:

• **Real-time movement.** The mobility of the mobile host is a real-time movement. Therefore, it is affected by many environment conditions. For example, the pre-planned travel route of a mobile host can be changed because of traffic jams or weather conditions. If there is a mobile task whose operations depend on the travel route of the mobile host, these operations can become invalid, or extra support is required. For example, a new route-map directory must be downloaded into the mobile host if the travel course is changed. Moreover, the movement of the mobile host can also depend on the objective of the mobile task. For example, an ambulance car wants to arrive at the accident scene by selecting the shortest route with fastest allowing speed, a bus must follow a strict time table on a bus-route, while a postman only wants to travel through each road once. During the movement, the mobile host can stop at some locations for some periods; therefore, the mobility of the mobile host includes both movement and non-movement intervals.

• **Change of locations**. The location of a mobile host changes dynamically and frequently in accordance with the speed and the direction of the movement. The faster the mobile host moves, the more frequently the location changes. The objective of mobile tasks can also specify the locations at which the mobile host must be, in order to carry out the mobile tasks. For example, a computer technician must come to customer locations to fix computer problems. A mobile support system must provide the utilities to manage the locations of mobile hosts (this demand is not needed in a distributed environment). Changes of locations can cause changes in the operating environments of the mobile hosts, for example network addresses, communication protocols, mobile services, or location dependent data [Ram+03, DK98].

The mobility of mobile hosts will have strong impact on the execution of transactions. The real-time movement of mobile hosts could pose timing constraints on the execution schedule of transactions. Furthermore, if mobile hosts change their locations frequently, additional time is required to reconfigure transaction application processes to the new environment conditions. Therefore, additional support is required for mobile transaction processing systems to cope with these challenges.

4.4.2.2 Wireless networks

Mobile hosts communicate to other hosts via wireless networks. Compared to wired networks, wireless networks are characterized by: lower bandwidth, unstable, disconnections, and ad-hoc connectivity [Sch02]. The characteristics of the wireless networks are described as follows:

• **Lower bandwidth.** The bandwidth of a wireless network is lower than a wired network. The wireless network does not have the capacity as the wired network. For example, a wireless network has bandwidth in the order of 10Kbps or a wireless local area network (WLAN) has bandwidth of 10Mbps; while gigabits (Gbps) are common in wired LAN [Sch02]. Therefore, it can take longer time for a mobile host to transfer the same amount of information via the wireless network than the wired network. Consequently, the wireless network introduces more overhead in transaction processing.

• **Unstable networks.** A wireless network has high error-rates, and the bandwidth of a wireless network is variable. Due to errors during data transmission, the same data packages are required to re-transmit many times, thus, extra overhead in communication and higher cost. Due to the varying bandwidth, it is hard to estimate the time required to completely transmit a data package from/to a mobile host. These problems will affect the data availability at the mobile hosts. As a result, the execution schedule of transactions at the mobile hosts can be delayed or aborted.

• **Disconnections.** Wireless networks pose disconnection problems. Disconnections in communication can interrupt or delay the execution processes of transactions. More seriously, on-going transactions could be aborted due to a disconnection. The disconnection in communication is categorized into two types: disconnection period and disconnection rate.

Disconnection period. The disconnection period indicates how long a mobile host is disconnected. While being disconnected, the mobile host will not be able to communicate to other hosts for sharing of data. If the mobile host holds vital shared data, it can block transaction processes on other hosts. Furthermore, the duration of a disconnected period of a mobile host is not always as planned, i.e., it can be longer than expected. The mobile transaction processing system must be able to continuously support transaction processing while the mobile host is being disconnected from the database servers by caching the needed data beforehand.

Disconnection rate. The disconnection rate indicates how often the wireless communication is interrupted within a predefined unit of time. The execution of transactions on a mobile host can be affected when an interruption occurs.

The more interruptions the many transactions are aborted or rollback. If the transactions on the mobile host are carrying out collaborative operations with other transactions on other mobile hosts, these collaborative activities can be suspended or aborted. To cope with this problem, the mobile transaction processing system must be able to support the mobile transactions to resume or recover from previous interrupted points.

- **Ad-hoc communication.** The wireless network technologies introduce a new way to support direct and nearby communications among mobile hosts, also called *any-to-any* or *mobile peer-to-peer* communication [Sch02, Rat+01]. For example, two mobile hosts can directly share information with the support of Bluetooth or infra-red technologies [PLZ05]. The characteristics of this peer-to-peer communication are: unstructured (i.e., ad-hoc), short-range, and mobility dependent [Rat+01]. Table 3.1 compares the communication ranges and bandwidth of different wireless technologies.

Table 3.1: Wireless communication technologies

Wireless technology	IEEE standard	Range (m)	Bandwidth
IrDA[2]	N/A	0.1-1	100kbps - 16Mbps
Bluetooth	IEEE 802.15.1	10-100	1Mbps
Wireless USB	IEEE 802.15.3[3]	1-10	2Mbps-480Mbps
Wi-Fi	IEEE 802.11	45-90	11Mbps-540Mbps
WiMAX	IEEE 802.16	2km-10km	75Mbps

4.4.2.3 Computing devices

There are many types of mobile computing devices such as mobile phones, laptop computers, or personal digital assistants (PDAs). Mobile devices are subject to be smaller and lighter than stationary computers. Consequently, mobile computers have limited energy supply, less storage capacity, and limited functionality compared to stationary computers. Furthermore, the mobile computers are easily damaged, i.e., less reliable. The characteristics of mobile computing devices are elaborated as follows:

- **Limited energy supply**. The operation of mobile computers heavily depends on the electrical power of batteries. This limited power supply is one of the major disadvantages of mobile computing devices. The energy consumption of a mobile device depends on the power of electronic equipments installed on the mobile device, for example types of hard disks or CPU. Moreover, the battery life also depends on the number of applications and the application types that operate on the mobile devices [FS99, KU99]. For example, a mobile phone can live up to five days but a laptop can only be able to operate for several hours; text processing applications consume less power than graphical applications. Transaction processes that are being carried out at a mobile host can be interrupted or re-scheduled if the mobile host is exhausting its power supply.
- **Limited storage capacity.** The storage capacity of a mobile computer (i.e., hard disks or memory) is much less than a stationary computer and is harder to be expanded. Therefore, a mobile host may not be able to store the necessary data that is required for its operations in disconnected mode [PS98, Mad+02]. Consequently, transaction processes on the mobile host will be delayed due to data unavailability, or require longer processing time due to frequent memory swapping operations.
- **Limited functionality.** The functionality of mobile devices is also limited in terms of the graphical user interface, the application functionalities, and the processing power. Therefore, a mobile host may be unable to perform some of transaction operations, or requires longer processing time to perform these operations. For example, a small PDA may only be able to view black and white pictures. Table 3.2 compares the configurations of several PDA types.

Table 3.2: Personal digital assistant device

PDA type	Size and weight (cm, gram)	Screen size (inch, color bits)	Processor type (MHz)
HP iPAQ Pocket PC hx2110	7.7 x 1.6 x 11.9, 164 g	3.5”, 16 bits	Intel XScale 312
ASUS MyPal A620BT	7.7 x 1.3 x 12.5, 141 g	3.5”, 16 bits	Intel XScale 400
Fujitsu Siemens Pocket LOOX 720	7.2 x 1.5 x 12.2, 170 g	3.6”, 16 bits	Intel XScale 520

- **Unreliable equipments**. The data stored at a mobile host can be lost if a catastrophic failure happens. This could heavily impact the durability property of transactions because of the losing of the committed results of transactions that

are stored at the mobile host. To avoid this problem, data stored at mobile hosts must be backed-up at stationary database servers as much and as soon as possible.

4.4.2.4 The behavior of mobile hosts in mobile environments

In mobile environments, mobile transactions are initiated [DHB97, KK00] and/or processed [WC99] at mobile hosts. The mobile hosts can participate in the mobile transaction execution processes in different ways. First, a mobile host can initiate a mobile transaction, submits the transaction to appropriate (non-mobile or mobile) hosts for processing, and receives the committed results. In this way, the mobile host plays arole as a terminal transaction client [GR93]. Second, a mobile host can take part in the actual transaction execution process, i.e., the entire or part of a mobile transaction is carried out by the mobile host. The mobile host plays a vital role in the transaction execution process. Therefore, we need to answer the following question: How do the characteristics of the mobile environments affect the behavior of the mobile host? The behavior of mobile hosts in mobile environments is categorized into two dimensions: *movement* and *operation* (see Figure 3.1).

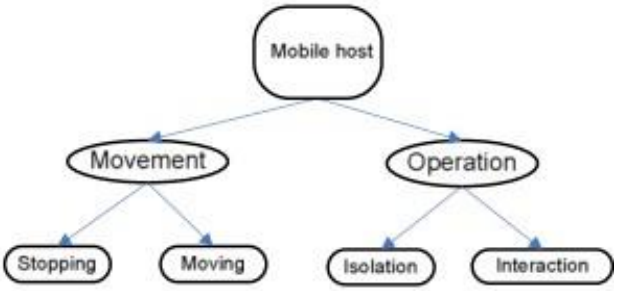


Figure 3.1: Behavior model for mobile hosts

First, the movement of the mobile host is affected by both the requirements of the mobile tasks and the environmental conditions [DK99, Sør+02]. Second, the operation of the mobile host depends on its internally designed capacity and externally associative factors. For example, the performance of computational operations depends on the available energy of the mobile host's battery, and the network operations rely on both the connectivity capacity of the mobile host and the provided network services. The behavior of mobile hosts is discussed in the following.

Movement of mobile hosts

The movement behavior of a mobile host describes the actual mobility states of the mobile host. While operating in mobile environments, the mobile host can be either in *stopping* or *moving* state. The two movement states are explained as follows:

- **Stopping.** A mobile host is said to be in stopping state either when its movement velocity is zero, or when the location of the mobile host is not considered changing within a period of time. For example, a bus stops at a bus-stop to pick up passengers, a salesman is selling products at a shopping centre, or two mobile hosts are always moving close to each other.

- **Moving**. A mobile host is in moving state either when its movement velocity has a value greater than zero, or when the location of the mobile host is considered changing over time. For example, a bus is moving along a road or a salesperson travels to several places during the day. While in moving state, the mobile host can continuously change its velocity and direction of movement.

4.4.3 Transaction processing in mobile environments

The main differences between the mobile environments and distributed environments are:
(1) Mobile computing hosts,
(2) Wireless networks.
Table compares the main different features between the distributed and mobile environments.

Distributed environments versus mobile environments

	Distributed environments	Mobile environments
Computing Hosts	Stationary sites Powerful computing capacity Reliable computing hosts	Mobile and non-mobile hosts Limited computing capacity of mobile hosts Less reliable computing hosts
Network connectivity	Wired and high-speed networks Reliable networks	Wireless, unstable and low speed networks. Unreliable, error-prone, frequent and long disconnection periods

The mobile hosts usually have less computing resources and capacity than stationary hosts. For example, a laptop computer has lower processing speed and smaller storage capacity than a desktop computer, and its operation might depend on the limited battery energy. Consequently, it takes longer time for a transaction to be processed at a mobile host. Moreover, mobile computers are easily damaged, i.e., less reliable. The results of committed transactions, which are stored at a mobile computer, can be lost if the mobile computer is damaged, i.e., the durability property of transactions may not be fully guaranteed. Therefore, the committed results of transactions in mobile environments should additionally be saved at the stationary hosts as in distributed environments. The movement of mobile hosts brings additional requirements and demands that the mobile transaction processing system must handle, for example hand-over processes [DHB97] or locally dependent data [DK99]. In distributed environments, these demands do not exist.

Mobile computing hosts communicate with other hosts via wireless networks. Compared to a wired network, a wireless network is usually less reliable, i.e., disconnections can occur frequently; has lower bandwidth, i.e., megabits versus gigabits; and is limited in communication range, i.e., mobile hosts must stay within limited distance to be connected. Because of these unique features of wireless networks, it can take longer time to download necessary data into the local storage devices at the mobile hosts; or due to disconnections, the mobile hosts will not be able to obtain the needed data. Consequently, transactions in mobile environments may experience long blocking periods and inconsistent data.

In mobile environments, transaction processing systems consist of both mobile and non-mobile hosts [SRA04], and can be divided into two different layers (see Figure 3.2). The distributed transaction processing layer corresponds to the execution of mobile transactions that are carried out on non-mobile hosts. The mobile transaction processing layer corresponds to the execution of mobile transactions that are carried out on a mobile host or distributed among mobile hosts. Due to the above distinguishing and challenging characteristics of mobile environments, transaction processing in mobile environments is more difficult than in distributed environments, especially in terms of concurrency control, data availability, and recovery mechanisms [Mur01].

4.4.4 Architecture of mobile transaction environments

In this section, we discuss the architecture of the mobile transaction environments. In general, the mobile transaction environments include three different components: mobile hosts (MH), mobile support stations (MSS) and fixed hosts where database servers (DB) reside [SRA04, Hir+01]. Figure 3.3 illustrates the mobile transaction environments.
A mobile environment is a geographical territory. The geographical territory is divided into a collection of areas called mobile cells. Wireless communications in each mobile cell is provided by a single low-power transmitter-receiver [Sch02]. There might be some areas in the mobile environments in which the wireless communication is not available. This could be caused by the limited service of the wireless communication providers or the structural of physical objects in the areas, for example concrete tunnels or remote islands. The geographical mobile environment, therefore, can be considered as a collection of mobile cells that are separated or overlapped with others. The size of mobile cells is not necessarily equal, due to the differences of operational power of the transmitter-receiver devices.

The wireless technologies that are provided in each mobile cell can be different, for example wireless LAN or wireless USB. As a consequence, network bandwidth, network latency, communication protocols and covered ranges are different among mobile cells. In each mobile cell, there is a special computing host called the mobile support station. The role of the mobile support station is to provide additional computing services to all the mobile hosts that currently reside in the mobile cell.

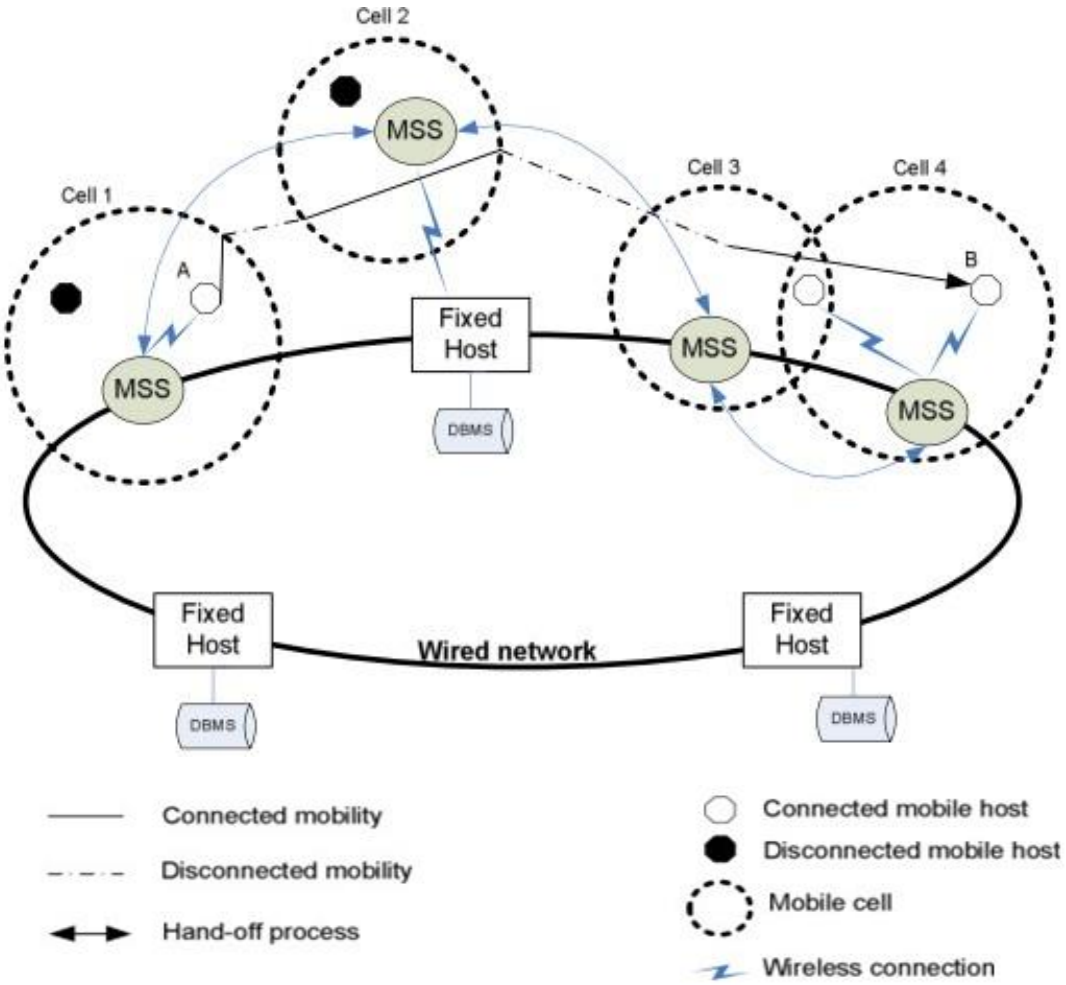


Figure 3.3: Mobile transaction environments

Mobile hosts are portable mobile computing devices which have the capability to cache a limited amount of information. Database servers are stationary computers that are connected via high speed wired-networks, and play roles as permanent data storage repositories. Shared data is distributed on these database servers. Mobile support stations (also called base stations) are stationary or mobile computers. Mobile support stations have higher processing power and data storage capacity than the mobile hosts. The role of the mobile support stations is to support mobile hosts communicating with other mobile hosts or database servers. Mobile hosts communicate with the mobile support stations via the wireless networks. Communications between the database servers and the mobile support stations are via wired networks or dedicated wireless connections.

Mobile hosts move in mobile environments while carry out mobile tasks. While being in a mobile cell, a mobile host can be either connected or disconnected with the mobile support station of this mobile cell. The mobile host may only connect to the mobile support station when there is a need for sharing of data. This will help to save the limited energy of the mobile host and to reduce the communication cost. On the other hand, because of the limitations of wireless networks, a mobile host may not always be able to establish a communication channel with the mobile support station. If a mobile host is in the area that is an intersection of two or more mobile cells, it can connect to any mobile support station.

The mobile hosts can move within one mobile cell or across a large area covered by several mobile cells. When a mobile host is leaving a mobile cell and entering a new mobile cell, the communication channel and other related information between the mobile host and the previous mobile support station will be transferred to the next mobile support station. This process is called *hand-over* or *hand-off* process [SRA04]. The new mobile support station at the new mobile cell will continue carrying out the support to the mobile host. However, it is not necessary that hand-

over processes must happen every time the mobile host enters a new mobile cell. For example, the mobile host can operate in an autonomous mode when the wireless network is not supported in the new mobile cell. Furthermore, a mobile host does not have to disconnect from the old mobile support station before it can connect to the new mobile support station. As shown in [CP98, TLP99], a mobile host can connect to a new mobile support station while connecting to the old mobile support station. The hand-off process can be planned beforehand if the travel route of the mobile host is known in advanced and strictly followed. Otherwise, the hand-off process can only be carried out after the mobile host has established a connection with the new mobile support station, i.e., after the new destination of the mobile host is known.

In Figure 3.3, there are four mobile cells in the mobile environments. Mobile cells one and two are separated, while mobile cells three and four are overlapped. A mobile host moves from position *A* in mobile cell one to position *B* in mobile cell four. The travel route of the mobile host passes through mobile cells two and three. When the mobile host is leaving cell one, it will enter a disconnected interval in the area between the mobile cells one and two. While in the mobile cell two, the mobile host will be supported by the mobile support station that is a dedicated mobile host. When the mobile host is in the mobile cell three, it may not connect to the mobile support station all the time. In the intersection region of the mobile cells three and four, the mobile host can connect to the mobile support station of either mobile cell three or mobile cell four. The hand-over processes occur when the mobile host moves from one mobile cell to another along the travel route.

4.4.5 Characteristics of mobile transactions

Transactions in mobile environments possess many challenging characteristics due to the characteristics of the mobile environments. In this section, we will discuss the characteristics of mobile transactions. The characteristics of mobile transactions are described as follows:

- **Mobility of transactions.** The execution of transactions in mobile environments is tightly coupled with the behavior of the mobile hosts. A mobile host can initiate mobile transactions or participate in the transaction execution processes. When a mobile host moves from one location to another, all the transactions that are being carried out at that mobile host will also move. Consequently, many computing activities associated with these transactions are moved or changed, for example handling hand-over processes, establishing new communication channels, or updating the routing tables. In other words, the mobility of transactions causes the movement of related transaction resources, controls, and services.
- **Long-lived transactions.** Transactions in mobile environments have longer life (i.e., long-lived) than traditional ACID transactions. This is due to the overheads that are caused by two aspects: the data availability and the execution interruptions.

Data availability. In mobile environments, the data availability at a mobile host can be affected by many factors. First, the movement of the mobile host causes the movement of related information. This will cause additional overhead to the transaction execution time. Second, the bandwidth of wireless networks is limited; therefore it will take longer time to obtain the necessary data. Third, the mobile computing devices have limitations in storage capacity; therefore, the mobile host may not able to cache the required information to support disconnected transaction processing. In addition, due to the unexpected disconnections of the wireless networks, a transaction will not be able to release the controls on shared data to transactions at other hosts as scheduled; this means that this transaction blocks the execution of other transactions.

***Execution interruptions*.** The execution of transactions can be interrupted while being carried out at the mobile host. The interruptions can be caused by either the surrounding environment conditions or the limitation of computing capacity of the mobile host. For example, a wireless network disconnection suddenly occurs during the execution of transactions, or the performance of the mobile host is slowing down due to heavy computing load. The interruptions can happen frequently and cause transactions to be suspended or aborted.

- **Adaptive transaction processing**. Due to the real-time movement of the mobile hosts, the limitations of the wireless networks, and the variation of the mobile resources, the execution plan of a transaction in mobile environments may not be as scheduled. Therefore, the mobile transaction processing system must have the ability to support adaptive transaction processing that includes: distributed and disconnected transaction processing.

Distributed transaction processing. Due to the limitations of processing capacity and resources, mobile hosts require additional support from other hosts to carry out transactions. For example, a transaction, which is initiated by a mobile host and accesses a large data set that is not cached at the mobile host, could be moved to stationary hosts for executing. This could reduce transaction processing time and avoid transferring a large amount of data through a slow wireless network, i.e., achieving higher throughput for the transaction processing system. Furthermore, the portable computing devices are easily damaged; therefore, the results of committed transactions must be saved at stationary database servers.

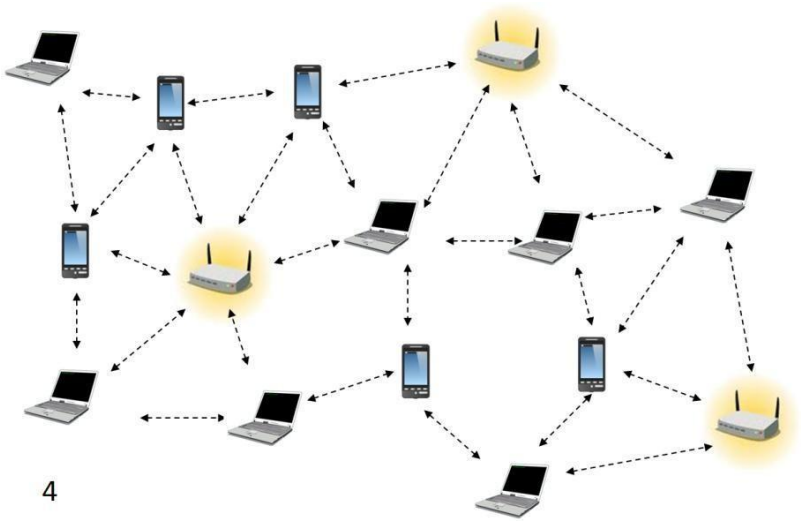
Disconnected transaction processing. A mobile host can be disconnected from the database servers for long periods; therefore, transactions that are executed at the mobile host may suffer from long blocking if the necessary data

is not available at the mobile host. To deal with this problem, the mobile transaction processing system should have the capacity to cache enough data so that it can carry out the transactions while being disconnected from the database servers.

- **Temporary data inconsistency**. Due to long disconnection periods, shared data among different mobile hosts may not be fully consistent all the time. For example, a transaction at a disconnected mobile host can modify a shared data item that is currently being read-only cached in a local storage of another disconnected mobile host. Data synchronization processes will be carried out when the disconnected mobile hosts reconnect to the database systems so that the data consistency of the database systems will be achieved.
- **Heterogeneous processing**. Many types of mobile devices can be involved in transaction execution processes. Interactions or communications among participating parties are carried out via the support of different types of wireless network technologies and protocols. Furthermore, different database systems are accessed during the execution of mobile transactions. All these factors contribute to the heterogeneous processing characteristic of mobile transactions.

Mobile Ad hoc Networks (MANETs): Overview, Properties of a MANET, spectrum of MANET, applications, routing and various routing algorithms, security inMANET's.

Mobile Ad hoc NETworks (MANETs) are wireless networks which are characterized by dynamictopologies and no fixed infrastructure. Each node in a MANET is a computer that may be required to act as both a host and a router and, as much, may be required to forward packets between nodes which cannot directly communicate with one another. Each MANET node has much smaller frequency spectrum requirements that that for a node in a fixed infrastructure network. A MANET is an autonomous collection of mobile users that communicate over relatively bandwidth constrained wireless links. Since the nodes are mobile, the network topology may change rapidly and unpredictably over time. The network is decentralized, where all network activity including discovering the topology and delivering messages must be executed by the nodes themselves, i.e., routing functionality will be incorporated into mobile nodes.



A mobile ad hoc network is a collection of wireless nodes that can dynamically be set up anywhere and anytime without using any pre-existing fixed network infrastructure.

MANET- Characteristics

- Dynamic network topology
- Bandwidth constraints and variable link capacity
- Energy constrained nodes
- Multi-hop communications
- Limited security
- Autonomous terminal
- Distributed operation
- Light-weight terminals

Need for Ad Hoc Networks

- ❖ Setting up of fixed access points and backbone infrastructure is not always viable

– Infrastructure may not be present in a disaster area or war zone
– Infrastructure may not be practical for short-range radios; Bluetooth (range ~ 10m)

- ❖ Ad hoc networks:

– Do not need backbone infrastructure support
– Are easy to deploy
– Useful when infrastructure is absent, destroyed or impractical

Properties of MANETs

- MANET enables fast establishment of networks. When anew network is to be established, the only requirement is to provide a new set of nodes with limited wireless communication range. A node has limited capability, that is, it can connect only to the nodes which are nearby. Hence it consumes limited power.
- A MANET node has the ability to discover a neighboring node and service. Using a service discovery protocol, a node discovers the service of a nearby node and communicates to a remote node in the MANET.
- MANET nodes have peer-to-peer connectivity among themselves.
- MANET nodes have independent computational, switching (or routing), and communication capabilities.
- The wireless connectivity range in MANETs includes only nearest node connectivity.
- The failure of an intermediate node results in greater latency in communicating with the remote server.
- Limited bandwidth available between two intermediate nodes becomes a constraint for the MANET. The node may have limited power and thus computations need to be energy-efficient.
- There is no access-point requirement in MANET. Only selected access points are provided for connection to other networks or other MANETs.
- MANET nodes can be the iPods, Palm handheld computers, Smartphones, PCs, smart labels, smart sensors, and automobile-embedded systems\
- MANET nodes can use different protocols, for example, IrDA, Bluetooth, ZigBee, 802.11, GSM, and TCP/IP.MANET node performs data caching, saving, and aggregation.
- MANET mobile device nodes interact seamlessly when they move with the nearby wireless nodes, sensor nodes, and embedded devices in automobiles so that the seamless connectivity is maintained between the devices.

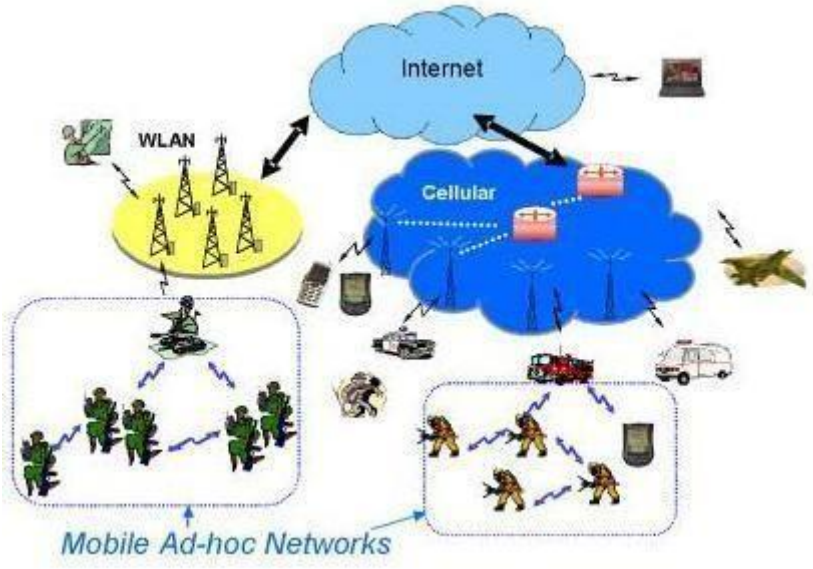
MANET challenges

To design a good wireless ad hoc network, various challenges have to be taken into account:

- Dynamic Topology: Nodes are free to move in an arbitrary fashion resulting in the topology changing arbitrarily. This characteristic demands dynamic configuration of the network.
- Limited security: Wireless networks are vulnerable to attack. Mobile ad hoc networks are more vulnerable as by design any node should be able to join or leave the network at any time. This requires flexibility and higher openness.
- Limited Bandwidth: Wireless networks in general are bandwidth limited. In an ad hoc network, it is all the more so because there is no backbone to handle or multiplex higher bandwidth
- Routing: Routing in a mobile ad hoc network is complex. This depends on many factors, including finding the routing path, selection of routers, topology, protocol etc.

Applications of MANETS

The set of applications for MANETs is diverse, ranging from small, static networks that are constrained by power sources, to large-scale, mobile, highly dynamic networks. The design of network protocols for these networks is a complex issue. Regardless of the application, MANETs need efficient distributed algorithms to determine network organization, link scheduling, and routing. Some of the main application areas of MANET's are:



- **Military battlefield**– soldiers, tanks, planes. Ad- hoc networking would allow the military to take advantage of commonplace network technology to maintain an information network between the soldiers, vehicles, and military information headquarters.

- **Sensor networks** – to monitor environmental conditions over a large area
- **Local level** – Ad hoc networks can autonomously link an instant and temporary multimedia network using notebook computers or palmtop computers to spread and share information among participants at e.g. conference or classroom. Another appropriate local level application might be in home networks where devices can communicate directly to exchange information.
- **Personal Area Network (PAN)** – pervasive computing i.e. to provide flexible connectivity between personal electronic devices or home appliances. Short-range MANET can simplify the intercommunication between various mobile devices (such as a PDA, a laptop, and a cellular phone). Tedious wired cables are replaced with wireless connections. Such an ad hoc network can also extend the access to the Internet or other networks by mechanisms e.g. Wireless LAN (WLAN), GPRS, and UMTS.
- **Vehicular Ad hoc Networks** – intelligent transportation i.e. to enable real time vehicle monitoring and adaptive traffic control
- **Civilian environments** – taxi cab network, meeting rooms, sports stadiums, boats, small aircraft
- **Emergency operations** – search and rescue, policing and fire fighting and to provide connectivity between distant devices where the network infrastructure is unavailable. Ad hoc can be used in emergency/rescue operations for disaster relief efforts, e.g. in fire, flood, or earthquake. Emergency rescue operations must take place where non- existing or damaged communications infrastructure and rapid deployment of a communication network is needed. Information is relayed from one rescue team member to another over a small hand held.

Routing in MANET's

Routing in Mobile Ad hoc networks is an important issue as these networks do not have fixed infrastructure and routing requires distributed and cooperative actions from all nodes in the network. MANET's provide point to point routing similar to Internet routing. The major difference between routing in MANET and regular internet is the route discovery mechanism. Internet routing protocols such as RIP or OSPF have relatively long converge times, which is acceptable for a wired network that has infrequent topology changes. However, a MANET has a rapid topology changes due to node mobility making the traditional internet routing protocols inappropriate. MANET-specific routing protocols have been proposed, that handle topology changes well, but they have large control overhead and are not scalable for large networks. Another major difference in the routing is the network address. In internet routing, the network address (IP address) is hierarchical containing a network ID and a computer ID on that network. In contrast, for most MANET's the network address is simply an ID of the node in the network and is not hierarchical. The routing protocol must use the entire address to decide the next hop.

Some of the fundamental differences between wired networks & ad-hocnetworks are:

- Asymmetric links: - Routing information collected for one direction is of no use for the other direction. Many routing algorithms for wired networks rely on a symmetric scenario.
- Redundant links: - In wired networks, some redundancy is present to survive link failures and this redundancy is controlled by a network administrator. In ad-hoc networks, nobody controls redundancy resulting in many redundant links up to the extreme of a complete meshed topology.
- Interference: - In wired networks, links exist only where a wire exists, and connections are planned by network administrators. But, in ad-hoc networks links come and go depending on transmission characteristics, one transmission might interfere with another and nodes might overhear the transmission of other nodes.
- Dynamic topology: - The mobile nodes might move in an arbitrary manner or medium characteristics might change. This result in frequent changes in topology, so snapshots are valid only for a very short period of time. So, in ad-hoc networks, routing tables must somehow reflect these frequent changes in topology and routing algorithms have to be adopted.

Summary of the difficulties faced for routing in ad-hoc networks

- Traditional routing algorithms known from wired networks will not work efficiently or fail completely. These algorithms have not been designed with a highly dynamic topology, asymmetric links, or interference in mind.
- Routing in wireless ad-hoc networks cannot rely on layer three knowledge alone. Information from lower layers concerning connectivity or interference can help routing algorithms to find a good path.
- Centralized approaches will not really work, because it takes too long to collect the current status and disseminate it again. Within this time the topology has already changed.
- Many nodes need routing capabilities. While there might be some without, at least one router has to be within the range of each node. Algorithms have to consider the limited battery power of these nodes.
- The notion of a connection with certain characteristics cannot work properly. Ad-hoc networks will be connectionless, because it is not possible to maintain a connection in a fast changing environment and to forward data following this connection. Nodes have to make local decisions for forwarding and send packets roughly toward the final destination.
- A last alternative to forward a packet across an unknown topology is flooding. This approach always works if the load is low, but it is very inefficient. A hop counter is needed in each packet to avoid looping, and the diameter of the ad-hoc network.

Types of MANET Routing Algorithms:

1. Based on the information used to build routing tables :
   - Shortest distance algorithms: algorithms that use distance information to build routing tables.
   - Link state algorithms: algorithms that use connectivity information to build a topology graph that is used to build routing tables.

2. Based on when routing tables are built:
   - Proactive algorithms: maintain routes to destinations even if they are not needed. Someof the examples are Destination Sequenced Distance Vector (DSDV), Wireless Routing Algorithm (WRP), Global State Routing (GSR), Source-tree Adaptive Routing (STAR), Cluster-Head Gateway Switch Routing (CGSR), Topology Broadcast Reverse Path Forwarding (TBRPF), Optimized Link State Routing (OLSR) etc.
     - ❖ Always maintain routes:- Little or no delay for route determination
     - ❖ Consume bandwidth to keep routes up-to-date
     - ❖ Maintain routes which may never be used
     - ❖ Advantages: low route latency, State information, QoS guarantee related to connection set-up or other real-time requirements
     - ❖ Disadvantages: high overhead (periodic updates) and route repair depends on update frequency

   - Reactive algorithms: maintain routes to destinations only when they are needed. Examples are Dynamic Source Routing (DSR), Ad hoc-On demand distance Vector (AODV), Temporally ordered Routing Algorithm (TORA), Associativity-Based Routing (ABR) etc
     - ❖ only obtain route information when needed
     - ❖ Advantages: no overhead from periodic update, scalability as long as there is only light traffic and low mobility.
     - ❖ Disadvantages: high route latency, route caching can reduce latency

   - Hybrid algorithms: maintain routes to nearby nodes even if they are not needed and maintain routes to far away nodes only when needed. Example is Zone Routing Protocol (ZRP).

Which approach achieves a better trade-off depends on the traffic and mobility patterns.

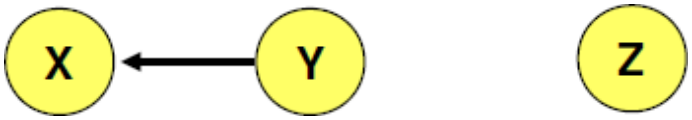
Destination sequence distance vector (DSDV)

Destination sequence distance vector (DSDV) routing is an example of proactive algorithms and an enhancement to distance vector routing for ad-hoc networks. Distance vector routing is used as routing information protocol (RIP) in wired networks. It performs extremely poorly with certain network changes due to the count-to-infinity problem. Each node exchanges its neighbor table periodically with its neighbors. Changes at one node in the network propagate slowly through the network. The strategies to avoid this problem which are used in fixed networks do not help in the case of wireless ad-hoc networks, due to the rapidly changing topology. This might create loops or unreachable regions within the network.

DSDV adds the concept of sequence numbers to the distance vector algorithm. Each routing advertisement comes with a sequence number. Within ad-hoc networks, advertisements may propagate along many paths. Sequence numbers help to apply the advertisements in correct order. This avoids the loops that are likely with the unchanged distance vector algorithm.

Each node maintains a routing table which stores next hop, cost metric towards each destination and a sequence number that is created by the destination itself. Each node periodically forwards routing table to neighbors. Each node increments and appends its sequence number when sending its local routing table. Each route is tagged with a sequence number; routes with greater sequence numbers are preferred. Each node advertises a monotonically increasing even sequence number for itself. When a node decides that a route is broken, it increments the sequence number of the route and advertises it with infinite metric. Destination advertises new sequence number.

When X receives information from Y about a route to Z,



- Let destination sequence number for Z at X be S(X), S(Y) is sent from Y
- If S(X) > S(Y), then X ignores the routing information received from Y
- If S(X) = S(Y), and cost of going through Y is smaller than the route known to X, then X sets Y as the next hop to Z
- If S(X) < S(Y), then X sets Y as the next hop to Z, and S(X) is updated to equal S(Y)

Besides being loop-free at all times, DSDV has low memory requirements and a quick convergence via triggered updates. Disadvantages of DSDV are, large routing overhead, usage of only bidirectional links and suffers from count to infinity problem.

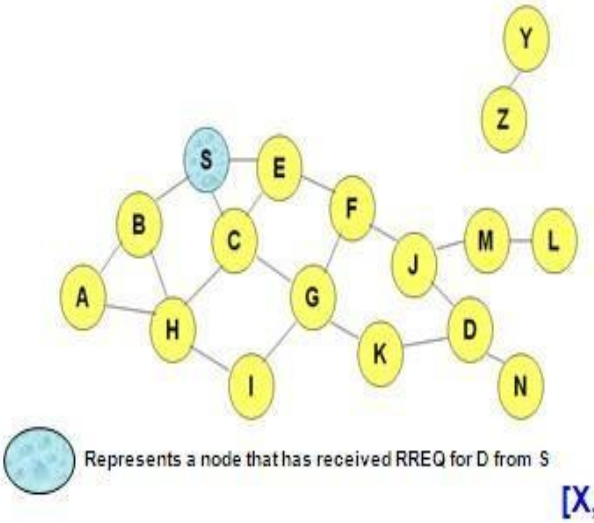
Dynamic Source Routing

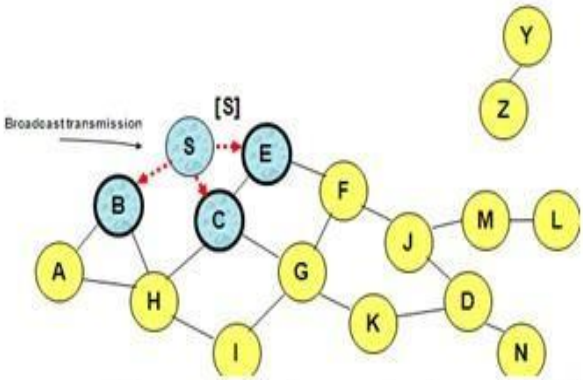
The Dynamic Source Routing protocol (DSR) is a simple and efficient routing protocol designed specifically for use in multi-hop wireless ad hoc networks of mobile nodes. DSR allows the network to be completely self-organizing and self-configuring, without the need for any existing network infrastructure or administration. The protocol is composed of the two main mechanisms of "Route Discovery" and "Route Maintenance", which work together to allow nodes to discover and maintain routes to arbitrary destinations in the ad hoc network. All aspects of the protocol operate entirely on-demand, allowing the routing packet overhead of DSR to scale automatically to only that needed to react to changes in the routes currently in use.

Route discovery. If the source does not have a route to the destination in its route cache, it broadcasts a route request (RREQ) message specifying the destination node for which the route is requested. The RREQ message includes a route record which specifies the sequence of nodes traversed by the message. When an intermediate node receives a RREQ, it checks to see if it is already in the route record. If it is, it drops the message. This is done to prevent routing loops. If the intermediate node had received the RREQ before, then it also drops the message. The intermediate node forwards the RREQ to the next hop according to the route specified in the header. When the destination receives the RREQ, it sends back a route reply message. If the destination has a route to the source in its route cache, then it can send a route response (RREP) message along this route. Otherwise, the RREP message can be sent along the reverse route back to the source. Intermediate nodes may also use their route cache to reply to RREQs. If an intermediate node has a route to the destination in its cache, then it can append the route to the route record in the RREQ, and send an RREP back to the source containing this route. This can help limit flooding of the RREQ. However, if the cached route is out-of-date, it can result in the source receiving stale routes.

Route maintenance. When a node detects a broken link while trying to forward a packet to the next hop, it sends a route error (RERR) message back to the source containing the link in error. When an RERR message is received, all routes containing the link in error are deleted at that node.

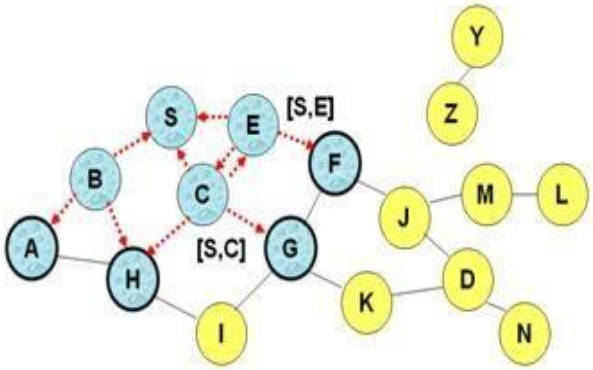
As an example, consider the following MANET, where a node S wants to send a packet to D, but does not know the route to D. So, it initiates a route discovery. Source node S floods Route Request (RREQ). Each node appends its own identifier when forwarding RREQ as shown below.



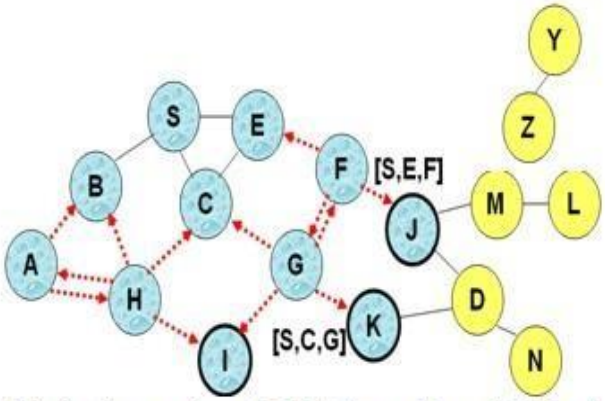


Represents transmission of RREQ

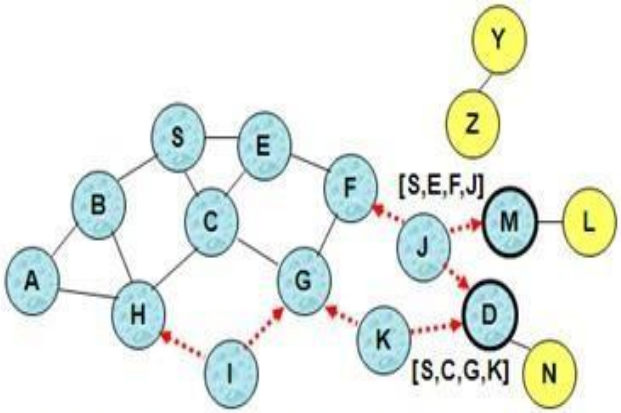
[X,Y] Represents list of identifiers appended to RREQ



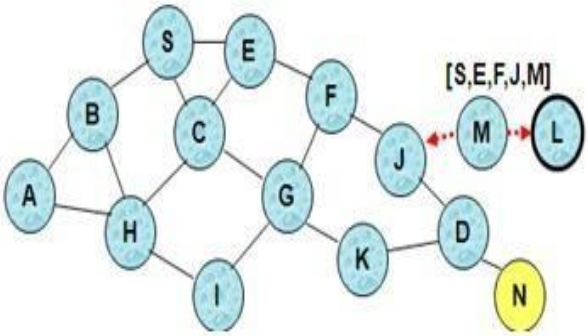
Node H receives packet RREQ from two neighbors: potential for collision



Node C receives RREQ from G and H, but does not forward it again, because node C has already forwarded RREQ once

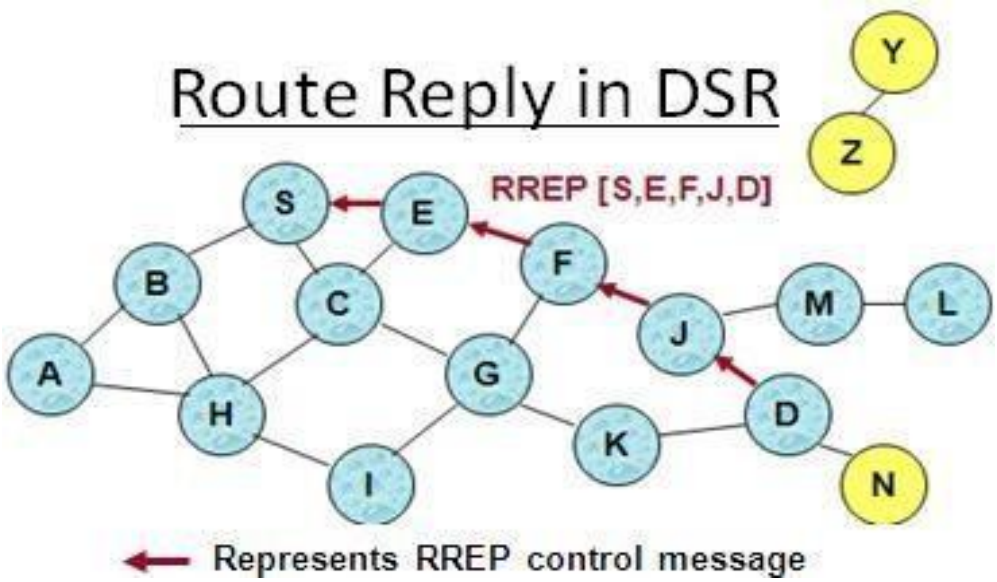


- Nodes J and K both broadcast RREQ to node D
- Since nodes J and K are hidden from each other, their transmissions may collide

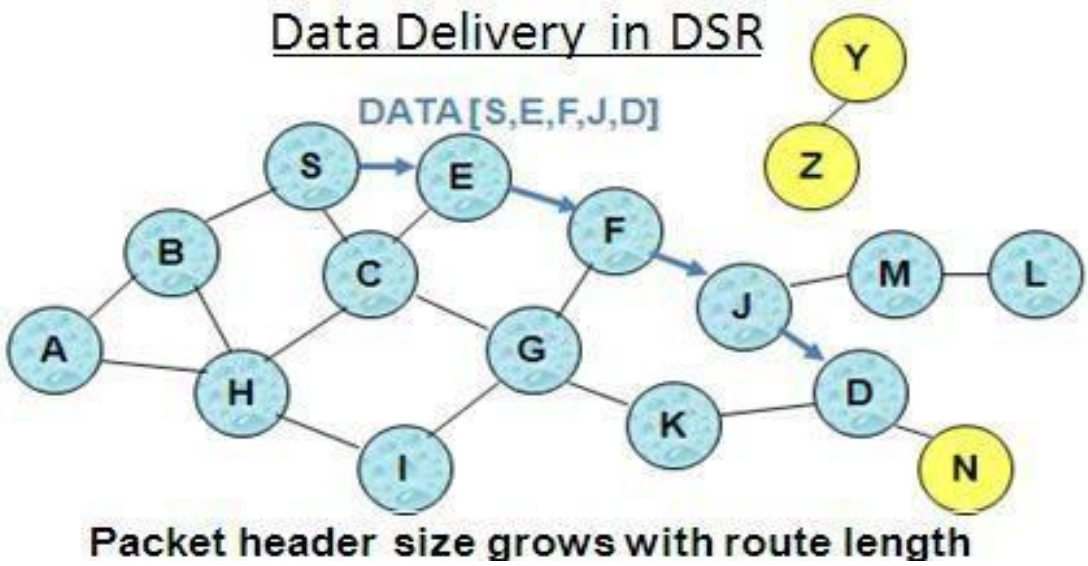


Node D does not forward RREQ, because node D is the intended target of the route discovery

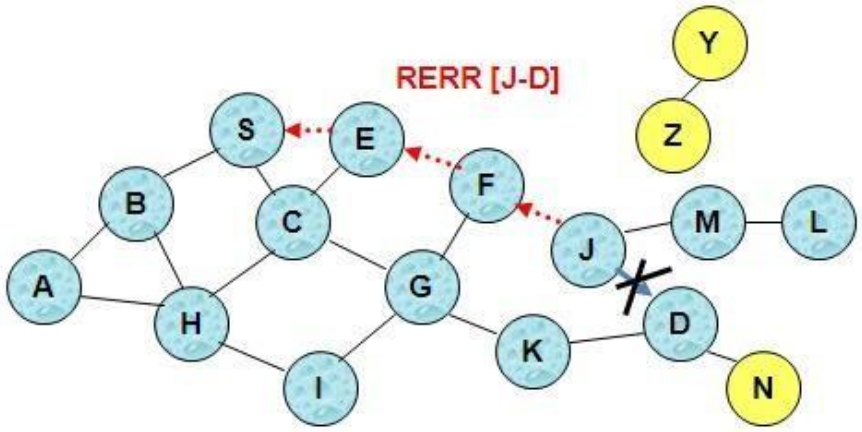
Destination D on receiving the first RREQ, sends a Route Reply (RREP). RREP is sent on a route obtained by reversing the route appended to received RREQ. RREP includes the route from S toD on which RREQ was received by node D.



Route Reply can be sent by reversing the route in Route Request (RREQ) only if links are guaranteed to be bi-directional. If Unidirectional (asymmetric) links are allowed, then RREP may need a route discovery from S to D. Node S on receiving RREP, caches the route included in the RREP. When node S sends a data packet to D, the entire route is included in the packet header {hence the name source routing}. Intermediate nodes use the source route included in a packet to determine to whom a packet should be forwarded.



J sends a route error to S along route J-F-E-S when its attempt to forward the data packet S (with route SEFJD) on J-D fails. Nodes hearing RERR update their route cache to remove link J-D



Advantages of DSR:

- Routes maintained only between nodes who need to communicate-- reduces overheadof route maintenance
- Route caching can further reduce route discovery overhead
- A single route discovery may yield many routes to the destination, due to intermediate nodes replying from local caches

Disadvantages of DSR:

- Packet header size grows with route length due to source routing
- Flood of route requests may potentially reach all nodes in the network
- Care must be taken to avoid collisions between route requests propagated by neighboring nodes -- insertion of random delays before forwarding RREQ
- Increased contention if too many route replies come back due to nodes replying using their local cache-- Route Reply Storm problem. Reply storm may be eased by preventing a node from sending RREP if it hears another RREP with a shorter route
- An intermediate node may send Route Reply using a stale cached route, thus polluting other caches

An optimization for DSR can be done called as Route Caching. Each node caches a new route it learns by any means. In the above example, When node S finds route [S,E,F,J,D] to node D, node S also learns route [S,E,F] to node F. When node K receives Route Request [S,C,G] destined for node, node K learns route [K,G,C,S] to node S. When node F forwards Route Reply RREP [S,E,F,J,D], node F learns route [F,J,D] to node D. When node E forwards Data [S,E,F,J,D] it learns route [E,F,J,D] to node D. A node may also learn a route when it overhears Data packets. Usage of Route cache can speed up route discovery and can also reduce propagation of route

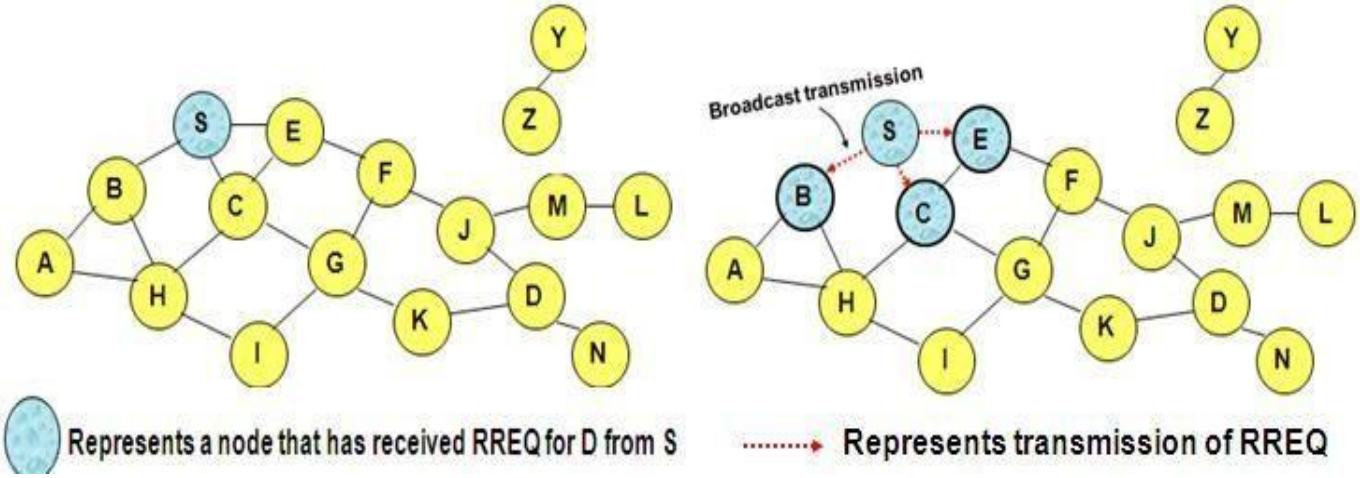
requests. The disadvantages are, stale caches can adversely affect performance. With passageof time and host mobility, cached routes may become invalid.

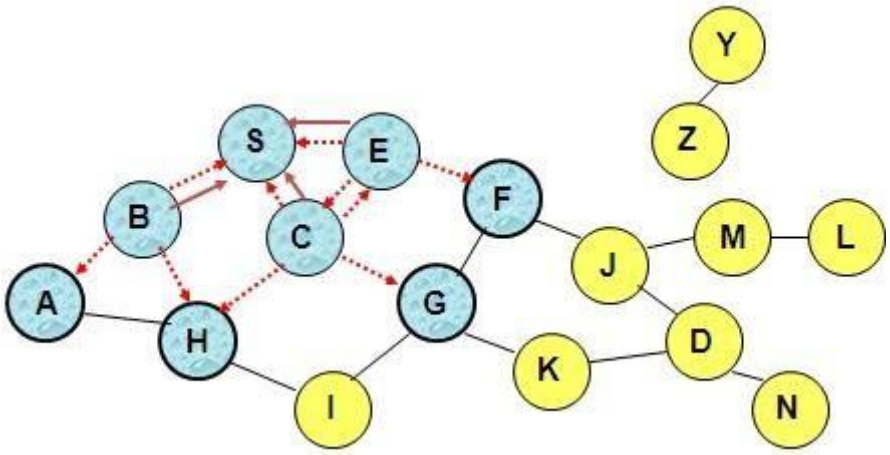
Ad Hoc On-Demand Distance Vector Routing (AODV)

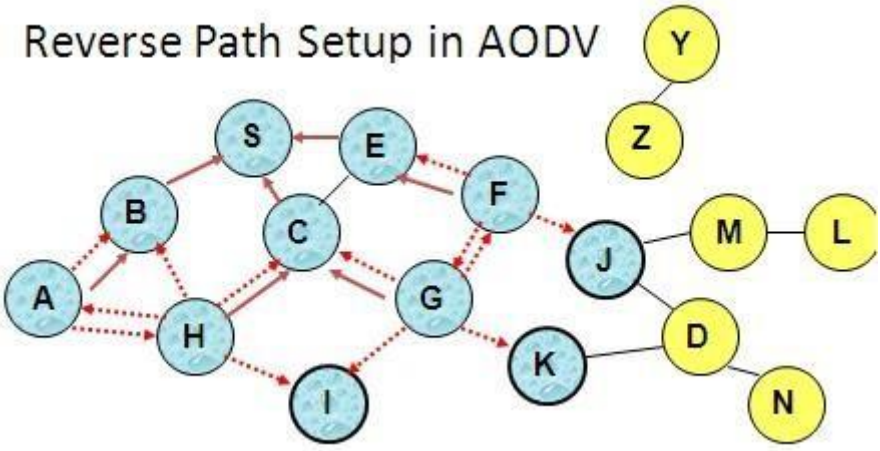
AODV is another reactive protocol as it reacts to changes and maintains only the active routes in the caches or tables for a pre-specified expiration time. Distance vector means a set of distant nodes, which defines the path to destination. AODV can be considered as a descendant of DSR and DSDV algorithms. It uses the same route discovery mechanism used by DSR. DSR includes source routes in packet headers and resulting large headers can sometimes degrade performance, particularly when data contents of a packet are small. AODV attempts to improve on DSR by maintaining routing tables at the nodes, so that data packets do not have to contain routes. AODV retains the desirable feature of DSR that routes are maintained only between nodes which need to communicate. However, as opposed to DSR, which uses source routing, AODV uses hop-by-hop routing by maintaining routing table entries at intermediate nodes.

Route Discovery. The route discovery process is initiated when a source needs a route to a destination and it does not have a route in its routing table. To initiate route discovery, the source floods the network with a RREQ packet specifying the destination for which the route is requested. When a node receives an RREQ packet, it checks to see whether it is the destinationor whether it has a route to the destination. If either case is true, the node generates an RREP packet, which is sent back to the source along the reverse path. Each node along the reverse path sets up a forward pointer to the node it received the RREP from. This sets up a forward path from the source to the destination. If the node is not the destination and does not have a route to the destination, it rebroadcasts the RREQ packet. At intermediate nodes duplicate RREQ packets are discarded. When the source node receives the first RREP, it can begin sending data to the destination. To determine the relative degree out-of-datedness of routes, each entry in the node routing table and all RREQ and RREP packets are tagged with a destination sequence number. A larger destination sequence number indicates a more current (or more recent) route. Upon receiving an RREQ or RREP packet, a node updates its routing information to set up the reverse or forward path, respectively, only if the route contained in the RREQ or RREP packet is more current than its own route.

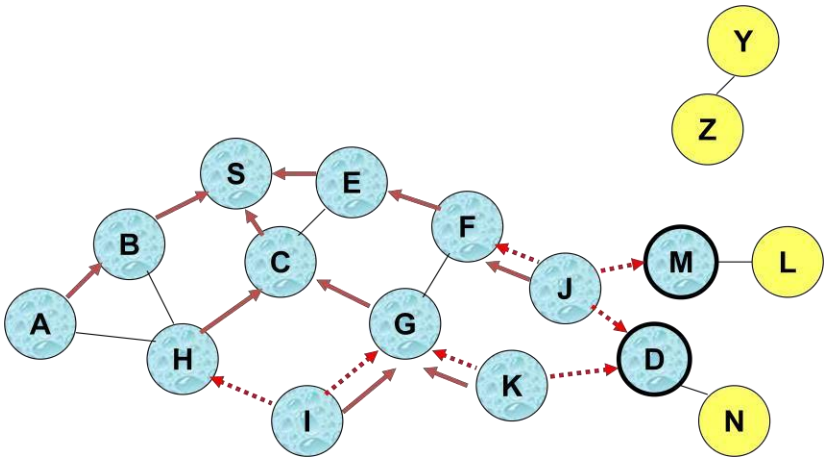
Route Maintenance. When a node detects a broken link while attempting to forward a packet to the next hop, it generates a RERR packet that is sent to all sources using the broken link. The RERR packet erases all routes using the link along the way. If a source receives a RERR packet and a route to the destination is still required, it initiates a new route discovery process. Routes are also deleted from the routing table if they are unused for a certain amount of time.

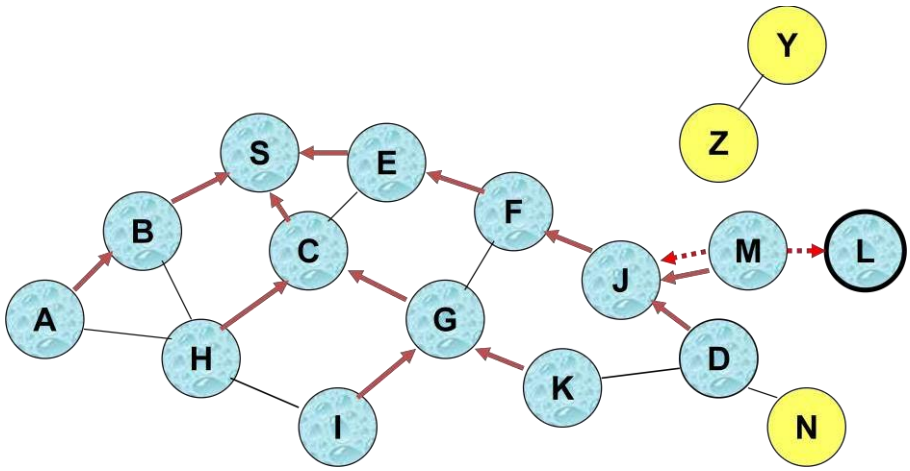




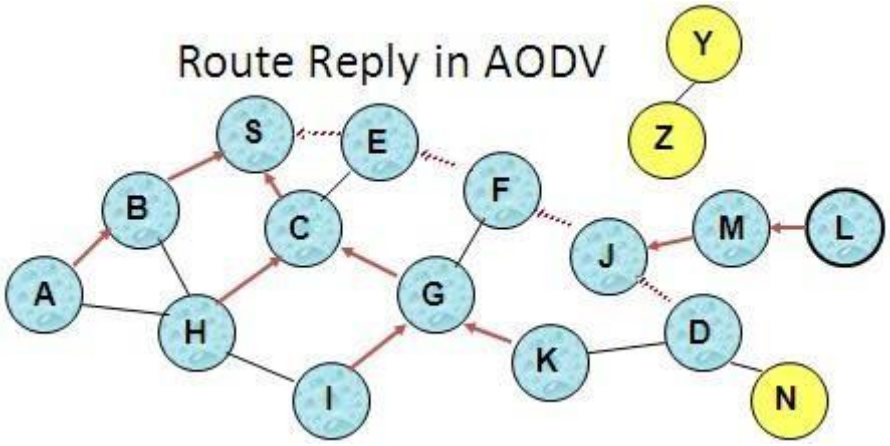


Node C receives RREQ from G and H, but does not forward it again, because node C has already forwarded RREQ once



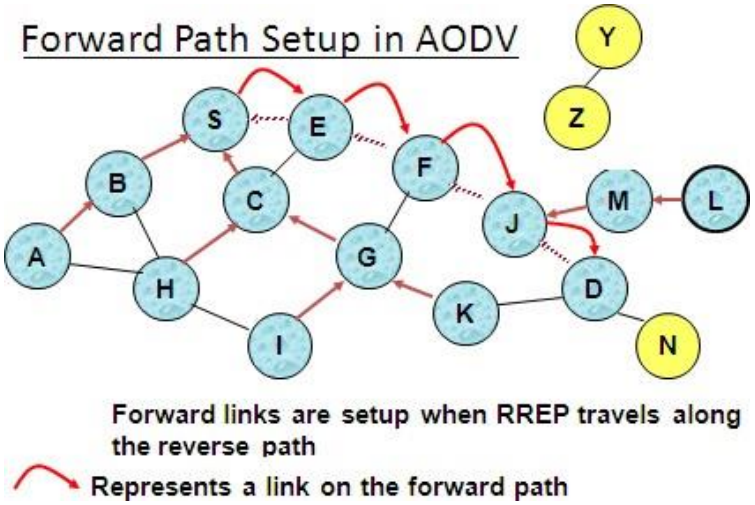


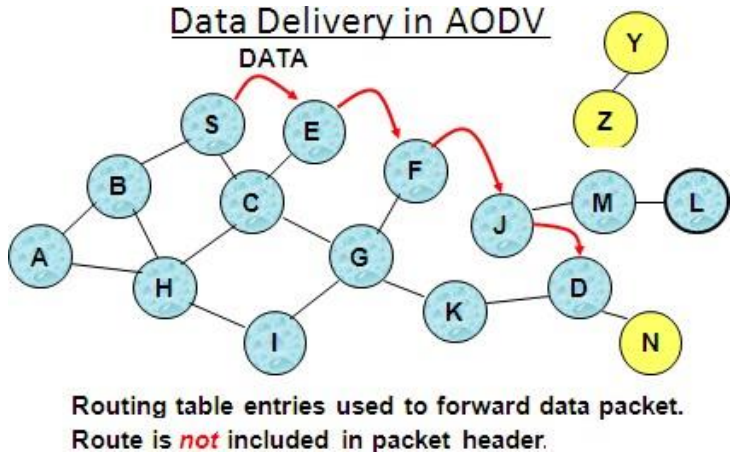
Node D does not forward RREQ, because node D is the intended target of the RREQ



Represents links on path taken by RREP

An intermediate node (not the destination) may also send a Route Reply (RREP) provided that it knows a more recent path than the one previously known to sender S. To determine whether the path known to an intermediate node is more recent, destination sequence numbers are used. The likelihood that an intermediate node will send a Route Reply when using AODV is not as high as DSR. A new Route Request by node S for a destination is assigned a higher destination sequence number. An intermediate node which knows a route, but with a smaller sequence number, cannot send Route Reply

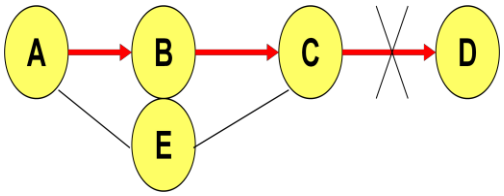




When node X is unable to forward packet P (from node S to node D) on link (X,Y), it generates a RERR message Node X increments the destination sequence number for D cached at node X. The incremented sequence number N is included in the RERR. When node S receives the RERR, it initiates a new route discovery for D using destination sequence number at least as large as

N. When node D receives the route request with destination sequence number N, node D will set its sequence number to N, unless it is already larger than N.

Sequence numbers are used in AODV to avoid using old/broken routes and to determine which route is newer. Also, it prevents formation of loops.



Assume that A does not know about failure of link C-D because RERR sent by C is lost.

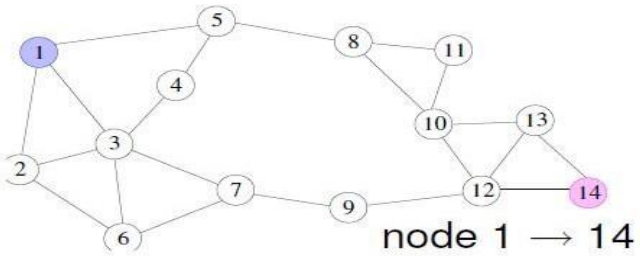
Now C performs a route discovery for D. Node A receives theRREQ (say, via path C-E-A)

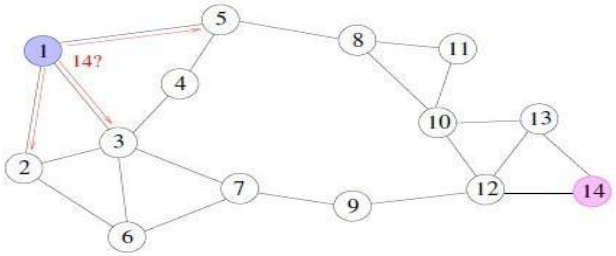
Node A will reply since A knows a route to D via node B resulting in a loop (for instance, C-E-A-B-C)

Neighboring nodes periodically exchange hello message and absence of hello message indicatesa link failure. When node X is unable to forward packet P (from node S to node D) on link (X,Y), it generates a **RERR message**. Node X increments the destination sequence number for D cached at node X. The incremented sequence number N is included in the RERR. When node S receives the RERR, it initiates a new route discovery for D using destination sequence numberat least as large as N. When node D receives the route request with destination sequence number N, node D will set its sequence number to N, unless it is already larger than N.

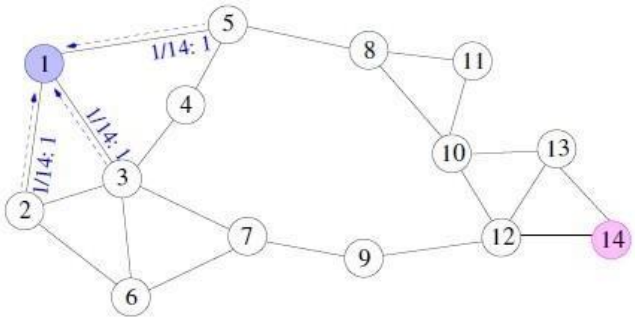
Another example for AODV protocol:

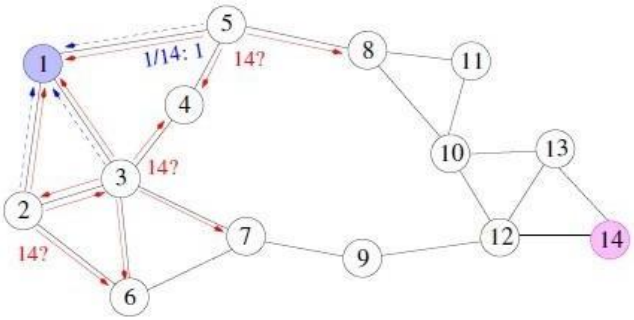
Assume node-1 want to send a msg to node-14 and does not know the route. So, it broadcasts (floods) route request message, shown in red.



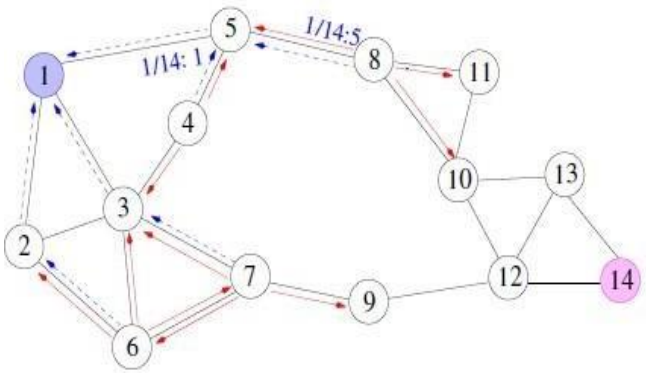


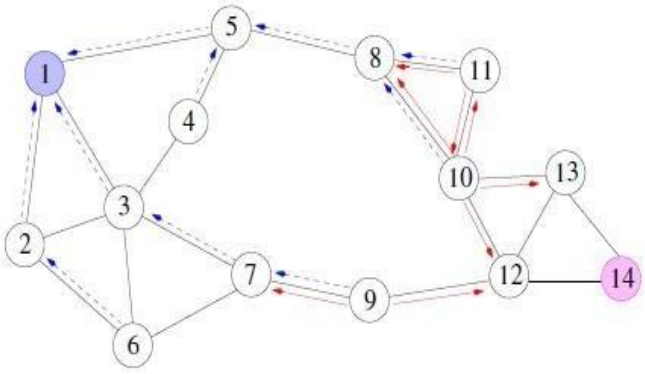
Node from which RREQ was received defines a reverse route to the source. (reverse routing table entries shown in blue).



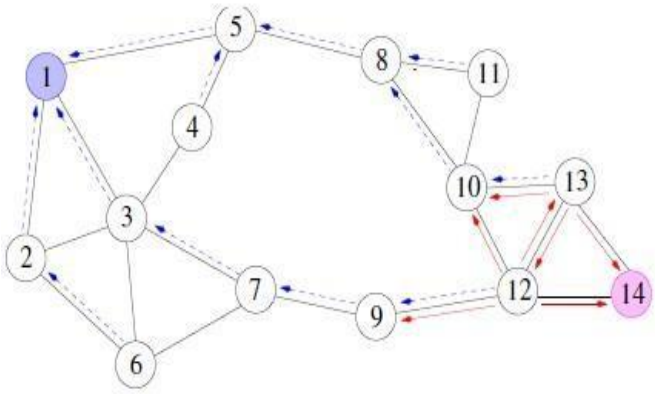


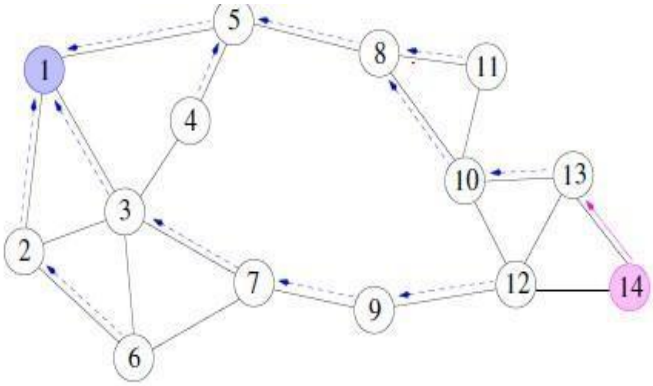
The route request is flooded through the network. Destination managed sequence number, ID prevent looping. Also, flooding is expensive and creates broadcast collision problem.



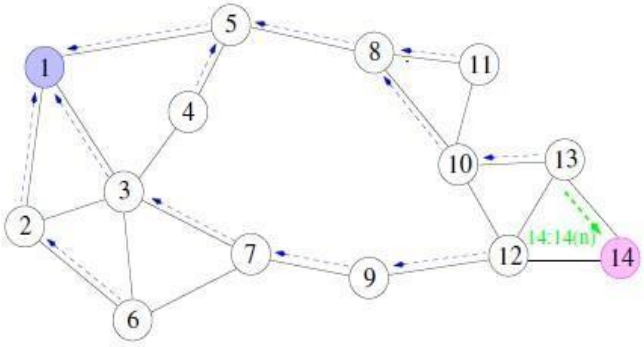


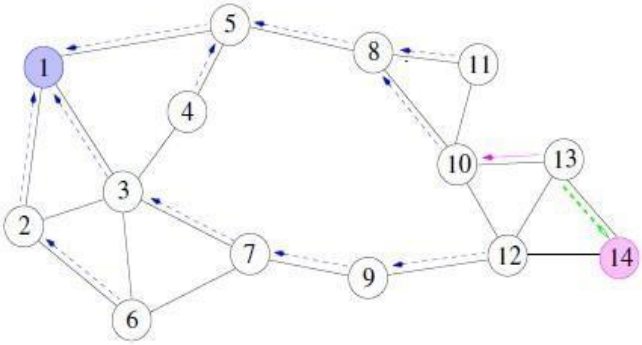
Route request arrives at the destination node-14. Upon receiving, destination sends route replyby setting a sequence number(shown in pink)



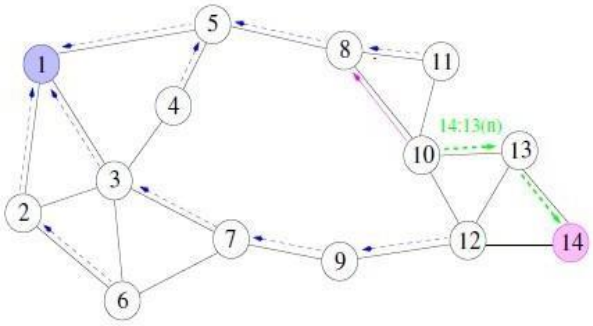


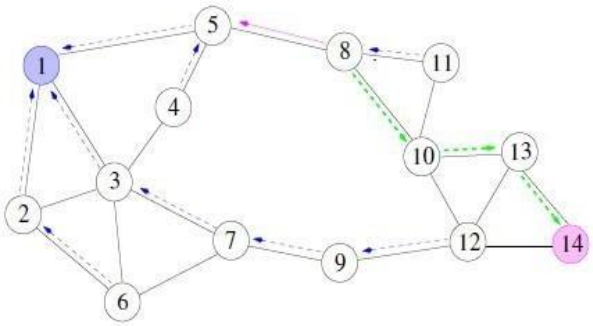
Routing table now contains forward route to the destination. Route reply follows reverse route back to the source.



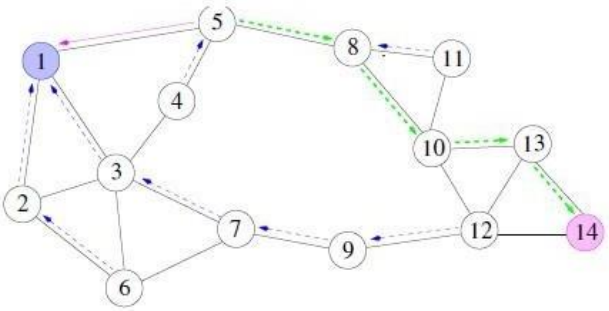


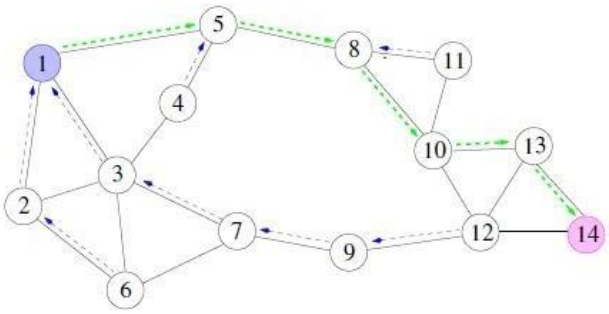
The route reply sets the forward table entries on its way back to the source.



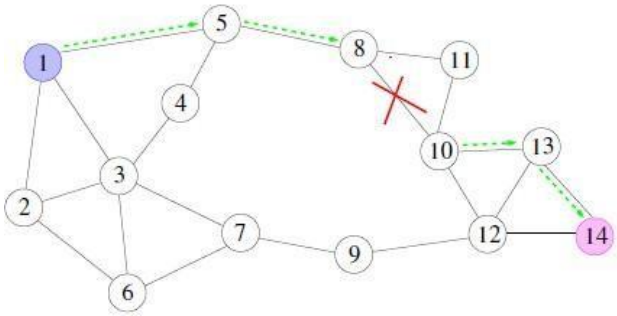


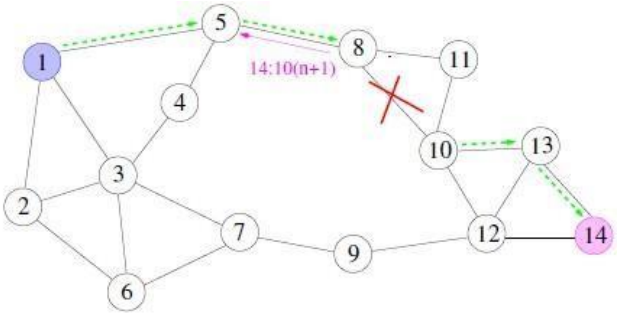
Once the route reply reaches the source, it adopts the destination sequence number. Traffic flows along the forward route. Forward route is refreshed and the reverse routes get timed out.



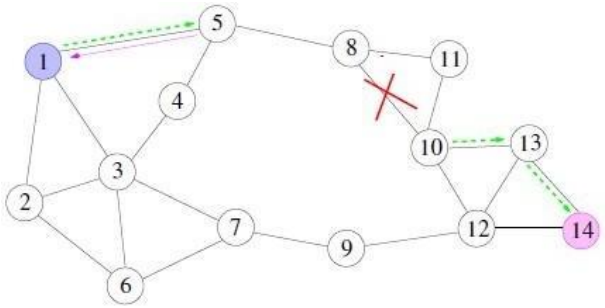


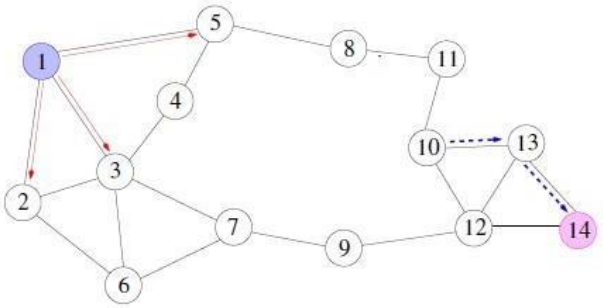
Suppose there has been a failure in one of the links. The node sends a return error message tothe source with incrementing the sequence number.





Once the source receives the route error, it re-initiates the route discoveryprocess.

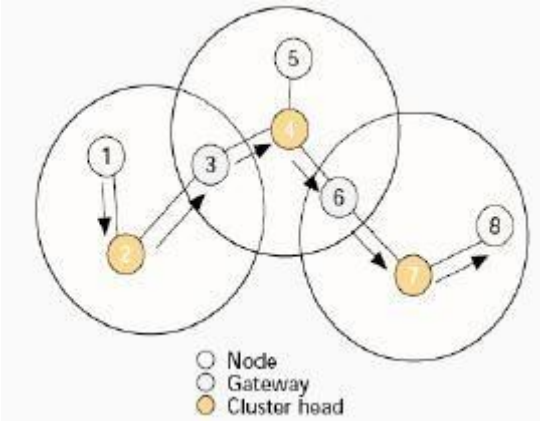




A routing table entry maintaining a reverse path is purged after a timeout interval. Timeout should be long enough to allow RREP to come back. A routing table entry maintaining a forward path is purged if not used for a active_route_timeout interval. If no data is being sent using a particular routing table entry, that entry will be deleted from the routing table (even if the route may actually still be valid).

Cluster-head Gateway Switch Routing (CGSR)

The cluster-head gateway switch routing (CGSR) is a hierarchical routing protocol. It is a proactiveprotocol. When a source routes the packets to destination, the routing tables are already available at the nodes. A cluster higher in hierarchy sends the packets to the cluster lower in hierarchy. Each cluster can have several daughters I and forms a tree-like structure in CGSR. CGSR forms a cluster structure. The nodes aggregate into clusters using an appropriate algorithm. The algorithm defines a cluster-head, the node used for connection to other clusters. It also defines a gateway node which provides switching (communication) between two or more cluster-heads. There will thus be three types of nodes— (i) internal nodes in a cluster which transmit and receive the messages and packets through a cluster-head, (ii) cluster-head in each cluster such that there is a cluster-head which dynamically schedules the route paths. It controls a group of ad-hoc hosts, monitors broadcasting within the cluster, and forwards the messages to another cluster-head, and (iii) gateway node to carry out transmission and reception of messages and packets between cluster-heads of two clusters.



The cluster structure leads to a higher performance of the routing protocol as compared to other protocols because it provides gateway switch-type traffic redirections and clusters provide an effective membership of nodes for connectivity.

CGSR works as follow:

- periodically, every nodes sends a hello message containing its ID and a monotonically increasing sequence number

- Using these messages, every cluster-head maintains a table containing the IDs of nodes belonging to it and their most recent sequence numbers.
- Cluster-heads exchange these tables with each other through gateways; eventually, each node will have an entry in the affiliation table of each cluster-head. This entry shows the node's ID & cluster-head of that node.
- Each cluster-head and each gateway maintains a routing table with an entry for every cluster-head that shows the next gateway on the shortest path to that cluster head.

Disadvantages:

- The same disadvantage common to all hierarchal algorithms related to cluster formation and maintenance.

Hierarchal State Routing (HSR)

A hierarchal link state routing protocol that solves the location management problem found in MMWN by using the logical subnets. A logical subnet is : a group of nodes that have common characteristics (e.g. the subnet of students, the subnet of profs , employees etc.). Nodes of the same subnet do not have to be close to each other in the physical distance.

HSR procedure:

1. Based on the physical distance, nodes are grouped into clusters that are supervised by cluster-heads. There are more than one level of clustering.
2. Every node has two addresses :
   I. a hierarchal-ID ,(HID), composed of the node's MAC address prefixed by the IDsof its parent clusters.
   II. a logical address in the form <subnet,host>.
3. Every logical subnet has a home agent, i.e. a node that keeps track of the HID of all members of that subnet.
4. The HIDs of the home agents are known to all the cluster-heads, and the cluster-head can translate the subnet part of the node's logical address to the HID of the corresponding home agent.
5. when a node moves to a new cluster, the head of the cluster detects it and informs the node's home agents about node's new HID.
6. When a home agent moves to a new cluster, the head of the cluster detects it and informs all other cluster-heads about the home agent's new HID.

To start a session:

1. The source node informs its cluster-head about the logical address of the destination node.

2. The cluster-head looks up the HID of the destination node's home agent and uses it to send query to the home agent asking about the destination's HID.
3. After knowing the destination's HID, the cluster-head uses its topology map to find a route to the destination's cluster-head.

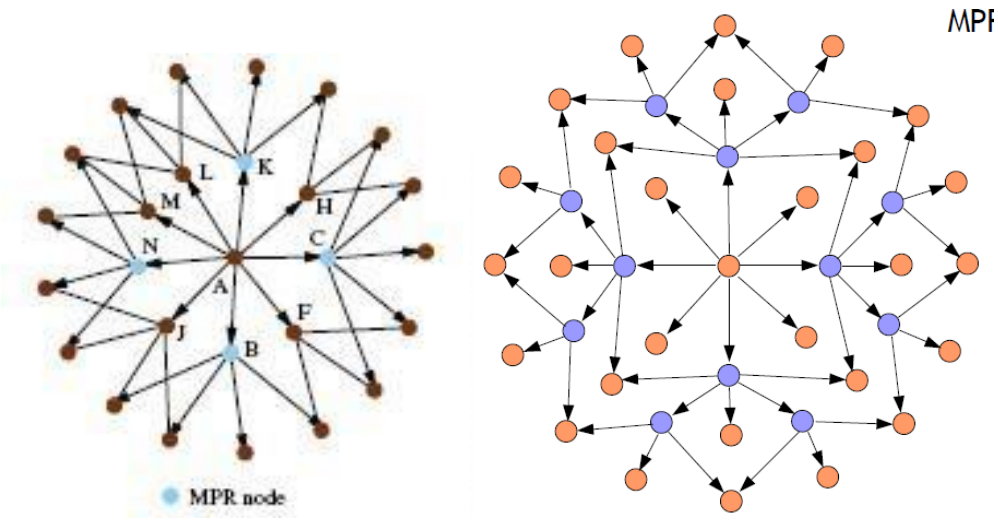
Disadvantages: cluster formation and maintenance.

Optimized Link State Routing Protocol

Optimized link state routing protocol (OLSR) has characteristics similar to those of link state flat routing table driven protocol, but in this case, only required updates are sent to the routing database. This reduces the overhead control packet size and numbers.

OSLR uses controlled flood to disseminate the link state information of each node.

- Every node creates a list of its one hop neighbors.
- Neighbor nodes exchange their lists with each other.
- Based on the received lists, each node creates its MPR.



The multipoint relays of each node, (MPR), is the minimal set of 1-hop nodes that covers all 2- hop points.

- The members of the MPR are the only nodes that can retransmit the link state informationin an attempt to limit the flood.

Security in MANET's

Securing wireless ad-hoc networks is a highly challenging issue. Understanding possible form of attacks is always the first step towards developing good security solutions. Security of communication in MANET is important for secure transmission of information. Absence of any central co-ordination mechanism and shared wireless medium makes MANET more vulnerable

to digital/cyber attacks than wired network there are a number of attacks that affect MANET. These attacks can be classified into two types:

1. External Attack: External attacks are carried out by nodes that do not belong to the network. It causes congestion sends false routing information or causes unavailability of services.

2. Internal Attack: Internal attacks are from compromised nodes that are part of the network. In an internal attack the malicious node from the network gains unauthorized access and impersonates as a genuine node. It can analyze traffic between other nodes and may participate in other network activities.

- **Denial of Service attack**: This attack aims to attack the availability of a node or the entire network. If the attack is successful the services will not be available. The attacker generally uses radio signal jamming and the battery exhaustion method.
- **Impersonation**: If the authentication mechanism is not properly implemented a malicious node can act as a genuine node and monitor the network traffic. It can also send fake routing packets, and gain access to some confidential information.
- **Eavesdropping:** This is a passive attack. The node simply observes the confidential information. This information can be later used by the malicious node. The secret information like location, public key, private key, password etc. can be fetched by eavesdropper.
- **Routing Attacks:** The malicious node makes routing services a target because it's an important service in MANETs. There are two flavors to this routing attack. One is attack on routing protocol and another is attack on packet forwarding or delivery mechanism. The first is aimed at blocking the propagation of routing information to a node. The latter is aimed at disturbing the packet delivery against a predefined path.
- **Black hole Attack:**: In this attack, an attacker advertises a zero metric for all destinations causing all nodes around it to route packets towards it.[9] A malicious node sends fake routing information, claiming that it has an optimum route and causes other good nodes to route data packets through the malicious one. A malicious node drops all packets that it receives instead of normally forwarding those packets. An attacker listen the requests in a flooding based protocol.
- **Wormhole Attack:** In a wormhole attack, an attacker receives packets at one point in the network, ―tunnels them to another point in the network, and then replays them into the network from that point. Routing can be disrupted when routing control message are tunnelled. This tunnel between two colluding attacks is known as a wormhole.
- **Replay Attack:** An attacker that performs a replay attack are retransmitted the valid data repeatedly to inject the network routing traffic that has been captured previously. This attack usually targets the freshness of routes, but can also be used to undermine poorly designed security solutions.

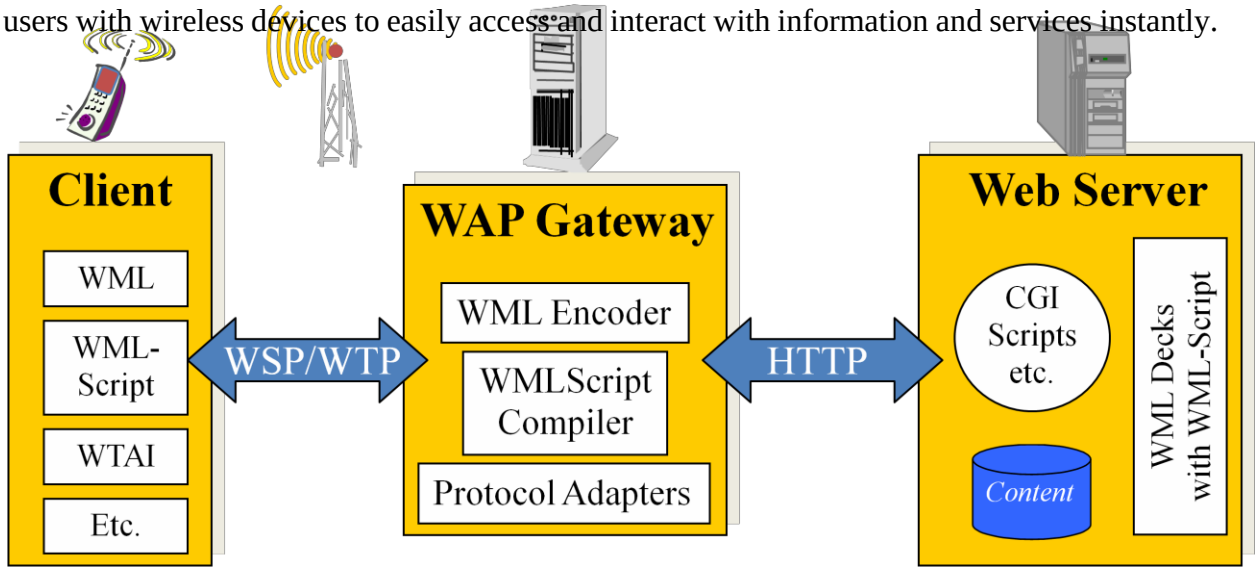
- **Jamming:** In jamming, attacker initially keep monitoring wireless medium in order to determine frequency at which destination node is receiving signal from sender. It thentransmit signal on that frequency so that error free receptor is hindered.
- **Man- in- the- middle attack:** An attacker sites between the sender and receiver and sniffs any information being sent between two nodes. In some cases, attacker may impersonate the sender to communicate with receiver or impersonate the receiver to reply to the sender.
- **Gray-hole attack**: This attack is also known as routing misbehavior attack which leads to dropping of messages. Gray-hole attack has two phases. In the first phase the node advertise itself as having a valid route to destination while in second phase, nodes drops intercepted packets with a certain probability.

Mobile Computing
Unit-5

Wireless Application Protocol (WAP)
Bluetooth, J2ME

Protocols and Tools: Wireless Application Protocol-WAP (Introduction. Protocol architecture, and treatment of protocols of all layers), Bluetooth ~~(User scenarios, physical layer, MAC layer, networking, security, link~~ management) and J2ME.

The Wireless Application Protocol (WAP) is an open, global specification that empowers mobile users with wireless devices to easily access and interact with information and services instantly.



WAP is a global standard and is not controlled by any single company. Ericsson, Nokia, Motorola, and Unwired Planet founded the **WAP Forum** in the summer of 1997 with the initial purpose of defining an industry-wide specification for developing applications over wireless communications networks. The WAP specifications define a set of protocols in application, session, transaction, security, and transport layers, which enable operators, manufacturers, and applications providers to meet the challenges in advanced wireless service differentiation and fast/flexible service creation.

All solutions must be:

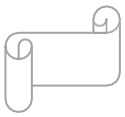
- **interoperable,** i.e., allowing terminals and software from different vendors to communicate with networks from different providers
- **scaleable,** i.e., protocols and services should scale with customer needs and number of customers
- **efficient,** i.e., provision of QoS suited to the characteristics of the wireless and mobile networks

- **reliable,** i.e., provision of a consistent and predictable platform for deploying services; and
- **secure,** i.e., preservation of the integrity of user data, protection of devices and services from security problems.

Why Choose WAP?

In the past, wireless Internet access has been limited by the capabilities of handheld devices and wireless networks.

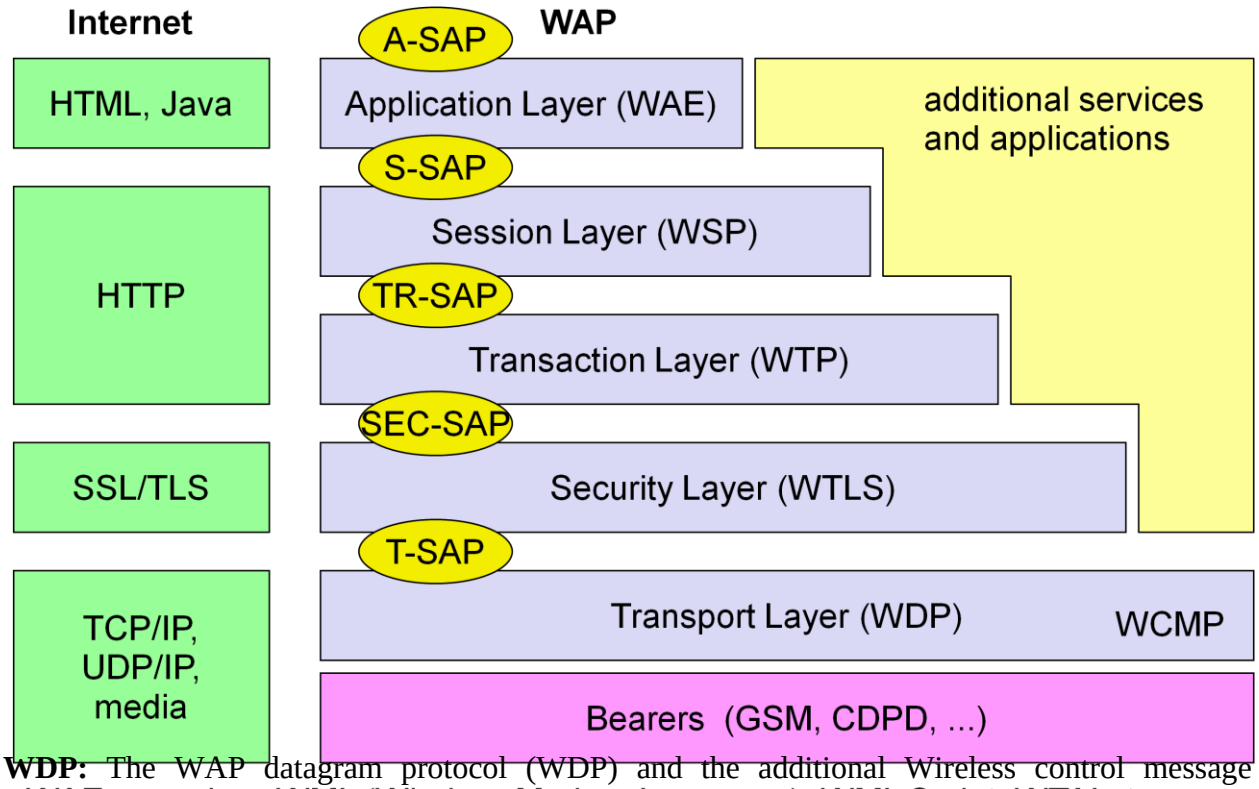
- WAP utilizes Internet standards such as XML, user datagram protocol (UDP), and Internet protocol (IP). Many of the protocols are based on Internet standards such as hypertext transfer protocol (HTTP) and TLS but have been optimized for the unique constraints of the wireless environment: low bandwidth, high latency, and less connection stability.
- Internet standards such as hypertext markup language (HTML), HTTP, TLS and transmission control protocol (TCP) are inefficient over mobile networks, requiring large amounts of mainly text-based data to be sent. Standard HTML content cannot be effectively displayed on the small-size screens of pocket-sized mobile phones and pagers.
- WAP utilizes binary transmission for greater compression of data and is optimized for long latency and low bandwidth. WAP sessions cope with intermittent coverage and can operate over a wide variety of wireless transports.
- WML and wireless markup language script (WML Script) are used to produce WAP content. They make optimum use of small displays, and navigation may be performed with one hand. WAP content is scalable from a two-line text display on a basic device to a full graphic screen on the latest smart phones and communicators.
- The lightweight WAP protocol stack is designed to minimize the required bandwidth and maximize the number of wireless network types that can deliver WAP content. Multiple networks will be targeted, with the additional aim of targeting multiple networks. These include global system for mobile communications (GSM) 900, 1,800, and 1,900 MHz; interim standard (IS)–136; digital European cordless communication (DECT); time-division multiple access (TDMA), personal communications service (PCS), FLEX, and code division multiple access (CDMA). All network technologies and bearers will also be supported, including short message service (SMS), USSD, circuit-switched cellular data (CSD), cellular digital packet data (CDPD), and general packet radio service (GPRS).
- As WAP is based on a scalable layered architecture, each layer can develop independently of the others. This makes it possible to introduce new bearers or to use new transport protocols without major changes in the other layers.



- WAP will provide multiple applications, for business and customer markets such as banking, corporate database access, and a messaging interface.

WAP Architecture

The following figure gives an overview of the WAP architecture, its protocols and components, and compares this architecture with the typical internet architecture when using the World Wide Web. The basis for transmission of data is formed by different **bearer services**. WAP does not specify bearer services, but uses existing data services and will integrate further services. Examples are message services, such as short message service (SMS) of GSM, circuit-switched data, such as high-speed circuit switched data (HSCSD) in GSM, or packet switched data, such as general packet radio service (GPRS) in GSM. Many other bearers are supported, such as CDPD, IS-136, PHS.



WAE comprises WML (Wireless Markup Language), WML Script, WTAI etc.

WDP: The WAP datagram protocol (WDP) and the additional Wireless control message protocol (WCMP) is the transport layer that sends and receives messages via any available bearer network, including SMS, USSD, CSD, CDPD, IS–136 packet data, and GPRS. The **transport layer**

service access point (T-SAP) is the common interface to be used by higher layers independent of the underlying network.

WTLS: The next higher layer, the security layer with its wireless transport layer security protocol WTLS offers its service at the **security SAP (SEC-SAP)**. WTLS is based on transport layer security (TLS, formerly SSL, secure sockets layer). WTLS has been optimized for use in wireless networks with narrow-band channels. It can offer data integrity, privacy, authentication, and (some) denial-of-service protection.

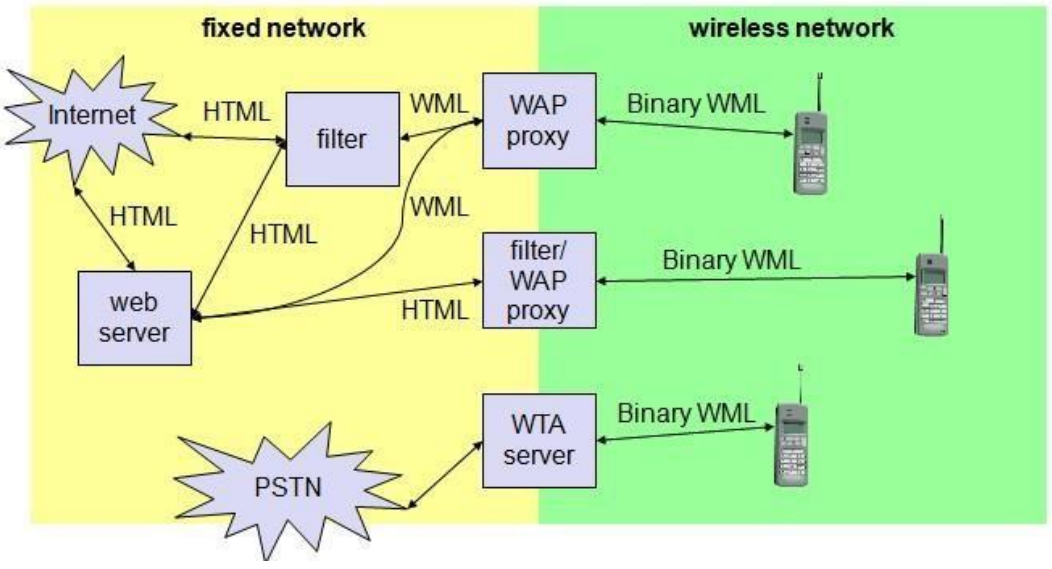
WTP: The WAP transaction protocol (WTP) layer provides transaction support, adding reliabilityto the datagram service provided by WDP at the **transaction SAP(TR-SAP)**.

WSP: The session layer with the wireless session protocol (WSP) currently offers two services at the session-SAP (S-SAP), one connection-oriented and one connectionless if used directly on top of WDP. A special service for browsing the web (WSP/B) has been defined that offers HTTP/1.1 functionality, long-lived session state, session suspend and resume, session migration and other features needed for wireless mobile access to the web.

WAE: The application layer with the wireless application environment (WAE) offers a framework for the integration of different www and mobile telephony applications.

Working of WAP

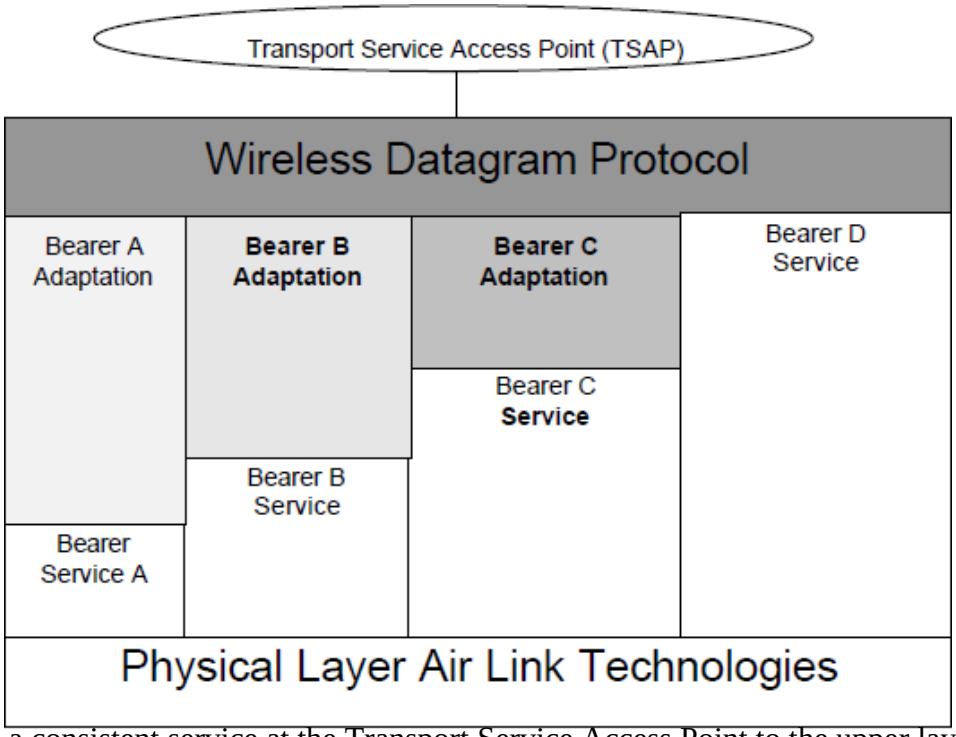
WAP does not always force all applications to use the whole protocol architecture. Applications can use only a part of the architecture. For example, if an application does not require security but needs the reliable transport of data, it can **directly** use a service of the transaction layer. Simple applications can directly use WDP.



Different scenarios are possible for the integration of WAP components into existing wireless and fixed networks. On the left side, different fixed networks, such as the traditional internet and the public switched telephone network (PSTN), are shown. One cannot change protocols and services of these existing networks so several new elements will be implemented between these networks and the WAP-enabled wireless, mobile devices in a wireless network on the right-hand side.

Wireless Datagram Protocol (WDP)

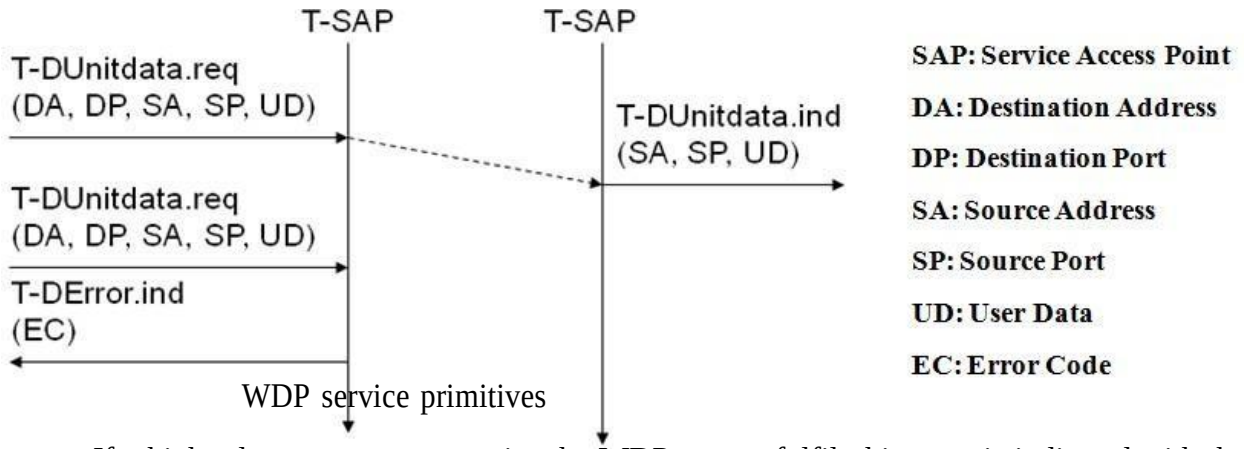
Wireless Datagram Protocol defines the movement of information from receiver to the sender and resembles the User Datagram Protocol in the Internet protocol suite.



Wireless Datagram Protocol Architecture

WDP offers a consistent service at the Transport Service Access Point to the upper layer protocol of WAP. This consistency of service allows for applications to operate transparently over different available bearer services. WDP can be mapped onto different bearers, with different characteristics. In order to optimise the protocol with respect to memory usage and radio transmission efficiency, the protocol performance over each bearer may vary.

WDP offers **source** and **destination port numbers** used for multiplexing and demultiplexing of data respectively. The service primitive to send a datagram is **TDUnitdata. req** with the **destination address (DA), destination port (DP), Source address (SA), source port (SP)**, and **user data (UD)** as mandatory parameters. Destination and source address are unique addresses for the receiver and sender of the user data. These could be MSISDNs (i.e., a telephone number), IP addresses, or any other unique identifiers. The **T-DUnitdata.ind** service primitive indicates the reception of data. Here destination address and port are only optional parameters.



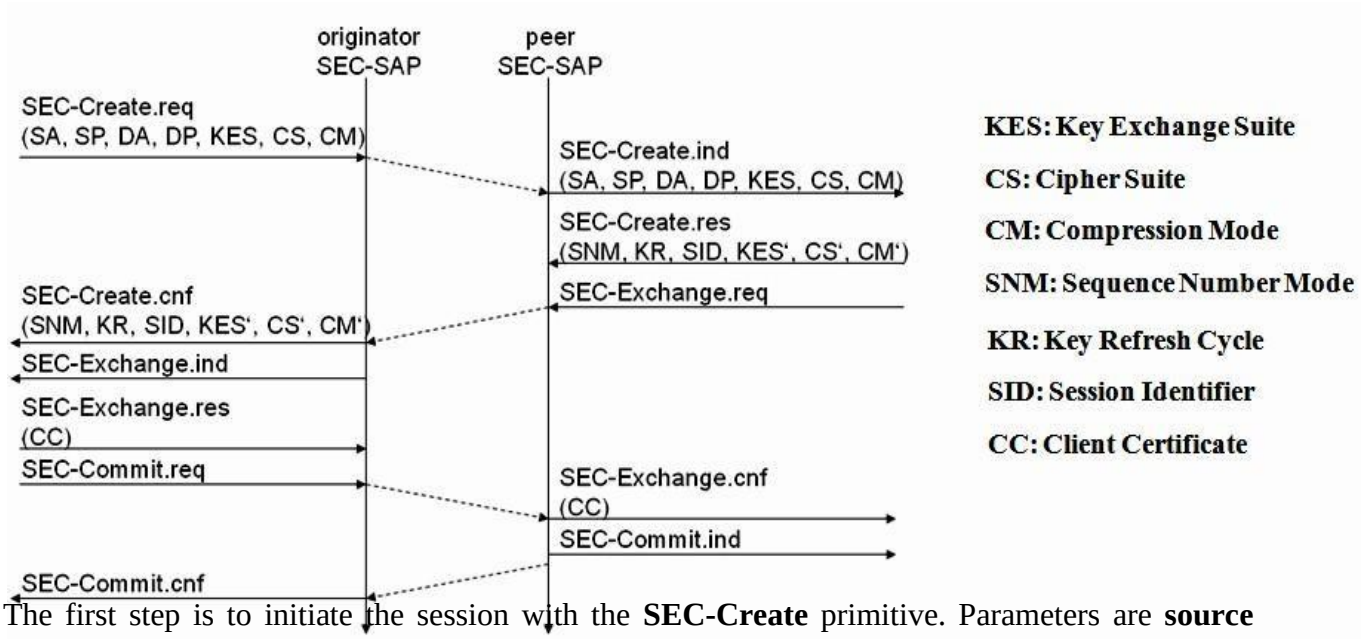
If a higher layer requests a service the WDP cannot fulfil, this error is indicated with the **T-DError.ind** service primitive. An **error code (EC)** is returned indicating the reason for the error to the higher layer. WDP is not allowed to use this primitive to indicate problems with the bearer service. It is only allowed to use the primitive to indicate local problems, such as a user data size that is too large. If any errors happen when WDP datagrams are sent from one WDP entity to another, the **wireless control message protocol (WCMP)** provides error handling mechanisms for WDP and should therefore be implemented. WCMP contains control messages that resemble the internet control message protocol messages and can also be used for diagnostic and informational purposes. WCMP can be used by WDP nodes and gateways to report errors.

Typical WCMP messages are **destination unreachable** (route, port, address unreachable), **parameter problem** (errors in the packet header), **message too big, reassembly failure**, or **echo request/reply**. An additional **WDP management entity** supports WDP and provides information about changes in the environment, which may influence the correct operation of WDP.

Wireless Transport Layer Security (WTLS)

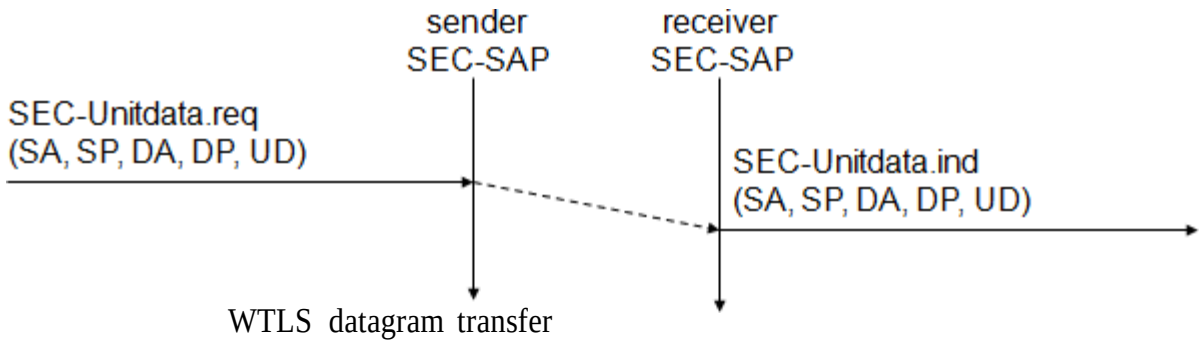
WTLS can provide different levels of security (for privacy, data integrity, and authentication) and has been optimized for low bandwidth, high-delay bearer networks. WTLS takes into account the low processing power and very limited memory capacity of the mobile devices for cryptographic algorithms. WTLS supports datagram and connection-oriented transport layer protocols. WTLS took over many features and mechanisms from TLS, but it has an optimized handshaking between the peers.

Before data can be exchanged via WTLS, a secure session has to be established. This session establishment consists of several steps: The following figure illustrates the sequence of service primitives needed for a so-called 'full handshake'.



The first step is to initiate the session with the **SEC-Create** primitive. Parameters are **source address (SA), source port (SP)** of the originator, **destinationaddress (DA), destination port (DP)** of the peer. The originator proposes **a key exchange suite (KES)** (e.g., RSA, DH, ECC), a **cipher suite (CS)** (e.g., DES, IDEA), and **a compression method (CM)**. The peer answers with parameters for the **sequence number mode (SNM)**, the **key refresh** cycle (**KR**) (i.e., how often keys are refreshed within this secure session), the **session identifier (SID)** (which is unique with each peer), and the selected **key exchange suite (KES'), cipher suite (CS'), compression method (CM')**. The peer also issues a **SEC-Exchange primitive**. This indicates that the peer wishes to perform public-key authentication with the client, i.e., the peer

requests a **client certificate (CC)** from the originator. The first step of the secure session creation, the negotiation of the security parameters and suites, is indicated on the originator's side, followed by the request for a certificate. The originator answers with its certificate and issues a **SEC-Commit.req** primitive. This primitive indicates that the handshake is completed for the originator's side and that the originator now wants to switch into the newly negotiated connection state. The certificate is delivered to the peer side and the SEC-Commit is indicated. The WTLS layer of the peer sends back a confirmation to the originator. This concludes the full handshake for secure session setup.

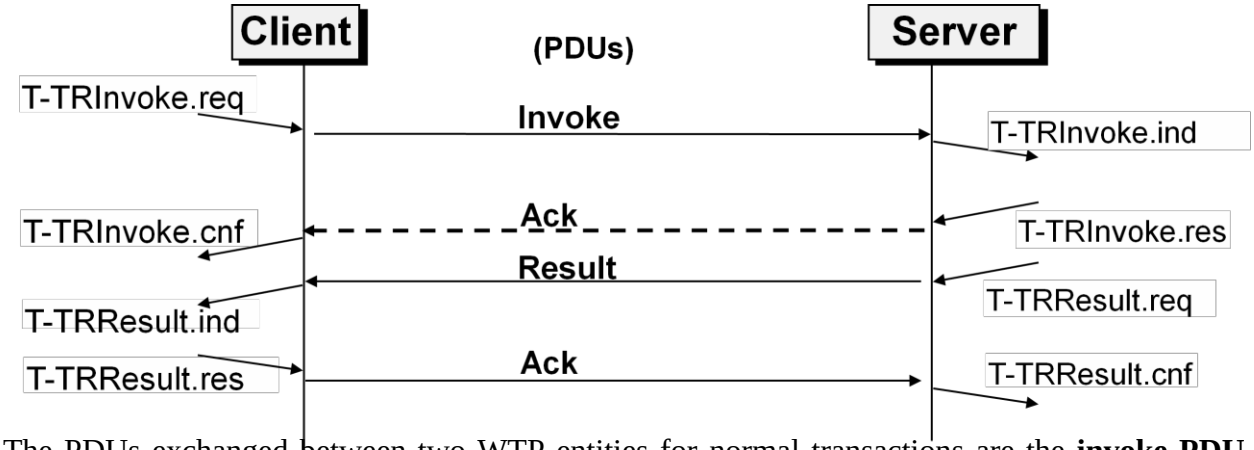


After setting up a secure connection between two peers, user data can be exchanged. This is done using the simple **SEC-Unitdata** primitive as shown in above figure. SEC-Unitdata has exactly the same function as T-DUnitdata on the WDP layer, namely it transfers a datagram between a sender and a receiver. This data transfer is still unreliable, but is now secure. This shows that WTLS can be easily plugged into the protocol stack on top of WDP.

Wireless Transaction Protocol (WTP)

WTP has been designed to run on very thin clients, such as mobile phones. WTP offers several advantages to higher layers, including an improved reliability over datagram services, improved efficiency over connection-oriented services, and support for transaction-oriented services such as web browsing. WTP offers many features to the higher layers. The basis is formed from three **classes of transaction service.** Class 0 provides unreliable message transfer without any result message. Classes 1 and 2 provide reliable message transfer, class 1 without, class 2 with, exactly one reliable result message (the typical request/response case). WTP achieves reliability using **duplicate removal, retransmission, acknowledgement**s and unique **transaction identifiers.**

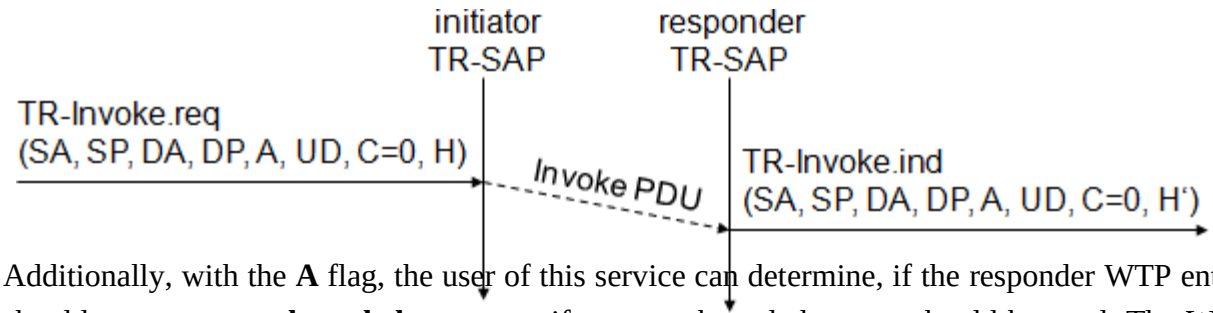
WTP allows for **asynchronous transactions, abort of transactions, concatenation of messages**, and can **report success or failure** of reliable messages (e.g., a server cannot handle the request). The three service primitives offered by WTP are **TR-Invoke** to initiate a new transaction, **TR-Result** to send back the result of a previously initiated transaction, and **TR-Abort** to abort an existing transaction.



The PDUs exchanged between two WTP entities for normal transactions are the **invoke PDU, ack PDU**, and **result PDU.** A special feature of WTP is its ability to provide a **user acknowledgement** or, alternatively, an **automatic acknowledgement** by the WTP entity.

WTP Class 0

Class 0 offers an unreliable transaction service without a result message. The transaction is stateless and cannot be aborted. The service is requested with the **TR-Invoke.req** primitive as shown below. Parameters are same as in WDP.

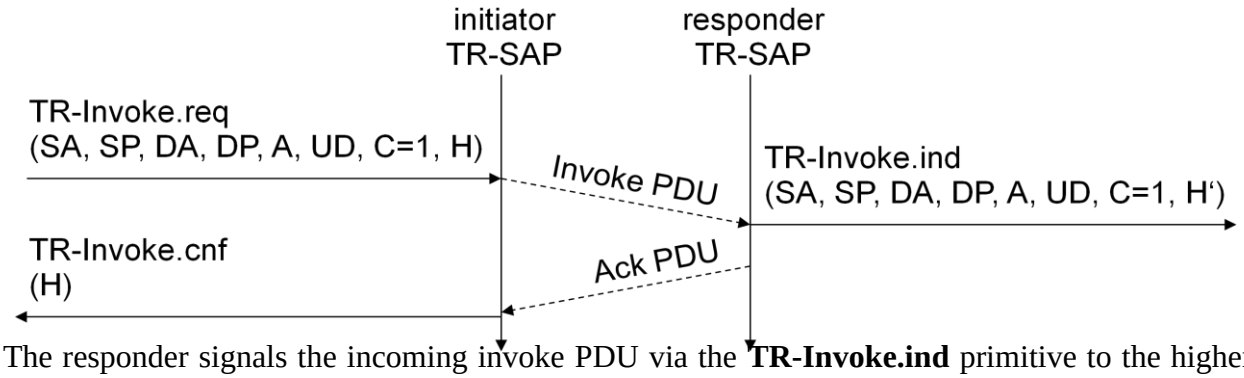


Additionally, with the **A** flag, the user of this service can determine, if the responder WTP entity should generate an **acknowledgement** or if a user acknowledgement should be used. The WTP layer will transmit the **user data (UD)** transparently to its destination. The class type **C** indicates here class 0. Finally, the transaction **handle H** provides a simple index to uniquely

identify the transaction and is an alias for the tuple (SA, SP, DA, DP), i.e., a socket pair, with onlylocal significance. The WTP entity at the initiator sends an invoke PDU which the responder receives. The WTP entity at the responder then generates a **TR-Invoke.ind** primitive with the same parameters as on the initiator's side, except for H' which is now the local handle for the transaction on the responder's side. WTP class 0 augments the transaction service with a simple datagram like service for occasional use by higher layers.

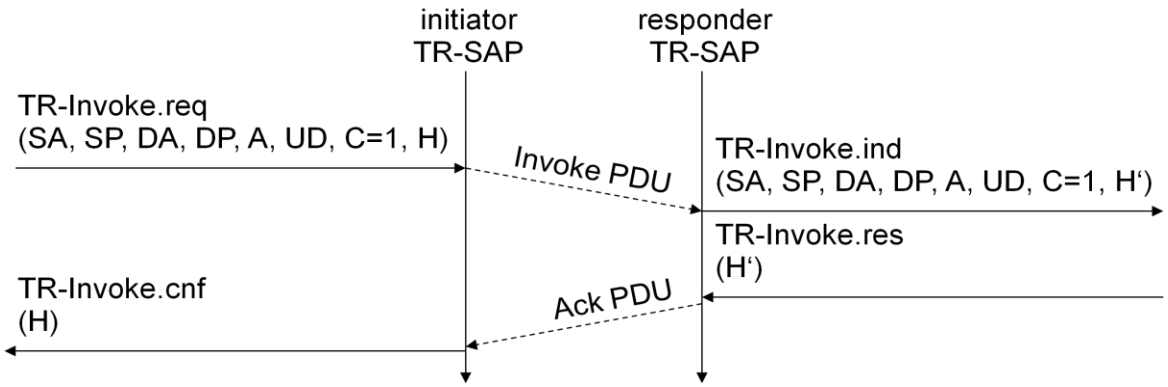
WTP Class 1

Class 1 offers a reliable transaction service but without a result message. Again, the initiator sends an invoke PDU after a **TR-Invoke.req** from a higher layer. This time, class equals '1', and no user acknowledgement has been selected as shown below.



The responder signals the incoming invoke PDU via the **TR-Invoke.ind** primitive to the higher layer and acknowledges automatically without user intervention. For the initiator the transaction ends with the reception of the acknowledgement. The responder keeps the transaction state for some time to be able to retransmit the acknowledgement if it receives the same invoke PDU again indicating a loss of the acknowledgement.

If a user of the WTP class 1 service on the initiator's side requests a user acknowledgement on the responder's side, the sequence diagram looks like the following figure.

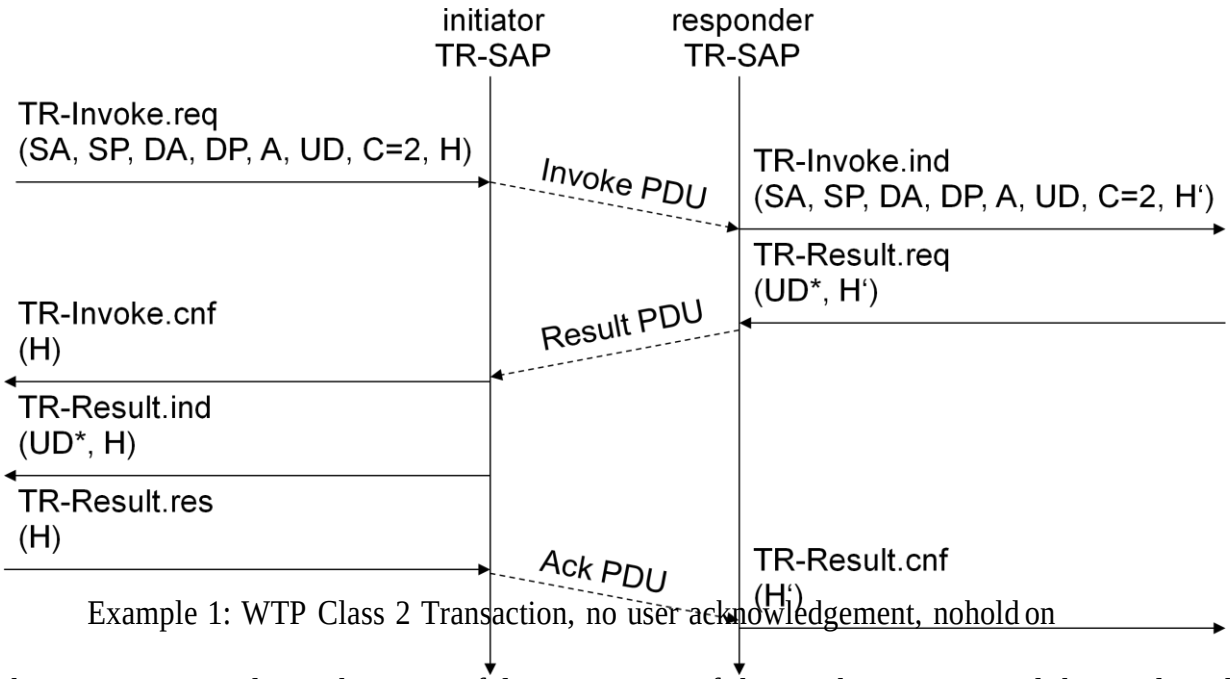


Now the WTP entity on the responder's side does not send an acknowledgement automatically, but waits for the **TR-Invoke.res** service primitive from the user. This service primitive must have the appropriate local handle H' for identification of the right transaction. The WTP entity can now send the ack PDU. Typical uses for this transaction class are reliable push services.

WTP Class 2

class 2 transaction service provides the classic reliable request/response transaction known from many client/server scenarios. Depending on user requirements, many different scenarios are possible for initiator/responder interaction. Three examples are presented below.

Example-1 scenario is shown below. A user on the initiator's side requests the service and the WTP entity sends the invoke PDU to the responder. The WTP entity on the responder's side indicates the request with the **TR-Invoke.ind** primitive to a user. The responder now waits for the processing of the request, the user on the responder's side can finally give the result UD* to the WTP entity on the responder side using **TR-Result.req.** The **result PDU** can now be sent back to the initiator, which implicitly acknowledges the invoke PDU.

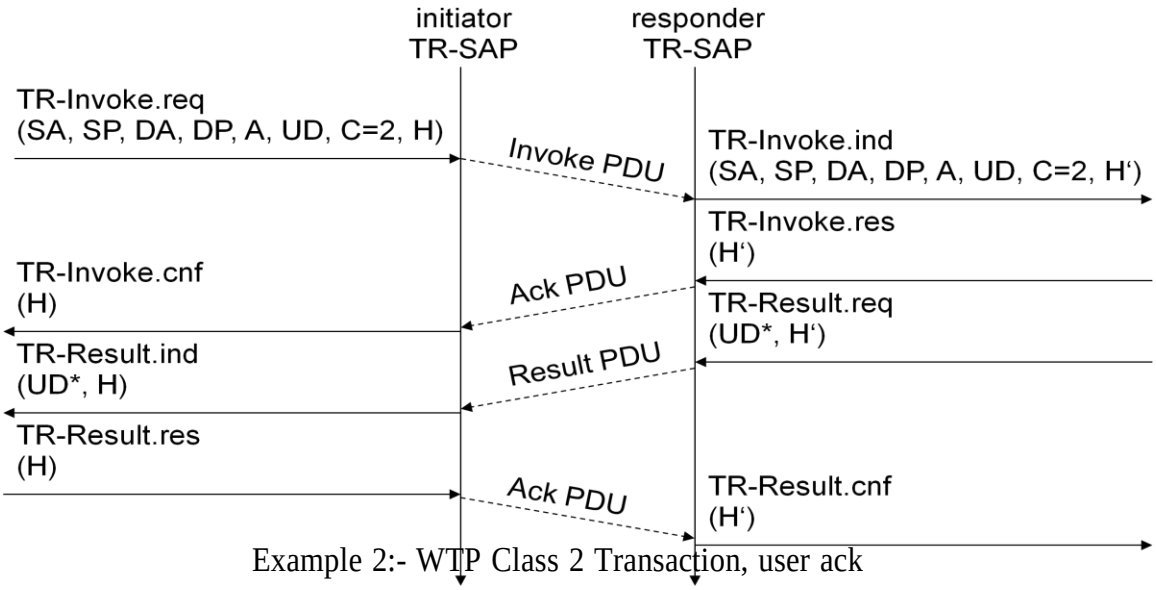


Example 1: WTP Class 2 Transaction, no user acknowledgement, nohold on

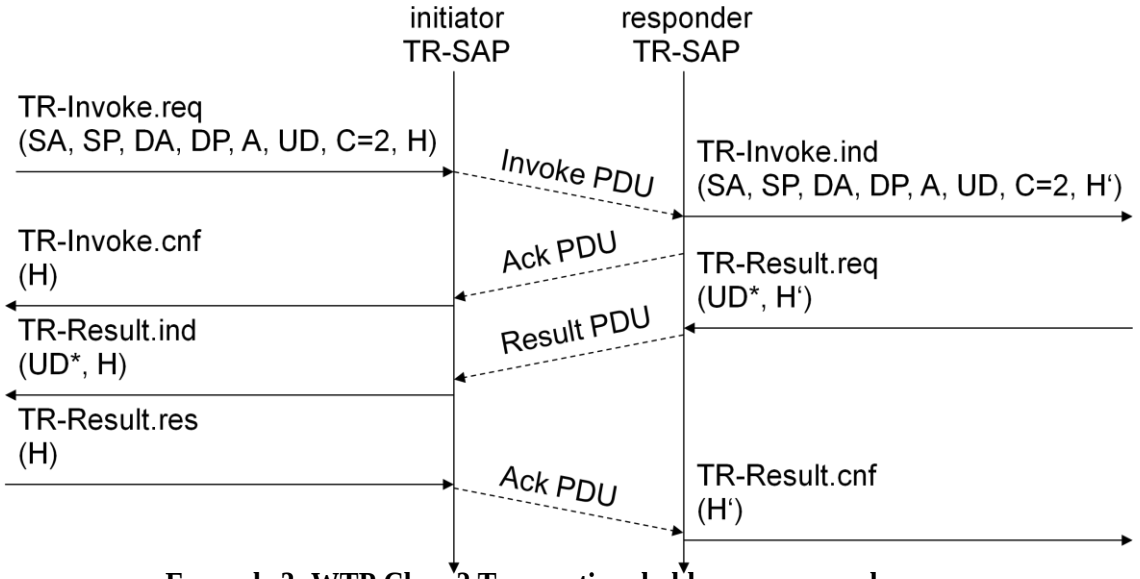
The initiator can indicate the successful transmission of the invoke message and the result with the two service primitives **TR-Invoke.cnf** and **TR-Result.ind**. A user may respond to this result with **TR-Result.res**. An acknowledgement PDU is then generated which finally triggers the **TR-Result.cnf** primitive on the responder's side. This example clearly shows the combination of

two reliable services (TR-Invoke and TR-Result) with an efficient data transmission/acknowledgement.

In example-2, the user on the responder's side now explicitly responds to the Invoke PDU using the **TR-Invoke.res** primitive, which triggers the **TR-Invoke.cnf** on the initiator's side via an **ack PDU**. The transmission of the result is also a confirmed service, as indicated by the next four service primitives. This service will likely be the most common in standard request/response scenarios as, e.g., distributed computing.



Example 2:- WTP Class 2 Transaction, user ack



Example 3: WTP Class 2 Transaction, hold on, no user ack

If the calculation of the result takes some time, the responder can put the initiator on "hold on"to prevent a retransmission of the invoke PDU as the initiator might assume packet loss if no result is sent back within a certain timeframe, which is shown above. After a time-out, the responder automatically generates an acknowledgement for the Invoke PDU. This shows the initiator that the responder is still alive and currently busy processing the request. After more time, the result PDU can be sent to the initiator.

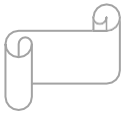
Wireless Session Protocol (WSP)

The **wireless session protocol (WSP)** has been designed to operate on top of the datagram service WDP or the transaction service WTP. WSP provides a shared state between a client and a server to optimize content transfer. WSP offers the following general features needed for content exchange between cooperating clients and servers:

- **Session management:** WSP introduces sessions that can be **established** from a client to a server and may be long lived. Sessions can also be **released** in an orderly manner. The capabilities of **suspending** and **resuming** a session are important to mobile applications.
- **Capability negotiation:** Clients and servers can agree upon a common level of protocol functionality during session establishment. Example parameters to negotiate are maximum client SDU size, maximum outstanding requests, protocol options, and server SDU size.
- **Content encoding:** WSP also defines the efficient binary encoding for the content it transfers. WSP offers content typing and composite objects, as explained for web browsing.

While WSP is a general-purpose session protocol, WAP has specified the **wireless session protocol/browsing (WSP/B)** which comprises protocols and services most suited for browsing-type applications, which offers the following features adapted to web browsing.

- **HTTP/1.1 functionality:** WSP/B supports the functions HTTP/1.1 offers, such as extensible request/reply methods, composite objects, and content type negotiation.
- **Exchange of session headers:** Client and server can exchange request/reply headers that remain constant over the lifetime of the session
- **pull data transfer:** Pulling data from a server is the traditional mechanism of the web. This is also supported by WSP/B using the request/response mechanism from HTTP/1.1. Additionally, WSP/B supports three push mechanisms for data transfer: a confirmed data push within an existing session context, a non-confirmed data push within an



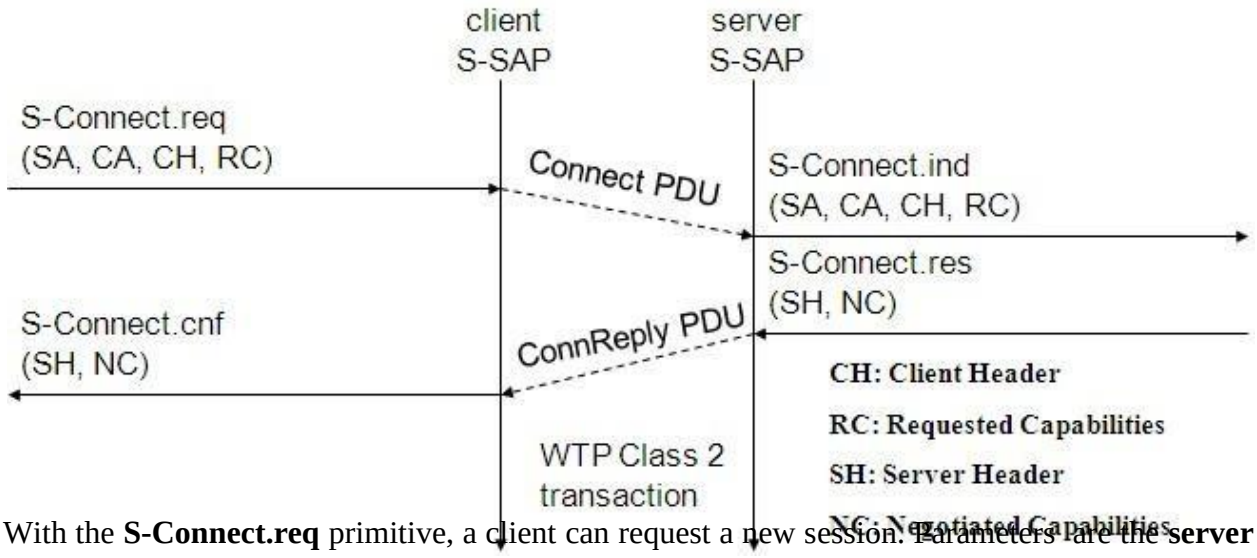
existing session context, and a non-confirmed data push without an existing session context.

- **Asynchronous requests:** Optionally, WSP/B supports a client that can send multiple requests to a server simultaneously. This improves efficiency for the requests andreplies can now be coalesced into fewer messages.

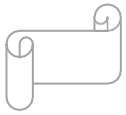
WSP/B over WTP

WSP/B uses the three service classes of WTP where, Class 0 is used for unconfirmed push, session resume, and session management. Confirmed push uses class 1, method invocation, session resume, and session management class 2.

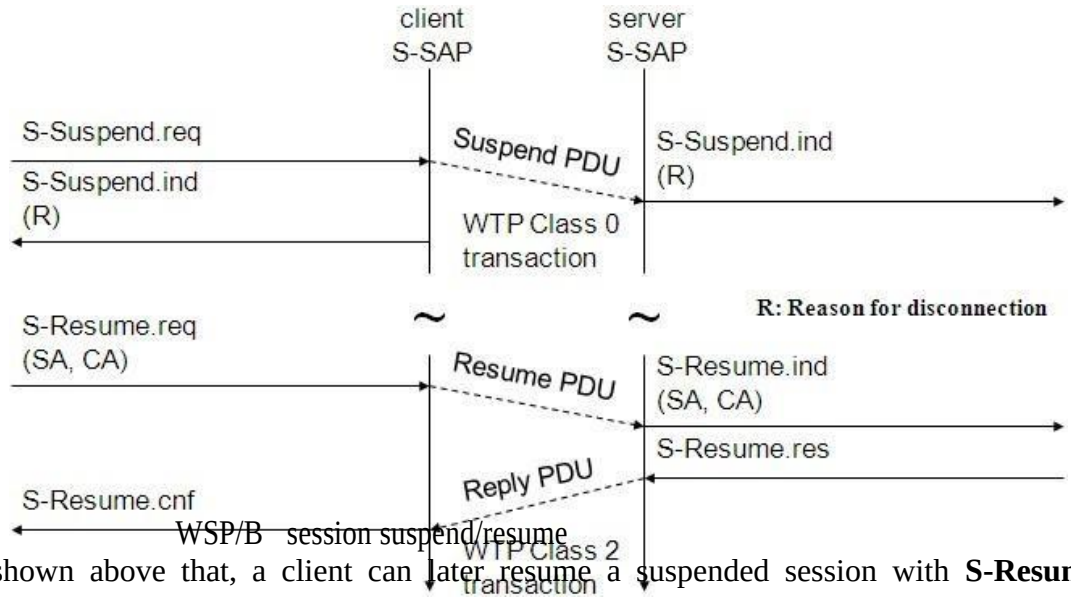
The first example of session establishment of WSP/B using WTS class 2 transactions is shown below:



With the **S-Connect.req** primitive, a client can request a new session. Parameters are the **server address (SA)**, the **client address (CA)**, and the optional **client header (CH)** and **requested capabilities (RC)**. The session layer directly uses the addressing scheme of the layerbelow. WTP transfers the **connect PDU** to the server S-SAP where an **S-Connect.ind** primitive indicates a new session. Parameters are the same, but now the capabilities are mandatory. If the server accepts the new session it answers with an **S-Connect.res**, parameters are an optional **server header (SH)** with the same function as the client header and the **negotiated capabilities (NC)** needed for capability negotiation. WTP now transfers the **connreply PDU** back to the client; **S-Connect.cnf** confirms the session establishment and includes the **server header** (if present) and the **negotiated capabilities** from the server. WSP/B includes several procedures to refuse a session or to abort session establishment.



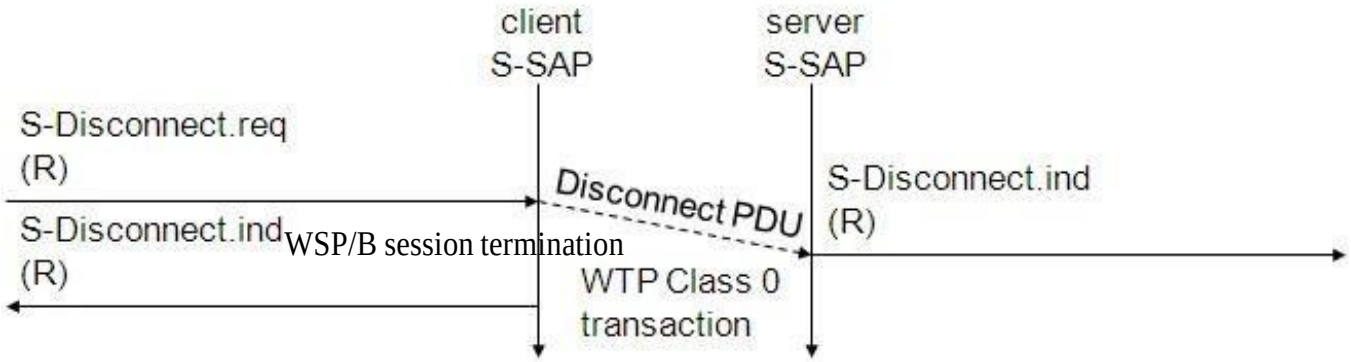
A very useful feature of WSP/B **session suspension** and **session resume** is shown below. A clientcan suspend the session because of several reasons. Session suspension will automatically abort all data transmission and freeze the current state of the session on the client and server side. A client suspends a session with **S-Suspend.req**, WTP transfers the **suspend PDU** to the server with a class 0 transaction, i.e., unconfirmed and unreliable. WSP/B will signal the suspension with **S-Suspend.ind** on the client and server side. The only parameter is the **reason R** for suspension. Reasons can be a user request or a suspension initiated by the service provider.



WSP/B session suspend/resume

Also shown above that, a client can later resume a suspended session with **S-Resume.req**. Parameters are **server address (SA)** and **client address (CA)**. Resuming a session is a confirmed operation. It is up to the server's operator how long this state is conserved.

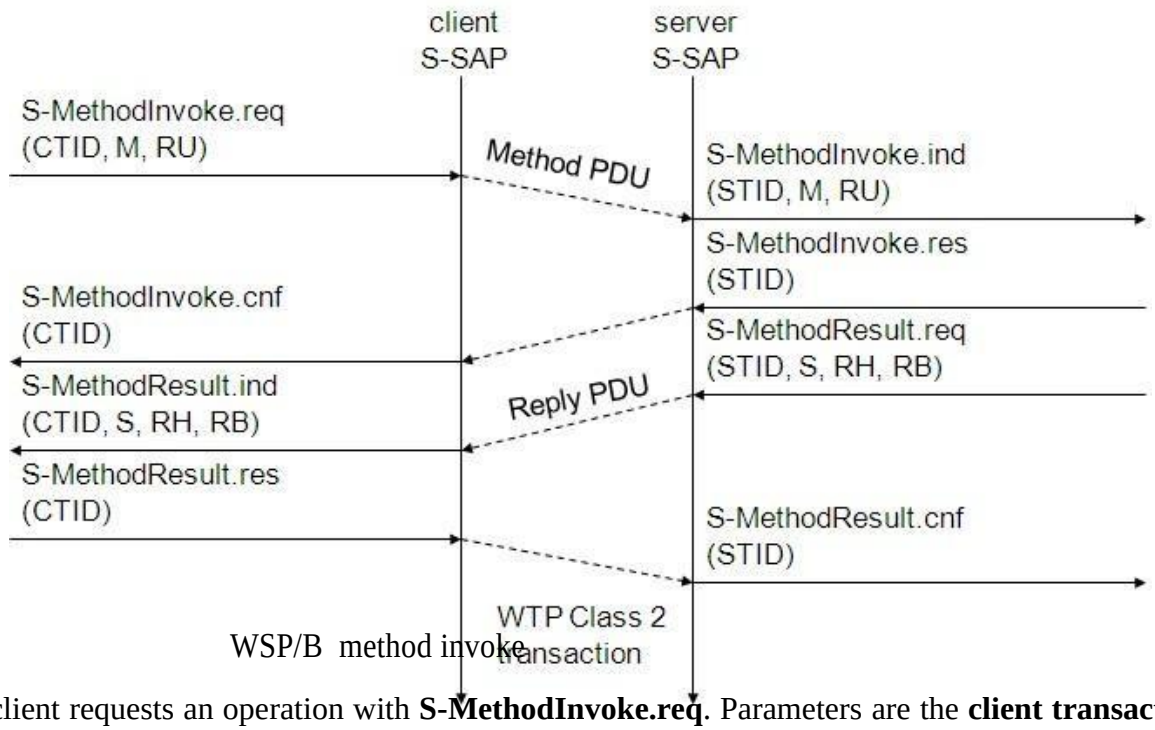
Terminating a session is done by using the **S-Disconnect.req** service primitive as shown below.



WSP/B session termination

This primitive aborts all current method or push transactions used to transfer data. Disconnection is indicated on both sides using **S-Disconnect.ind**. The **reason R** for disconnection can be, e.g., network error, protocol error, peer request, congestion, and maximum SDU size exceeded.

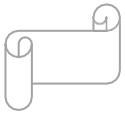
The **S-MethodInvoke** primitive is used to request that an operation is executed by the server. The result, if any, is sent back using the **S-MethodResult** primitive as shown below:



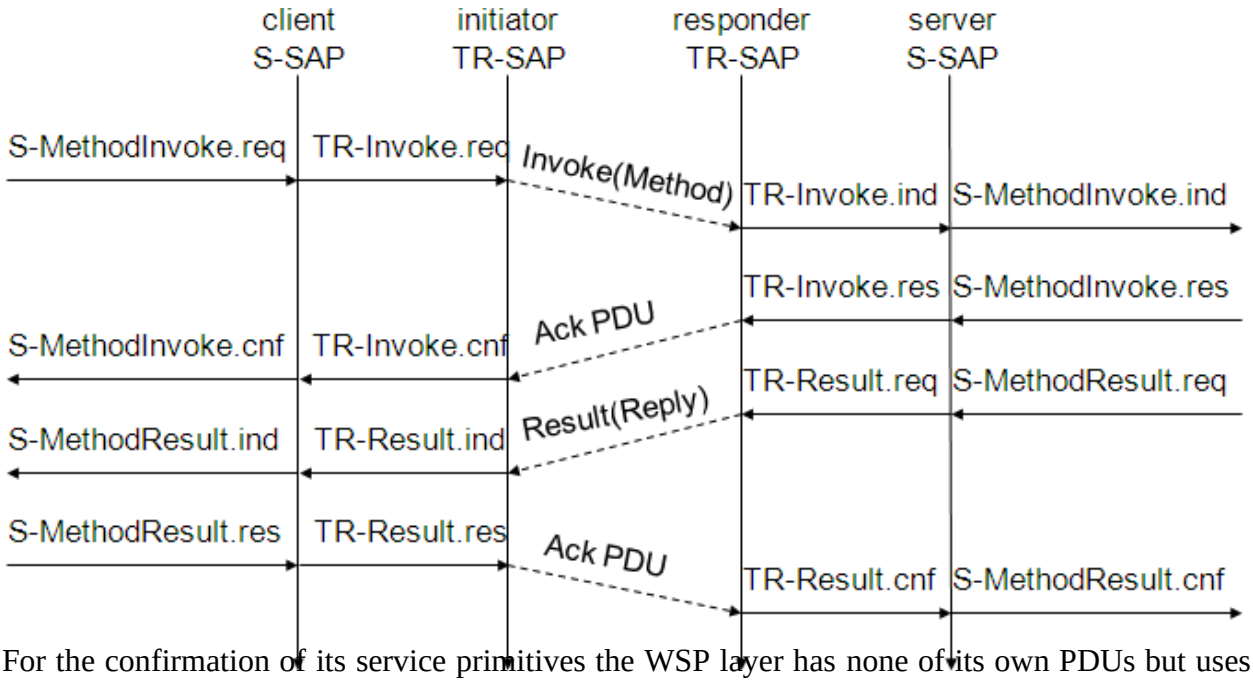
WSP/B method invoke

A client requests an operation with **S-MethodInvoke.req**. Parameters are the **client transaction identifier CTID** to distinguish between pending transactions, the **method M** identifying the requested operation at the server, and the **request URI** (Uniform Resource Identifier **RU**. The WTP class 2 transaction service now transports the **method PDU** to the server. A method PDU can be either a get PDU or a post PDU.

On the server's side, **S-MethodInvoke.ind** indicates the request. In this case, a **server transaction identifier STID** distinguishes between pending transactions. The server confirms the request, so WSP/B does not generate a new PDU but relies on the lower WTP layer. Similarly, the result of the request is sent back to the client using the **SMethodResult** primitive. Additional parameters are now the **status (S)**, the **response header (RH)**, and the **response body (RB)**.

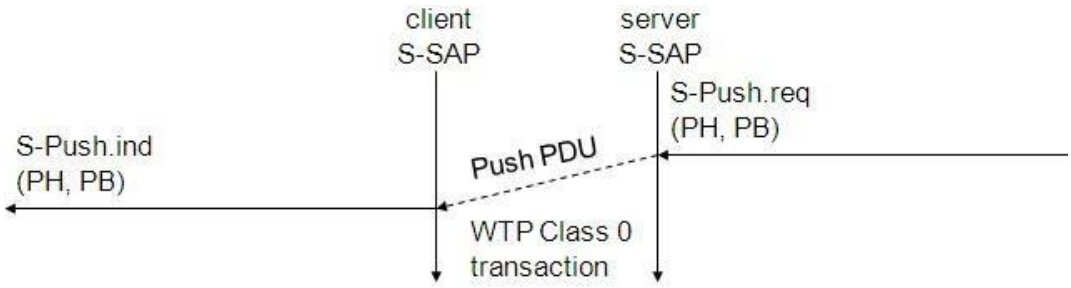


WSP does not introduce PDUs or service primitives just for the sake of symmetric and aesthetic protocol architecture. The following figure shows how WSP (thus also WSP/B) uses the underlying WTP services for its purposes. The **S-MethodInvoke.req** primitive triggers the **TR-Invoke.req** primitive, the parameters of the WSP layer are the user data of the WTP layer. The **invoke PDU** of the WTP layer carries the **method PDU** of the WSP layer inside.



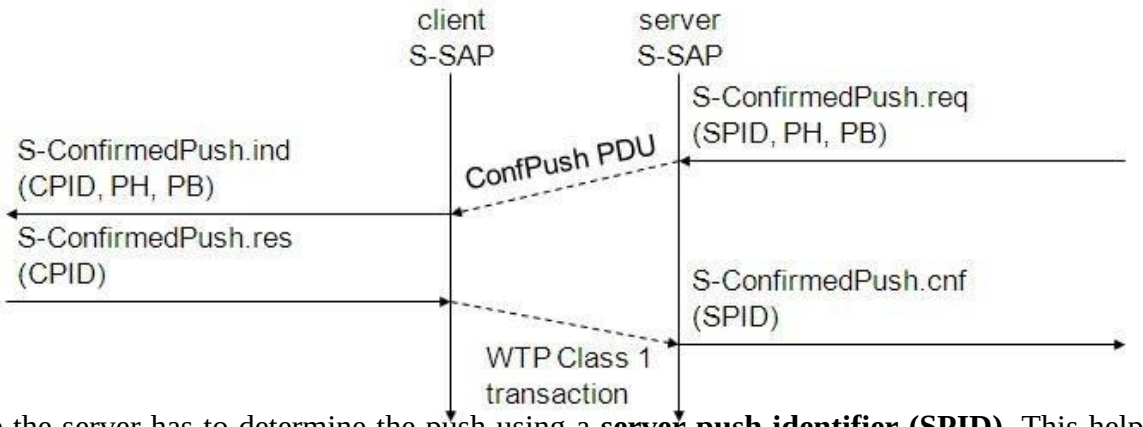
For the confirmation of its service primitives the WSP layer has none of its own PDUs but uses the **acknowledgement PDUs** of the WTP layer. **S-MethodInvoke.res** triggers **TR-Invoke.res**, the **ack PDU** is transferred to the initiator, here **TR-Invoke.cnf** confirms the invoke service and triggers the **S-MethodInvoke.cnf** primitive which confirms the method invocation service. This mingling of layers saves a lot of redundant data flow but still allows a separation of the tasks between the two layers.

With the help of push primitives, a server can push data towards a client if allowed. The simplest push mechanism is the non-confirmed push as shown below.



The server sends unsolicited data with the **S-Push.req** primitive to the client. Parameters are the **push header (PH)** and the **push body (PB)** again, these are the header and the body known from HTTP. The unreliable, unconfirmed WTP class 0 transaction service transfers the **push PDU** to the client where **S-Push.ind** indicates the push event.

A more reliable push service offers the **S-ConfirmedPush** primitive as shown below.

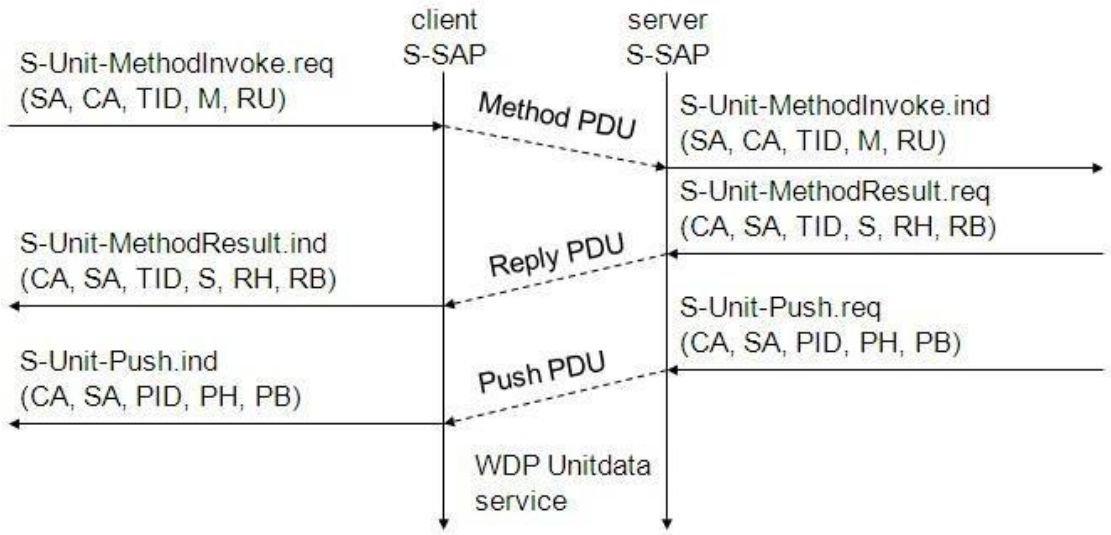


Here the server has to determine the push using a **server push identifier (SPID)**. This helps to distinguish between different pending pushes. The reliable WTP class 1 transaction service is now used to transfer the **confpush PDU** to the client. On the client's side a **client push identifier (CPID)** is used to distinguish between different pending pushes.

WSP/B as connectionless session service

WSP/B could be run on top of the connectionless, unreliable WDP service. As an alternative to WDP, WTLS can always be used if security is required. The service primitives are directly mapped onto each other. The following figure shows the three service primitives available for connectionless session service: **S-Unit-MethodInvoke.req** to request an operation on a server, **S-Unit-MethodResult.req** to return results to a client, and **S-Unit-Push.req** to push data onto a client. Transfer of the PDUs (**method, reply** and **push**) is done with the help of the standard unreliable datagram transfer service of WDP.

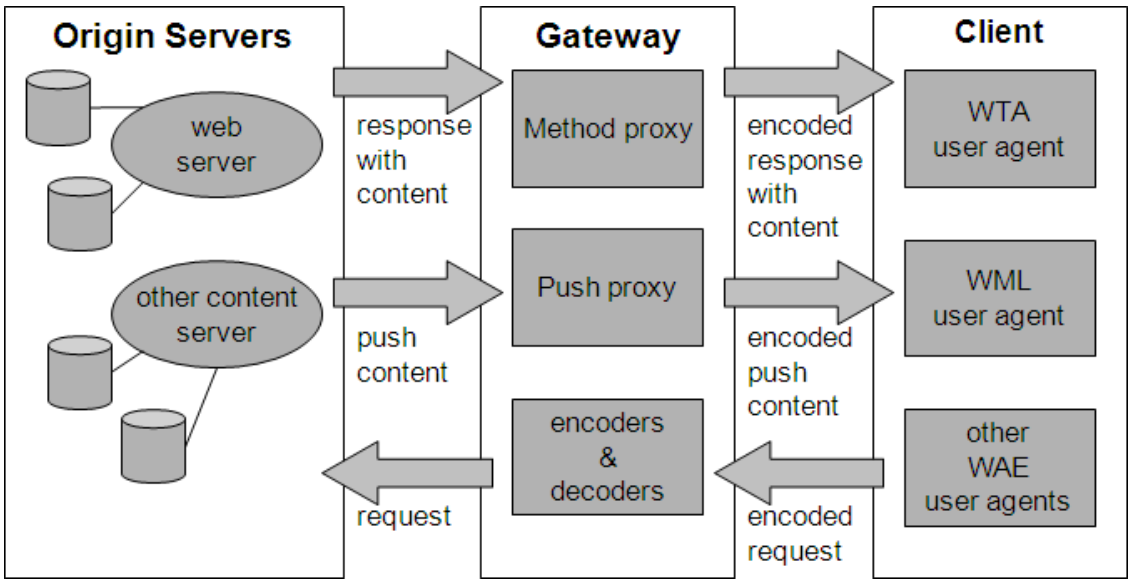
Besides the server address **(SA)**, the client address **(CA)**, the method **(M)**, and the request URI **(RU)**, the user of the **S-Unit-MethodInvoke.req** primitive can determine a transaction identifier **(TID)** to distinguish between different transactions on the user level. TID is communicated transparently from service user to service user.



The function of the **S-Unit-MethodResult** primitive remains the same as explained above: the **status (S), response header (RH)**, and **response body (RB)** represent the result of the operation. The **S-Unit-Push** primitive has the parameters **client address (CA)**, **server address (SA), push identifier (PID), push header (PH)**, and **push body (PB)**.

Wireless application environment (WAE)

The main idea behind the wireless application environment (WAE) is to create a general-purpose application environment based mainly on existing technologies and philosophies of the world wide web. One global goal of the WAE is to minimize over-the-air traffic and resource consumption on the handheld device, which is reflected in the logical model shown below:



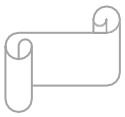
A **client** issues an encoded request for an operation on a remote server. Encoding is necessary to minimize data sent over the air and to save resources on the handheld device. Decoders in a **gateway** now translate this encoded request into a standard request as understood by the **origin servers**. This could be a request to get a web page to set up a call. The gateway transfers this request to the appropriate origin server as if it came from a standard client. Origin servers could be standard web servers running HTTP and generating content using scripts, providing pages using a database, or applying any other (proprietary) technology.

The origin servers will respond to the request. The gateway now encodes the response and its content (if there is any) and transfers the encoded response with the content to the client. The WAE logical model not only includes this standard request/response scheme, but it also includes push services. Then an origin server pushes content to the gateway. The gateway encodes the pushed content and transmits the encoded push content to the client. Several user agents can reside within a client. User agents include such items as: browsers, phonebooks, message editors etc. WAE does not specify the number of user agents or their functionality, but assumes a basic **WML user agent** that supports WML, WMLscript, or both (i.e., a 'WML browser'). However, one more user agent has been specified with its fundamental services, the **WTA user agent**. This user agent handles access to, and interaction with, mobile telephone features (such as call control). As over time many vendor dependent user agents may develop, the standard defines **a user agent profile (UAProf)**, which describes the capabilities of a user agent.

Wireless Markup Language (WML)

The **wireless markup language (WML)** is based on the standard HTML known from the www and on HDML. WML is specified as an XML document type. Several constraints of wireless handheld devices had to be taken into account, when designing WML.

WML follows a deck and card metaphor. A WML document is made up of multiple **cards**. Cards can be grouped together into a **deck**. A WML deck is similar to an HTML page, in that it is identified by a URL and is the unit of content transmission. A user navigates with the WML browser through a series of WML cards, reviews the contents, enters requested data, makes choices etc. The WML browser fetches decks as required from origin servers. Either these decks can be static files on the server or they can be dynamically generated.WML describes the intent of interaction in an abstract manner. The user agent on a handheld device has to decide how to



best present all elements of a card. This presentation depends much on the capabilities of the device.

WML includes several basic features:

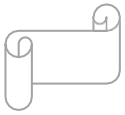
- **Text and images:** WML gives, as do other mark-up languages, hints how text and images can be presented to a user
- **User interaction:** WML supports different elements for user input. Examples are: text entry controls for text or password entry, option selections or controls for task invocation. **Navigation:** As with HTML browsers, WML offers a history mechanism with navigation through the browsing history, hyperlinks and other intercard navigation elements.
- **Context management:** WML allows for saving the state between different decks without server interaction, i.e., variable state can last longer than a single deck, and so state can be shared across different decks.

WML script

WMLScript complements to WML and provides a general scripting capability in the WAP architecture. While all WML content is static (after loading on the client), WMLScript offers several capabilities not supported by WML:

- **Validity check of user input:** before user input is sent to a server, WMLScript can check the validity and save bandwidth and latency in case of an error.
- **Access to device facilities:** WMLScript offers functions to access hardware components and software functions of the device.
- **Local user interaction:** Without introducing round-trip delays, WMLScript can directly and locally interact with a user, show messages or prompt for input.
- **Extensions to the device software:** With the help of WMLScript a device can be configuredand new functionality can be added even after deployment.

WMLScript is based on JavaScript, but adapted to the wireless environment. WMLScript is event-based, i.e., a script may be invoked in response to certain user or environment events. WMLScript also has full access to the state model of WML, i.e., WMLScript can set and read WML variables. WMLScript provides many features known from standard programming languages such as functions, expressions, or while, if, for, return etc. The WAP Forum has specified several **standard libraries** for WMLScript (WAP Forum, 2000i). These libraries provide access to the core functionality of a WAP client so they, must be available in the client's scripting environment. The six libraries defined are **Lang, Float, String, URL, WML browser** and **Dialogs.**



BLUETOOTH

"Bluetooth" was the nickname of Harald Blåtland II, king of Denmark from 940 to 981, who united all of Denmark and part of Norway under his rule. **Bluetooth** is a proprietary open wireless technology standard for exchanging data over short distances (using short wavelength radio transmissions in the ISM band from 2400-2480 MHz) from fixed and mobile devices, creating personal area networks (PANs) with high levels of security. The Bluetooth technology aims at so-called **ad-hoc piconets**, which are local area networks with a very limited coverage and without the need for an infrastructure.

Bluetooth Features

- Bluetooth is wireless and automatic. You don't have to keep track of cables, connectors, and connections, and you don't need to do anything special to initiate communications. Devices find each other automatically and start conversing without user input, expect where authentication is required; for example, users must log in to use their email accounts.
- Bluetooth is inexpensive. Market analysts peg the cost to incorporate Bluetooth technology into a PDA, cell phone, or other product at a minimum cost.
- The ISM band that Bluetooth uses is regulated, but unlicensed. Governments have converged on a single standard, so it's possible to use the same devices virtually wherever you travel, and you don't need to obtain legal permission in advance to begin using the technology.
- Bluetooth handles both data and voice. Its ability to handle both kinds of transmissions simultaneously makes possible such innovations as a mobile hands-free headset for voice with applications that print to fax, and that synchronize the address books on your PDA, your laptop, and your cell phone.
- Signals are omni-directional and can pass through walls and briefcases. Communicating devices don't need to be aligned and don't need an unobstructed line of sight like infrared.
- Bluetooth uses frequency hopping. Its spread spectrum approach greatly reduces the risk that communications will be intercepted.

Bluetooth Applications

- File transfer.
- Ad-hoc networking: Communicating devices can spontaneously form a community of networks that persists only as long as it's needed

- Device synchronization: Seamless connectivity among PDAs, computers, and mobile phones allows applications to update information on multiple devices automatically when data on any one device changes.
- Peripheral connectivity.
- Car kits: Hands-free packages enable users to access phones and other devices without taking their hands off the steering wheel
- Mobile payments: Your Bluetooth-enabled phone can communicate with a Bluetooth-enabled vending machine to buy a can of Diet Pepsi, and put the charge on your phone bill.

The 802.11b protocol is designed to connect relatively large devices with lots of power and speed, such as desktops and laptops, where devices communicate at up to 11 Mbit/sec, at greater distances (up to 300 feet, or 100 meters). By contrast, Bluetooth is designed to connect small devices like PDAs, mobile phones, and peripherals at slower speeds (1 Mbit/sec), within a shorter range (30 feet, or 10 meters), which reduces power requirements. Another major difference is that 802.11b wasn't designed for voice communications, while any Bluetooth connection can support both data and voice communications.

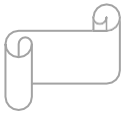
User scenarios

Many different user scenarios can be imagined for wireless piconets or WPANs:

Connection of peripheral devices: Today, most devices are connected to a desktop computer via wires (e.g., keyboard, mouse, joystick, headset, speakers). This type of connection has several disadvantages: each device has its own type of cable, different plugs are needed, wires block office space. In a wireless network, no wires are needed for data transmission. However, batteries now have to replace the power supply, as the wires not only transfer data but also supply the peripheral devices with power.

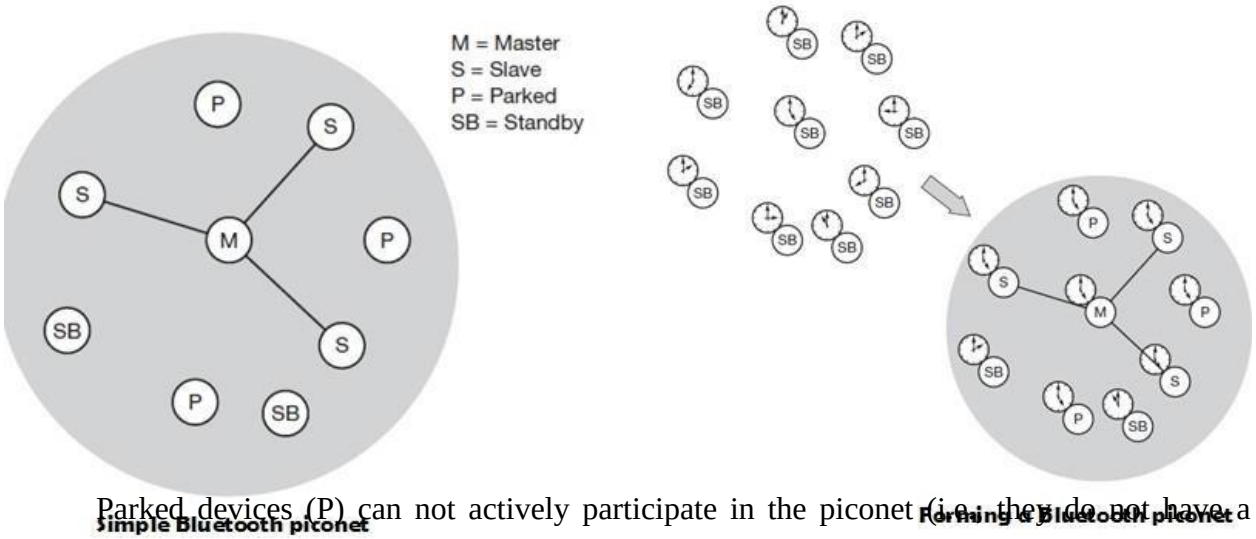
Support of ad-hoc networking: Imagine several people coming together, discussing issues, exchanging data (schedules, sales figures etc.). For instance, students might join a lecture, with the teacher distributing data to their personal digital assistants (PDAs). Wireless networks can support this type of interaction; small devices might not have WLAN adapters following the IEEE 802.11 standard, but cheaper Bluetooth chips built in.

Bridging of networks: Using wireless piconets, a mobile phone can be connected to a PDA or laptop in a simple way. Mobile phones will not have full WLAN adapters built in, but could have a Bluetooth chip. The mobile phone can then act as a bridge between the local piconet and, e.g., the global GSM network.



Networking in Bluetooth

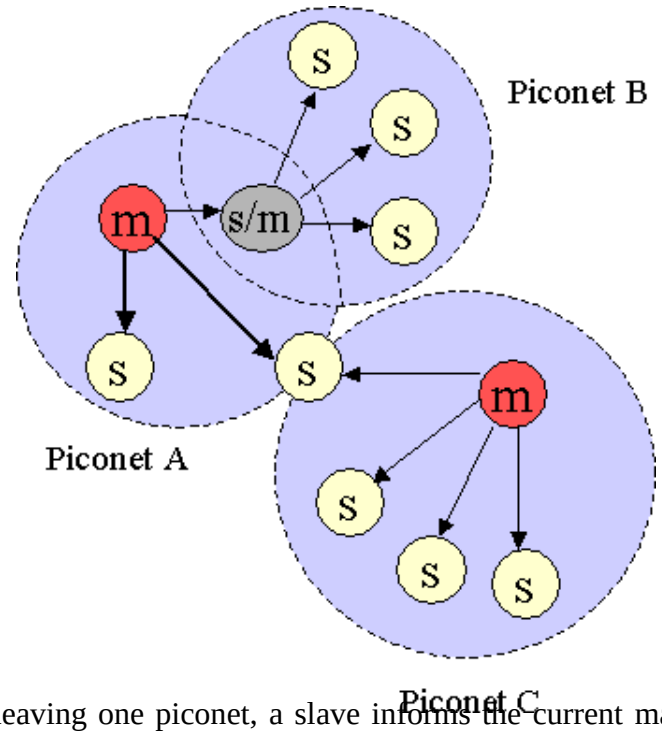
Bluetooth operates on 79 channels in the 2.4 GHz band with 1 MHz carrier spacing. Each device performs frequency hopping with 1,600 hops/s in a pseudo random fashion. A piconet is a collection of Bluetooth devices which are synchronized to the same hopping sequence. One device in the piconet can act as **master** (M), all other devices connected to the master must actas **slaves** (S). The master determines the hopping pattern in the piconet and the slaves have to synchronize to this pattern. Each piconet has a unique hopping pattern. If a device wants to participate it has to synchronize to this. A typical piconet is shown below:



Parked devices (P) can not actively participate in the piconet (i.e., they do not have a connection), but are known and can be reactivated within some milliseconds. Devices in stand-by (SB) do not participate in the piconet. Each piconet has exactly one master and up to seven simultaneous slaves. More than 200 devices can be parked. The first step in forming a piconet involves a master sending its clock and device ID. All the Bluetooth devices have the same capability to become a master or a slave and two or three devices are sufficient to form a piconet. The unit establishing the piconet automatically becomes the master, all other devices will be slaves. The hopping pattern is determined by the device ID, a 48-bit worldwide unique identifier.

The phase in the hopping pattern is determined by the master's clock. After adjusting the internal clock according to the master a device may participate in the piconet. All active devices are assigned a 3-bit **active member address** (AMA). All parked devices use an 8-bit **parked member address** (PMA). Devices in stand-by do not need an address.

A device in one piconet can communicate to another device in another piconet, forming a **scatternet**. A master in one piconet may be a slave in another piconet. Both piconets use a different hopping sequence, always determined by the master of the piconet. Bluetooth applies **FH-CDMA** for separation of piconets. A collision occurs if two or more piconets use the same carrier frequency at the same time. This will probably happen as the hopping sequences are not coordinated. If a device wants to participate in more than one piconet, it has to synchronize to the hopping sequence of the piconet it wants to take part in. If a device acts as slave in one piconet, it simply starts to synchronize with the hopping sequence of the piconet it wants to join. After synchronization, it acts as a slave in this piconet and no longer participates in its former piconet. To enable synchronization, a slave has to know the identity of the master that determines the hopping sequence of a piconet. Before leaving one piconet, a slave informs the current master that it will be unavailable for a certain amount of time. The remaining devices in the piconet continue to communicate as usual.



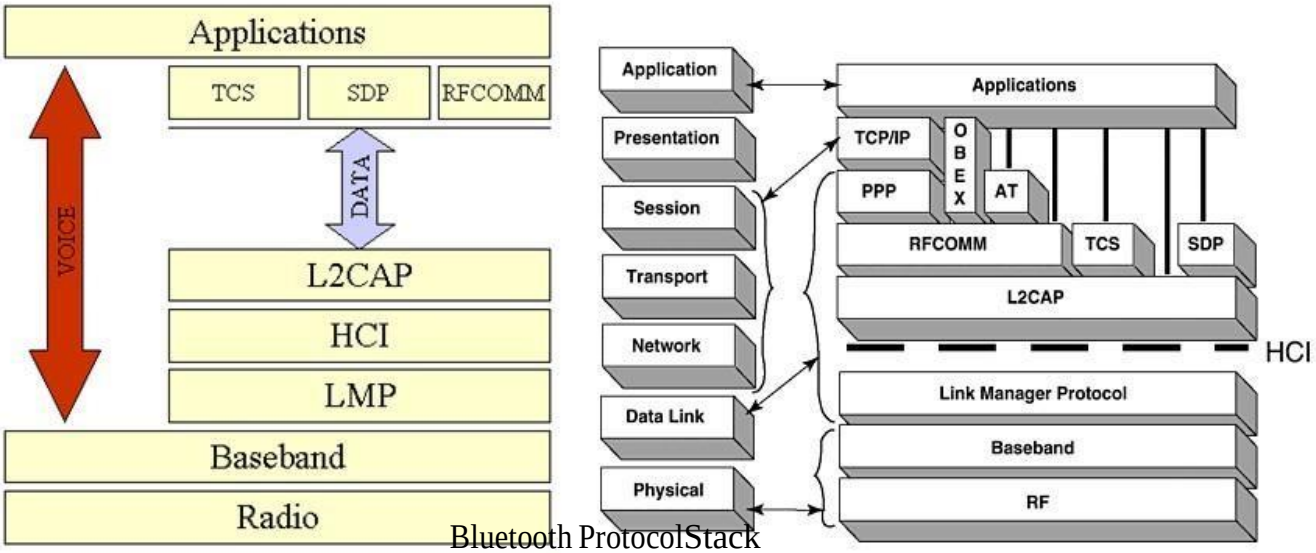
Bluetooth Protocol Stack

The Bluetooth protocol stack can be divided into a **core specification**, which describes the protocols from physical layer to the data link control together with management functions, and **profile specifications** describing many protocols and functions needed to adapt the wireless Bluetooth technology to legacy and new applications.

A high-level view of the architecture is shown. The responsibilities of the layers in this stack are as follows:

- The radio layer is the physical wireless connection. To avoid interference with other devices that communicate in the ISM band, the modulation is based on fast frequency hopping. Bluetooth divides the 2.4 GHz frequency band into 79 channels 1 MHz apart (from 2.402 to 2.480 GHz), and uses this spread spectrum to hop from one channel to another, up to 1600

times a second. The standard wavelength range is 10 cm to 10 m, and can be extended to100 m by increasing transmission power.



Bluetooth ProtocolStack

- The baseband layer is responsible for controlling and sending data packets over the radio link. It provides transmission channels for both data and voice. The baseband layer maintains Synchronous Connection-Oriented (SCO) links for voice and Asynchronous Connectionless (ACL) links for data. SCO packets are never retransmitted but ACL packets are, to ensure data integrity. SCO links are point-to-point symmetric connections, where time slots are reserved to guarantee timely transmission. A slave device is allowed to respond during the time slot immediately following an SCO transmission from the master. A master can support up to three SCO links to a single slave or to multiple slaves, and a single slave can support up to two SCO links to different slaves. Data transmissions on ACL links, on the other hand, are established on a per-slot basis (using slots not reserved for SCO links). ACL links support point-to-multipoint transmissions. After an ACL transmission from the master, only a slave addressed specifically may respond during the next time slot; if no device is addressed, the message is treated as a broadcast.
- The Link Manager Protocol (LMP) uses the links set up by the baseband to establish connections and manage piconets. Responsibilities of the LMP also include authentication and security services, and monitoring of service quality.
- The Host Controller Interface (HCI) is the dividing line between software and hardware. The L2CAP and layers above it are currently implemented in software, and the LMP and lower layers are in hardware. The HCI is the driver interface for the physical bus that connects these two components. The HCI may not be required. The L2CAP may be accessed directly

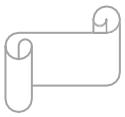
by the application, or through certain support protocols provided to ease the burden on application programmers.

- The Logical Link Control and Adaptation Protocol (L2CAP) receives application data and adapts it to the Bluetooth format. Quality of Service (QoS) parameters are exchanged at this layer.

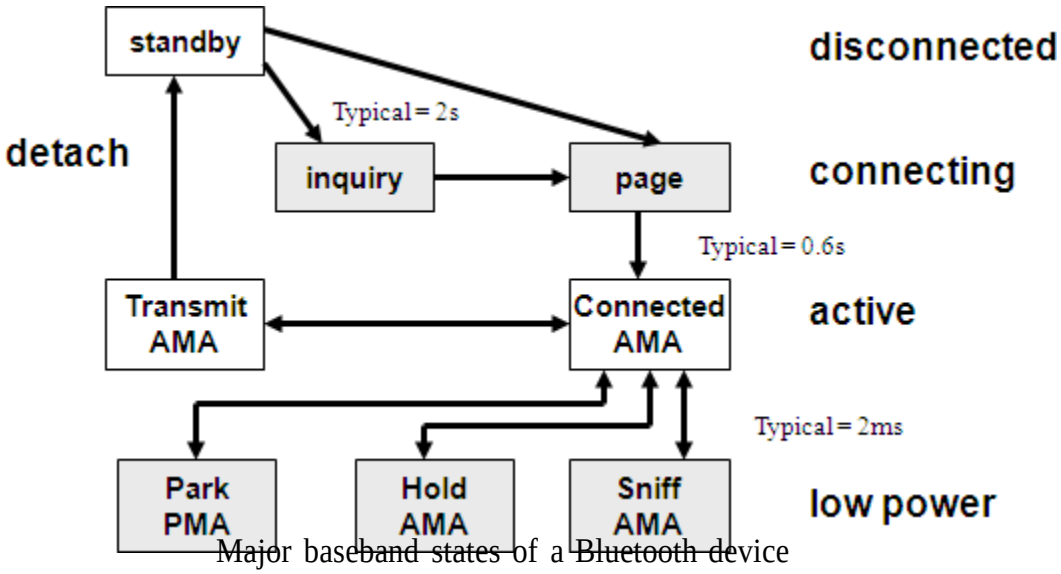
Link Manager Protocol

The link manager protocol (LMP) manages various aspects of the radio link between a master and a slave and the current parameter setting of the devices. LMP enhances baseband functionality, but higher layers can still directly access the baseband. The following groups of functions are covered by the LMP:

- **Authentication, pairing, and encryption**: Although basic authentication is handled in the baseband, LMP has to control the exchange of random numbers and signed responses. LMP is not directly involved in the encryption process, but sets the encryption mode (no encryption, point-to-point, or broadcast), key size, and random speed.
- **Synchronization:** Precise synchronization is of major importance within a Bluetooth network. The clock offset is updated each time a packet is received from the master.
- **Capability negotiation:** Not only the version of the LMP can be exchanged but also information about the supported features. Not all Bluetooth devices will support all features that are described in the standard, so devices have to agree the usage of, e.g., multi-slot packets, encryption, SCO links, voice encoding, park/sniff/hold mode, HV2/HV3 packets etc.
- **Quality of service negotiation:** Different parameters control the QoS of a Bluetooth device at these lower layers. The poll interval, i.e., the maximum time between transmissions from a master to a particular slave, controls the latency and transfer capacity. A master can also limit the number of slots available for slaves' answers to increase its own bandwidth.
- **Power control:** A Bluetooth device can measure the received signal strength. Depending on this signal level the device can direct the sender of the measured signal to increase or decrease its transmit power.
- **Link supervision:** LMP has to control the activity of a link, it may set up new SCO links, or itmay declare the failure of a link.
- **State and transmission mode change:** Devices might switch the master/slave role, detach themselves from a connection, or change the operating mode



Bluetooth defines several low-power states for a device. The following figure shows the major states of a Bluetooth device and typical transitions. Every device, which is currently not participating in a piconet (and not switched off), is in **standby** mode. This is a low-power mode where only the native clock is running. The next step towards the **inquiry** mode can happen in two different ways. Either a device wants to establish a piconet or a device just wants to listen to see if something is going on.



Major baseband states of a Bluetooth device

- A device wants to establish a piconet: A user of the device wants to scan for other devices in the radio range. The device starts the inquiry procedure by sending an inquiry access code (IAC) that is common to all Bluetooth devices. The IAC is broadcast over 32 so-called wake-up carriers in turn.
- Devices in standby that listen periodically: Devices in standby may enter the inquiry mode periodically to search for IAC messages on the wake-up carriers. As soon as a device detects an inquiry it returns a packet containing its device address and timing information required by the master to initiate a connection. From that moment on, the device acts as slave.

If the inquiry was successful, a device enters the page mode. The inquiry phase is not coordinated, so it may take a while before the inquiry is successful. After a while, a Bluetooth device sees all the devices in its radio range.

During the **page** state two different roles are defined. After finding all required devices the master is able to set up connections to each device, i.e., setting up a piconet. As soon as a device synchronizes to the hopping pattern of the piconet it also enters the connection state. The connection state comprises the active state and the low power states: **park, sniff,** and **hold.** In the **active** state the slave participates in the piconet by listening, transmitting, and receiving. ACL and SCO links can be used. A master periodically synchronizes with all slaves. All devices being active must have the 3-bit **active member address** (AMA). To save battery power, a Bluetooth device can go into one of three low power states:

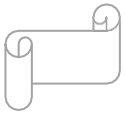
- **Sniff state:** The sniff state has the highest power consumption of the low power states. Here, the device listens to the piconet at a reduced rate (not on every other slot as is the case in the active state). The interval for listening into the medium can be programed and is application dependent. The master designates a reduced number of slots for transmission to slaves in sniff state. However, the device keeps its AMA.
- **Hold state:** The device does not release its AMA but stops ACL transmission. A slave may still exchange SCO packets. If there is no activity in the piconet, the slave may either reduce power consumption or participate in another piconet.
- **Park state:** In this state the device has the lowest duty cycle and the lowest power consumption. The device releases its AMA and receives a parked member address (PMA). The device is still a member of the piconet, but gives room for another device to become active (AMA is only 3 bit, PMA 8 bit). Parked devices are still FH synchronized and wake up at certain beacon intervals for re-synchronization. All PDUs sent to parked slaves are broadcast.

L2CAP

The logical link control and adaptation protocol (L2CAP) is a data link control protocol on top of the baseband layer offering logical channels between Bluetooth devices with QoS properties. L2CAP is available for ACLs only.

L2CAP provides three different types of logical channels that are transported via the ACL between master and slave:

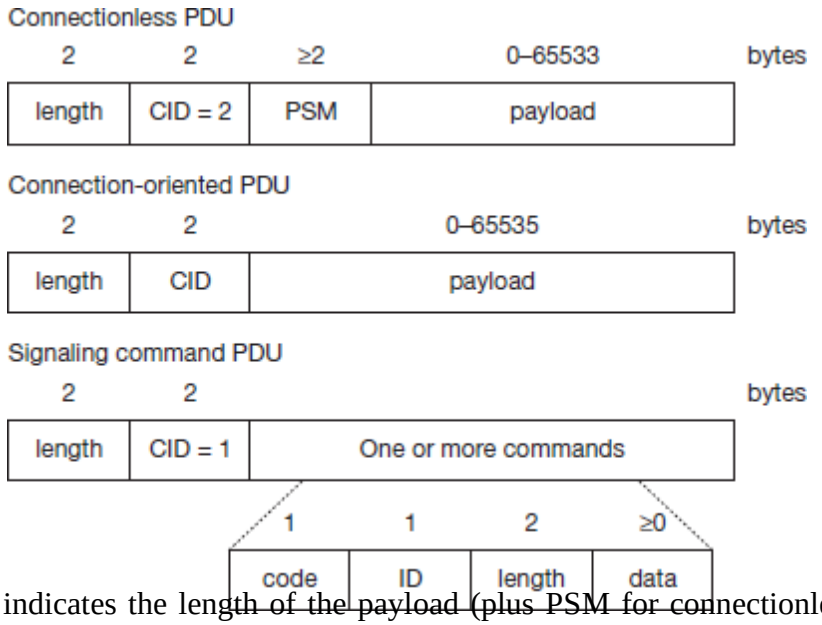
- Connectionless: These unidirectional channels are typically used for broadcasts from a master to its slave(s).
- Connection-oriented: Each channel of this type is bi-directional and supports QoS flow specifications for each direction. These flow specs follow RFC 1363 and define average/peak data rate, maximum burst size, latency, and jitter.



- Signaling: This third type of logical channel is used to exchanging signaling messages between L2CAP entities.

Each channel can be identified by its **channel identifier (CID)**. Signaling channels always use a CID value of 1, a CID value of 2 is reserved for connectionless channels. For connection-oriented channels a unique CID (>= 64) is dynamically assigned at each end of the channel to identify the connection.

The following figure shows the three packet types belonging to the three logical channel types.

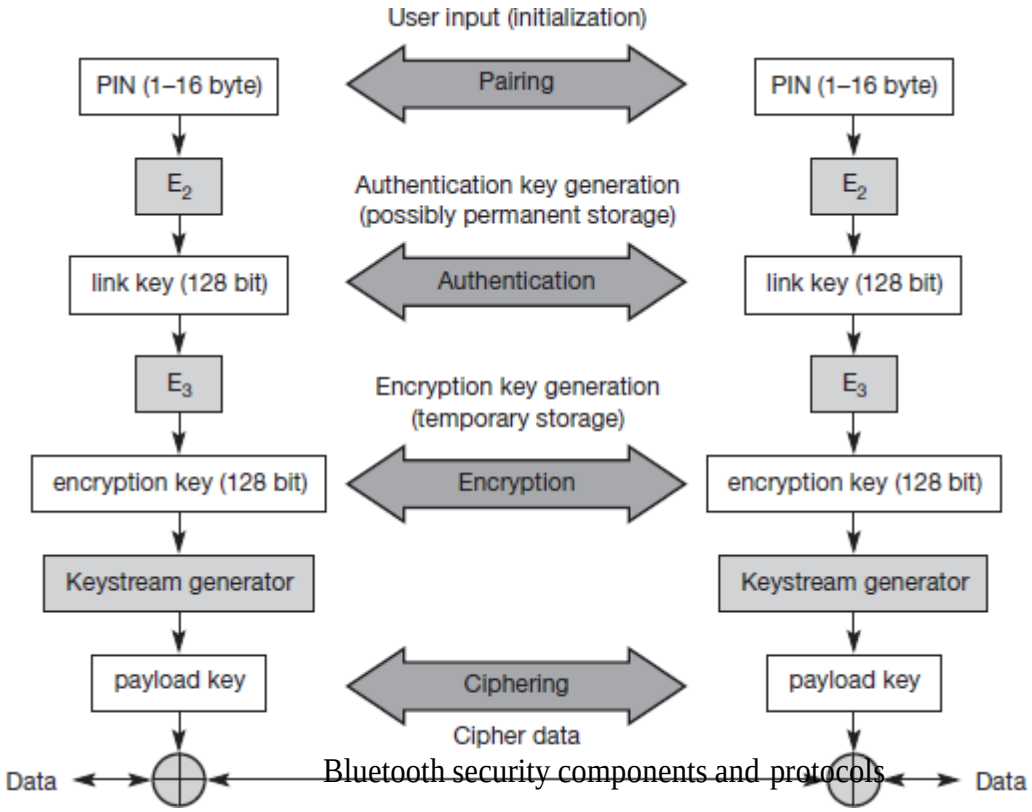


The **length** field indicates the length of the payload (plus PSM for connectionless PDUs). The **CID** has the multiplexing/demultiplexing function. For connectionless PDUs a **protocol/service multiplexor (PSM)** field is needed to identify the higher layer recipient for the payload. For connection-oriented PDUs the CID already fulfills this function. Several PSM values have been defined, e.g., 1 (SDP), 3 (RFCOMM), 5 (TCS-BIN). Values above 4096 can be assigned dynamically. The payload of the signaling PDU contains one or more **commands**. Each command has its own **code** (e.g., for command reject, connection request, disconnection response etc.) and an **ID** that matches a request with its reply. The **length** field indicates the length of the **data** field for this command.

Besides protocol multiplexing, flow specification, and group management, the L2CAP layer also provides segmentation and reassembly functions. Depending on the baseband capabilities, large packets have to be chopped into smaller segments.

Security

The main security features offered by Bluetooth include a challenge response routine for authentication, a stream cipher for encryption, and a session key generation. Each connection may require a one-way, two-way, or no authentication using the challenge-response routine. The security algorithms use the public identity of a device, a secret private user key, and an internally generated random key as input parameters. For each transaction, a new random number is generated on the Bluetooth chip. Key management is left to higher layer software. The following figure shows several steps in the security architecture of Bluetooth.



Bluetooth security components and protocols

The first step, called **pairing**, is necessary if two Bluetooth devices have never met before. To set up trust between the two devices a user can enter a secret PIN into both devices. This PIN can have a length of up to 16 byte. Based on the PIN, the device address, and random numbers, several keys can be computed which can be used as link key for **authentication**. The authentication is a challenge-response process based on the link key, a random number generated by a verifier (the device that requests authentication), and the device address of the claimat (the device that is authenticated).

Based on the link key, and again a random number an encryption key is generated during the **encryption** stage of the security architecture. This key has a maximum size of 128 bits and can be individually generated for each transmission. Based on the encryption key, the device address and the current clock a payload key is generated for ciphering user data. The payload key is a stream of pseudo-random bits. The **ciphering** process is a simple XOR of the user data and the payload key.

All Bluetooth-enabled devices must implement the Generic Access Profile, which contains all the Bluetooth protocols and possible devices. This profile defines a security model that includes three security modes:

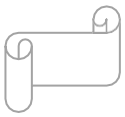
- Mode 1 is an insecure mode of operation. No security procedures are initiated.
- Mode 2 is known as service-level enforced security. When devices operate in this mode, no security procedures are initiated before the channel is established. This mode enables applications to have different access policies and run them in parallel.
- Mode 3 is known as link-level enforced security. In this mode, security procedures are initiated before link setup is complete.

Though Bluetooth offers a better security than WER in 802.11, it has several limitations. The PIN's are often fixed and some keys are permanently stored on the devices. The quality of the random number generators has not been specified.

SDP

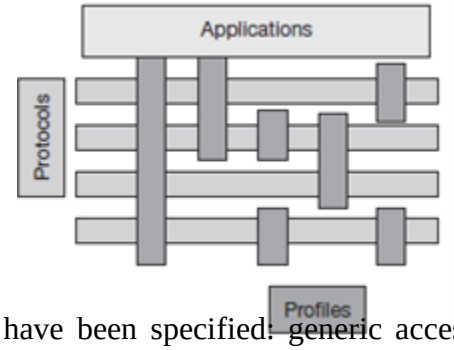
To find new services available in the radio proximity, Bluetooth defined the **service discovery protocol (SDP)**. SDP defines only the discovery of services, not their usage. Discovered services can be cached and gradual discovery is possible. All the information an SDP server has about a service is contained in a **service record**. This consists of a list of service attributes and is identified by a 32-bit service record handle.

A service attribute consists of an attribute ID and an attribute value. The 16-bit attribute ID distinguishes each service attribute from other service attributes within a service record. The attribute ID also identifies the semantics of the associated attribute value. The attribute value can be an integer, a UUID (universally unique identifier), a string, a Boolean, a URL (uniform resource locator) etc.



Bluetooth Profiles

Bluetooth profiles are intended to ensure interoperability among Bluetooth-enabled devices and applications from different manufacturers and vendors. A profile defines the roles and capabilities for specific types of applications. **Profiles** represent default solutions for a certain usage model. They use a selection of protocols and parameter set to form a basis for interoperability. Protocols can be seen as horizontal layers while profiles are vertical slices as shown below:



The following **basic profiles** have been specified: generic access, service discovery, cordless telephony, intercom, serial port, headset, dialup networking, fax, LAN access, generic object exchange, object push, file transfer, and synchronization. **Additional profiles** are: advanced audio distribution, PAN, audio video remote control, basic printing, basic imaging, extended service discovery, generic audio video distribution, hands-free, and hardcopy cable replacement. Some of the profiles are given below:

- The Generic Access Profile defines connection procedures, device discovery, and link management. It also defines procedures related to use of different security models and common format requirements for parameters accessible on the user interface level. At a minimum all Bluetooth devices must support this profile.
- The Service Discovery Application and Profile defines the features and procedures for an application in a Bluetooth device to discover services registered in other Bluetooth devices, and retrieves information related to the services.
- The Serial Port Profile defines the requirements for Bluetooth devices that need to set up connections that emulate serial cables and use the RFCOMM protocol.
- The LAN Access Profile defines how Bluetooth devices can access the services of a LAN using PPP, and shows how PPP mechanisms can be used to form a network consisting of Bluetooth devices.

- The Synchronization Profile defines the application requirements for Bluetooth devices that need to synchronize data on two or more devices.

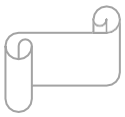
Java 2 Micro Edition (J2ME)

Sun Microsystems defines J2ME as "a highly optimized Java run-time environment targeting a wide range of consumer products, including pagers, cellular phones, screen-phones, digital set-top boxes and car navigation systems." J2ME brings the cross-platform functionality of the Java language to smaller devices, allowing mobile wireless devices to share applications. Java 2 Micro Edition maintains the qualities that Java technology has become known for:

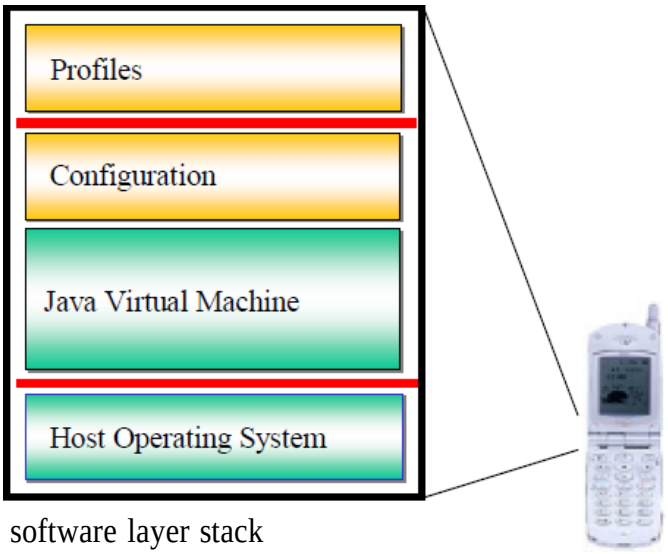
- built-in consistency across products in terms of running anywhere, anytime, on any device
- the power of a high-level object-oriented programming language with a large developer base;
- portability of code;
- safe network delivery; and
- upward scalability with J2SE and J2EE

While connected consumer devices such as cell phones, pagers, personal organizers and set-top boxes have many things in common, they are also diverse in form, function and features. Information appliances tend to be special-purpose, limited-function devices. To address this diversity, an essential requirement for J2ME is not only small size but also modularity and customizability. The J2ME architecture is modular and scalable so that it can support the kinds of flexible deployment demanded by the consumer and embedded markets. To support this kind of customizability and extensibility, two essential concepts are defined by J2ME:

- Configuration. A J2ME configuration defines a minimum platform for a "horizontal" category or grouping of devices, each with similar requirements on total memory budget and processing power. A configuration defines the Java language and virtual machine features and minimum class libraries that a device manufacturer or a content provider can expect to be available on all devices of the same category.
- Profile. A J2ME profile is layered on top of (and thus extends) a configuration. A profile addresses the specific demands of a certain "vertical" market segment or device family. The main goal of a profile is to guarantee interoperability within a certain vertical device family or domain by defining a standard Java platform for that market. Profiles typically include



class libraries that are far more domain-specific than the class libraries provided in a configuration.



J2ME software layer stack

Configurations

A configuration is a subset of profile. A configuration defines a Java platform for a "horizontal" category or grouping of devices with similar requirements on total memory budget and other hardware capabilities. More specifically, a configuration:

- specifies the Java programming language features supported,
- specifies the Java virtual machine features supported,
- specifies the basic Java libraries and APIs supported.

To avoid fragmentation, there will be a very limited number of J2ME configurations. Currently, the goal is to define two standard J2ME configurations:

- **Connected, Limited Device Configuration (CLDC)**. The market consisting of personal, mobile, connected information devices is served by the CLDC. This configuration includes some new classes, not drawn from the J2SE APIs, designed specifically to fit the needs of small-footprint devices. It is used specifically with the KVM for 16-bit or 32-bit devices with limited amounts of memory. This is the configuration (and the virtual machine) used for developing small J2ME applications.
- **Connected Device Configuration (CDC)**. The market consisting of shared, fixed, connected information devices is served by the Connected Device Configuration (CDC). To ensure upward compatibility between configurations, the CDC shall be a superset of the CLDC. This

is used with the C virtual machine (CVM) and is used for 32-bit architectures requiring more than 2 MB of memory.

MIDP Mobile Information Device Profile	PDAP Personal Digitial Assistant Profile	Personal Profile
		Personal Basis Profile
		Foundation Profile
CLDC Connected, Limited Device Configuration		CDC Connected Device Configuration
J2ME Java 2, Micro Edition		

J2ME Universe

Profiles

The J2ME framework provides the concept of a profile to make it possible to define Java platforms for specific vertical markets. Profiles can serve two distinct portability requirements:

- A profile provides a complete toolkit for implementing applications for a particular kind of device, such as a pager, set-top box, cell phone, washing machine, or interactive electronic toy.
- A profile may also be created to support a significant, coherent group of applications that might be hosted on several categories of devices.

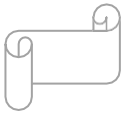
Foundation profile contains APIs of J2SE without GUIs. PersonalProfile is profile for embedded devices. Two profiles have been defined for J2ME and are built on CLDC: KJava and Mobile Information Device Profile (MIDP). These profiles are geared toward smaller devices.

MIDP 3.0 is the latest profile version, which is a profile for special-featured phones and handheld devices. It provides improved UI's, UI extensibility and interoperability between the devices. It supports multiple network interfaces in a device, IPv6, large display devices and high performance games. Development tools are used to develop MIDP applications. MIDP applications are composed of two parts:

- JAR File – Contains all of the classes and resources used by the application
- JAD File – Application descriptor, describes how to run the MIDP application

	Source Package	Sets of Java class libraries
1	Java.lang	standard java types and classes for String, Integer, Math, Thread, Security and Exception
2	Java.io	Standard java types and classes for input and output streams
3	Java.util	A set of classes such as Timers, Calenders, Dates, Hashtables, Vectors and others
4	Javax.microedition.rms	A record management system (RMS) API to retrieve and save data and limited querying capability
5	Javax.microedition.pim	Personal information management API (optional), access the device's address book
6	Javax.microedition.pki	Secure connections authenticate API's

MIDP source packages and sets of Java class libraries



K Virtual Machine

The KVM is a compact, portable Java virtual machine specifically designed from the ground up for small, resource-constrained devices. The high-level design goalfor the KVM was to create the smallest possible "complete" Java virtual machine that would maintain all the central aspects of the Java programming language, but would run in a resource-constrained device with only a few hundred kilobytes total memory budget. More specifically, the KVM was designed to be:

- small, with a static memory footprint of the virtual machine core in the range of 40 kilobytes to 80 kilobytes (depending on compilation options and the target platform,)
- clean, well-commented, and highly portable,
- modular and customizable, as "complete" and "fast" as possible without sacrificing the other design goals.

The "K" in KVM stands for "kilo." It was so named because its memory budget is measured in kilobytes (whereas desktop systems are measured in megabytes). KVM is suitable for 16/32-bit RISC/CISC microprocessors with a total memory budget of no more than a few hundred kilobytes (potentially less than 128 kilobytes). This typically applies to digital cellular phones, pagers, personal organizers, and small retail payment terminals.