Unit III

Chapter 3

Generations of Mobile Communication Technologies

Syllabus

First Generation wireless networks, Second Generation (2G) wireless cellular networks, Major 2G standards, 2.5G Wireless Networks, Third Generation 3G wireless networks, Fourth Generation 4G wireless networks, Fifth Generation 5G wireless networks.

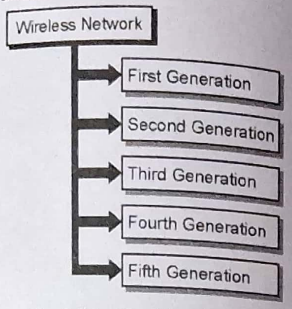
Chapter Contents

3.1 Wireless Network Generations :

- The cellular systems are classified into different evolutionary generations from first generation (1G) to fifth generation (5G).
- The **first generation wireless networks** are based on analog technology and they are used only for analog voice services.
- The **second generation wireless systems (2G)** employ digital modulation and advanced call processing capabilities.
- Typical examples include Global System for Mobile (GSM), cordless telephone (CT2) etc.
- The **third generation wireless systems (3G)** are developed to provide universal access throughout the world.
- They have used broadband ISDN to provide access to information networks like internet, communications using Voice Over Internet Protocol (VoIP), voice-activated calls etc.
- The **fourth generation wireless systems (4G)** are currently under deployment but continue to evolve.
- The next generation cellular networks have been designed to support high speed data communications traffic in addition to the voice calls.
- The new technologies and standards are being implemented so that the wireless networks can replace the fiber optic or copper cables.
- The wireless networks are used as replacement for wires within offices, buildings, homes with the use of Wireless Local Area Networks **(WLANs).**
- The **Bluetooth** modem standard can connect several devices with invisible wireless connections within a person's personal workspace.
- It was conceived as a wireless alternative to RS232 cables.
- WLANs and Bluetooth use low power levels. They don't need a license for spectrum use.
- They are used for adhoc wireless communication of voice and data anywhere in the world.
- The cellular systems are classified into three different evolution of generations.

3.1.1 Classification of Wireless Generations :

- Fig 3.1.1 shows classification of wireless network generations.



(G-3046) **Fig. 3.1.1 : Classification of wireless network generations**

3.2 First Generation Wireless Networks :

- The first generation of cellular telephony was suitable only for voice communication using analog signals.
- Now cellular technology is in the fourth generation.
- One of the important first generation mobile system used in North America is AMPS.
- The first generation of wireless mobile system was implemented in 1980's.
- The modulation scheme used was frequency modulation (FM).
- Long form of AMPS is Advanced Mobile Phone System.
- It is one of the leading analog cellular system in North America.
- It makes use of FDMA (Frequency Division Multiple Access) to separate channels in a link.

Frequency bands :

- AMPS uses the ISM 800-MHz band for its operation.
- It uses two separate channels for forward i.e. base station to mobile station and for reverse i.e. from mobile station to base station communication.

- The frequency bands allotted for the forward and reverse communication are as follows :
- Reverse Communication : 824 MHz to 849 MHz.
- Forward Communication : 869 MHz to 894 MHz.
- Each band has been divided into 832 channels. But two providers are allowed to share an area. That each provider is allowed to use 416 channels in each cell.
- Out of 416, 21 channels are used for control and the remaining 395 channels for information.

Transmission :

- AMPS makes use of FSK and FM systems for modulation. FM stands for frequency modulation while FSK is frequency shift keying.
- FM is used for the modulation of voice signals whereas FSK is used for the control channels.
- The coverage area of first generation systems was 2100 square km.
- The other 1G technologies developed to provide only analog voice communication were Nordic Mobile Telephone (NMT) and Total Access Communication System (TACS).
- 1G technology was developed only for providing the voice communication but **paging networks** also are considered as 1G technology.
- Pager system provides only one way messaging.

3.2.1 Features of First Generation :

Table 3.2.1 : Features of First Generation :

Sr. No.	Feature	Value / Description
1.	Generation	1G (1970 – 1984)
2.	Technology	Analog cellular
3.	Standard	AMPS
4.	Switching	Circuit switching
5.	Frequency band	824-894 MHz
6.	Modulation	FM
7.	Data speed	2.4 kbps
8.	Multiplexing	FDMA
9.	Core network	PSTN
10.	Service	Only voice or only message

3.2.2 Drawbacks of 1G System :

1. Poor voice quality.
2. No security.

3.3 Second Generation Networks :

- The second generation of cellular telephony was developed in order to improve the quality of communication.
- The second generation was designed for digital voice.
- 2G networks began to emerge around 1980's but their actual implementation started by 1990's.
- The second generation (2G) cellular systems provide more features as compared to the first generation (1G) systems.
- As stated earlier, the first generation cellular systems were based on analog transmission of voice on FDMA / FDD and analog FM.
- The second generation (2G) cellular systems are based on the digital modulation formats along with the TDMA/FDD and the CDMA/FDD multiple access techniques.
- The second generation mobile systems are digital systems.

3.3.1 Types of 2G Standards :

- Following are some popular second generation standards.
  1. Global System Mobile (GSM)
  2. Interim Standard 136 (IS-136)
  3. Pacific Digital Cellular (PDC)
  4. Interim standard 95 CDMA (IS-95)
- Out of these, the first three i.e. GSM, IS-136 and PDC are **TDMA** standard whereas, IS-95 is the **CDMA** standard.
- In the three TDMA standards, time axis is subdivided into many time slots.
- Each user is allotted its own time slot to transmit its data. These time slots form TDMA frames.
- The cellular traffic congestion is minimized by this technology.

1. GSM Standard :

- Out of all the 2G standards, GSM is the most popular standard.
- It is a TDMA based standard which uses eight time slots for users with each user assigned a bandwidth of 200 KHz.
- It is deployed in Europe, Asia, Australia, South America and few parts of the U.S. for cellular as well as PCS applications.

2. Interim Standard IS-136 :

- This standard is also called as North American Digital Cellular(NADC) or US Digital Cellular (USDC).
- It is a TDMA standard which uses three time slots for users with each one assigned a channel bandwidth of 30 MHz.
- It is deployed in Australia, South America and North America for cellular as well as PCS applications.

3. Pacific Digital Cellular PDC :

- It is a digital TDMA standard which is used by over 50 million people in Japan.
- It is similar to the IS-136 standard.

4. Interim Standard IS-95 :

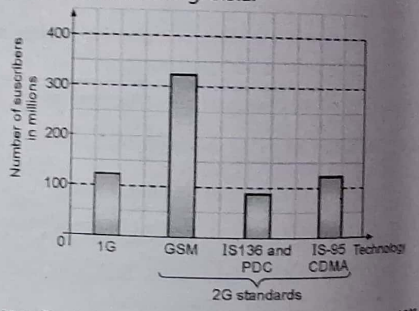
- The interim standard IS-95 is a very popular CDMA based 2G standard which is also called as **cdmaOne**.
- In the CDMA standard (IS-95) every subscriber is assigned an n-bit code and assures a high degree of security than other access methods.
- IS-95 standard does not use TDMA. Hence it allows up to 64 users to transmit simultaneously.
- All these users are coded in an orthogonal manner and they transmit simultaneously over a 1.25 MHz channel.
- It is deployed in Japan, Korea, China, South America and Australia for cellular as well as PCS applications.
- The simultaneous transmission from several users can be differentiated from each other by a unique code word that has been assigned to each subscriber.
- These 2G standards are the first set of digital standards that make use of digital modulation and digital signal processing techniques in the handsets as well as the base stations.
- The handset manufactures are producing handsets, that support both TDMA and CDMA standards for cellular and PCS applications.

3.3.2 Features of TDMA Standard :

- A few important features of the TDMA are :
  1. Operates at faster data rates.
  2. Utilizes the spectrum efficiently.
  3. Needs synchronization for its operation.
  4. It has guard intervals for better operation.
  5. The TDMA standards use half duplex methods.

3.3.3 Graphical Comparison of 1G and 2G :

- A simple comparison with respect to the number of mobile users for the 1G and 2G standards in 2001 is shown in Fig. 3.3.1.



(G-2544) **Fig. 3.3.1 : Growth of number of users in late 2001**

- The GSM-2G standard had the highest number of users in 2001.
- 2G networks are designed for development of conventional mobile service.
- In-order to provide fixed telephone service to businesses in the developing nations and the residential areas the newer cellular systems are being installed.

- In the countries that have poor telecommunication infrastructure, it is cost effective to provide **plain old telephone service (POTS)** with the cellular phones.
- Advantage of using 2G network over 1G is that digital encryption of phone conversation is possible.
- With the help of 2G technologies the spectrum efficiency has increased by three folds as compared to 1G technology.
- With the development in GSM, several countries like Japan and US decided to abandon the IS-136 and PDC standard.

3.3.4 Technical Specifications of 2G :

- Table 3.3.1 gives the technical specifications of GSM, CDMA and IS-136/PDC 2G standards.

Table 3.3.1 : Technical specifications of leading 2G technologies

Sr. No.	Parameter	GSM	IS-136/PDC	IS-95 CDMA
1.	Duplexing	FDD	FDD	FDD
2.	Multiple Access	TDMA	TDMA	CDMA
3.	Uplink frequencies	890-915 MHz (Europe) 1850-1910 MHz (US PCS)	800 MHz, 1500 MHz (Japan) 1850-1910 MHz (US PCS)	824-849 MHz (US cellular) 1850-1910 MHz (US PCS))
4.	Downlink frequencies	935-960 MHz (Europe) 1930-1990 MHz (US PCS)	869-894 MHz (US cellular) 1930-1990 MHz (US PCS) 800 MHz, 1500 MHz (Japan)	869-894 MHz (US Cellular) 1930-1990 MHz (US PCS)
5.	Carrier separation	200 kHz	30 kHz (IS-136) 25 kHz for PDC	1.25 MHz
6.	Voice channels per carrier	8	3	64
7.	Channel Data Rate	270.833 kbps	48.6 kbps (IS-136) 42 kbps for PDC	1.2288 Mchips/sec
8.	Modulation scheme	GMSK	DQPSK	QPSK
9.	Physical channel bandwidth	200 kHz	30 KHz	1.25 MHz
10.	Typical power radiated	1000/125 mW	600/200 mw	600 mW
11.	Number of users	8	6	20 to 35

- 2G networks are designed for development of conventional mobile service.
- In-order to provide fixed telephone service to businesses in the developing nations and the residential areas the newer cellular systems are being installed.
- In the countries that have poor telecommunication infrastructure, it is cost effective to provide **plain old telephone service (POTS)** with the cellular phones.
- Advantage of using 2G network over 1G is that digital encryption of phone conversation is possible.

- With the help of 2G technologies the spectrum efficiency has increased by three folds as compared to 1G technology.
- With the development in GSM, several countries like Japan and US decided to abandon the IS-136 and PDC standard.

3.3.5 Features of 2G Systems :

- Some of the important features of the 2G-mobile systems are as follows :

Sr. No.	Feature	Value / Description
1.	Generation	2G (1990)
2.	Technology	Digital Cellular Technology
3.	Standard	CDMA, TDMA and GSM
4.	Switching	Circuit / packet switching
5.	Frequency band	850 - 1900 MHz (GSM)
6.	Data speed	9.6 kbps.
7.	Multiplexing	CDMA, TDMA
8.	Modulation	GMSK
9.	Core network	PSTN
10.	Services	Digital voice, Data and SMS facility
11.	Handoff	Horizontal

Performance :

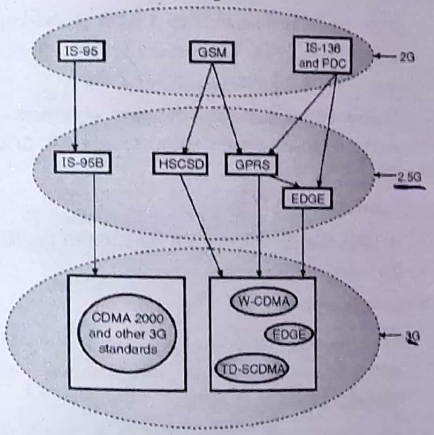
- Although 2G systems provided a huge improvement over 1G and increased the number of subscribers, the standards of 2G systems were poor.
- This is because, 2G systems were unable to handle complex data and they could not use the available bandwidth efficiently.
- Therefore the 2G standards were upgraded first to 2.5G standards and subsequently to the 3G standards.

3.4 Evolution From 2G to 2.5G Networks :

Why 2.5G ?

- The 2G digital standards were being used widely by wireless carriers for cellular and PCS applications since the 1990s.
- However these standards use circuit-switched data modems that limit the data users to a signal circuit-switched voice channel.
- Hence, in 2G standards, the data transmission rates are generally limited to the data throughout rate of an individual user.
- Each 2G standard specifies a different coding scheme and error protection algorithm for data transmission and voice transmissions.
- All the 2G networks only support single user data rates of the order of 10 kbps.
- Such a low data rate is not sufficient for sending email rapidly and for Internet browsing applications.
- With low data rates, the 2G standards can support limited internet browsing and short messaging service (SMS) with the help of switched circuit method.
- The **short messaging service (SMS)** is a an important feature of GSM.
- The subscribers can send short real time messages to the other subscribers in the same network by dialing the receiver's mobile phone number.
- In order to make the 2G standards support the increased throughput data rate requirement for modern Internet applications, new standards have been developed.
- These new standards are upgraded upon the existing 2G standards and they are called as 2.5G technology.
- They support the higher data rates by allowing the use of existing 2G handsets with some modifications and by using new base station add-ons and subscriber unit software upgrades.

- Hence the 2.5 standards can support for web browsing, e-mail traffic, mobile commerce (m-commerce), and location-based mobile services.
- The 2.5G technologies also support a popular new web browsing format language, called **Wireless Application Protocol (WAP).**
- WAP allows the web pages to be observed in a compressed format because it has been designed for small, portable hand held devices.
- For a specific wireless network the 2.5G upgrade must be compatible with the earlier 2G technology at the base station.
- A wide range of 2.5G standards have been developed in order to upgrade each 2G technology (IS-136, GSM, CDMA) for faster Internet data rates.
- Fig. 3.4.1 shows various upgrade paths for different 2G technologies.

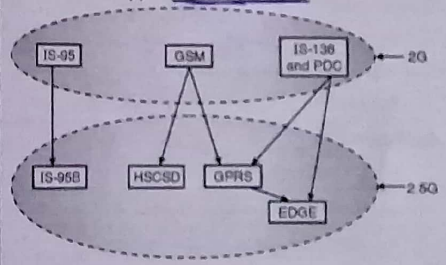


(G-1584) Fig. 3.4.1 : Various upgrade paths for 2G technologies

- In other words it shows how a 2G standard evolves into 2.5G and then into a 3G standard.
- Fig. 3.4.1 shows that CDMA IS-95 evolves into IS-95B as 2.5G standard which further evolved into CDMA 2000 as the 3G standard.
- On the other side the GSM standard evolves into HSCSD and GPRS as 2.5G standards which further evolved into W-CDMA, EDGE or and TD-SCDMA standards as shown.
- Finally IS-136 and PDC standards followed the evolution route of either GPRS--- W-CDMA, EDGE or and TD-SCDMA or GPRS --- EDGE----W-CDMA as shown in Fig. 3.4.1.
- It is necessary to choose an appropriate upgrade path for any 2G standard.
- That means the 2.5 standard corresponding to the GSM standard must have a backward compatibility with the GSM air interface standard.
- This would ensure there is no need of complete change in the base station equipments.
- Hence all the 2.5G standards provide a marginal improvement in the data rates over corresponding 2G standard.

3.4.1 Evolution for 2.5G TDMA Standards :

- The three 2G standards that are based on TDMA technique are : GSM, IS - 136 and PDC.
- As shown in Fig. 3.4.2 three different upgrade paths are developed for the GSM, out of which two also support IS - 136 and PDC.



(G-1584(a)) Fig. 3.4.2 : Evolution path for GSM to 2.5G

- Following are the three upgrade solutions for 2G TDMA standards :
  1. High speed circuit switched Data (HSCSD)
  2. General Packer Radio Service (GPRS)
  3. Enhanced Data rates for GSM Evolution (EDGE)
- These improvements increase the data rates of the 2G standards significantly and also support development of Internet ready cell phones.

3.4.1.1 HSCSD for 2.5G GSM :

Principle :

- HSCSD is a high speed circuit switched data technique.
- It allows a single mobile user to use consecutive time slots in the GSM standard.
- In GSM, one user is allowed to use only one TDMA slot, but in HSCSD a user can use more than one consecutive TDMA slots to improve the data rates.
- This standard relaxes the error control algorithm in GSM to further increase the data rates.
- The amount of increase in data rates depends on the number of consecutive time slots used by a user.
- With four consecutive TDMA slots allotted to a single user the HSCSD can increase the raw data rates up to 57.6 kbps which is a remarkable improvement over 9.6 kbps data rate of GSM.

Implementation :

- HSCSD can be employed if the service provider implements a simple software change at the GSM base station.

Applications :

- HSCSD can be used for :
  1. Dedicated streaming Internet access
  2. Real time interactive web sessions.

3.4.1.2 GPRS for 2.5G GSM and IS-136 :

- GPRS is a packet based technique which could be the next step in evolution of GSM as well as IS-136 and PDC standards as shown in Fig. 3.4.2.

Principle :

- GPRS operates by supporting a multiple user network sharing of individual channels and time slots.
- This is different than the principle of HSCSD. Due to this technique, GPRS can support many more users than HSCSD but in **a burst manner** (not continuous manner).
- The GPRS standard provides a packet network on dedicated GSM or IS-136 radio channels.

Air Interface :

- The modulation formats specified in the original 2G TDMA standards (GSM and IS-136) are retained in GPRS.
- But it uses a completely redefined air interface as compared to GSM or IS-136) for better handling of data.
- GPRS has dedicated radio channels and particular time slots that allow an always on access to the network.
- The GPRS subscribers are instructed automatically to tune to the above mentioned channels and time slots.

Data rates :

- If all the eight time slots of a GSM channel are dedicated to GPRS, it is possible for an individual user to achieve a data rate of 171.2 kbps.
- However these data rates decrease with increase in the number of users trying to use the GPRS network.

Error correction :

- In GPRS the applications are required to provide their own error correction schemes.

GPRS implementation :

- GPRS implementation was easy because the GSM operator needs to install new routers and gateways for the base station.
- A new software that redefines the base station interface and time slots, also needs to be installed at standard the base station.
- A new RF base station hardware is also required.

Applications :

- Some important applications of GPRS are :
  1. Non-real time Internet applications.
  2. Retrieval of e-mails, faxes
  3. Asymmetric web browsing (fast downloading and less uploading).

Table 3.4.1 : Technical features of the emerging 2.5G and 3G data communications standards

Sr. No.	Wireless Data Technology	Generation	Channel Bandwidth	Duplex method	Need of new spectrum	Change in infrastructure	Need of a new handset
1.	HSCSD	2.5G	200 kHz	FDD	No	Needs software upgrade at the base station.	Yes
2.	GPRS	2.5G	200 kHz	FDD	No	Needs new packet overlay in addition to routers and gateways.	Yes
3.	EDGE	2.5G	200 kHz	FDD	No	Needs new transceiver at the base station and software upgrades to base station controller and base station.	Yes
4.	W-CDMA	3G	5 MHz	FDD	Yes	Needs entirely new base stations.	Yes
5.	IS-95B	2.5G	1.25 MHz	FDD	No	Needs new software in base station controller.	Yes
6.	CDMA 2000 1xRTT	3G	1.25 MHz	FDD	No	Needs new software in backbone and new channel cards at base station. They also need to build a new packet service node.	Yes
7.	CDMA 2000 1xEV (DO and DV)	3G	1.25 MHz	FDD	No	It needs software and digital card upgrade on the 1xRTT networks.	Yes
8.	CDMA 2000 3xRTT	3G	3.75 MHz	FDD	May be	They need backbone modifications and new channel cards at base station.	Yes

3.4.1.3 EDGE for 2.5G GSM and IS-136 :

- The long form of EDGE is Enhanced Data Rates for GSM or Global Evolution.
- This is a more advanced upgrade to both GSM and IS 136 2G standards as shown in Fig. 3.4.2.
- The development of EDGE provides a common evolution path for both GSM and IS-136 which eventually learn to 3G high speed data rates.

Modulation format :

- EDGE uses a new modulation format called 8-PSK (Octal PSK) in addition to GSM standard GMSK modulation.
- There are nine different air interface formats that the EDGE technology offers to the user.
- These formats are rapidly selectable and are known as multiple modulation and coding schemes (**MCSs**). Each MCS can use the **GMSK** for low data rate and **8-PSK** for high data rates.
- EDGE is sometimes also called as the Enhanced GPRS or EGPRS.

Range of coverage :

- The range of coverage in EDGE is smaller than that as compared to GPRS and the other systems, due to the following reasons :
  1. EDGE supports higher data rates.
  2. EDGE allows a relatively lower power.

Incremental redundancy :

- For each time slot in GSM, the EDGE technology uses the 8- PSK modulation.
- This allows each user (in that time slot) to adaptively determine the best MCS setting for itself.
- This adaptive capability of selecting The best possible air interface is defined as the **incremental redundancy.**
- In the incremental redundancy, the packets are first sent with maximum error protection and maximum data rate throughput.
- The error protection and data rate throughput are reduced progressively for the subsequent packets, until the wireless link faces an unacceptable outage or delay.
- The previously acceptable air interface is then restored with the help of rapid feedback between the base station and user.

Advantages of incremental redundancy :

- Following are some of the benefits of incremental redundancy :
  1. Each radio link uses minimum overheads.
  2. It maximizes the user capacity.
  3. It provides an acceptable link quality for each user.

Data rate :

- We can obtain a raw peak throughput data rate of 547.2 kbps if EDGE uses the 8-PSK modulation without any error protection, with all the eight time slots in GSM dedicated to a single user.
- However the practical maximum raw throughput data rate in EDGE is restricted to 384 kbps for a single dedicated user on a single GSM channel. Due to EDGE slotting schemes, network contention issues and error coding requirements.
- EDGE can provide the throughput data rates of several Mbps to individual users, by combining the capacity of different radio channels.

EDGE Implementation :

- EDGE requires a few hardware as well as Software upgrades at the existing GSM on IS-136 base station.
- It requires new transceivers at the base stations and software upgrades at the base station and base station controller.

3.4.2 IS-95B for 2.5G CDMA :

- The 2G TDMA standards (GSM and IS - 136) have multiple evolutionary paths.
- But 2G CDMA standard (IS-95 or cdmaOne) has only one evolutionary path which is from IS-95 to IS-95B.
- IS-95B, like GPRS is being used all over the world.

Principle :

- IS-95B dedicates multiple orthogonal user channels (Walsh functions) to different users simultaneously, similar to IS-95.
- It uses high speed packer and circuit switched data on a common CDMA channel.

Data rates :

- The maximum data rate of IS-95 is 9.6 kbps which was improved to 14.4 kbps in IS-95A which is further improved to 115.2 kbps to a dedicated user (8 x 14.4 kbps).
- This is achieved by allowing a dedicated user to command eight different user Walsh codes simultaneously and in parallel.
- However, due to the slotting techniques used by IS-95B standard, the maximum practical data rate at the most equal to 64 kbps.

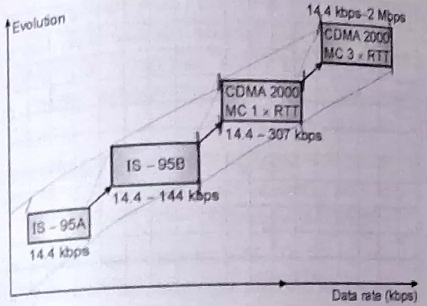
Better link quality :

- IS-95B standard improves the link quality experienced by the users to a great extent.
- This improvement results due to the **hard hand-off procedure** specified by the IS-95B standard.
- In hard hand-off procedure, subscriber units are allowed to search different radio channels in the network, without instructions from the switch.

- Due to this the subscriber units can tune rapidly to different base stations to maintain a high link quality.
- The link quality in IS-95 is inferior, because it used the soft handoff procedure.

Applications :

- IS-95B standard, supports 64 simultaneous subscribers and it is used for the medium data rare **(MDR)** services.
- Fig. 3.4.3 illustrates the evolution path for the only 2G CDMA standard.



(G-2546) **Fig. 3.4.3 : Evolutionary path for IS-95A (cdmaOne) standard**

3.5 Third Generation (3G) Networks :

- The third generation of wireless mobile communication systems have been developed to meet the International Mobile Telecommunication - 2000 (IMT - 2000) specifications which are defined by International Telecommunications Union (ITU).
- The 3G systems have evolved due to the need for high speed, fast data transmission and better quality of service (QoS).
- The 3G systems were launched in 2001 and it provides the network for transporting rich **multimedia** contents.
- The 3G systems use circuit switching technology for voice calls/SMS facility, whereas they use the packet switching for the high speed data.
- The well known examples of 3G systems are :
  1. W-CDMA.
  2. CDMA – 2000 3. TD - SCDMA
- 3G systems are compatible with the other cellular standards like CDMA, GSM and TDMA.
- The frequency range used by the 3G standards is 2100 MHz and it has a bandwidth of 15-20 MHz.
- 3G standards facilitate the users to use high speed internet services, as well as video chatting.
- The international roaming has become possible due to 3G standard.
- Universal Mobile Telecommunications System (UMTS) was adopted by Europe which uses W-CDMA as its standard.
- UMTS is based on the GSM infrastructure. Hence UMTS is the most popular 3G technology.

3.5.1 Features of Third Generation :

- Some of the important features of 3G mobile systems are as follows :

Sr. No.	Feature	Value / Description
1.	Generation	3G (2001)
2.	Technology	Broadband/IP, FDD, TDD
3.	Standards	CDMA, W-CDMA, UMTS.
4.	Switching	Circuit/Packet switching
5.	Frequency band	1.6 GHz to 2.5 GHz
6.	Data speed	2 Mbps
7.	Multiplexing	CDMA
8.	Core network	Packet network
9.	Services	High speed data, voice, video
10.	Handoff	Horizontal

- Aim of the third generation technology is to combine the services such as internet, telephony and multimedia into a single device.

3.5.2 Services Provided by 3G Systems :

1. **Voice :**
   - 3G system will provide good speech quality as compared to the telephone network.
2. **Messaging :**
   - E-mail attachments are allowed in 3G system.

3. MMSC multimedia messaging services :

- 3G supports MMS which are designed for rich text, icons, logos, animated clips etc.

4. Medium multimedia :

- In 3G system, for web surfing, games, location based maps its downstream data rate is suitable.

5. Interactive high multimedia :

- 3G supports for this service which is used for high quality videophones, videoconferencing etc.

3.5.3 Advantages of 3G Networks :

- The third generation of wireless networks (3G systems) provide a wireless access, that the earlier networks could never provide in ways that have never been possible earlier.
- Some of the advantages of the 3G networks are as follows :
  1. They provide data rates of multiple mbps for Internet access.
  2. Communication using Voice over Internet Protocol (VoIP) is possible.
  3. Voice activated calls.
  4. Very high networks capacity.
  5. The access to data networks is "always-on" type.
- The 3G equipment are designed in such a way that the users can receive live music, conduct interactive web sessions and have simultaneous voice and data access with multiple parties at the same time using a single mobile handset, either moving or stationed at one place.
- In order to facilitate the operation of 3G systems, the International Telecommunications Union (ITU) prepared a plan called as **International Mobile Telephone 2000 (IMT-2000)**.
- IMT 2000 standard uses a global frequency band in the 2000 MHz range that supports a wireless communication standard for all countries throughout the world.
- This ITU plan has been successful in initiating an active debate and technical analysis for new high speed mobile telephone solutions in comparison to the 2G.
- However, the hope for a single worldwide standard has not materialized, because the worldwide user community remains divided between two standards : GSM/IS-136/PDC and CDMA IS-95.
- The cdma2000, a 3G standard was eventually developed by following the path : IS-95, IS-95B and cdma 2000 as shown in Fig. 3.4.1.
- Several variants of CDMA 2000 are currently being developed, that are all based on the fundamentals of IS-95 and IS-95B technologies.
- The evolution of the 2G TDMA standards (GSM, IS-136 and PDC systems) has eventually lead to the Wideband CDMA (W-CDMA), or Universal Mobile Telecommunication Service (UMTS) as shown in Fig. 3.4.1.
- W-CDMA is based on the network fundamentals of GSM, as well as the merged versions of GSM and IS-136 through EDGE.
- The ITU 200 MT - 2000 standard organizations are separated into two organizations, according to the two 3G camps :
  1. 3GPP and
  2. 3GPP2
- The 3GPP stands for 3G Partnership project for wideband CDMA standard and it is backward compatible with GSM and IS-136/PDG 2G standards.
- 3GPP2 stands for 3G Partnership Project for cdma 2000 standard which is backward compatible to IS-95.

3.6 Fourth Generation (4G) :

- The 4G wireless systems were designed to fulfill the requirements of International Mobile Telecommunications Advanced (IMT-A) using IP (Internet Protocol) for all the services.
- 4G LTE which means fourth generation **long term evolution**.
- It offers the users faster, more reliable mobile broadband Internet for the devices such as the smartphones, the tablets and the laptops.

- 4G LTE is very fast (10 times faster than the 3G network). It provides extremely high voice quality.
- It is possible to download large files, very fast using the wireless 4G LTE network.
- In 4G systems, an advanced radio interface is used with Orthogonal Frequency Division Multiplexing (OFDM), Multiple Input Multiple Output (MIMO) and the link adaptation technologies.
- 4G standards also includes **Long Term Evolution (LTE)** and IEEE 802.16 (Wi-Max).
- The 4G systems provide very high data rates as compared to 3G. But the major problem with 4G systems is security because of its IP address system.

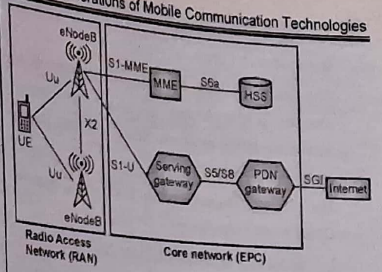
3.6.1 Features of 4G Systems :

- The important features of 4G mobile systems are as follows :

Sr. No.	Feature	Value / Description
1.	Generation	4G (2010)
2.	Technology	IP-Broadband, Wi-Fi, MIMO
3.	Standard	Wi Max and LTE
4.	Switching	Packet switching
5.	Frequency band	2 GHz – 8 GHz
6.	Data speed	50 Mbps
7.	Multiplexing	MC-CDMA and OFDM
8.	Core network	Internet
9.	Service	Dynamic Information Access
10.	Handoff	Vertical

3.6.2 4G (LTE) Network Structure (Architecture) : 100Mbps

- Fig. 3.6.1 shows the network structure of LTE (4G), where LTE stands for Long Term Evolution.
- 4G network architecture is simple which contains single type of access point i.e eNode B or Base station. One or more cells are supplied by eNode B or BS. CDMA Multiplexing Ultra wide band. Carrier frequency - 15MHz



MME : Mobile Management Entity
HSS : Home Subscriber Server
PDN : Packet Data Network
UE : User Equipment

(G-2590) Fig. 3.6.1 : 4G architecture

- As shown in Fig. 3.6.1, the 4G (LTE) network consists of :
  1. Radio Access Network (RAN)
  2. Core network (EPC)
  3. Radio interface

1. Radio Access Network :

- The radio access network is also known as EUTRAN or Evolved Universal Mobile Telecommunications System.
- The Radio Access Network consists of an LTE mobile terminal, radio interface and eNodeB.
- **LTE Mobile Terminals :** LTE mobile terminals are the mobile phones and other devices which support the LTE.
- **Radio Interface :** Radio interface are the radio links that connect the LTE mobile terminals and eNodeB.
- **eNodeB :** E-UTRAN Node B or eNodeBs are located all over the network of the mobile operator and they connect the LTE mobile terminal to the core network via radio interface S1.

Functions of eNodeB :

- The functions of eNodeB are as given below :
  1. Scheduling / Radio resource allocation.
  2. Retransmission control.
  3. Physical layer functions.
  4. Air interface communication.

2. Core network (EPC) :

- The Enhanced Packet Core (EPC) developed for the 4G is also known as System architecture evolution (SAE) which is based on the packet switched transmission.
- The LTE Core Network is the brain of 4G system. Core network consists of :
  1. Mobility management Entity (MME)
  2. Serving gateway (S-GW)
  3. PDN (Packet data network) gateway (P-GW)
  4. HSS (Home subscriber server)

1. **Mobility management entity (MME) :** The function of MME is to handle the signalling of messages, tracking, security and paging of mobile terminals.
2. **Serving gateway (S-GW) :** The serving gateway is connected to RAN (Radio access network) through S1 interface. It serves as a router for forwarding the data packets between the user equipment and the PDN gateway.
3. **PDN (Packet data network) gateway (P-GW) :** PDN (Packet Data Network) Gateway connects the EPC network with internet through SGI interface. PDN routes traffic to and from the PDN networks.
4. **HSS (Home subscriber server) :** HSS (Home Subscriber Server) is the database of all mobile users, which contains all data of subscriber. It is also responsible for authentication and call and session setup.

Functions of core network :

- Functions of the core network are as follows :
  1. Charging and subscriber management.
  2. Mobility management.
  3. Provision of quality of service.
  4. Policy control of user data flows.
  5. Connection to other external networks.

3. Radio Interface :

- Refer Fig. 3.6.1 to understand different radio interfaces in LTE.
- **Interface x2 :** This is an interface, which connects different base stations. The important information required for the coordination of transmission in neighboring cells can be exchanged through this interface.
- **Interface S1 :** RAN is connected to the core network through interface S1.
- **Interface Uu :** This is an interface, which connects UE (user equipment) and eNodeB.
- **Interface S1-U :** This is an interface for the user plane between an E-UTRAN and S-GW. It provides GTP tunnel per carrier.
- **Interface S1-MME :** This is an interface for the control plane between an E-UTRAN and MME.
- **Interface S6a :** This is an interface for the control plane between an HSS and MME. It exchanges user subscription and authentication information.
- **Interface S5/S8 :** This is an interface defined for the control and user planes between an S-GW and P-GW.
- **Interface SGI :** This is an interface defined for the control and user planes between an P-GW and Internet.

3.6.3 Specifications of 4G LTE :

- Following are some important specifications of 4G LTE :
  1. Peak data rates : Downlink - 1 Gbps; uplink - 300 Mbps.
  2. Spectrum efficiency : 3 times greater than LTE.
  3. Speed 10 times faster than the 3G network.
  4. Peak spectrum efficiency : Downlink - 30 bps / Hz; uplink - 15 bps / Hz.
  5. 4G LTE is flexible and reliable.
  6. It is easy to standardize and affordable.

3.6.4 Advantages of 4G :

1. Improved voice quality.
2. High data transfer speed.
3. Uninterrupted connectivity.
4. Good coverage.
5. It provides security, privacy and safety.
6. Affordable service.

3.6.5 Disadvantages of 4G :

- A 4G network needs more number of antennas and transmitters.
- Therefore, user will experience poor battery life on their mobile devices.

3.6.6 Applications of 4G :

- The 4G is developed to support the QoS and data rate requirements of the advanced applications such as :
  1. Wireless broadband access.
  2. Multimedia Messaging Service (MMS).
  3. Video chat.
  4. Mobile TV.
  5. HDTV.
  5. Digital Video Broadcasting (DVB).
  7. Voice and data.
  8. Other services which need large bandwidth.

3.7 5G and Above Wireless Networks :

- The 4G technology has now been deployed and the research for the next generation named as 5G has already begun.
- It is considered to be the next major phase of mobile telecommunication standard after 4G.
- The 5G standard will be made commercially available by 2020. This standard is way beyond just the faster data speeds or faster mobile devices.
- 5G will provide an access to high and low speed data services. It will involve combination of existing and evolving systems.

3.7.1 Why 5G ?

- Fifth generation technology is useful because of the following reasons :
  1. It provides very High speed, high capacity, and low cost per bit.
  2. It supports interactive multimedia, voice, video, Internet, and other broadband services, more effectively.
  3. 5G technology offers Global access and service portability.
  4. It can provide the high quality services due to high error tolerance.
  5. 5G technology uses remote management so that user can get better and fast solution.
  6. With the use of 5G technology the uploading and downloading speed will be very high.

3.7.2 Features of 5G :

- Following are the most important features of 5G :
  1. Ubiquitous connectivity.
  2. It provides Speed up to 10 Gbit/s.
  3. The 5G technology supports virtual private network.
  4. The uploading and downloading speed of 5G technology is high.
  5. 5G network is very fast and reliable.
  6. Larger data volume per unit area (i.e. high system spectral efficiency).
  7. High capacity to allow more devices connectivity concurrently and instantaneously.
  8. Lower battery consumption.
  9. It provides better connectivity irrespective of the geographic region.
  10. Larger number of supporting devices.

3.7.3 Features of Fifth Generation :

Table 3.7.1 : Features of fifth generation

Sr. No.	Feature	Value / Description
1.	Generation	5G (2020)
2.	Technology	WWW, IPv6
3.	Standard	Yet to be finalized
4.	Switching	Packet
5.	Frequency	15 GHz

Sr. No.	Feature	Value / Description
6.	Data speed	> 1 Gbps
7.	Multiplexing	MC-CDMA, LAS-CDMA, OFDM
8.	Core network	Internet
9.	Services	Interactive multimedia, Voice over IP, Virtual reality, Augmented reality, IOT etc.
10.	Handoff	Horizontal and vertical

3.7.4 Expectations in 5G Network :

- In future, 5G will be necessary worldwide because of the increasing traffic rates of data, voice, and video streaming.
- 5G technology will have the capability to share the data everywhere, every time, by everyone.
- The 5G technology is expected to cater the following requirements :
  1. 10-100 times higher data rate.
  2. 10 times longer battery life for low power devices.
  3. 10-100 times higher number of connected devices.
  4. 5-times reduced end to end latency.
  5. 1000 times higher mobile data volume per area.

3.7.5 Differences Between 4G and 5G :

- The key differences between 4G and 5G network architecture are as follows :
  1. Latency, speed and bandwidth
  2. Potential download speeds
  3. Base stations
  4. OFDM encoding
  5. Cell density

1. Latency, speed and bandwidth :

- The biggest difference between 4G and 5G networks is latency.
- The latency promised by 5G is low latency under 5 milliseconds, while 4G latency ranges from 60 ms to 98 ms.
- Low latency requires higher bandwidth and indicates that the speed of 5G networks is higher than that of the 4G networks.
- In addition, due to lower latency there is an increase in the download speeds.

2. Potential download speeds :

- The download speeds in 5G are much higher than those obtainable in the 4G networks.
- The download speeds in 4G can be up to 1 Gbps, and 5G's goal is to increase that tenfold for maximum download speeds of 10 Gbps.

3. Base stations :

- Another important difference between 4G and 5G is the base station required to transmit signals.
- The 4G transmits signals from cell towers. However, 5G uses small cell technology, due to its faster speeds and mm Wave frequency bands.
- Due to high frequency carriers deployed in 5G the base stations are of very small size (size of a pizza box) and present at multiple locations.
- 5G networks will still use cell towers when it uses the lower frequency spectrums.
- When used in the mm wave range (above 6 GHz) the carriers must deploy small cells in various areas.
- However, the mm wave has weaker signals that travel across shorter distances.
- Therefore, small cell stations must be placed more frequently in 5G-capable areas to ensure the signals reach users and businesses.

4. OFDM encoding :

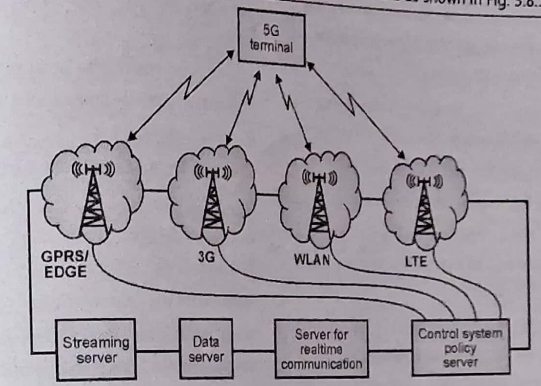
- OFDM is used to split different wireless signals into separate channels to avoid interference, which also provides greater bandwidth.
- OFDM encodes data on different frequencies. Therefore, this can enhance 4G and 5G download speeds.
- These networks would have their own signal channels rather than a shared one between them.
- Hence, 4G uses 20 MHz channels, while 5G will use 100 MHz to 800 MHz channels.

5. Cell density :

- Due to small cell technology, 5G can provide more cell density and enhance network capacity.
- While 4G also promised this, it could never completely met its promises.
- 5G will hopefully succeed where its predecessor falls short.
- The 5G networks will be denser, which means they have more capacity to support more users and connected devices, which leads to increased mobile device and connection capacity.

3.8 Functional Architecture of 5G :

- The system model of 5G is entirely IP based model designed for the wireless and mobile networks is as shown in Fig. 3.8.1.



(GT-44) **Fig. 3.8.1 : 5G structure based on combining the current and future networks**

- Fig. 3.8.1 shows the 5G structure based on combining the current and future networks.
- As shown in Fig. 3.8.1, the system consists of a main user terminal and a number of independent and autonomous radio access technologies such as GPRS/EDGE, 3G, WLAN and LTE etc.
- Each radio technology in the architecture of 5G is considered as the IP link for the outside internet world.
- The IP technology is designed entirely to make sure the sufficient data control for suitable routing of IP packets related to a certain application connections i.e. sessions between client applications and servers somewhere on the Internet.

3.9 Technologies Used in 5G :

- To make 5G and beyond networks a reality, many advanced ideas have been proposed and analyzed in recent years.
- The major key enabling technologies that have been considered for 5G and 6G systems are as given below :
  1. Small cells
  2. Beamforming
  3. Millimeter (mm) waves
  4. Device centric architecture
  5. Full-duplex technology
  6. Massive MIMO
  7. NOMA

3.9.1 Small Cell Concept :

Definition :

- A small cell is a broad term used to describe a miniature radio access point (AP) or wireless network base station with a low radio frequency (RF) power output, footprint and range.

Functions :

- The functions of small cells are to enhance cellular network coverage and capacity in areas such as densely populated city centers where user demands are the highest.
- They also significantly improved signal penetration and ensure superior coverage.

Working of small cells :

- Generally the operator-controlled small cells are deployed indoors or outdoors in a licensed, shared or unlicensed spectrum.
- These are primarily the miniature base stations and have very small physical size.
- The macro-cells are the tall cell towers that are installed on the highway or on top of buildings.
- On the other hand, the small cells are much smaller, about the size of a pizza box and are low power devices.
- Because they are small and compact, they are installed every few blocks to complement macro-cells in densely populated areas.
- The mobile operators that run the carrier networks attach small cells to physical structures, such as the sides of buildings, streetlights and signs.
- The small cells are essential for the transmission of data to and from different wireless devices.
- The small cells used in 5G networks get their connection via a macrocell, before sending data from one small cell to another in a relay. This helps carry signals over much larger distances.
- Small cells enhance macro communications networks, by boosting coverage in a specific area by adding targeted capacity.
- They also support new services and deliver enhanced user experience by enabling high-speed wireless Broadband.

Spectrum and range :

- Small cells transfer data using low-, mid- and high-band spectrum.
- They also use multiple spectral frequencies for boosting network bandwidth, increasing data transmission and for improving speed.
- The range of small cells is limited from 10 meters to 2 kilometers depending on the type of small cell.

Types of small cells :

- There are different types of small cells.
- They have different ranges, form factor and power levels.
- Out of them, the largest small cell units in used in the urban, outdoor applications and the smallest for indoor applications.
- The three types of small cells are the following :
  1. Femto-cells
  2. Pico cells and
  3. Micro cells

1. Femto cells :

- These devices are similar to wireless routers and have a typical maximum range of 10 m.
- Femto cells can only accommodate a few users at a time and are usually used for indoor applications.

2. Picocells :

- Pico cells can have an extensive range of up to 200 m.
- Therefore, they can support a maximum of 100 users and are typically installed in larger indoor areas, such as hospitals, airports and train stations.

3. Microcells :

- These small cells have a range of up to 2 km. Microcells are usually attached to traffic lights and street signs.

Advantages of small cells :

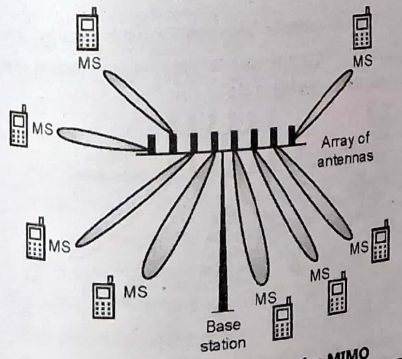
- Mobile operators and their customers experience the following benefits from small cells.
  1. They are quick and easy to install.
  2. They can be deployed without highly technical skill sets.
  3. Small cells improve coverage.
  4. They boost capacity in densely populated urban areas.
  5. Small cells are cheaper than macro cells.
  6. They are highly flexible.
  7. They can extend the battery life of mobile devices.

Limitations and Challenges of small cells :

- There are certain limitations and challenges for the small cells.
- Some of them are as follows :
  1. Operators must ensure they have backhaul – either fibre or wireless – at the site.
  2. There must be an appropriate power source and operators must have physical access to the site.
  3. This may need complicated or lengthy negotiations with the landowner.
  4. One of the biggest challenges within the world of 5G small cells is coming up with a set of industry standards and definitions for this emerging technology.

3.9.2 Massive MIMO :

- Massive MIMO is the new wireless access technology in 5G, in both sub-6 GHz and mm Wave bands.
- The concept of massive MIMO has been demonstrated in Fig. 3.9.1.



(G-3013) **Fig. 3.9.1 : Concept of massive MIMO**

- Massive MIMO is a multi-user MIMO (multiple-input multiple-output) technology that can provide uniformly good service to wireless terminals in high-mobility environments.
- In massive MIMO the base stations are equipped with arrays of many antennas.
- These antenna arrays are used to serve many terminals simultaneously, using the same time-frequency resource.
- The word "massive" indicates that a large number of antennas are being used and does not refer to the physical size.
- The rectangular antenna arrays use the 2 GHz frequency band and contain around 200 dual polarized elements.
- The antennas are half-wavelength-spaced and have a size of about 1.5 x 0.75 meters large.
- The highly directional antennas in the array have an adjustable beam pattern and each beam can be precisely designed to serve a specific mobile station.
- Thus massive MIMO uses the beamforming technique for its operation.
- The use of precise beams reduces the amount of transmitted power per mobile customer and avoids interference between adjacent mobile stations.
- Massive MIMO is implemented in two topologies i.e. FDD and TDD. In FDD, different frequencies are used on downlink (DL) and uplink (UL).
- In TDD, same frequency is used on downlink and uplink. TDD benefits due to channel reciprocity as channel estimation using UL (Uplink) can be utilized for downlink beamforming and hence leads to less overhead.
- Massive MIMO uses TDD mode for its operation and the downlink beamforming makes use of the uplink-downlink reciprocity of radio propagation.
- That means, the base station array uses channel estimates obtained from uplink pilots transmitted by the terminals to learn the channel in both directions.

- This makes Massive MIMO entirely scalable with respect to the number of base station antennas.
- Base stations in Massive MIMO operate autonomously, with no sharing of payload data or channel state information with other cells.

Advantages of Massive MIMO :

- As a key building block of 5G, massive MIMO has a number of benefits for both network operators and end users as follows :

1. **Improved spectral efficiency :**
   - The massive MIMO technology improves spectral efficiency.
   - That means, it can improve network capacity for the same amount of spectrum.
   - This enables the operators to maximize their investments in this expensive resource.
2. **Supports larger number of users :**
   - Massive MIMO, in conjunction with beamforming technology enables highly targeted use of spectrum, supporting a larger number of users in the cell.
3. **Improves end-user experience :**
   - Due to the highly targeted use of spectrum and reduced interference, massive MIMO improves end-user experience in densely populated areas.
4. **Higher connection reliability and reduced interference :**
   - Other potential benefits are, higher connection reliability and increased resistance to interference or intentional jamming.
5. **Improved indoor coverage :**
   - Massive MIMO networks are more responsive to devices transmitting at higher frequencies. This improves the indoor coverage.

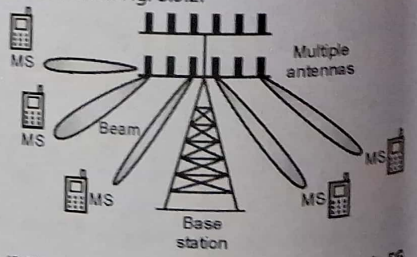
Disadvantages of Massive MIMO :

- Following are the drawbacks or disadvantages of Massive MIMO :

1. Massive MIMO units are costlier than the traditional radio units.
2. The antenna design is more complex.
3. Use of FDD can lead to feedback overhead. Hence TDD is preferred.
4. Spacing between antennas is very small which requires packing of entire hardware into a small space.
5. Massive MIMO requires complex signal processing algorithms.
6. It is sensitive to beam alignment with the mobile stations, as extremely narrow beams are being used.

3.9.3 Beamforming :

- Today's mobile users need faster data speeds and more reliable service. The next generation of wireless networks-5G promises to deliver that, and much more.
- Right now, though, 5G is still in the planning stages, and companies and industry groups are working together to figure out exactly what it will be.
- But they all agree on one matter : As the number of mobile users and their demand for data rises, 5G must handle far more traffic at much higher speeds than the base stations that make up today's cellular networks.
- Beamforming is one of the burgeoning technologies that will help get us there.
- Beamforming is the ability of the base station to adapt the radiation pattern of the antenna.
- Beamforming helps the base station to find a suitable route for delivering data to the mobile user.
- It also reduces interference with nearby users along the route due to very narrow beams used as shown in Fig. 3.9.2.



(G-3012) **Fig. 3.9.2 : Concept of beamforming in 5G**

- As shown in Fig. 3.9.2, the beamforming combined with the massive MIMO consists of antenna arrays at the base station with each antenna having a very sharp adjustable beam.
- It is a traffic-signaling system for cellular base stations that identifies the most efficient data-delivery route to a particular user, and it reduces interference for nearby users in the process.
- Depending on the situation and the technology, there are several ways to implement it.
- Beamforming is a technique that constructs the antenna radiation pattern as shown in Fig. 3.9.2.
- Beamforming has several advantages for 5G networks and beyond.
- Depending upon the situation, beamforming technology can be implemented in several different ways in future networks.

Beamforming with massive MIMO :

- For massive MIMO systems, beamforming helps with increasing spectrum efficiency, and for millimeter waves, it helps in boosting data rate.
- In massive MIMO systems, the base station can send data to the user from various paths, and beamforming here determines the packet movement and arrival time which allows more users to send data simultaneously.

Beamforming with millimeter waves :

- Since the millimeter waves cannot penetrate through obstacles and do not propagate to longer distances due to a shorter wavelength, beamforming here helps to send concentrated beams towards the users.
- Thus, beamforming helps a user to receive a strong signal without interference with other users.

Beam management :

- The beam management is nothing but a procedure with a set of phases like,

1. Beam sweeping
2. Beam measurements
3. Beam determination
4. Beam reporting
5. Beam failure recovery

Types of beamforming :

- There are 3 types of beamforming used in 5G :

1. Analog beamforming
2. Digital beamforming
3. Hybrid beamforming

Benefits of beamforming :

- The beamforming alongwith massive MIMO systems has the following advantages :

1. Higher energy efficiency,
2. Improvement in spectral efficiency,
3. Higher system security, and
4. Applicability for mm-wave bands.

Drawbacks of beamforming :

- Following are the disadvantages of Beamforming technique :

1. Higher hardware complexity due to use of multiple antennas and other hardware systems.
2. The beamforming system needs to use advanced high processing DSP chip for the mathematical algorithms in the design.
3. Higher cost due to the use of increased hardware resources and advanced DSP chip.
4. Higher power requirement due to use of more resources. Hence battery in beamforming system drains faster.

3.9.4 Millimeter Waves :

Definition :

- Millimeter Wave (also known as mmWave) can loosely be defined as electromagnetic waves with frequency range from 30 GHz to 300 GHz because this corresponds to the wavelengths of 1 mm to 10 mm.

Principle :

- This frequency range corresponds to the highest band of Radio Frequency Bands, namely Extremely High Frequency (EHF) band.

- The frequencies below 6 GHz are generally used for cellular communication, while frequencies above 6 GHz are used for other services like medical imaging, microwave remote sensing, amateur radio, terahertz computing, and radio astronomy.
- Due to the massive increase in data traffic the radio frequency spectrum has become congested which results in limited bandwidth for a user, causing a slower and unreliable connection.
- One solution to this problem is to use frequencies above 6 GHz for wireless communication.
- The frequencies above 6 GHz have never been used for wireless communication, and there has been a lot of research going on with broadcasting millimeter waves.
- Millimeter waves are thus very high frequency radio waves which find their applications in the 5G and other advanced networks.
- The main advantage of mm Waves is that they can provide bandwidth ten times higher than that of the entire 4G cellular band.
- These high-frequency waves are used in some satellite application, but it has never been used for mobile broadband.
- Since millimeter has a smaller wavelength, they are not suitable for long-range applications.
- Another problem with millimeter waves is that they cannot penetrate buildings and obstacles, and they tend to get absorbed by rain.

Benefits of mm wave technology :

- Some of the important advantages of this technology are as follows :
  1. It can transmit a large amount of information with low latency.
  2. Higher bandwidth.
  3. Improved sensor resolution.
  4. Reduced latency.
  5. Higher transmission speed.
  6. Equipments and components are of smaller size.

Disadvantages of millimeter wave :

- Compared with microwave, using of millimeter wave also has the following disadvantages :
  1. Higher costs required to manufacture greater precision hardware and components with smaller size.
  2. Lower sensitivity in a receiving system due to lesser energy collected by the smaller size antenna.
  3. Due to significant attenuation of extremely high frequencies waves, the millimeter waves cannot be used for long distance applications.

Applications :

- The applications of millimeter wave can be classified to following categories : communication, sensor, studies and experiments, and weapons systems.

3.9.5 NOMA :

- Non-orthogonal multiple access (NOMA) is one of the most promising multiple access techniques in next-generation wireless communications.
- As compared to the orthogonal frequency division multiple access (OFDMA), which is the orthogonal multiple access (OMA) technique, NOMA has the following advantages :
  1. Enhanced spectrum efficiency,
  2. Reduced latency with high reliability, and
  3. Massive connectivity.
- The NOMA system serves multiple users using the same resources in terms of time, frequency, and space domain.
- NOMA can achieve a higher data rate compared to OMA. However, the required processing of NOMA can be higher than OMA.
- In recent years, non-orthogonal multiple access (NOMA) schemes have received significant attention for the fifth generation (5G) cellular networks.

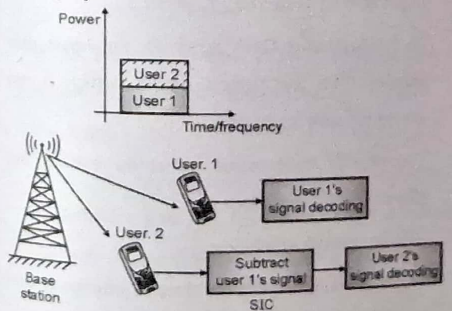
- This is due to the ability of **NOMA** to serve multiple users using the same time and frequency resources.
- Recent studies demonstrate that NOMA has the potential to used in various 5G communication applications such as, Machine-to-Machine (M2M) communications and the Internet-of-Things (IoT).
- NOMA can accommodate many more users via non-orthogonal resource allocation.

How does NOMA work ?

- In NOMA, each user operates in the same band and at the same time where they are distinguished by their power levels.
- NOMA uses superposition coding (SC) at the transmitter such that the successive interference cancellation (SIC) receiver can separate the users both in the uplink and in the downlink channels.
- A **non orthogonal** transmission is not sensitive to errors of synchronization.

Block diagram :

- The block diagram of NOMA is as shown in Fig. 3.9.3 which is a multiple access technique employed in 5G cellular wireless network.



(G-2793) Fig. 3.9.3 : NOMA in downlink

- The main function of NOMA is to serve multiple UEs (User Equipments) using a single 5G-NB (Node B or Base Station) on same time/frequency resources.
- There are two main techniques employed in NOMA for multiple access :

1. **Power domain :**
   - In power domain, NOMA achieves multiplexing based on different power levels.
2. **Code domain :**
   - In the code domain, NOMA achieves multiplexing based on different codes.

Transmit side :

- Refer Fig. 3.9.3. A special coding called **superposition coding (SC)** is used in NOMA at the transmitter end.
- NOMA assigns different power levels to different users.
- As shown in Fig. 3.9.3, the base station transmits superposed signals to User-1 and User-2.
- Here, User-2 uses high gain and User-1 uses low gain as shown, in order to successfully receive signals with different power levels.
- Hence, User-2 is known as a strong user while User-1 is known as a weak user.

Receive side :

- At each receiver, NOMA uses **SIC** (Successive interference cancellation) technique to retrieve data.
- On the receiving side, User-2 (strong user) subtracts signal of user-1 through SIC and later decodes its own signal.
- Whereas, User-1 (weak user) treats signal of User-2 as noise and decodes its own signal directly as shown in Fig. 3.9.3.

Type of modulation :

- **5G** technology implements quadrature phase shift keying (QPSK) as the lowest order **modulation** format.

- It has a demerit of providing the slowest data throughput and the advantage of providing the most robust link which can be **used** when signal levels are low or when interference is high.

Classification of NOMA :

- We divide NOMA schemes into two categories :
  1. Power-domain multiplexing and
  2. Code-domain multiplexing.
- Power-domain NOMA attains multiplexing in power domain and code-domain NOMA achieves multiplexing in code domain.
- One way of implementing power-domain NOMA is utilizing superposition coding (SC) at the transmitter and successive interference cancellation (SIC) at the receiver.
- SC allows the transmitter to transmit multiple users' information at the same time.
- To decode the superposed information at each receiver, SIC technique can be used.
- The main idea of SIC is that user signals are successively decoded.
- After one user's signal is decoded, it is subtracted from the combined signal before the next user's signal is decoded.

Resource management in NOMA networks :

- One of the main challenges in NOMA networks is to attain a compromise between the bandwidth efficiency and the energy efficiency.
- This can be achieved with the help of following resource management techniques :
  1. Intelligent control of the PA of the superimposed signals.
  2. Dynamic scheduling of the users for the sub-channels.
  3. Forming spatially correlated clusters.

Implementation challenges of NOMA :

- Although NOMA is a promising candidate for 5G and beyond, there are several implementation challenges to be tackled, such as
  1. Error Propagation in SIC
  2. Channel Estimation Error and Complexity for NOMA
  3. Security Provisioning for NOMA
  4. Maintaining the Sustainability of NOMA With RF Wireless Power Transfer

Performance enhancement :

- It is possible to improve the performance of NOMA by integrating it with various effective wireless communications techniques, such as cooperative communications, multiple-input multiple-output (MIMO), beam forming, space-time coding, network coding, etc.

Advantages of NOMA :

- By invoking the SC technique, the BS transmits the superposition coded signals of all users.
- Then, the channel gains of the users are sorted in the increasing or decreasing order.
- In the traditional OMA schemes, strongest user benefits from the highest power, which is not always the case for NOMA.
- The NOMA transmission schemes exhibit the following main advantages :

1. **High bandwidth efficiency :**
   - NOMA exhibits a high bandwidth efficiency and hence improves the system's throughput, because it allows multiple users to exploit each resource.
2. **Fairness :**
   - An important feature of NOMA is that it allocates more power to the weak users.

- Therefore, NOMA can guarantee an attractive tradeoff between the fairness among users in terms of their throughput.
- NOMA uses sophisticated techniques of maintaining fairness, such as the intelligent PA policies and the cooperative NOMA scheme.

3. **Ultrahigh connectivity :**
   - The future 5G system is expected to support the connection of billions of smart devices in the IoT.
   - The NOMA can efficiently solve this nontrivial task, with the help of its non-orthogonal characteristics.
   - The convectional OMA requires the same number of frequency/time RBs as the number of devices.
   - However, NOMA is capable of serving them by using less RBs.
   - In short, it offers massive connectivity by serving more uses simultaneously.
4. **Compatibility :**
   - Due to the mature status of SC and SIC techniques, NOMA can be compatible with the existing multiple access techniques.
   - In fact, it can be invoked as an "add-on" technique for any existing OMA techniques.
5. **Flexibility :**
   - With an appropriate integration of coding, modulation, and subcarrier allocation, the NOMA systems can yield a low-complexity, high flexibility design.
6. **Lower latency :**
   - NOMA offers lower latency due to simultaneous transmission all the time rather than dedicated scheduled time slot.
7. **Higher QoS :**
   - It has a better QoS (Quality of Service) to all the users due to its flexible power control algorithms.
   - It helps in increasing cell-edge throughput and better user experience at cell-edges.
8. **Enhanced performance :**
   - The NOMA along with MIMO (Multiple Input Multiple Output) delivers an enhanced performance.

Drawbacks of NOMA :

- Following are the disadvantages of NOMA :

1. **Higher complexity and low energy efficiency :**
   - Each of the users within the cluster need to decode information of all the other users even one having worst channel gains.
   - This leads to complexity in the receiver. Moreover energy consumption is higher.
2. **Erroneous detection :**
   - If error occurs in single user due to SIC, decoding of all the other users information will be erroneous.
   - This limits maximum number of users.
3. **Limit on active number of users :**
   - For the power domain concept in NOMA to work properly at the receiver, there should be an adequate channel gain difference between the users.
   - This limits effective number of user pairs served by clusters.
4. **Sensitivity to gain information :**
   - Each user needs to feedback channel gain information back to Base Station.
   - Hence, NOMA performance depends on the correctness of this information.

Higher security threats :

- Although NOMA technique offers a number of advantages, the enhanced information sensing ability of more users, leads to higher security and privacy threat.

5. Possibility of interference :

- Since the principle of NOMA is to allow multiple users to be superimposed on the same resource, this leads to interference for such systems.

3.10 Specifications of 5G :

- Some of the important specifications of 5G network are as follows :
  1. Up to 10 Gbps data rate which is 10 to 100 times the speed of 4G and 4.5G networks.
  2. 1-millisecond latency.
  3. 1000 times higher bandwidth per unit area.
  4. Up to 100 fold number of connected devices per unit area as compared with 4G LTE.
  5. 99.999 % availability.
  6. 100 % coverage.
  7. 90 % reduction in network energy usage.
  8. Up to 10-year battery life for low power IoT device.

3.10.1 Advantages of 5G Technology :

- Following are the advantages of 5G :
  1. 5G technology can gather all networks on one platform.
  2. It is more effective and efficient.
  3. 5G technology will provide a huge broadcasting data (in Gigabit), which will support more than 60,000 connections.
  4. 5G is compatible with the previous generations.
  5. It can offer uniform, uninterrupted, and consistent connectivity across the world.

3.10.2 Disadvantages of 5G Technology :

- Following are some drawbacks of 5G technology :
  1. 5G technology is still under process and research on its capability is going on.
  2. Claimed speed of this technology seems difficult to achieve because of the incompetent technological support in most parts of the world.
  3. Developing infrastructure of 5G is costly.
  4. Security and privacy issue are yet to be resolved.

3.10.3 Applications of 5G :

- 5G technology can be used in the following applications :
  1. Entertainment and multimedia
  2. Internet of Things – Connecting everything
  3. Smart Home
  4. Logistics and shipping
  5. Smart cities
  6. Smart farming
  7. Healthcare and mission critical applications
  8. Drone operation
  9. Security and surveillance

3.10.4 Comparison of Various Mobile System Generations :

- Table 3.10.1 shows the comparison of Various Mobile System Generations.

Table 3.10.1 : Comparison of various mobile system generations

Sr. No.	Feature	Generation				
		1G	2G	3G	4G	5G
1.	Generation	First	Second	Third	Fourth	Fifth
2.	Year of introduction	1970	1990	2001	2010	2020
3.	Technology	Analog cellular	Digital cellular	Broadband, IP, FDD, TDD	IP-broadband Wi-Fi, MIMO	IPv6
4.	Standard	AMPS	CDMA, TDMA, GSM	CDMA, UMTS, W-CDMA	Wi-Max and LTE	Yet to be finalized
5.	Switching	Circuit	Circuit / Packet	Circuit/Packet	Packet	packet
6.	Frequency band	824-894 MHz	850-1900 MHz	1.6-2.5 GHz	2-8 GHz	15 GHz
7.	Data speed	2.4 kbps	9.6 kbps	2 Mbps	50 Mbps	Higher than 1Gbps
8.	Multiplexing	FDMA	CDMA, TDMA	CDMA	MC-CDMA OFDM	MC-CDMA, LAS-CDMA, OFDM
9.	Core network	PSTN	PSTN	Packet Network	Internet	Internet
10.	Services	Only voice or only message	Digital voice, Data, SMS	High speed data, Voice, Video	Dynamic Information Access.	Interactive multimedia, Voice over IP

Review Questions

Q. 1 Explain evolution of wireless communication in detail.

Q. 2 Explain the first generation of wireless communication and state its features.

Q. 3 Explain the second generation of wireless communication and state its features.

Q. 4 Explain the evolution of wireless networks from 2G to 2.5G.

Q. 5 Explain the third generation of wireless communication and state its features.

Q. 6 State the advantages and disadvantages of 3G standards.

Q. 7 State the services provided and advantages of the third generation of wireless communication.

Q. 8 Explain the fourth generation of wireless communication and state its features, advantages, disadvantages and applications.

Q. 9 State the specifications of 4G LTE.

Q. 10 Elaborate the frequency bands used in LTE technology.

Q. 11 Explain LTE network architecture.

Q. 12 Explain the LTE frame structure in detail.

Q. 13 Write a short note on : the fifth generation of wireless communication.

Q. 14 Discuss the requirements of 5G.

Q. 15 Explain open wireless 5G architecture

Q. 16 Write short notes on : Disruptive technologies for 5G.

Q. 17 List out various challenges for 5G service.

Q. 18 Compare the five generations of wireless communication.

❑❑❑

Unit IV

Chapter 4

Mobile Network Layer

Syllabus

Mobile IP : Goals, Assumptions and requirements, Entities and Terminology, IP packet delivery, Agent advertisement and discovery, Registration, Tunneling and Encapsulation, Optimizations, Reverse tunneling.

IPv6 : DHCP, **AdHoc networks** : Routing, **Proactive protocol** : DSDV, **Reactive Routing Protocols** : DSR, AODV, Hybrid routing-ZRP, **Multicast Routing** : ODMRP, Vehicular Ad Hoc networks (VANET), MANET versus VANET Security.

Chapter Contents

4.1 Mobile IP :

- Mobile IP is the extension of IP protocol. It has been developed for the mobile and personal computers such as notebook.
- Mobile IP allows the mobile computers to get connected to the Internet at any location.

4.1.1 Goals, Assumptions and Requirements :

- Mobile computing is clearly the paradigm of the future.
- Hundreds of millions of users use the internet for global data communication. Then, why can't we simply use a mobile computer in the internet?
- The reason is that in such a case your mobile computer will not receive a single packet as soon as you leave your home network and reconnect your computer (wireless or wired) at another place.
- The home network is the network your computer is configured for.
- We need to know the routing mechanisms on the internet to know the reason for the above mentioned problem.

Routing on the internet :

- The stepwise routing process on the internet is as follows :
- A host sends an IP packet. The header of this packet contains a destination address and some other fields.
- The destination address determines the receiver of the packet, as well as the physical subnet of the receiver.
- Consider the destination address 130.13.42.69. It indicates that the receiver is to be connected to the physical subnet with the network prefix 130.13.42 unless CIDR is used.
- Routers in the internet read the destination addresses of the arriving packets and forward them according to their internal look-up tables.
- In order to avoid very long routing tables, only prefixes are stored.
- Otherwise, each router would have to store the addresses of all computers in the internet, which is not feasible.
- As long as the receiver is within its physical subnet, the packets with this prefix can be delivered to it.
- As soon as it leaves the home network, the physical subnet address no more remains the same and it is not possible to deliver packets to it.
- This is why it is not possible to simply connect a mobile computer to the internet and a host needs a so-called topologically correct address.

4.1.1.1 Quick 'Solutions' :

- One quick solution to this problem can be to assign the mobile computer a new, topologically correct IP address every time it moves.
- Many users do it with the help of DHCP.
- So, as soon as a mobile computer moves to a new location assign it a new IP address.
- However the problem with this quick solution is that nobody knows about this new address.
- It is next to impossible to find a mobile host on the internet which has just changed its address.

4.1.1.2 Requirements :

- Since the quick 'solutions' fail to work, it was necessary to design a more general architecture.
- The mobility in the internet was enabled when many field trials were carried out to finally lead to mobile IP as a standard.
- Following requirements accompanied the development of the standard :
  1. Compatibility
  2. Transparency
  3. Scalability and efficiency
  4. Security

1. Compatibility :

- There is a huge installed base of Internet computers. These computers are running TCP/IP and connected to the internet.

- Therefore it is not possible for a new standard like mobile IP to introduce changes for applications or network protocols already in use.
- People do not want to change applications and operating systems just for mobility.
- Therefore, it is necessary to either integrate Mobile IP into the existing operating systems or at least work with them.
- Also the routers within the internet should not necessarily require other software to successfully work with mobile IP.
- For this, Mobile IP must remain compatible with all lower layers used for the standard IP (used for non-mobile computers).
- Mobile IP must use the same interfaces and mechanisms to access the lower layers as IP does and should not require special media or MAC/LLC protocols.
- And last but not the least, end-systems using a mobile IP implementation should be able to communicate with fixed systems without mobile IP.
- That means even with Mobile IP, users should be still able to access all the other servers and systems in the internet.
- For this to happen, we need to use the same address format and routing mechanisms.

2. Transparency :

- The meaning of transparency is that the mobility should remain 'invisible' for many higher layer protocols and applications.
- With may be some interruption in service, the higher layers should continue to work even if the mobile computer has changed location.
- For TCP this means that the computer must keep its IP address as explained above.
- If the duration of interruption of the connectivity is short, then TCP connections survive the change of the attachment point of mobile to the internet.
- Since most of today's applications have not been designed for use in mobile environments, the only effects of mobility on them should be a higher delay and lower bandwidth.
- However, some applications such as cost-based routing or video compression need to be 'mobility aware'.
- It is possible to use different networks, therefore, the software could choose the cheapest one.
- In case of a video application if it knows that only a low bandwidth connection is currently available, it could use a suitable compression scheme.
- We need to develop additional mechanisms to inform these applications about mobility.

3. Scalability and efficiency :

- When a new mechanism is introduced to the internet it should not adversely affect its efficiency.
- That means too many new messages flooding the whole network should not get generated due to enhancing IP for mobility.
- Special care should be taken by taking the lower bandwidth of wireless links into consideration.
- Many mobile systems have a wireless link to an internet attachment point. Therefore, only a few additional packets will be necessary between a mobile system and a node in the network.
- Looking at the number of computers connected to the internet and at the growth rates of mobile communication, it is expected that a variety of devices will be connected to the internet as mobile components.
- These mobile components could be cars, trucks, mobile phones, every seat in every plane around the world etc.
- Each one of them needs mobile IP to move between different networks.
- Therefore it is important for a mobile IP to be scalable due to a large number of participants in the whole internet, worldwide.

4. Security :

- Mobility is accompanied with many security problems.

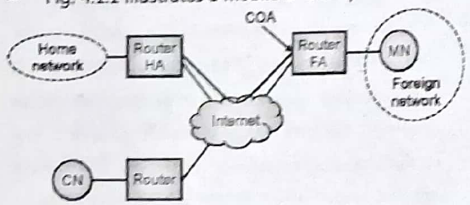
- Therefore it is essential that all the messages related to the management of Mobile IP should be authenticated.
- The IP layer must be sure that if it forwards a packet to a mobile host then this host will receive the packet.
- However, the IP layer only guarantees that the IP address of the receiver is correct, and nothing else.

Goal of mobile IP :

- The goal of mobile IP is to support the end-system mobility while maintaining scalability, efficiency, and compatibility in all respects with existing applications and Internet protocols.

4.1.2 Entities and Terminology :

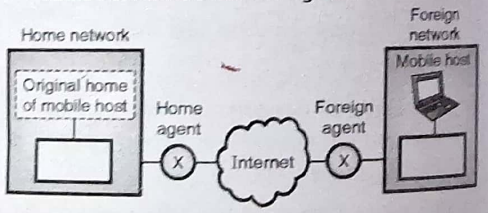
- In this section we will define different entities and terminologies related to mobile IP.
- Fig. 4.1.1 illustrates a mobile IP example network.



(G-3068) **Fig. 4.1.1 : Mobile IP example network**

- In Fig. 4.1.1 a CN (correspondent node) is connected via a router to the internet.
- The home network and the foreign network are also connected to the internet via routers.
- The HA (home agent) is implemented on the router that connects the home network to the internet.
- Similarly, a FA (foreign agent) is implemented on the router that connects the foreign network to the internet.
- The MN (mobile network) is currently in the foreign network.
- There is a tunnel formed to deliver packets to the MN which starts at the HA (home agent) and ends at the FA.

1. **Mobile node (MN) :**
   - A mobile node is defined as an end-system or router which can change its point of attachment to the internet with the help of mobile IP.
2. **Correspondent node (CN) :**
   - CN is defined as the partner that is needed for communication with MN.
   - The CN can be either a fixed node or mobile node.
3. **Home network :**
   - The home network is defined as the subnet the MN belongs according to its IP address. The mobile IP support is not required within the home network.
4. **Foreign network :**
   - We can define the foreign network as the current subnet the MN visits and which is not the home network.
5. **Home agent (HA) :**
   - A home agent is basically a router attached to the home network of a mobile host.
   - When a remote host sends a packet to the mobile host, the home agent acts on behalf of the mobile host, receives the packet and sends it to the foreign agent.
   - The position of home agent with respect to the home network is shown in Fig. 4.1.2.



(G-2257) **Fig. 4.1.2 : Home agent and foreign agent**

- The HA provides several services for the MN.
- The tunnel for packets toward the MN starts at the HA. The HA maintains a location registry, i.e., it is informed of the MN's location by the current COA.

6. **Foreign agent (FA) :**
   - A foreign agent is a router connected to the foreign network. The packets sent by the home agent are received by the foreign agent and delivers them to the mobile host.

- The position of the foreign agent with respect to the foreign network is as shown in Fig. 4.1.2.
- Sometimes, a mobile host itself can act as foreign agent.
- The FA provides different services as mentioned below to the MN when MN visits the foreign network.
  1. The FA can make the COA (defined below), act as tunnel endpoint and forward packets to the MN.
  2. The FA can work as the default router for the MN.
  3. FAs can also provide security services.

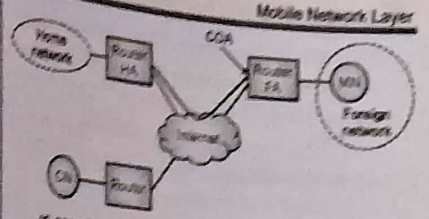
1. **Care-of address (COA) :**
   - The COA is used to define the current location (address) of the MN from an IP point of view.
   - All IP packets sent to the IP address of MN are not directly delivered to MN but they are delivered to the COA.
   - Packets are delivered to the MN by using a tunnel, as explained later.
   - The COA is the endpoint of this tunnel i.e., the address where packets exit the tunnel.
   - Following are two different possibilities for the location of the COA :
     1. **Foreign agent COA :** The COA could be located at the FA, i.e., the COA is an IP address of the FA.

        The FA is the tunnel end-point and forwards packets to the MN. Many MN using the FA can share this COA as common COA.
     2. **Co-located COA :** The COA is called as co-located COA if the MN temporarily acquires an additional IP address which acts as COA.

        This address is topologically correct, and the tunnel endpoint is at the MN.

4.1.3 IP Packet Delivery :

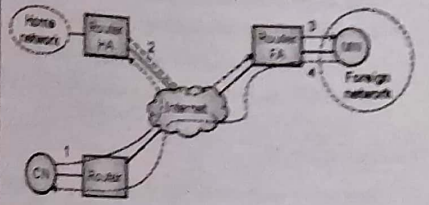
- Consider the example mobile network of Fig. 4.1.3, which consists of a home network, a home agent (home router), a foreign network along with a foreign agent, Internet, a correspondent node CN along with its router.



(G-3069) **Fig. 4.1.3 : Example network for mobile IP**

IP Packet delivery :

- Fig. 4.1.4 demonstrates the packet delivery to and from the MN using an example network.



(G-3069) **Fig. 4.1.4 : Packet delivery to and from the mobile node**

From CN to MN :

- A correspondent node CN wants to send an IP packet to the mobile node MN. A CN can be either a fixed or mobile node.
- One of the requirements of mobile IP is that it should hide the mobility of the MN.
- This is because, CN need not know anything about the MN's current location and it should send the packet as usual to the IP address of MN as shown as step 1 in Fig. 4.1.4.
- This means that in the IP packet sent by CN, the destination address is MN and the source address is CN.
- The Internet, which does not have any information about the current location of MN, routes the IP packet to the router responsible for the home network of MN (i.e. the home agent).
- This routing is carried out using the standard routing mechanisms of the Internet.
- The HA now receives the IP packet for the MN. The HA knows that MN is currently not in its home network.

- Therefore, the IP packet is not forwarded into the subnet as usual, but it is encapsulated and tunneled to the COA (care of address) as shown in Fig. 4.1.4.
- A new header is put in front of the old IP header that shows the COA as new destination and HA as source of the encapsulated packet as shown in step 2.
- The foreign agent decapsulates the IP packet, to remove the additional header, and forwards the original IP packet with CN as source and MN as destination to the MN as shown in step 3).
- Thus the MN receives the packet with the same sender and receiver address as it would have done in the home network.

From MN to CN :

- Now let us understand the process of sending IP packets from the MN to the CN.
- As shown in step 4 of Fig. 4.1.4, the MN sends the IP packet as usual with its own fixed IP address as source and CN's address as destination.
- The FA acts as default router and forwards this IP packet in the same way as it would do for any other node in the foreign network.
- If CN is a fixed node the routing process takes place in the fixed Internet as usual.
- However, if CN is also a mobile node residing in a foreign network, then the same mechanisms described in steps 1 through 3 would apply now in the other direction.

4.1.4 Agent Discovery :

- The first problem faced by an MN after moving into a foreign network is how to find a foreign agent.
- How does the MN discover that it has moved ?
- For this purpose mobile IP describes the following two methods :
  1. Agent advertisement and
  2. Agent solicitation
- Both these are actually router discovery methods plus extensions.

4.1.4.1 Agent Advertisement :

- For the agent advertisement method, the foreign agents and home agents advertise their presence periodically with the help of special agent advertisement messages.
- These advertisement messages are in the form of beacon broadcast into the subnet.
- For these advertisements ICMP messages are used with some mobility extensions.
- Fig. 4.1.5 shows the agent advertisement packet with the extension for mobility.

0 7	8 15	16 23	24 31
Type	Code	Checksum	
#addresses	add. size	Lifetime	
Router address 1			
Preference level 1			
Router address 2			
Preference level 2			
. . . .			

Type = 16	Length	Sequence number							
Registration lifetime		R	B	H	F	M	G	r	T	reserved
COA 1									
COA 2									
. . . .									

(G-3115) **Fig. 4.1.5 : Agent advertisement packet + mobility extension**

- There are two parts of this packet. The upper part is the ICMP packet while the lower part represents the extension needed for mobility.
- Fig. 4.1.5 does not show the fields necessary on lower layers for the agent advertisement.
- The packets should reach the mobile nodes with the appropriate link layer address.
- It is necessary to set the TTL field of the IP packet to 1 for all advertisements to avoid forwarding them.
- The IP destination address can be set to 224.0.0.1, if it is the multicast address for all systems on a link.

- The IP destination address can be set to 255.255.255.255 if it is the broadcast address.

Different fields in ICMP packet :

- Different fields in the ICMP part are defined as follows :
  1. Type field = 9,
  2. The code field = 0 if the agent also routes traffic from non-mobile nodes, or 16, if it does not route anything other than mobile traffic.
  3. #addresses field shows the number of addresses advertised with this packet.
  4. Lifetime field contents denote the length of time for which this advertisement is valid.
  5. Preference level is associated with each address which helps a node to choose the router which is the most appropriate one to get a new node.

Extension for mobility fields :

- The lower part of agent advertisement packet is the extension for mobility.
- It has the following fields :
  1. Type field = 16,
  2. Length field content depends on the number of COAs provided with the message and equals 6 + 4 * (number of addresses).
  3. The sequence number field shows the total number of advertisements sent by an agent since initialization.
  4. The registration lifetime field is used by the agent to specify the maximum lifetime in seconds a node can request during registration.

Bits specifying the characteristics of agent :

- The following bits are used to specify the characteristics of an agent in detail.
  1. The R bit (registration) : It shows, if a registration with this agent is required even when using a co-located COA at the MN.
  2. B bit : The agent can set the B bit if it is currently too busy to accept new registrations.
  1. The H and F bits : These two bits denote if the agent offers services as a home agent (H) or foreign agent (F) on the link where the advertisement has been sent.
  2. M and G bits : These two bits specify the method of encapsulation used for the tunnel. M specifies minimal encapsulation and G generic routing encapsulation.
  3. The r bit : This bit specifies the use of header compression which is set to zero and must be ignored.
  4. The T bit : This bit indicates that reverse tunnelling is supported by the FA.
- The fields that follow the T bit contain the COAs advertised e.g. COA 1, COA 2 etc.
- A foreign agent that sets the F bit must advertise at least one COA.
- A MN in a subnet can receive agent advertisements from either its home agent or a foreign agent.
- This is one way for the MN to discover its location

4.1.4.2 Agent Solicitation :

- The mobile node must send agent solicitations if the agent advertisements are not present or the inter-arrival time is too high, and an MN has not received a COA by other means, e.g. DHCP.
- The agent solicitations are based on RFC 1256.
- It must be ensured that these solicitation messages do not flood the network even when a MN searches for an FA endlessly sending out solicitation messages.
- As soon as a mobile node enters a new network, it can send out three solicitations, one per second.
- Note that in highly dynamic wireless networks with moving MNs, even one second intervals between solicitation messages might be too long.
- Many packets will be lost before an MN even gets a new address, without additional mechanisms.
- If a node does not receive any answer to its solicitations, then it should decrease the rate of solicitations exponentially, until the maximum interval between solicitations (typically one minute) is reached.

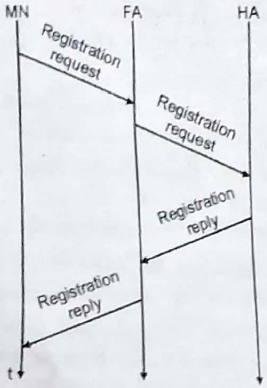
- It is possible to discover a new agent anytime, not just if the MN is not connected to a new agent.
- Let us consider a situation where an MN is looking for a better connection but is still sending via the old path.
- The MN goes through the steps of advertisements or solicitations after which it can now receive a COA, either one for an FA or a co-located COA.
- The MN knows its current location (either home network or foreign network) and the capabilities of the agent.
- As the next step, the MN should register with the HA (the **registration** process) if the MN is in a foreign network as described below.

4.1.5 Registration :

- After receiving a COA, the MN needs to register with the HA.
- The purpose of the registration is to inform the HA of the current location of MN so that HA can correctly forward packets to MN.
- It is possible to carry out registration in the following two different ways depending on the location of the COA.
  1. If the COA is at the FA
  2. If the COA is co-located

1. If the COA is at the FA :

- Fig. 4.1.6 (a) demonstrates the registration process if the COA is at the FA.

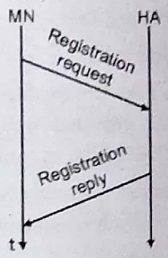


(G-3116) **Fig. 4.1.6(a) : Registration of a mobile node via the FA**

- As the first step the MN sends its registration request containing the COA) to the FA (see Fig. 4.1.6(a)). The FA forwards this request to the HA.
- The HA now sets up a mobility binding that contains the mobile node's home IP address and the current COA.
- In addition to this, the mobility binding also contains the lifetime of the registration which is negotiated during the registration process.
- Registration will expire automatically after the lifetime and will be deleted. Therefore, an MN should reregister before expiration.
- This mechanism is essential to avoid mobility bindings which are not in use anymore.
- After setting up the mobility binding, the HA sends a registration reply message back to the FA which forwards it to the MN as shown in Fig. 4.1.6 (a).

2. If the COA is co-located :

- Fig. 4.1.6 (b) illustrates the registration process if the COA is co-located.

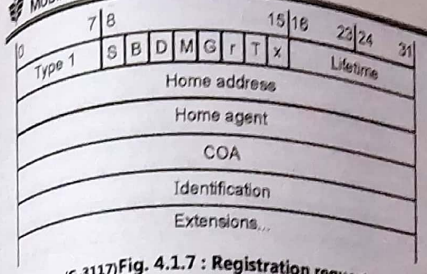


(G-3116) **Fig. 4.1.6(b) : Registration of a mobile node directly with the HA**

- The MN sends the registration request message directly to the HA.
- And in response, the HA can send the registration reply message directly to HA as shown.
- This is also the registration procedure for MNs that are returning to their home network.

Registration request :

- The registration request packet is as shown in Fig. 4.1.7.



(G-3117) **Fig. 4.1.7 : Registration request**

- For registration requests , the UDP packets are used.
- UDP is used because it has low overheads and better performance as compared to TCP in wireless networks.
- In the registration request the IP source address is set to the interface address of the MN, while, the IP destination address is set to that of the FA or HA depending on the location of the COA.
- The UDP destination port is set to 434.
- Different fields used in registration request message are set as follows (Fig. 4.1.7) :

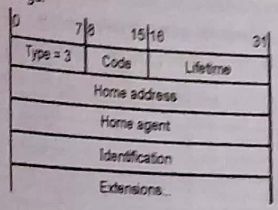
1. **Type** : This field is set to 1 for registration request.
2. **S bit** : The S bit is used by an MN to specify if it wants the HA to retain prior mobility bindings. This allows for simultaneous bindings.
3. **B bit** : If this bit is set then it indicates that an MN also wants to receive the broadcast packets which have been received by the HA in the home network.
4. **D bit** : The D bit indicates that if an MN uses a co-located COA, it also takes care of the decapsulation at the tunnel endpoint.
5. **M and G bits** : The bits M and G denote the use of minimal encapsulation or generic routing encapsulation, respectively.
6. **T bit** : The T bit indicates reverse tunnelling.
7. **r and x bits** : These bits are set to 0.
8. **Lifetime** : This field denotes the validity of the registration in seconds.

- **Home address** : this address is the fixed IP address of the MN.
- **Home agent** : This is the IP address of the HA.
- **COA** : Represents the tunnel endpoint.
- **Identification** : This is a 64 bit field which is generated by the MN to identify a request and match it with registration replies.
- It is also used for protection against replay attacks of registrations.

9. **Extensions** : This field must at least contain parameters for authentication.

Registration reply :

- Fig 4.1.8 shows the registration reply message. UDP packets are used for the registration reply message.



(G-3118) **Fig. 4.1.8 : Registration reply**

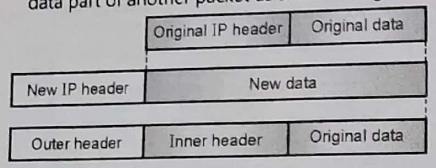
- Various fields in registration reply message are as follows :

1. **Type** : The type field set to 3.
2. **Code** : this field indicates the result of the registration request.
3. **Lifetime** : This field shows how many seconds the successful registration is valid.
4. **Home address** : Home address is the addresses of the MN.
5. **Home agent** : Home agent is the addresses of the HA.
6. **Identification** : This 64-bit field is used to match registration requests with replies.
7. **Extensions** : This field must at least contain parameters for authentication.

4.1.6 Tunneling and Encapsulation :

- The following discussion we will describe the mechanisms used to forward IP packets between the HA and the COA, as shown in step 2 of Fig. 4.1.4.

- A **tunnel** is used to establish a virtual pipe to send data packets between a tunnel entry and a tunnel endpoint.
- Packets entering a tunnel are forwarded inside the tunnel and leave the tunnel without any change.
- Thus, we can define tunneling as sending a packet through a tunnel without changing it.
- Tunneling, is achieved by using encapsulation.
- **Encapsulation** is the mechanism in which a packet consisting of packet header and data is taken and put into the data part of a new packet as shown in Fig. 4.1.9.
- Decapsulation is exactly the reverse process of encapsulation where, a packet is taken out of the data part of another packet as shown in Fig. 4.1.9.



(G-3070) **Fig. 4.1.9 : IP encapsulation**

- Encapsulation is performed when a packet is transferred from a higher protocol layer to a lower layer.
- Whereas, decapsulation is performed when a packet is transferred from a lower layer to a higher layer.
- Here these functions are used within the same layer.
- This mechanism is illustrated in Fig. 4.1.9 and shows exactly what the HA at the tunnel entry does.
- The HA takes the original packet with the MN as destination, puts it into the data part of a new packet and sets the new IP header in such a way that the packet is routed to the COA.
- The new header is also called the **outer header** for obvious reasons.
- Additionally, there is an **inner header** which can be identical to the original header as this is the case for IP-in-IP encapsulation, or the inner header can be computed during encapsulation.

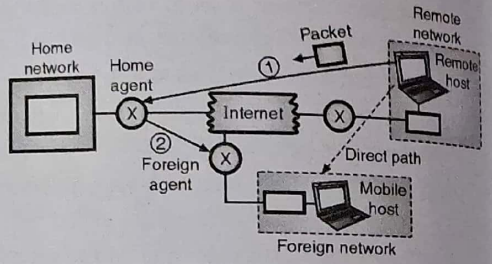
4.1.7 Optimizations :

Inefficiency in Mobile IP :

- The communication done with the help of mobile IP can be moderately to severely inefficient.
- The case of moderate inefficiency is also called as the **triangle routing, or dog leg routing** whereas the case of severe inefficiency is also referred to as **Double Crossing** or **2X.**

Triangle Routing or Dog Leg Routing :

- A triangle routing or dog leg routing is a case of moderate inefficiency.
- It occurs when a remote host wants to send a packet to the mobile host which is not attached to its own (remote) network.
- This situation is illustrated in Fig. 4.1.10.



(G-2261) **Fig. 4.1.10 : Triangle routing**

- In this situation as well, if a mobile host wants to send a packet to a remote host it can do so directly without any loss of efficiency.
- But when a remote host wants to send a packet to a mobile host the packet has to first travel to the home agent and then to the mobile host as shown in Fig. 4.1.10.
- Thus the packet has to travel along two sides of a triangle instead of only one which is the direct path shown by a dotted line in Fig. 4.1.10.

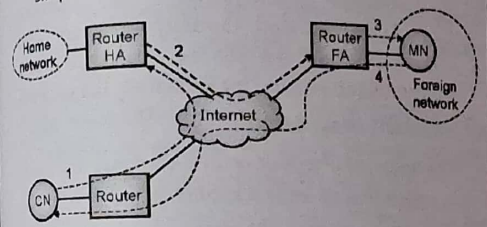
Remedy :

- Binding the care-of-address to the home address of mobile could be one of the solutions to the problem of inefficiency.
- That means when the home agent receives the first packet from the remote host and sends it to the foreign agent it should also send an **update binding packet** to the remote host.

- By doing this it is ensured that all the future packets to this mobile host can be sent to the **care-of-address** rather than **home** address.
- The remote host can save this information in a cache.
- However this remedy also has an inherent flaw. The **cache entry** would become outdated as the mobile host moves to a new network.
- To avoid this the **home agent** must send a **warning packet** to the remote host to inform that the mobile host has moved to a new network.

4.1.8 Reverse Tunnelling :

- Refer Fig. 4.1.11. At first glance, you may find the return path from the MN to the CN to be quite simple.



(G-3069) **Fig. 4.1.11 : Packet delivery to and from the mobile node**

- You might feel that the MN can directly send its packets to the CN without any problems.
- In this packet transaction, the destination address in the packets is that of CN.
- However there are following problems associated with this simple solution :
  1. Firewalls
  2. Multicasting
  3. TTL
- Let us discuss all of them.

1. **Firewalls :**

- Almost all companies and many other institutions use a firewall to secure their internal networks (intranet) connected to the internet.
- Therefore, all data to and from the intranet has to pass through the firewall.
- In addition to many other functions, one can set up firewalls to filter out malicious addresses from an administrator's point of view.
- Very often firewalls will only allow packets with topologically correct addresses to pass through them.
- However, MN still sending packets with its fixed IP address as source is not topologically correct in a foreign network.
- Firewalls often filter those packets coming from outside that contain a source address of the internal network computers.
- Now this means that, an MN cannot send a packet to a computer residing in its home network.
- Thus due to the firewall the destination address as well as the source message matters for forwarding IP packets.
- Some more complications arise when private addresses are used inside the intranet and they are translated into global addresses when communicating with the internet.
- Many companies use the network address translation (NAT) to hide internal resources such as routers, computers, printers etc. and to use only some globally available addresses.
- Problems invariably arise when using NAT together with mobile IP.

2. **Multi-cast :**

- Reverse tunnels are needed so as to ensure that the MN can participate in a multicast group.
- The nodes in the home network can participate in a multi-cast group, but, a MN in a foreign network cannot transmit multi-cast packets.
- The foreign network may not even provide the technical infrastructure required for multi-cast communication.

3. **TTL :**

- Consider that a MN is sending packets with a certain TTL when it is still in its home network.
- If the TTL is low enough then, no packet is transmitted outside a certain region.

- Now the MN now moves to a foreign network. So, this TTL might be too low and the packets will not reach the same nodes as before.
- So a MN may have to change the TTL value while moving and mobile IP is not transparent if a user has to adjust the TTL while moving.
- A reverse tunnel is needed because it represents only one hop, irrespective of how many hops are really needed from the foreign network to the home network.

Problems with reverse tunneling :

- The reverse tunneling creates a triangular routing problem in the reverse direction.
- Therefore, all packets sent by MN to a CN go through the HA.
- RFC 3024 does not provide any solution for this reverse triangular routing, because we don't know if the CN can decapsulate packets.
- Note that mobile IP should work together with all traditional, non-mobile IP nodes.
- Therefore, we cannot assume that a CN can be a tunnel endpoint.
- There are also several security issues related to the reverse tunneling which have not been really solved up to now.

4.2 IPv6 (Next Generation IP) :

- IPv6 is the next generation Internet Protocol designed as the next step of the IP version 4.
- IPv6 was designed to enable high-performance and larger address space.
- This was achieved by overcoming many of the weaknesses of IPv4 protocol and by adding several new features.

4.2.1 Advantages of IPv6 :

1. Improved header format :

- IPv6 uses an improved header format. In its header format the options are separated from the base header.
- These options are inserted when needed, between the base header and upper layer data.
- The routing process is simplified due to this modification.
- The speed of the routing process increases and the routing time is reduced.

2. Larger address space :

- IPv6 has 128-bit address, which is 4 times wider in bits is compared to IPv4's 32-bit address space. So there is a large increase in the address space.

Address space of IPv6 = (2^{128})

3. New options :

- IPv6 has increased functionality due to the addition of entirely new options that are absent in IPv4.

4. More security :

- IPv6 includes security in the basic specification.
- It includes encryption of packets (ESP : Encapsulated Security Payload) and authentication of the sender of packets (AH : Authentication Header) for enhancing the security.

5. Possibility of extension :

- The design of IPv6 is done in such a way that there is a possibility of extension of protocol if required.

6. Support to resource allocation :

- To implement better support for real time traffic (such as video conference), IPv6 includes flow label in the specification.
- With flow label mechanism, routers can recognize to which end-to-end flow the given packet belongs to.

7. Plug and play :

- IPv6 includes plug and play in the standard specification. It therefore must be easier for novice users to connect their machines to the network, it will be done automatically.

8. Clearer specification and optimization :

- IPv6 follows good practices of IPv4, and omits flaws / obsolete items of IPv4.

4.2.2 IPv6 in Mobile Computing :

- Mobile IP was originally designed for IPv4. However, IPv6 is more suitable for mobile computing applications.

- This is because, several mechanisms that are specified separately for mobility support come free in IPv6.
- For example one issue is security with regard to authentication, which is now a required feature for all IPv6 nodes.
- There is no need of any special mechanisms as add-ons for securing mobile IP registration.
- Every IPv6 node has the built in address autoconfiguration – the mechanisms for acquiring a COA in communication.
- Neighbor discovery is a mandatory mechanism for every node in the IPv6 specification.
- Therefore, special foreign agents are no longer needed to advertise services.
- Due to the inbuilt features of autoconfiguration and neighbor discovery, every mobile node can create or obtain a topologically correct address for the current point of attachment.
- As every IPv6 node is able to send binding updates to another node, the MN can send its current COA directly to the CN and HA.
- These mechanisms are an integral part of IPv6. It is also possible to facilitate a soft handover with IPv6.
- The MN sends its new COA to the old router servicing the MN at the old COA.
- In response to this, the old router encapsulates all incoming packets for the MN and forwards them to the new COA.
- Thus, mobile IP in IPv6 networks would require very few additional mechanisms of a CN, MN, and HA.
- The FA is no more needed with IPv6. A CN should only be able to process binding updates, that is, to create or to update an entry in the routing cache.
- That means, the MN itself should be able to carry out the following tasks :
  1. To decapsulate packets,
  2. To detect when it needs a new COA, and
  3. To determine when to send binding updates to the HA and CN.
- A HA must be able to encapsulate packets.
- However, IPv6 does not solve any firewall or privacy problems. Additional mechanisms on higher layers are needed for this.

4.3 Host Configuration : DHCP :

- DHCP (Dynamic Host Configuration Protocol) is the first client server application program that is used after a host is booted.
- Thus it works as a bootstrap when the host is booted and is to be connected to the Internet, but does not know its IP address.
- A computer that makes use of the TCP/IP suite must know its IP address.
- Along with its IP address it must also know the following information :
  1. Subnet mask of the computer.
  2. IP address of the router, so that it can communicate with other networks.
  3. IP address of the name server so that it can use the names instead of addresses.
- All this information can be saved in a configuration file and accessed by computer when booting takes place. This is known as host configuration process.
- But what will happen if the workstation is diskless or the computer is with a disc but it is being booted for the first time.
- If a computer is diskless, then it is possible to store the operating system and networking software in the ROM.
- But this information is not known to the manufacture and therefore cannot be stored in ROM.
- This information is dependent on the configuration of individual machine and it defines which network the machine is connected to.

4.3.1 Previously used Protocols :

- Now a days DHCP has become the formal protocol for host configuration. But the two protocols which were used earlier for the same purpose were RARP and BOOTP.

- RARP is Reverse Address Resolution Protocol and BOOTP stands for Bootstrap protocol.

4.3.2 DHCP :

- The Dynamic Host Configuration Protocol (DHCP) was devices by IETF in order to make the configuration automatic.
- Thus DHCP does not require an administrator to add an entry for each computer, to the database that a server uses.
- Instead, in DHCP a mechanism is provided for any computer to join a new network and obtain an IP address automatically with no manual intervention. This is known a plug and play networking.
- Thus DHCP allows the use of computers that, run server software as well as computers that run client software.
- When a computer that runs client software is shifted to a new network, it can use DHCP to obtain configuration information automatically.
- DHCP assigns a permanent address to a non-mobile computer that run server software.
- This address will not change when the computer **reboots.**
- To accommodate both type of computers, DHCP makes use of a client server approach.
- When a computer boots, it will broadcasts a DHCP Request. In response a server sends a DHCP Replay. An administrator can configure a DHCP server to have two types of addresses.
- First is the permanent address that are assigned to server computers, and second type is a pool of addresses which can be assigned on the basis of demand, when a computer boots and sends a request to DHCP.
- The DHCP find the configuration information by accessing its database If the database contains a specific entry for the computer then the server returns the information from the entry.
- However if there is no such entry exists for the computer, then the server chooses the next IP address from the pool and assigns it to the computer.

What is DHCP ?

- DHCP, as the name suggests, is a protocol used for dynamically configuring the hosts on a network, such as workstations, personal computers and printers.
- DHCP can help in assigning various types of information such as routing information, directory-services information and default web server and mail servers.
- However, the most important and commonly used information for which DHCP is used is the IP address and subnet mask information.
- DHCP was primarily designed for managing the network and the clients automatically. With DHCP, it is not necessary to configure the network and client information manually for individual hosts.
- In addition, DHCP can coexist with statically configured hosts with fixed IP addresses. DHCP can also carry out the allocation of certain configuration information to a host on a permanent basis.
- This protocol provides a four point information (IP address, subnet mask, IP address of router, IP address of name server) to a diskless computer or to a computer which is booted for the first time.
- It is a client / server protocol which is backward compatible to the BOOTP.

4.3.3 Advantages of DHCP :

- The use of DHCP on a network offers the following advantages :

1. It sets free the network administrator from the duties of setting up the configuration information, such as the IP address, the subnet mask, and the routing tables, manually. The DHCP simplifies network administration by doing these tasks automatically.
2. Avoids this and the sometimes the same IP address is assigned to two different hosts. The DHCP avoids this and the consequent malfunctioning of both the hosts from happening.

3. If the DHCP was not used, then the movement of computers from one network to another requires must be reconfigured. With DHCP, you can move the computers to different subnets or networks without the need to reconfigure them. In such situations, DHCP takes care of IP address assignment and other configuration details.
4. Mobile computers, such as laptops and palmtops, can easily get connected to different networks. They don't require reconfiguration any more as they get their configuration information from the DHCP server.
5. DHCP allocates IP addresses from a pool of IP addresses. In addition, when a computer gets disconnected, its released IP address is returned to the resource pool. Therefore, the possibility of having unused IP addresses are minimized.

4.3.4 Components of DHCP :

- The use of DHCP on a network requires the following three components :

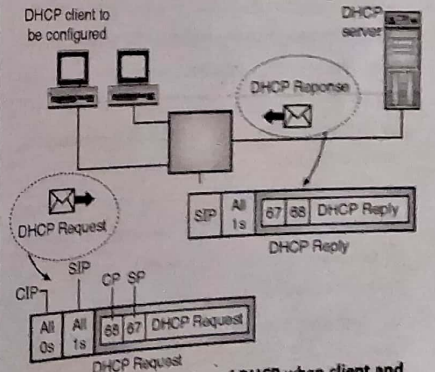
1. **DHCP server :**
   - It assigns the IP address and other information to the clients when they request for the information.
2. **DHCP client :**
   - It communicates with the DHCP server to get the desired information regarding its configuration. This communication can take place when the computer starts.
   - The user of the DHCP client can also initiate a DHCP client request to the DHCP server to renew its information.
3. **DHCP relay agent :**
   - It is used to relay (forward) client requests to the DHCP server. This is required when the DHCP server is yet to assign the client an IP address.
   - Without an IP address, a client cannot use IP routing on its own.
   - A DHCP relay agent helps the client to communicate with the DHCP server when the client does not have an IP address.
   - When a client starts, it has an IP address of 0.0.0.0. It sends a broadcast message containing its MAC address and the computer name.
   - In response the DHCP server sends an offer message that contains the MAC address of the client, the IP address offered to that client, the lease period for which the IP address will remain valid and its own IP address.
   - The lease period is the time duration for which a client can use the IP address that has been assigned to it by the DHCP server.
   - You can configure a DHCP server to set the lease time.
   - When the client receives the IP address, it accepts the offer and then broadcasts the message that it has accepted the offer.

4.3.5 DHCP Operation :

- We will discuss the DHCP operation under two different operating conditions :
  1. DHCP client and server on the same network.
  2. DHCP client and server on different networks.

Operation on the same network :

- This situation is not a very common one. But sometimes the DHCP client and server happen to be on the same network as shown in Fig. 4.3.1.



(G-1789) **Fig. 4.3.1 : Operation of DHCP when client and server are on the same network**

- The operation takes place as follows :
  1. The DHCP server sends a passive open command on port 67 of UDP and waits for clients response.
  2. The DHCP client sends an active open command on port 68 of UDP. This message is encapsulated in the UDP datagram with port 67 as destination port and port 68 as the source port. The UDP datagram is then encapsulated in an IP datagram. Note that the client at this time does not know its own IP address (i.e. the source address) and the server's IP address (destination address). Therefore the client uses an all zero address as source address and an all one address as destination address.
  3. The server responds to this message by sending either a broadcast or a unicast message using port 67. It uses port 68 as the destination port. Broadcast address is used only for those system which do not allow the bypassing of ARP.

4.4 Introduction to Wireless Network :

Definition :

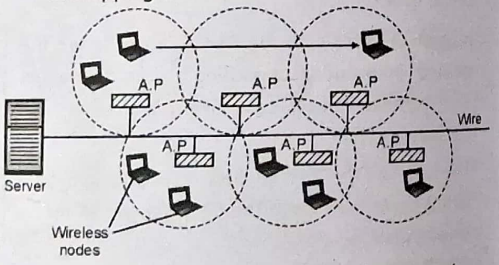
- A wireless network is defined as the collection of computers, servers, terminals and various other components, connected to each other by **wireless links** instead of connecting wires.
- Wireless networks use **radio waves** to connect one device in the network to the others.
- All the devices in a wireless network can be moved within the **range** or **coverage** area of the network.
- This makes the wireless networks extremely portable.
- In wireless networks **air** is used as a **medium** to transfer the data.
- A computer needs to have the **wireless networkcard** to be a part of a wireless network.
- It will have the same functionality, as that of a wired computer but its speed will reduce.
- A wireless network facilitates two or more terminals like laptops, Personal Digital Assistants (PDA) etc. to communicate without any hard-wired link.
- In a wireless network, a user can move around within a specified geographical area (range) without getting disconnected from the network.
- Thus wireless networks allow its users to have **mobility.**
- The best **example** of a wireless network in our day to day life is the cellular phone network.

Applications :

- A few applications of wireless networks are :
  1. Telecommunications.
  2. Satellite communication.
  3. Police wireless.
  4. Domestic and office applications.
  5. Data sharing applications.
  6. To provide a fast and inexpensive internet connectivity.

An Example of Wireless Network :

- Fig. 4.4.1 shows an example of a wireless network.
- Several **Access Points (AP)** are connected to a server through a cable. Each A.P. will connect the wireless nodes within its coverage area to the server.
- The wireless network of Fig. 4.4.1 has many overlapping **radio cells.**



(O-922) **Fig. 4.4.1 : An example of wireless network**

- A **radio cell** is defined as the coverage area of an access point A.P.
- Each radio cell provides wireless connectivity to nodes present inside its coverage area.

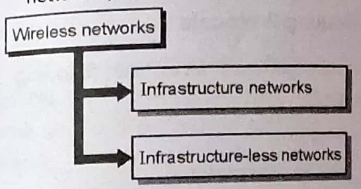
Frequency :

- Each radio cell in a wireless network operates at a **different frequency** and covers a specific area.

- Due to different operating frequencies, the nodes in adjacent radio cells will not interfere with each other.
- The distributed A.P.s are connected to each other by wires. They form a distributed system.
- The nodes located within a radio cell can communicate with each other through the A.P. of that cell.

4.4.1 Classification of Wireless Networks :

- Wireless networks offers flexibility of connection between the users at different places and they also provide network extension to any area or building without any physical wired connection.
- Such connections are classified into two types. They are as follows :
  1. Infrastructure networks.
  2. Infrastructure-less networks (Ad-hoc networks).

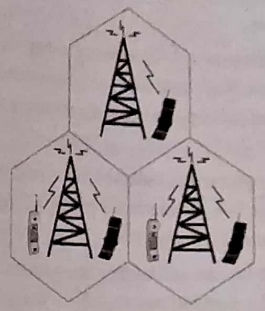


(G-2024) **Fig. 4.4.2 : Classification of wireless networks**

4.4.2 Infrastructure Networks :

Definition :

- It is a network with pre-constructed infrastructure made up of fixed network nodes and gateways where network services are delivered through pre-constructed infrastructures.
- Infrastructure based networks consists of fixed base stations called as access points and they are connected by wires. When mobile node is inside the communication range they can communicate with BS via wireless link.
- When mobile node is not in the communication range of base station then it makes connection with other base stationfor communication.
- In this type of networks, access point is a central controller for each device.
- Installation of such infrastructure is too expensive and it is sometimes technically impossible for some remote localities.
- In this type of networks all devices on a wireless network communicate with each other through a wireless router (Access point). Fig. 4.4.3 shows an infrastructure based network.



(G-1880) **Fig. 4.4.3 : Infrastructure network**

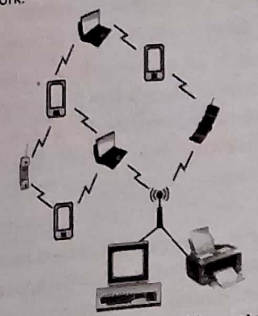
Examples :

- Cellular networks, satellite networks and wireless local area networks are examples of infrastructure networks.

4.4.3 Infrastructure-less Networks / Ad-hoc Networks :

Definition :

- A Wireless ad hoc network is a group of independent terminals or nodes which communicate with each other by forming a multi-hop radio network.
- It maintains a connectivity in decentralized manner. Fig. 4.4.4 shows infrastructure-less network.



(G-1881) **Fig. 4.4.4 : Wireless Ad hoc network**

- Nodes in this networks have to compete with some effects of radio communication such as interference, noise and fading etc. as the communication between the nodes takes place over the wireless (Radio) links,
- The links in wireless ad hoc network have less bandwidth as compared to that of a wired network.
- In wireless ad hoc networks or simply an ad hoc network, each node acts as a router and a host and the network control is distributed among the nodes.
- Infrastructure-less network is a group of self-configurable, autonomous, self-organizing nodes connected by wireless links.
- The nodes can move freely from one to another place by changing the topology regularly because there is no predefined infrastructure. In this type of networks, communication between nodes takes place using multi-hop communication.
- In Ad hoc network, the network formation takes place dynamically through the collaboration of an random set of independent nodes.
- Due to decentralized network, topology discovery and delivery of message should be executed by nodes themselves.
- The nodes can join or leave the network at any time due to free node mobility.
- Wireless ad hoc networks allows short range and long range communication ranging from wireless mobile networks to wireless sensor networks.
- In wireless Ad hoc network, the network topology is highly dynamic due to mobility of nodes.
- The nodes in ad hoc network can communicate with each other directly within their wireless range whereas they can communicate with each other by using multi-hop radio network beyond their wireless range.

Examples :

- Examples of wireless ad hoc network are Wireless Sensor Networks (WSN), Mobile Ad hoc Networks (MANETs), Wireless Mesh Networks (WMN) and Vehicular Ad hoc Networks (VANETs) etc.

4.5 Issues and Challenges in Ad hoc Networks :

- To design an adhoc wireless networks, the following are major challenges and issues which affect the design, operation and performance of an ad hoc wireless system :
  1. Routing protocols.
  2. Multicasting.
  3. Transport layer protocol.
  4. Pricing schemes.
  5. Medium access scheme.
  6. Self-organisation.
  7. Security.
  8. Quality of service.
  9. Management of energy.
  10. Scalability.
  11. Discovery of service and addressing.
  12. Deployment considerations.

4.5.1 Routing Protocols :

- A routing protocol faces the following major challenges :
  1. Mobility.
  2. Bandwidth constraint.
  3. Error prone and shared channel.
  4. Location dependent contention.
  5. Other resource constraints.

1. Mobility :

- The mobility associated with the nodes is one of the most important properties of ad hoc wireless networks. The mobility of nodes is associated with problems such as frequent path breaks, packet collisions, transient loops, stale routing information and difficulty in resource reservation.
- A good routing protocol should solve all these problems related to mobility.

2. Bandwidth constraint :

- The total available bandwidth of a wireless link is decided by the number of nodes and the traffic they handle. Thus the bandwidth available for every node is only a fraction of the total bandwidth.

3. Error-prone and shared channel :

- The wireless channels are more prone to errors than the wired channels. In other words the bit error rate (BER) in a wireless channel is very high (typically 10^{-5} to 10^{-3}) as compared to that in wired channels (typically 10^{-12} to 10^{-9}).
- The design of routing protocols should take this fact into consideration.
- The efficiency consideration of the routing protocol can be improved by considering the state of the wireless link, path loss and signal to noise ratio.

4. Location-dependent contention :

- The load on the wireless channel is directly dependent on the number of nodes present in a given area. With increase in the number of nodes the contention for the channel becomes high.
- The high contention for the channel will give rise to a large number of collisions and a wastage of bandwidth.
- In order to overcome this problem, a routing protocol must have a built-in mechanisms to distribute the network load uniformly across the network.

5. Other resource constraints :

- The capability of a routing protocol is also affected by the other constraints such as computing power, battery power and buffer storage.

Challenges for dynamic routing protocols :

- The ad hoc wireless networks do not have any infrastructure (they don't have the access points or base stations).
- Therefore computing the proper and efficient routes from source to destination is a bigger challenge for such networks.
- The mobility of nodes and frequent changes in network leads to unstable and improper routes.
- Hence, it is necessary to design dynamic routing protocols which can take care of continuously changing network configuration of wireless networks.

4.5.2 Multicasting :

- Multicasting is very important in applications like emergency search-and-rescue operations and military communication.
- In such applications, nodes form different groups need to carry out the point-to-multipoint and multipoint-to-multipoint voice and data communication.
- The multicast routing in wireless networks become challenging due to the factors like the mobility of nodes, power source and bandwidth.
- The multicast protocols designed for the traditional wired networks do not perform well in ad hoc wireless networks because these are tree based protocols.
- A tree-based multicast structure is highly unstable and need to be readjusted frequently to include broken links.
- We cannot use the global routing structure such as the link-state table because it results in high control overhead.
- If we use the single-link connectivity among the nodes in a multicast group it results in a tree-shaped multicast routing topology. Due to the frequent tree breaks, this topology provides high multicast efficiency with low packet delivery ratio.
- If multiple links are provided among the nodes in an ad hoc wireless network then we get a mesh-shaped structure. The mesh-shaped multicast routing structure may work well in a high-mobility environment.

Major problems in multicasting :

- The major problems in designing the multicast routing protocols are as follows :
  1. Robustness
  2. Efficiency
  3. Control overheads
  4. Quality of service
  5. Efficient group management
  6. Scalability
  7. Security

1. Robustness :

- The multicast routing protocol must be robust. That means it should be able to recover and reconfigure itself quickly from the link breaks induced by the node mobility.

2. Efficiency :

- The multicast routing protocol will be highly efficiency if it makes a minimum number of transmissions to deliver a data packet to all the group members.

3. Control overheads :

- A multicast routing protocol should generate a minimal control overhead for the multicast session so as to save the scarce bandwidth availability in ad hoc wireless networks.

4. Quality of service :

- The multicast routing protocols need the QoS support, since the data transferred in a multicast session in most cases is time-sensitive.

5. Efficient group management :

- Group management is the process of accepting multicast session members and maintaining the connectivity among them until the session expires.
- It is necessary to perform the group management with minimal exchange of control messages.

6. Scalability :

- The multicast routing protocol should be able to scale itself according to the network size specially for the networks with a large number of nodes.

7. Security :

- An authentication of session members is necessary in military communications in order to prevent the non-members from gaining unauthorized information.

4.6 Routing in Ad-hoc Network :

- Routing is an activity, which connects calls from source to destination in telecommunication networks.
- Routing plays an important role in construction, design and operation of the network.
- Ad-hoc networks are wireless networks where multi-hop links are used for communication of nodes.
- Routing is challenging task in Ad-hoc because there is constant change in topology of network due to mobility of nodes.
- To accomplish this task various protocols have been developed.

4.6.1 Design Issues / Challenges in Routing Protocol :

- The major design issues in ad-hoc networks are :
  1. Mobility of nodes.
  2. Error-prone shared broadcast radio channel.
  3. Resource constraints.
  4. Bandwidth constraints.
  5. Hidden and exposed terminal problem.

1. Mobility of nodes :

- In ad-hoc wireless networks, routing protocols should perform effective and efficient management of mobility of nodes.
- Due to movement of nodes, network topology becomes highly dynamic which results in the frequent path breaks in an ongoing session.
- Disturbance occurs in the network due to movement of end nodes or intermediate nodes in the path.
- The routing protocols of wired network cannot be used in adhoc wireless networks.

2. Error-prone shared broadcast radio channel :

- The routing protocols in Ad-hoc network should be able to search path with minimum congestion.
- A unique challenge in ad-hoc wireless networks is the broadcast nature of the radio channel.
- Time-varying characteristics of wireless links are link-capacity and link-error probability.
- The ad-hoc wireless routing protocols should interacts with MAC layer in order to find alternate routes through better link quality.

- In ad hoc network, the collision of data and control packets takes place at the time of transmission. This problem comes under hidden terminal problem.

3. Resource constraints :

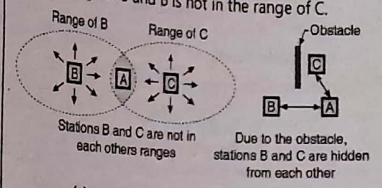
- Two main and limited resources in an ad-hoc wireless network are processing power and battery life.
- The devices used in ad-hoc network requires portability and hence they have weight and size constraints.
- The nodes in ad hoc network becomes bulky and less portable by increasing the battery power and processing ability.
- These resource problems should be managed in routing protocols in ad hoc network.

4. Bandwidth constraints :

- Large amount of bandwidth is available in wired networks due to the advent of fiber optics and the utilization of WDM technologies.
- A wireless networks provides less data rate as compared to wired network because of limited radio band.
- To overcome these bandwidth constraints such routing protocols are required that use optimum bandwidth by keeping overhead as low as possible.
- The limited bandwidth availability creates limitation on routing protocol in order to maintain the topological information.
- Due to dynamic topology, it is very difficult to maintain topological information at each node that occupies more control overhead that results in wastage of bandwidth.

5. Hidden terminal problem :

- The hidden station problem occurs when a station may not be aware that some other station is transmitting because of either range problem or some obstacle.
- In this situation collision may occur but may not be detected.
- The hidden station problem is illustrated in Fig. 4.6.1. Refer Fig. 4.6.1(a) which shows three wireless stations A, B and C.
- The transmission ranges of stations-B and C have been shown by the two ovals on left and right respectively which shows that station-C is not in the range of B and B is not in the range of C.



(a) (b)

(G-2098) **Fig. 4.6.1 : Hidden station problem**

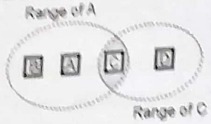
- However station-A is in the range of both B and C. So A can hear signals transmitted by B and C.
- Refer Fig. 4.6.1(a) where station-B is transmitting to station A.
- Now if station-C checks the medium to see if anyone is transmitting, it will not hear station B because it is out of range. So station-C will come to a wrong conclusion that no one is transmitting and so it can start transmitting to station A.
- If station-C starts transmitting, it will create a collision at station-A and will wipe out the frames from station-B.
- This problem in which a station is not able to detect an already transmitting other station which is too far away is called as the **hidden station problem.**
- In this example it is said that stations-B and C are hidden from each other with respect to station-A.
- Now consider Fig. 4.6.1(b) which shows the hidden station problem occurring due to an obstacle.

Note : Due to hidden station problem, the possibility of collision increases and the capacity network will reduce.

6. Exposed terminal problem :

- We have discussed the problem of hidden station. The **exposedstation problem** is a similar problem.

- In this problem, a station refrains from using the common medium even when no other station is using it (i.e. the channel is actually free).
- In order to understand this concept clearly refer Fig. 4.6.2 where A is the sending station and B is the destination. A is sending data to B.
- Station C wants to send its data to station D and it is possible to do so without interfering in the communication between A and B.
- As shown in Fig. 4.6.2, station C is in the range of station A. In other words C is exposed to A.



(0-918) Fig. 4.6.2 : Exposed station problem

- Therefore C listens to what A is transmitting and decides to refrain itself from sending its message to D.
- This causes wastage of channel capacity.

4.6.2 Characteristics/ Goals of Routing Protocols :

- The routing protocols for Adhoc wireless network should have following characteristics :

1. **Fully distributed :**
- It should be fully distributed. The centralized routing involves the risk of single point failure. The distributed routing is more fault-tolerant as compared to centralized routing.

2. **Adaptive to frequent in topology :**
- The routing protocol should be adaptive to frequent topology changes.

3. **Connection set up time :**
- Route calculation and maintenance in the network should involve minimum number of nodes. Minimum connection set up time is required because each node in the network must have quick access to routes.

4. **Localized :**
- The routing protocol should be localized as a huge state propagation control overhead is involved in a global state maintenance.

5. **Loop free :**
- The routing protocols in ad hoc network should be loop free and free from stale routes.

6. **Reliable transmission :**
- The packet transmission should be reliable, in order to reduce message loss and to prevent the occurrence of stale routes,.

7. **Quick convergence of routes :**
- The convergence should be fast. Once the network topology becomes stable, it should cover best possible routes.

8. **Use of limited resources :**
- It should use the limited sources like bandwidth, computing power, memory and battery power.

9. **Information storage :**
- Each node in the ad hoc network should try to store information regarding the stable local topology.

10. **Quality of service :**
- The routing protocol should provide a certain level of quality of service and provide support for the time-sensitive traffic.

4.6.3 Requirements of Routing Protocols :

- The important requirements of routing protocols are as follows :

1. **Minimum route acquisition delay :**
- It is the delay associated with the route acquisition for a node that does not have a route to a particular destination. This delay should be as small as possible. This delay depends on the size of the network and the network load.

2. **Quick route recognition :**
- The topology of wireless networks changes in an unpredictable manner. Therefore the routing protocol should be able to quickly reconfigure the route in order to minimize the path breaks and subsequent packet losses.

3. **Loop-free routing :**
- To provide a loop-free routing is a fundamental requirement of any routing protocol because it avoids any wastage of network bandwidth.
- In ad hoc wireless networks, transient loops may be formed due to the random movement of nodes.
- A routing protocol should detect such transient routing loops and take the necessary actions.

4. **Distributed routing approach :**
- An ad hoc wireless network is a distributed wireless network and therefore it needs the use of distributed routing approaches to reduce the consumption of bandwidth.

5. **Minimum control overhead :**
- The required number of control packets exchanged to find a new route and to maintain the existing routes should be minimized in order to minimize the bandwidth consumption and reduce their collisions with data packets, thereby improving the network throughput.

6. **Scalability :**
- Scalability of the routing protocol is defined as its ability to adapt or scale well (i.e., perform efficiently) in a network with a large number of nodes.
- This will be possible if the control overhead are minimized and the routing protocol adapts to the network size.

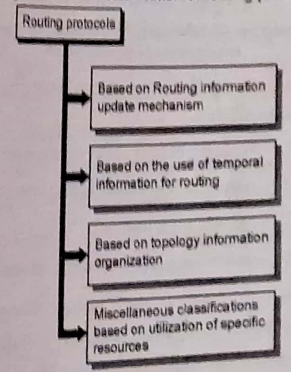
7. **Provisioning of QoS :**
- The routing protocol must provide a certain level of QoS as per the demand by the nodes or the category of service.
- The QoS parameters are bandwidth, delay, jitter, packet delivery ratio and throughput.

8. **Support for time-sensitive traffic :**
- The routing protocol needs to support the tactical communications and time-sensitive applications such as video streaming.
- The routing protocol must support both hard real-time and soft real-time traffic.

9. **Security and privacy :**
- The routing protocol in ad hoc wireless networks should be capable of handling the threats and vulnerabilities.
- It must avoid the security attacks like resource consumption, denial-of-service, impersonation, etc. against an ad hoc wireless network.

4.7 Classification of Routing Protocols :

- Fig. 4.7.1 shows classification of routing protocols.

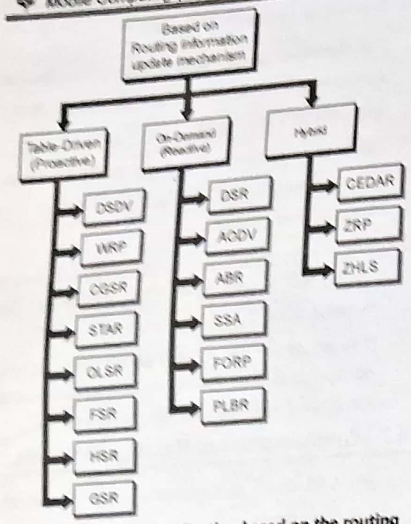


(0-961(a)) Fig. 4.7.1 : Classification of routing protocols

- Routing protocols for ad-hoc wireless network broadly classified into four categories based on :
  1. Routing information update mechanism.
  2. Use of temporal information for routing.
  3. Routing topology.
  4. Utilization of specific resources.

4.7.1 Based on the Routing Information Update Mechanism :

- Fig. 4.7.2 shows the classification routing protocols based on the Routing Information update Mechanism.



(O-961(a)) **Fig. 4.7.2 : Classification based on the routing information update mechanism**

- We can classify these protocols further into the following types :
  1. Pro-active or table-driven routing protocols
  2. Reactive or on-demand routing protocols
  3. Hybrid routing protocols

1. Pro-active or table-driven routing protocols :

- In this protocol, each node maintains the information of network topology. Information is maintained in the form of routing tables by exchanging routing information periodically.
- Following are pro-active or table-driven routing protocols :
  1. DSDV (Destination Sequenced Distance Vector Routing)
  2. WRP (Wireless Routing Protocol)
  3. CGSR (Cluster Head Gateway Switch Routing)
  4. STAR (Source Tree Adaptive Routing)
  5. OLSR (Optimized Link State Routing)
  6. FSR (Fisheye State Routing)
  7. HSR (Hierarchical State Routing)
  8. GSR (Greedy Perimeter Stateless Routing)

2. Reactive or on-demand routing protocols :

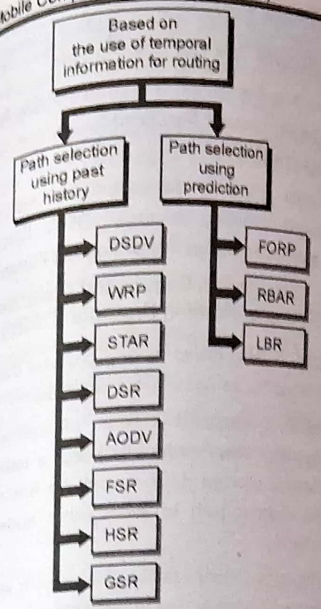
- These protocols do not maintain topology information.
- When required these protocols obtain a path by using a connection establishment process.
- Following are reactive or on-demand routing protocols :
  1. DSR (Dynamic source routing)
  2. AODV (Ad-hoc on demand distance vector routing)
  3. ABR (Associativity based routing)
  4. SSA (Signal stability based adaptive routing)
  5. FORP (Flow oriented routing protocol)
  6. PLBR (Preferred link based routing)

3. Hybrid routing protocols :

- These protocols combine the best features of proactive and reactive protocols.
- Table driven approach is used for routing within a particular geographical region.
- If nodes located beyond the geographical zone, on demand approach is used.
- Following are hybrid routing protocols :
  1. CEDAR (Core extraction distributed Adhoc routing)
  2. ZRP (Zone routing protocol)
  3. ZHLS (Zone-based hierarchical link state routing)

4.7.2 Based on the use of Temporal Information for Routing :

- Fig. 4.7.3 shows the classification routing protocols based on the use of temporal information for routing.



(O-961(c)) **Fig. 4.7.3 : Classification based on the use of temporal information for routing**

- Due to frequent path breaks and highly dynamic topology, use of temporal information is needed.
- These protocols are classified into two groups.
  1. Using past temporal information
  2. Using future temporal information

Using past temporal information :

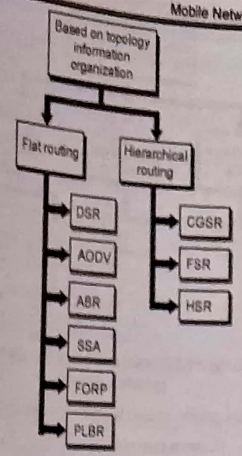
- In order to make appropriate routing decisions, these protocols use information of past status of the links or link status at the time of routing.

Using future temporal information :

- In order to make appropriate routing decisions, these protocols use information about expected future status of the wireless links.
- FORP (Flow Oriented Routing Protocol), RBAR (Receiver Based Auto rotate Protocol) and LBR (Load Balanced Ad-hoc Routing) are the protocols that use future temporal information.

4.7.3 Based on the Routing Topology :

- Fig. 4.7.4 shows the classification routing protocols based on the routing topology.



(O-961(d)) **Fig. 4.7.4 : Classification based on the routing topology**

- Ad-hoc wireless networks can use either flat or hierarchical topology due to their relatively small number of nodes.

Flat topology :

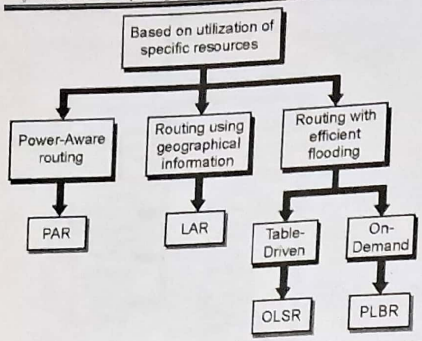
- These protocols use a flat addressing mechanisms similar to mechanisms used in 802.3 LANs.
- Flat topology assumes the presence of globally unique addressing method for nodes.

Hierarchical topology :

- Protocols under this category makes use of logical hierarchy and an associated addressing scheme.
- This topology is based on hop distance or geographical information.
- CGSR (Cluster Head Gateway Switch Routing), PAR (Preferred Ad-hoc Routing) and LAR (Location Aided Routing) are examples of protocols belonging to the hierarchical topology

4.7.4 Based on the Utilization of Specific Resources :

- Fig. 4.7.5 shows the classification routing protocols based on theutilization of specific resources.



(O-961(e)) **Fig. 4.7.5 : Classification based on utilization of specific resources**

- These protocols are classified into two groups.
  1. Power-aware routing
  2. Geographical information associated routing

Power-aware routing :

- Aim of routing protocol under this category is to minimize important resources such as battery power consumption in ad-hoc network.
- Appropriate routing decisions are based on reducing the power consumption either globally or locally.
- PAR (Preferred ad-hoc routing) is an example of Power-aware routing.

Geographical information associated routing :

- The protocols under this category will reduce the control overhead by utilizing the available geographical informationand improve the routing performance.
- LAR (Location Aided Routing) is an example of geographical information associated routing.

4.8 Table Driven Routing Protocols :

- These protocols are also known as **proactive protocols** because they maintain the routing information before it is required.
- Table driven protocols maintain topology information at each node in the form of table. These tables are frequently updated in order to maintain consistent and accurate network state information.
- Examples of table driven routing protocols are DSDV, WRP, HSR, GSR, FSR, FSLS etc.
- The proactive protocols are not suitable for large networks because they need to maintain table for each node.
- In the following subsection we will discuss proactive routing protocol : DSDV (Destination Sequenced Distance Vector Routing Protocol).

4.8.1 Destination Sequenced Distance Vector Routing Protocol (DSDV) :

- DSDV is table driven routing protocol that is first protocol for ad-hoc wireless networks.
- DSDV is improved version of Bellman Ford algorithm where each node keeps a table, which contains shortest distance and the first node on the shortest path to every other node in the network.
- The node includes table updates with increasing sequence number tags to prevent loops, to answer the count to infinity problem and faster convergence.
- In DSDV, routes to all destinations are readily available at every node at all times as it is table driven protocol.
- In order to maintain an up to date record of the network topology, the tables are exchanged between neighbors at regular intervals.
- If node observes major change in local topology, then tables are forwarded to neighboring nodes.
- There are two types of table updates :
  1. Incremental updates.
  2. Full dumps.

1. Incremental updates :

- An incremental update takes single Network Data Packet Unit (NDPU). Incremental updates are usedwhen there is no significant change observed in topology

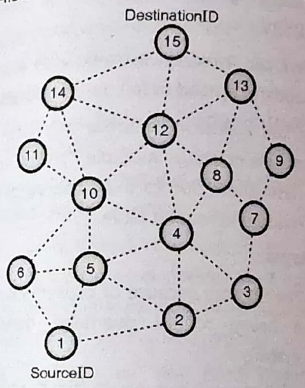
2. Full dumps :

- Full dump can take multiple NDPUs. A full dump is done when the local topology changes significantly or when an incremental update needs more than single NDPU.

- Destination nodes initiate the table update with new sequence number. Initiated sequence number is always greater than previous one.
- Once updated table is obtained, nodes update its tables based on received data or hold it for some time to select the best metric.
- Metric can be the smallest number of hops received from multiple versions of the same update table from various neighboring nodes.
- A node may forward or decline the table based on the sequence number of the table update.

Route establishment in DSDV :

- Fig. 4.8.1 shows route establishment in DSDV.



(a) Topology graph of the network

Dest	NextNode	Dist	SeqNo
2	2	1	22
3	2	2	26
4	5	2	32
5	5	1	134
6	6	1	144
7	2	3	162
8	5	3	162
9	2	4	186
10	6	2	142
11	6	3	176
12	5	3	190
13	5	4	198
14	6	3	214
15	5	4	256

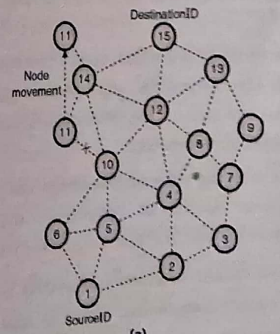
(b) Routing table for node 1

(G-1690) **Fig. 4.8.1 : Route establishment in DSDV**

- In Fig. 4.8.1(a), node 1 is the source node and node 15 is the destination node.
- The route already exists as shown in Fig. 4.8.1(b) because all the nodes maintains global topology information.
- The routing table of source node 1 shows that smallest route to the destination node 15 exists through node 5 and its minimum distance is 4 hops as shown in Fig. 4.8.1(b).

Route maintenance in DSDV :

- The protocol handles the reconfiguration of path used by on-going data transfer in the following way :
- The last node of broken link begin a table update message with the weight of broken link assigned to ∞ and with a sequence number larger than the registered sequence number for that destination node.
- Once, a node get an update table with weight '∞', each node immediately circulate it to its adjacent nodes to broadcast the broken link data to the entire network.
- Hence, breaking of single link leads to the propagation of table update information to the entire network.
- A node assigns an odd sequence number to the link break update to distinguish it from the even sequence number generated by the destination node.
- Consider the case when node 11 moves from its current location, is as shown in Fig. 4.8.2(a).



(a)

Fig. 4.8.2 (Contd...)

Dest	NextNode	Dist	SeqNo
2	2	1	22
3	2	2	26
4	5	2	32
5	5	1	134
6	6	1	144
7	2	3	162
8	5	3	162
9	2	4	186
10	6	2	142
11	5	4	180
12	5	3	190
13	5	4	198
14	6	3	214
15	5	4	256

(b)

(G-1691) **Fig. 4.8.2 : Route maintenance in DSDV**

- When an adjacent node observe the link break, it establishes all the paths passing through the broken link with distance as '∞'.
- For example, when node 10 is aware of the link failure, it sets the path to node 11 as ∞ and transmits its routing table to its neighboring nodes.
- The neighboring nodes finding important changes in their routing tables retransmit it to their neighbors. In this way, broken link information spreads all over the network.
- Node 1 also establish the distance to node 11 as '∞'.
- When node 14 gets a table update message from node 11, it informs the neighbors about shortest distance to node 11. This information is circulated throughout the network.
- After receiving new update message with higher sequence number, all nodes saves the new distance to node 11 in their corresponding tables.
- Fig. 4.8.2(b) shows updated table at node 1, where the current distance from node 1 to node 11 is increased from 3 to 4 hops.

Advantages :

1. The availability of paths to all destinations in network always shows that less delay is required in the path set up process.
2. DSDV protocol guarantees loop free path.
3. The method of incremental updates with sequence number labels, makes the existing wired network protocols adaptable to ad hoc wireless networks.
4. With few modificationin wired network protocol can be applied to adhoc wireless networks.
5. Count to infinity problem is reduced in DSDV.
6. We can avoid extra traffic with incremental updates instead of full dump updates.
7. DSDV maintains only the best path instead of maintaining multiple paths to every destination. Due to this, amount of space is reduced in routing table.

Disadvantages :

1. The updates because of broken links lead to a heavy control overhead during high mobility.
2. DSDV doesn't support multipath routing.
3. The small network with high mobility or large network with less mobility can totally block the existing bandwidth.
4. DSDV protocol suffers from too much control overhead which is proportional to the number of nodes in the network, this is not scalable in ad hoc network. It has restricted bandwidth and network topologies are highly dynamic.
5. To obtain information about a specific destination node, a node has to wait for table update message sent by same destination node. This wait could results in stale routing information at nodes.
6. It is difficult to maintain routing table for larger network.

4.9 On-Demand Routing Protocol :

On demand routing protocol is also called as **reactive protocols.**

On demand routing protocol perform path finding procedure and exchange of routing information only when a path is required by a node to make communication with a destination.

If there is no communication between nodes, they don't maintain routing information or activity hence these protocols are known as **reactive** protocols.

If one node wants to send packet to other node this protocol finds the route in on-demand manner and it creates connection in order to transmit and receive the packet.

In the following subsection we will discuss on-demand routing protocols : AODV and DSR.

4.9.1 Dynamic Source Routing (DSR) Protocol:

Dynamic Source Routing Protocol (DSR) is an on demand routing protocol.

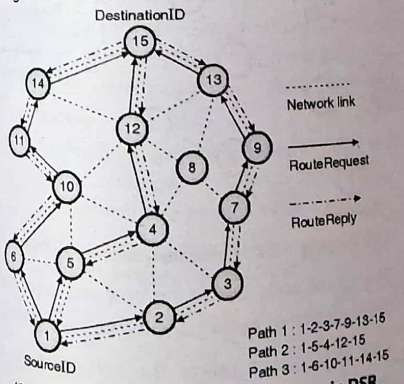
DSR is designed to control the bandwidth consumed by control packets in ad-hoc networks by removing the periodic table update messages needed in the table driven method.

DSR protocol do not need periodic packet transmissions as DSR is beacon-less.

A node uses periodic beacon packet to inform its presence to the neighboring node.

Route establishment in DSR :

Fig. 4.9.1 shows route establishment in DSR.



(G-1692) **Fig. 4.9.1 : Route establishment in DSR**

- During the route establishment phase, DSR establish a route by flooding RouteRequest packets in the network.
- After receiving RouteRequest packet, destination node responds to the source node by sending a RouteReply packet.
- RouteReply packet carries the route traversed by the received RouteRequest packet.
- Let a source node do not have a route to the destination.
- The sender node initiates RouteRequest packet when it has data packet to be sent to destination.
- This RouteRequest packet is flooded into the entire network.
- After receiving RouteRequest packet, each node retransmits the RouteRequest packet to its neighbors ifit has not forwarded already or if the node itself is not destination node provided Time to Live (TTL) counter of packet is not exceeded.
- Each RouteRequest packet carries sequence number generated by the source node and its traversed path.
- Before forwarding RouteRequest packet node checks the sequence number of the packet.
- The packet is forwarded only if it is not duplicate RouteRequest packet.
- The sequence number on the packet is used to avoid loop formations and to prevent more transmission of the similar RouteRequest packets by intermediate node, which receives RouteRequest packet through many routes.
- Hence, all nodes except destination node forward a RouteRequest packet during the route establishment phase.
- A destination node after receiving first RouteRequest packet, responds to the source node through the reverse path traversed by RouteRequest packet as shown in Fig. 4.9.1.
- As shown in Fig. 4.9.1, to obtain a path for destination node 15, source node 1 initiates RouteRequest packet.

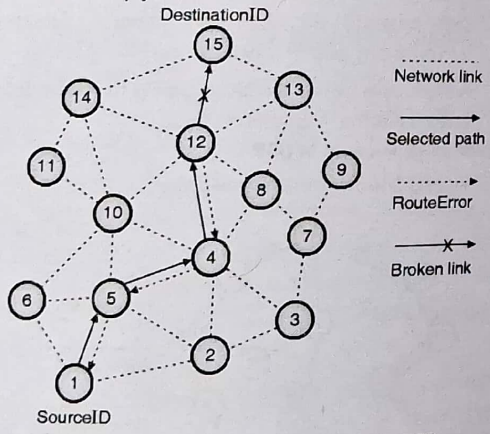
- DSR protocol uses a route cache, which stores all possible data obtained from the source route contained in a data packet.
- Nodes can know about the neighboring nodes routes traversed by data packets if nodes are operated in the promiscuous mode. (The mode in which node can receive the packet which are neither transmits nor addressed to itself).
- This route cache is also useful during the route establishment phase.
- If RouteRequest packet is received by intermediate node, it has route to the destination node in its route cache.
- Thenintermediate node respond to the source node by sending RouteReply with all route data from the source to the destination node.

Optimizations :

- In order to improve the performance of DSR protocol, many optimization techniques have been proposed.
- DSR protocol uses route cache at intermediate nodes. The route cache is settled with routes that can be removed from the information contained in data packets that get forwarded.
- The intermediate nodes use this cache information to reply to the source node when they receive a RouteRequest packet.
- It also uses cache information if they have a route to the corresponding destination.
- An intermediate node discover about breaks in route when it operates in the promiscuous mode.
- Thus obtained information is useful to update the route cache so that the active routes kept in the route cache do not use such broken links.
- The affected node initiates RouteRequest packet at the time of network partition.
- An exponential back off algorithm is used to prevent RouteRequest flooding in the network when the destination node is in another disjoint set.
- DSR allows piggybacking of a data packet on the RouteRequest so that a data packet can be transmitted along with RouteRequest packet.
- In DSR, route construction phase becomes simple without optimization. If intermediate nodes are not redundant, they flood RouteRequest packet.
- As shown in Fig. 4.9.1, after getting the RouteRequest packet from node 1 all its adjacent nodes such as node 2, node 5 and node 6 forward RouteRequest packet. Node 4 gets RouteRequest packet from node 2 and node 5.
- Node 4 sends the first RouteRequest packet it gets from either node 2 or node 5 and rejects the other duplication or redundant RouteRequest packet.
- The RouteRequest is circulated until it reaches the destination node that initiates the RouteReply.
- If intermediate nodes are allowed to begin RouteReply packets, the Source node may receive multiple replies.

Route maintenance in DSR :

- If the node 10 in Fig. 4.9.2 has a path to the destination node through node 14, it sends the RouteReply to the source node.



(G-1693) Fig. 4.9.2 : Route maintenance in DSR

- The source node selects the best and latest route and uses that route for transmitting data packets.
- Each data packet carries the entire path to its destination node.

When intermediate node is out of path, it causes link break (E.g. link between node 12 and 15).

The neighboring node of failed link generates the RouteError message to inform the source node.

The source node restarts the route establishment process.

After receiving RouteError message, the cached entries at intermediate node and source node are removed.

If a wireless link fails due to movement of edge node (Node 1 and node 15), the source node again starts the route discovery process.

Advantages :

1. DSR protocol uses reactive approach that eliminates the need of periodic flood in the network with table update messages required in a table driven method.
2. A path is established only when it is needed and hence they need to find paths to all other nodes in the network.
3. The route cache information is utilized efficiently in order to minimize the control overhead by intermediate nodes.

Disadvantages :

1. Route maintenance method is not able to repair broken link locally.
2. During the route reconstruction phase, stale route cache information may cause inconsistency.
3. The connection set up delay is higher as compared to table driven protocols.
4. Though DSR protocol works well in static and low mobility environments, the performance degrades with increasing mobility.
5. The routing overhead involved in DSR due to source routing method is directly proportional to the path length.

4.9.2 Ad-hoc on Demand Distance Vector Routing Protocol (AODV) :

AODV routing protocol uses on demand method for discovering routes. i.e. route is created only when it is needed by source node for sending data packets.

- AODV protocol makes use of destination sequence number to identify the latest path.

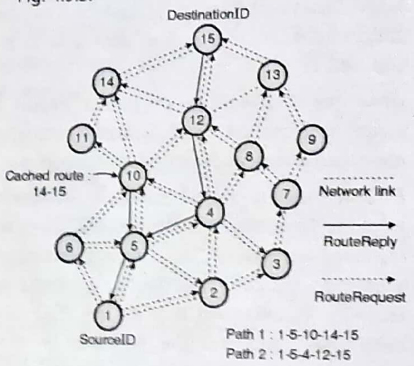
Difference between AODV and Dynamic Source Routing (DSR) :

- DSR uses source routing in which a data packet contains the complete path to be traversed.
- In AODV, the source nodes and intermediate nodes keep the next hop information related to each flow for data packet transmission.
- In reactive protocol, the source node flood the RouteRequest packet in the network when there is no route available for the desired destination.
- The source node can obtain many routes of different destination from a single RouteRequest.
- AODV uses destination sequence number (DestSeqNum) to obtain up to date path to the destination,.
- When the DestSeqNum of current packet is greater than the previous DestSeqNum stored at node, then a node updates its path information.
- A RouteRequest packet contains destination identifier (DestID), source identifier (SICID), source sequence number (SrcSeqNum) and destination sequence number (DestSeqNum), broadcast identifier (BcastID) and time to live (TTL) field. DestSeqNum indicates the freshness of route accepted by source node.
- When an intermediate node gets a RouteRequest packet, it either sends it or makes a RouteReply if it has valid routeto the destination.
- The route validity at the intermediate node is decided by comparing sequence number at intermediate node with the DestSeqNum in the RouteRequest packet.
- If a RouteRequest is received many times, which is indicated by the broadcast identifier and source identifier then the duplicate RouteRequest packets are rejected.
- All intermediate nodes having valid routes to the destination or destination node itself are allowed to send RouteRequest packets to the source node.

- At the time of sending a RouteRequest packet, every intermediate node dothe entry of previous node address and its broadcast identifier (BcastID).
- A timer is used to delete this entry, if RouteReply is not received before the time expires.
- As AODV does not employ source routing of data packet, timer helps in storing an active path at the intermediate node.
- When a node receives a RouteReply packet, data about the previous node from which the packet was received is also stored in order to forward the data packet to the next node.

Route establishment in AODV :

- The route establishment in AODV is as shown in Fig. 4.9.3.

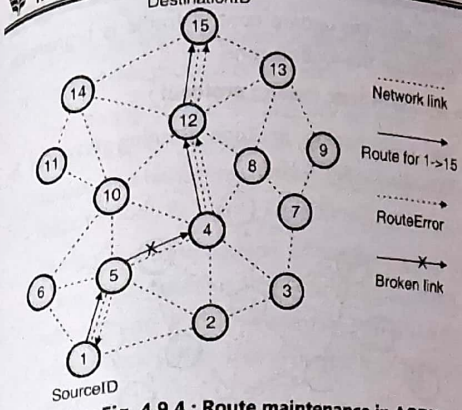


(G-1688) **Fig. 4.9.3 : Route establishment in AODV**

- As shown in Fig. 4.9.3, source node 1 starts a path discovery process by initiating RouteRequest to the destination node 15 in the network.
- Node 1 assumes that RouteRequest consists of the destination sequence number as 3 and source sequence number as 1.
- When nodes 2, 5 and 6 gets RouteRequest packet, these nodes checks their routes to the destination node.
- If the routes to node 15 do not exist, they further send RouteRequest packet to their neighbouring node.
- As shown in Fig. 4.9.3 node 3, node 4 and node 10 are the neighbours of node 2, node 5 and node 6 respectively.
- It is assumed that node 3 and node 10 have already existing routes to destination node 15. The route exists as 10-14-15 and 3-7-9-13-15 respectively.
- If the DestSeqNum at node 10 is 4 and at node 3 is 1, then only node 10 is permitted to reply along with the cached route to the source node 1.
- This is because intermediate node 3 has an older route to destination node 15 as compared to the route available at the source node 1.
- The DestSeqNum at node 3 is 1, whereas source node 1 has DestSeqNum 3.
- At the same time, node 10 has recent route with DestSeqNum 4 to the destination.
- If the RouteRequest reaches the destination node 15 through the path 4-12-15 or any other alternative route the destination node 15 sends a RouteReply to the source node 1.
- In this case multiple RouteReply packets reach the source node.
- All intermediate nodes after receiving RouteReply packet update their route tables with the latest DestSeqNum.
- All intermediate nodes also update the routing information if it leads shortest path between source and destination node.
- AODV does not repair a damaged path locally.

Route maintenance in AODV :

- The end nodes (i.e. source and destination nodes) are notified when a wireless link breaks, that is determined by monitoring periodical beacons or through the link level acknowledgements.
- When the source node learns about the path break. Then it restarts the route to the destination node if required by the higher layers.
- If path break is found at intermediate node, the intermediate node update the source and destination nodes by sending unsolicited RouteReply with the value of hop count as '∞'.
- As shown in Fig. 4.9.4, if the path breaks between the nodes 4 and 5, both the nodes initiate Route Error messages to inform their end nodes about the link break.



(G-1689) **Fig. 4.9.4 : Route maintenance in AODV**

- The end node removes the related entries from their tables.
- The source restarts the path finding process with new broadcast identifier and previous DestSeqNum.

Advantages :

1. In this protocol, paths are established on demand and DestSeqNum are used to discover the latest route to destination.
2. The connection set up delay is less in AODV.
3. It can handle highly dynamic behavior of Ad-hoc networks due to its reactive nature.

Disadvantages :

1. If the source sequence number is old and intermediate nodes have higher but not the latest DestSeqNum then the intermediate nodes can lead to inconsistent routes. This results in stale entries in the table.
2. Several RouteReply packets can result in heavy control overhead.
3. The routing information cannot be reused.
4. The periodic beaconing can results in unnecessary bandwidth consumption.

4.9.3 Comparison of DSDV, DSR and AODV :

- Table 4.9.1 shows the comparison of routing protocols in Ad hoc wireless networks.

Table 4.9.1 : Comparison of DSDV, DSR and AODV

Sr. No	Parameter	DSDV	DSR	AODV
1.	Protocol Scheme	Table driven	On-demand	On-demand
2.	Loop free	Yes	Yes	Yes
3.	Multiple routes	No	Yes	No
4.	Routing over head	Medium	Low	High
5.	Hop count	Medium	Very high	Normal
6.	Security	No	No	No
7.	Periodic broadcast	Yes	No	Possible
8.	Congestion	High	Low	Medium
9.	Multicast capability	No	No	Yes
10.	Routes maintained in	Route table	Route cache	Route table
11.	QoS support	No	No	No
12.	Use of sequence number	Yes	No	Yes
13.	Routing metric	Shortest distance	Shortest path	Shortest path

4.10 Hybrid Routing Protocols :

- In the hybrid routing protocols, each node maintains the network topology information up to m hops. Following are the Hybrid Routing Protocols :
  1. Core extraction distributed Ad-hoc routing (CEDAR).
  2. Zone routing protocol (ZRP).
  3. Zone-based hierarchical link state routing (ZHLS).
- In the following subsection, we will discuss the working of Zone Routing Protocol (ZRP).

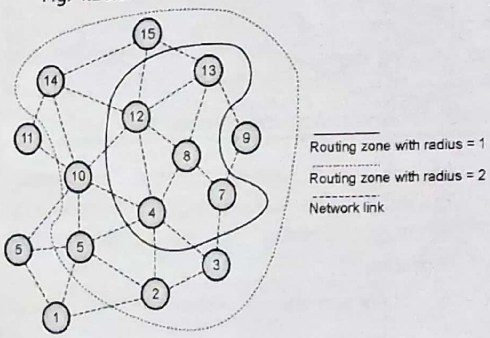
4.10.1 Zone Routing Protocol (ZRP) :

- Zone Routing Protocol (ZRP) is a hybrid routing protocol that effectively combines the best qualities of proactive and reactive routing protocols.

- The concept behind this protocol is to use a proactive routing protocol within a limited zone in the r-hop neighborhood of each node and use a reactive routing protocol for nodes beyond this zone.
- There are two types of protocols :
  1. Intra-zone routing protocol (IARP).
  2. Inter-zone routing protocol (IERP).
- An **intra-zone routing protocol (IARP)** is the protocol used in the region where a particular node uses proactive routing and **Inter-zone routing protocol (IERP)** is the reactive routing protocol used beyond the region.

Routing zone :

- The routing zone of a given node is a subset of the network. In the routing zone, all nodes are reachable within less than or equal to zone radius hops.
- The concept of the routing zone is as shown in Fig. 4.10.1.

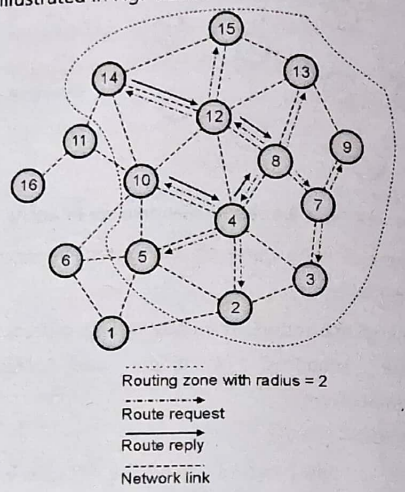


(O-1021) Fig. 4.10.1 : Routing zone for node 8 in ZRP

- Fig. 4.10.1 shows the routing zones of node 8, with radius of 1 hop and 2 hops.
- The nodes 4, 7, 12, and 13 are inner nodes with the zone radius r = 2, while nodes 2, 3, 5, 9, 10 and 15 are peripheral nodes. The nodes with the shortest distance equal to the zone radius.
- Within the routing zone of node, node keeps the information about routes to all nodes by exchanging periodic route update packets. This is responsibility of IARP.
- Therefore the update control traffic is higherwith the larger the routing zone.

Path finding in zone routing protocol :

- The path finding in Zone Routing Protocol is illustrated in Fig. 4.10.2.



(O-1022) Fig. 4.10.2 : Finding a path between node 8 and node 16

- The responsibility of the IERP is to find the paths to the nodes that are not within the routing zone.
- IERP efficiently uses the information available at every node's routing zone.
- In Fig. 4.10.2, if a source node 8 has packets to be sent to the destination node 15, the source node will check whether node 15 is within its zone or not.
- The node 8 will deliver the packet directly if the destination belongs to its own zone.
- Otherwise, node 8 will transmits the RouteRequest packet to its peripheral nodes. To deliver packets directly to the border nodes unicast routing is used.
- As shown in Fig. 4.10.2, node 8 sends the RouteRequests to border nodes 2, 3, 5, 7, 9, 10, 13, 14, and 15.
- If any peripheral node discovers node 15 to be located within its routing zone, then that node sends a RouteReply packet back to the source node 8 indicating the path.

- Otherwise, the node retransmits the RouteRequest packet to the peripheral nodes. Until node 15 is located this process continues.
- Two nodes 10 and 14 find the information about node 16 to be available in their intra-zone routing tables, and hence they initiate Route Reply packets back to node 8.
- During the propagation of RouteRequest, each node that forwards the RouteRequest adds its address into it. For delivery of the Route Reply packet back to the source node,this information is used.
- The path-finding process can result in several Route Reply packets reaching the source; in such case, the source node can select the best path between them.
- The selection of the best path depends on the shortest path, least delay path, etc.
- In an active path if an intermediate node detects a broken link, it will perform a local path reconfiguration in which the broken link is bypassed by a short alternate path connecting the ends of the broken link.
- To inform the source node about the path change a path update message is sent to it.
- This results in a sub-optimal path between two end points. However, it achieves quick reconfiguration in case of link failures.
- After a number of local reconfigurations, the sender reinitiates the global path-finding process in order to get an optimal path.

Advantage :

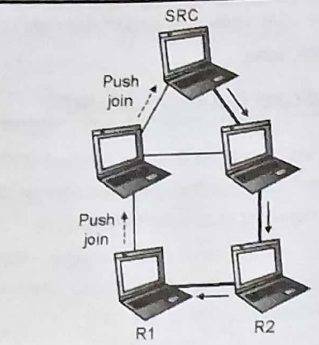
- ZRP reduces the control overhead as compared to on-demand approaches and table-driven approaches.

Disadvantages :

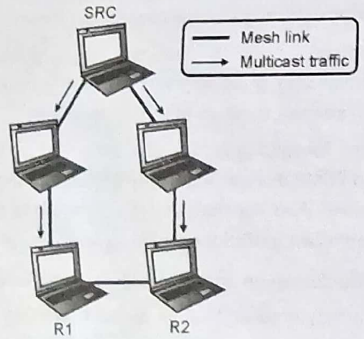
1. ZRP generates higher control overhead than the previous schemes in the absence of a query control. The query control should make sure that redundant or duplicate RouteRequests are not forwarded.
2. The performance of packet depends on the zone radius.

4.11 Multicast Routing : ODMRP :

- On-Demand Multicast Routing Protocol (ODMRP) is a multicast routing protocol designed for ad-hoc networks with mobile hosts.
- ODMRP is a mesh-based, rather than a conventional tree-based, multicast scheme and uses a Forwarding Group concept.
- ODMRP is a flooding-based multicast routing protocol for mobile ad-hoc networks.
- Unlike the pure flooding scheme, data is not flooded throughout the network in ODMRP.
- Instead, data is flooded only throughout forwarding group, which is continually maintained by periodic flooding of control messages.
- The forwarding group, which was first introduced in FGMP (Forwarding Group Multicast Protocol), is a set of ad-hoc nodes specially chosen to forward multicast traffic for a particular multicast group.
- The formation and maintenance of this forwarding group ensures that all these forwarding group nodes provide at least one path from the multicast sender to all receivers.
- To establish and maintain such forwarding group, ODMRP depends on the following operations :
  1. Multicast sender advertisement, and
  2. JOIN-TABLE broadcast by multicast receivers.
- When a multicast sender has data to send, it starts the periodic broadcast of JOIN-REQUEST messages.
- These JOIN-REQUEST messages are flooded throughout the mobile ad-hoc network.
- Each node, upon receiving the JOIN-REQUEST message, will update its unicast routing table with the address of the node from which the JOIN-REQUEST message is received.
- With this routing table, the unicast path back to the multicast sender is known.



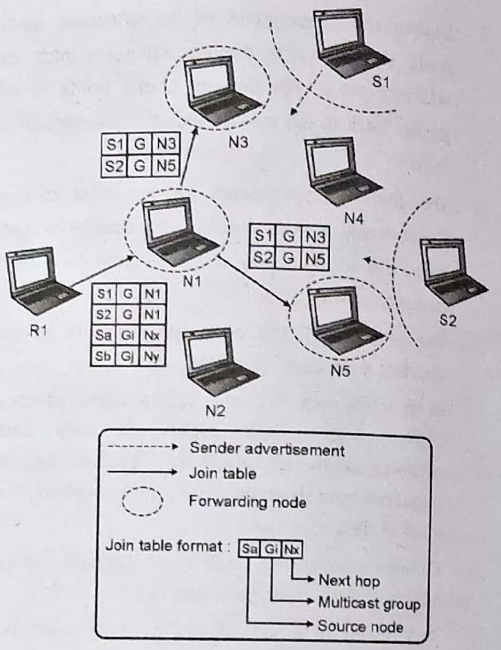
(G-3146)**(a) When R1 detects that the packets from SRC are not received on the shortest path, push join messages are sent**



(G-3147)**(b) The new shortest path is now included in the mesh**

Fig. 4.11.1 : The use of push join messages in CAMP to maintain the multicast mesh

- When a multicast receiver receives the JOIN-REQUEST message, it will update its member table with the address of the multicast sender and periodically broadcast JOIN-TABLE messages.
- The JOIN TABLE message contains the list of all multicast senders known to that receiver and also the next-hop nodes towards those multicast senders.
- These next-hop information are readily available from the unicast routing table.
- Only the node listed as the nest hop in the JOIN-TABLE message will process the JOIN-TABLE message.
- These nodes will become forwarding group nodes and create the new JOIN-TABLE with the next-hop information from its own message cache.
- The newly created JOIN-TABLE will be broadcast further.
- Eventually JOIN TABLE information will be propagated back to all multicast senders and all nodes along the way from each receiver to each sender will be included in the forwarding group, as illustrated by Fig. 4.11.2.



(G-3148)**Fig. 4.11.2 : ODMRP forwarding group information using periodic sender advertisements and join table broadcasts**

- By having all these periodic messages, the forwarding group will be continually refreshed.
- A timeout mechanism is then used to remove stale forwarding group nodes.
- In addition, multiple paths are generally available through forwarding group so that when a link is broken, data packets are still forwarded along these alternate paths.

4.12 Vehicular Ad-hoc Networks (VANETs) :

- VANETs are an envision of the intelligent transportation system (ITS).

There are two types in which vehicles communicate with each other in VANETs are :

1. Intervehicle communication.
2. Vehicle to roadside infrastructure communication.

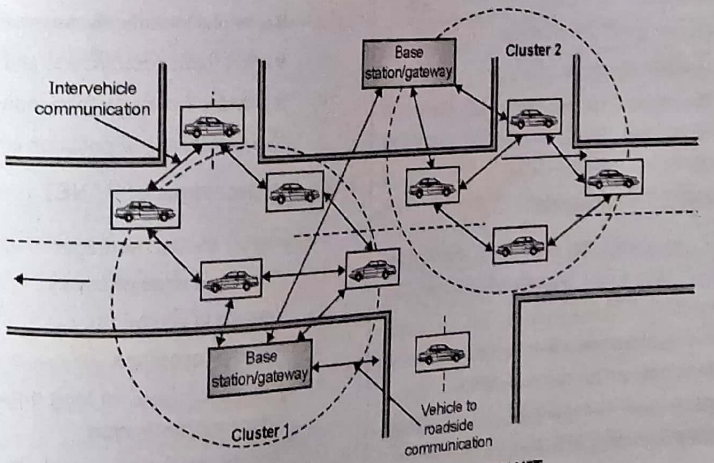
- Vehicular Ad-hoc networks are based on short-range wireless communication between vehicles.
- VANET is made up of the components such as vehicles, restricted vehicle movements, high mobility, and time-varying vehicle density.
- An important advantage of VANETs over MANETs is that most of the vehicles offers sufficient computational and power resources which results in reducing the need for complicated energy-aware algorithms.
- The main goal of VANETs is to offer safe and more efficient roads in future with communicating timely information to drivers and concerned authorities.
- Many manufacturers in automobile industry have already developed system models, which allows vehicles to communicate with their surroundings using wireless media.
- This has given the motivation to the research communities to design and develop protocols and standards for VANETs.

4.12.1 Characteristics of VANETs :

- Following are the characteristics of VANETs :
  1. Geographically constrained topology
  2. Partitioning and large-scale
  3. Self-organization
  4. Unpredictability
  5. Power consumption
  6. Node reliability
  7. Channel capacity
  8. Vehicle density
  9. Vehicle mobility.

4.12.2 Network Architecture of VANET :

- Fig. 4.12.1 shows a typical VANETs architecture.
- To provide safety and other information services to vehicle users, vehicle-to-vehicle and vehicle-to-road-side base station / gateway communication is possible as shown in Fig. 4.12.1.



(O-950) **Fig. 4.12.1 : Network architecture of VANET**

- A cluster is formed by a group of vehicles together and they distribute information between themselves as well as to other clusters and base stations.
- Each vehicle in the VANET is equipped with :
  1. Computing device.
  2. Short-range wireless interface.
  3. Global positioning system (GPS) receiver.
  4. The function of GPS receiver is to provide location, speed, current time and direction of the vehicle.
- Car manufacturers are improving cars with sensors that will help drivers to park and provide GPS compasses as standard equipment on luxury cars.
- Full addition of on-board software and hardware computing facilities with wireless communications and environmental sensors can be achieved within a decade.
- In VANET, in a local database each vehicle stores data about itself and other vehicles.
- The record of information in this database are transmitted at regular intervals.
- In a database, a record consists of :
  1. Vehicle identification.
  2. Position in the form of latitude and longitude.
  3. Current speed of the vehicle.
  4. Direction of vehicle.
  5. Timestamps corresponding to when this record was first generated and when this record was received.

4.12.3 VANET Technologies :

- VANET can incorporate networking technologies like Wi-Fi (IEEE 802.11 b/g), WiMAX (IEEE 802.16), and Bluetooth (IEEE 802.15).
- Wi-Fi is used for vehicle-to-vehicle as well as vehicle-to-base station communication.
- WiMAX is used for the formation of wireless backbone connecting different base stations.
- Bluetooth is used for intravehicle communication as well as communication with nearest neighbors.
- Dedicated short range communications(DSRC) was conceived to :
  1. Provide an architecture for nodes within a vehicular network to communicate with each other.
  2. Provide an architecture for nodes within a vehicular network to communicate with the infrastructure.
- In DSRC, subsequently specialized as WAVE, GPS-enabled vehicles are equipped with on-board entities which can communicate with each other to broadcast the information through vehicle-to-vehicle communications.
- The operating frequency band of DSRC / WAVE in United States is 5.9 GHz and 5.8 GHz for Japan and Europe.
- The bandwidth allocated for vehicle communications is 75 MHz based on line-of-sight with a range of up to 1 km and vehicle speeds of up to 140 kmph.

4.12.4 Advantages of VANET :

- Following are the advantages of VANET :
  1. It provides public safety.
  2. It is helpful in traffic management.
  3. It offers traffic coordination and assistance.
  4. It provides Traveller information support.
  5. Measurement of air pollution emission.

4.12.5 Disadvantages of VANET :

- Following are the disadvantages of VANET :
  1. Flooding in route discovery.
  2. Wastage of bandwidth.
  3. Network congestion.
  4. Poor performance for long distance between source and destination.
  5. External sources required for destination location.

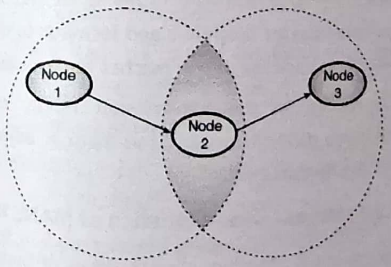
4.12.6 Applications of VANETs :

Following are the important applications of VANETs :

1. Delivery of message and file.
2. Internet connectivity.
3. Communication-based longitudinal control.
4. Cooperative assistance systems.
5. Safety services.
6. Traffic monitoring and management services.
7. Other applications.

4.13 Mobile Ad-hoc Networks (MANETs) :

- A MANET is a collection of mobile nodes which are independent. Mobile nodes in MANET communicate with each other through radio waves.
- The mobile nodes can directly communicate with each other if and only if they are in radio range, whereas other mobile nodes requires an intermediate node to route the packets.
- For communication with each other, each mobile node has wireless interface.
- These networks are distributed and can work at any place without taking the help of any fixed infrastructure.
- Fig. 4.13.1 shows an example of mobile Ad-hoc network that contains 3 mobile nodes.



(G-1882) **Fig. 4.13.1 : MANET**

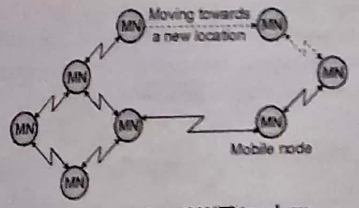
- Node 1 and 3 are not in the range of each other. Node 2 can be used to forward packets between node 1 and node 3. Node 2 acts as a router and three nodes together form an ad-hoc network.

Definition :

- A MANET is a **self configuring** network in which mobile routers are connected by wireless links.
- As a MANET is decentralized topology, network organization and message delivery should be executed by node themselves.

4.13.1 MANET Topology :

- The dynamic collection of mobile nodes creating short-lived networks with the absence of fixed infrastructure are Mobile Ad-hoc Network (MANET).
- Every node in MANET has wireless transmitter and receiver with proper antenna. All nodes acts as routers connected by wireless links.
- In MANET, nodes can move freely and can be organized in an arbitrary manner.
- MANET is an autonomous system of mobile terminals, mobile nodes or mobile station acting as routers which are interconnected by links.
- In MANET network management and communication tasks are performed in distributed manner.
- There may be change in MANET topology time to time as nodes moves or adjust themselves during their transmission and reception parameters.
- MANET node has no base stations i.e. MANET is an infrastructure-less network. Formation and operation of a MANET is as shown in Fig. 4.13.2.

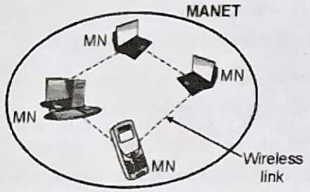


(G-2555) **Fig. 4.13.2 : MANET topology**

- From source node to destination node data packets are transmitted in store-and-forward mechanism in peer-to-peer multihop nodes, which acts as routers.
- With the help of few routers network operate in either extension of an infrastructure network or standalone.

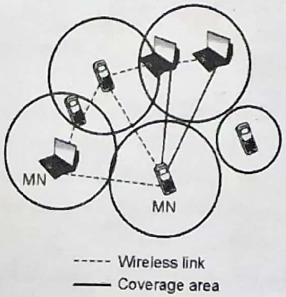
4.13.2 Network Architecture of MANETs :

- MANET topology is decentralized and dynamic and it changes with the capability and possibility of modes moving randomly.
- The use of MANET is common in emergencies like wars and disasters and in daily life such as conference settings and university campus.
- MANET is used because of its collaboration and efficient communication without costly network infrastructure.
- Fig. 4.13.3(a) and (b) shows a single hop and multihop architecture of MANET respectively.



(O-945) **Fig. 4.13.3(a) : MANET single hop architecture**

- MANET is a set of mobile networks (MNs) like mobile phones, laptops, desktop with wireless interface ability. MNs communicate with each other by means of air as a communication media.
- In single hop communication, all hosts are present in a single coverage area as shown in Fig. 4.13.3(a) and the communication between the host to host is direct.
- In multihop communication, the hosts communicate with each other by using intermediate hosts such as in internet communications as shown in Fig. 4.13.3(b).



(O-946) **Fig. 4.13.3(b) : MANET multihop architecture**

- In this type of communication, there are a number of coverage areas intersecting each other.
- Sometimes some hosts can be isolated which cannot be in any one of the coverage area as shown in Fig. 4.13.3(b).
- The hosts set up their own network dynamically. They establish network without relying on the infrastructure support or central administration. They forward the data in multi-hop fashion.
- A network can change constantly as the mobile devices form a MANET without infrastructure.
- The devices can move freely in the network and the devices can leave or join the network at any time.
- When last device leave the network, the network disappears.
- Within a group of two or more people communication can takes place.
- The communication group can be formed for one or many communication sessions. During the communication the group can remain unchanged or can change constantly.
- Each node in MANET acts as a router as well as a host.
- When the node is executing some host/network based application programs, the node acts as a host.
- When the node acts as a router it receives data packets from the neighbors and forwards to other neighbours and the packet reaches to destination.
- Depending on the transmission range, node discovers its neighbors and use them as adjacent routers for data forwarding.

4.13.3 Features / Characteristics of MANETs :

- Some of the important characteristics of MANETs are as follows :

1. **Distributed operation :**
   - As there is no central control on network operations, control of network is distributed among the nodes.

2. **Dynamic topology :**
   - Nodes in MANET move freely in network with different speeds, hence network topology may change randomly. They establish their own network.
3. **Multihoprouting :**
   - When a node sends information to other node in MANET which is not in communication range, the packet should be forwarded via one or more intermediate nodes.
4. **Acts as autonomous terminal :**
   - In MANET, each mobile node is an independent node, which acts as a host and a router.
5. **Light-weight terminal :**
   - In MANET, nodes are mobile with less capability of CPU, less memory size and low storage of power.
6. **Shared physical medium :**
   - In MANET, access to channel cannot be restricted because in MANET the wireless communication medium is accessible to any node.
7. **Spontaneous established networks :**
   - MANET can be established spontaneously.
8. **Adaptive and self-organizing :**
   - MANETs are adaptive and self-organizing.
9. **Communication between various devices :**
   - They can accommodate communication between various devices.
10. **Direct communication :**
    - There can be direct communication between neighbouring devices.
11. **Communication :**
    - MANETs can use single or multi-hop communication.

4.13.4 Advantages of MANET :

- Some of the advantages of MANET are as follows :
  1. MANET is scalable as it accommodates additional nodes in the network.
  2. MANET is less expensive as compared to wired network.
  3. There is no need of central network administration in MANET. Due to this, MANET is self configuring network.
  4. MANET is robust due to decentralized administration.
  5. Set up of MANET can be done at any place and at any time.

4.13.5 Limitations / Disadvantages of MANET :

- The disadvantages of MANET are as follows :
  1. Lack of centralized network administration.
  2. In MANET there is no system of verification of user's identity before allowing data access.
  3. Can't define physical boundary of the network.

4.13.6 Applications of MANET :

1. In military applications to keep updated information network between the soldiers, vehicles etc.
2. In collaborative work, to exchange information on a given project in business.
3. In emergency services, to search and rescue operation, supporting doctors and nurses in hospitals.
4. In commercial sector e.g. in fire, flood or earthquake.
5. Indefence applications.
6. Education through internet.

4.13.7 Comparison of MANET and VANET :

- Table 4.13.1 shows the comparison of MANET and VANET.

Table 4.13.1 : Comparison of MANET and VANET

Sr. No	Parameter	MANET	VANET
1.	Mobility	Low	High
2.	Production cost	Less	More

Sr. No	Parameter	MANET	VANET
3.	Change in network topology	Slow	Fast and frequent
4.	Density of node	Sparse	Dense
5.	Transmission range	Upto 100 m	Upto 500 m
6.	Bandwidth	100 kps	1000 kps
7.	Node life time	Depends on power resource	Depends on lifetime of vehicle
8.	Addressing scheme	Attribute based	Location based
9.	Multi-hop routing	Available	Weakly available
10.	Reliability	Medium	High
11.	Moving pattern of node	Random	Regular

Review Questions

Q. 1 What is mobile IP ?

Q. 2 How does the addressing in stationary host take place ?

Q. 3 Explain home agent and foreign agent in mobile IP.

Q. 4 Explain the communication between mobile host and remote host.

Q. 5 Write short note on routing in Ad hoc network.

Q. 6 State the characteristics of routing protocol for Ad hoc networks.

Q. 7 Give classification of routing protocols.

Q. 8 Write short note on AODV.

Q. 9 Explain the connection establishment and data transfer phase in the following routing protocols with suitable diagram : 1. AODV 2. DSDV

Q. 10 Write short note on DSDV.

Q. 11 Write short note on DSR.

Q. 12 What are the major issues a routing protocol ?

Q. 13 What are the problems faced in hidden terminal ?

Q. 14 What are the problems faced in exposed terminal ?

Q. 15 What are the characteristics of an ideal routing protocol ?

Q. 16 Write a note on classification of routing protocol.

Q. 17 Write a note on DSDV and state it's advantages.

Q. 18 Explain dynamic source routing protocol (DSR). Also mention its advantages and disadvantages.

Q. 19 Give an example of hybrid routing protocol. Explain.

Q. 20 Define and explain MANET.

Q. 21 Explain in brief mobile ad-hoc networks (MANETs).

Q. 22 Explain the network architecture of MANET.

Q. 23 What are the characteristics of MANET ?

Q. 24 What are the design challenges in MANET ?

Q. 25 Explain the advantages, disadvantages and applications of MANET.

Q. 26 Explain in brief types of MANET.

Q. 27 Write a short note on : Vehicular Ad-hoc networks (VANETs).

Q. 28 Explain the characteristics of VANET.

Q. 29 Compare MANET and VANET.

Q. 30 Draw and explain the architecture of VANET.

Q. 31 Enlist and explain the applications of VANETs.

□□□

Unit V

Chapter 5

Mobile Transport Layer

Syllabus

Traditional TCP : Congestion control, Slow start, Fast retransmit / fast recovery, Implications on mobility, Indirect TCP, Snooping TCP, Mobile TCP, Fast retransmit / fast recovery, Transmission / time-out freezing, Selective retransmission, Transaction oriented TCP. **Support for Mobility :** File systems : Consistency, Examples. **World Wide Web :** Hypertext transfer protocol, Hypertext markup language, Some approaches that might help wireless access, System architectures.

Wireless Application Protocol : Architecture, Wireless datagram protocol, Wireless transport layer security, Wireless transaction protocol, Wireless session protocol, Wireless application environment, Wireless markup language, WML script, Wireless telephony application, Examples Stacks with WAP, Mobile databases, Mobile agents.

Chapter Contents

5.1 Mobile Transport Layer :

- If mobility support is to be provided for applications then supporting mobility only on lower layers up to the network layer is not enough.
- Because, in case of the internet, most applications rely on a transport layer, such as TCP or UDP.
- The transport layer performs the following two functions in the internet :
  1. Checksumming over user data and
  2. Multiplexing / demultiplexing of data from/to applications.
- The network layer can only address a host, however, ports in UDP or TCP allow addressing of dedicated applications.
- The connectionless UDP offers only the addressing, and nothing more than that.
- Therefore we will concentrate on TCP.
- UDP is a connectionless protocol which does not give certain guarantees about reliable data delivery.
- However, TCP is a much more complex protocol which needs special mechanisms to be useful in mobile environments.
- Mobility support in IP (such as mobile IP) is already enough for UDP to work.
- The main differences between UDP and TCP are as follows :
  1. TCP offers connections between two applications.
  2. TCP can guarantee in-order delivery or reliable data transmission.
- TCP has been designed with network friendly built-in mechanisms.
- If there is a loss of packets, then TCP assumes that there is network internal congestion and slows down the transmission rate.
- UDP requires that the reality and in-order delivery be handled by the applications.
- The behavior of UDP is not network friendly because it does not reduce the transmission rate in case of congestion.
- This would worsen the congestion.
- The following section we will discuss various mechanisms within TCP that play an important role when using TCP for mobility.

5.2 Traditional TCP :

- There are several mechanisms of the transmission control protocol (TCP) that influence the efficiency of TCP in a mobile environment.
- Some of them are as follows :
  1. Congestion control
  2. Slow start
  3. Fast retransmit / fast recovery

5.2.1 Congestion Control :

- The traditional TCP has been designed for fixed networks with fixed end-systems.
- Hardware like network adapters, fiber optics, copper wires, special hardware for routers etc., is used to enable data transmission.
- This hardware in traditional TCP works without introducing transmission errors.
- If the software is good enough, then it will not drop packets.
- Therefore, we conclude that no packet (on its way from a sender to a receiver) is lost in a fixed network, because of hardware or software errors.
- Thus, the possible reason for a packet loss in a fixed network is a state of congestion which is a temporary overload at some point in the transmission path.
- Congestion can take place even in carefully designed fixed networks.
- The congestion takes place in the following situation : the packet buffers of a router are full of packets and the router is unable to forward the packets fast enough.
- This happens when the sum of the input rates of packets destined for one output link is higher than the capacity of the output link.
- In the situation of congestion the router can do only one thing ; to drop packets.

A dropped packet is lost for the receiver, and it notices a gap in the packet stream.

The receiver does not directly tell the sender which packet is missing, but it continues to acknowledge all in-sequence packets up to the missing one.

Thus no acknowledgement is sent to the sender for the lost packet.

The sender notices this missing acknowledgement for the lost packet and assumes that a packet is lost due to congestion.

The sender understands that retransmitting the missing packet and continuing at full sending rate would now be incorrect, because this might increase the congestion.

Thus traditional TCP mitigates congestion by slowing down the transmission rate dramatically.

All other TCP connections experiencing the same congestion react in the same manner so the congestion is soon resolved.

The use of UDP instead of TCP is not a solution, because the throughput is higher compared to a TCP connection only at the beginning.

As soon as everyone starts using UDP, this advantage is lost, the congestion is standard and data the users experience an unpredictable transmission quality .

5.2.2 Slow Start :

- As discussed earlier, TCP's reacts to a missing acknowledgement drastically in order to get rid of congestion quickly.
- This TCP behavior in response to the detection of congestion is called **slow start**.
- In traditional TCP, the sender always calculates a **congestion window** for a receiver.
- The size of the congestion window in the beginning is one segment (TCP packet).
- The sender sends one packet and waits for acknowledgement.
- On receiving the acknowledgement for this packet, the sender increases the congestion window by one, and sends two packets (congestion window = 2).
- On receiving the two corresponding acknowledgements, the sender again adds 2 to the congestion window (congestion window = 4) and sends four packets.
- In this way this scheme doubles the congestion window every time the acknowledgements come back, which takes one round trip time (RTT).
- Thus the congestion window size increases exponentially in the slow start mechanism.
- However it is dangerous to double the congestion window each time because the steps might become too large.
- The exponential growth of congestion window stops automatically at the **congestion threshold**.
- As soon as the congestion window size reaches the congestion threshold, the sender further increases the transmission rate linearly by adding 1 to the congestion window each time the acknowledgements come back.
- This linear increase continues until a time-out at the sender occurs due to a missing acknowledgement, or until the sender detects a gap in transmitted data because of continuous acknowledgements for the same packet.
- In either case the sender will set value of the congestion threshold to half of the current congestion window.

5.2.3 Fast Retransmit / Fast Recovery :

- As stated earlier, there are two things that lead to a reduction of the congestion threshold.
- One of them is a sender receiving continuous acknowledgements for the same packet.
- This gives the sender information about two things.
- One of them is that the receiver got all packets up to the acknowledged packet in sequence.
- The second thing it shows is that the receiver continuously receives something from the sender.
- And therefore, the gap in the packet stream is not due to severe congestion, but only due to a simple packet loss due to a transmission error.

- In such situation, the sender can retransmit the missing packet(s) before the timer expires. This behavior is called **fast retransmit**.
- The receipt of acknowledgements shows that there is no congestion to justify a slow start and the sender need not change the size of the current congestion window.
- The sender performs a **fast recovery** from the packet loss. This mechanism can improve the efficiency of TCP dramatically.
- One more reason to activate slow start is a time-out due to a missing acknowledgement.
- In such situation the TCP with fast retransmit/fast recovery interprets this congestion in the network and activates the slow start mechanism.

5.2.4 Implications on Mobility :

- Though, slow start is a very useful mechanisms in fixed networks, it decreases the efficiency of TCP drastically if we use it together with mobile receivers or senders.
- This is because the use of slow start takes place under the wrong assumptions.
- If an acknowledgement is missing, then, TCP concludes a congestion situation.
- However, in networks with mobile and wireless end systems, the packet loss may happen due to some other reason than congestion.
- Error rates on wireless links are much higher as compared to fixed fiber or copper links which may lead to the packet loss.
- Packet loss in wireless mobile networks is much more common and cannot always be compensated for by retransmissions (ARQ) or error correction (FEC).
- For example, trying to retransmit on layer 2 could trigger TCP retransmission if it takes too long.
- That creates a problem that Layer 2 transmits the same packet twice over a bad link.
- Can we detect these duplicates on layer 2 ? The answer is no. It is not possible because more and more connections use end-to-end encryption, making it impossible to look at the packet.
- Another reason for packet loss could be the mobility itself. For example in many situations a soft handover from one access point to another is not possible resulting in packet loss.
- When using mobile IP, there could still be some packets in transit to the old foreign agent while the mobile node moves to the new foreign agent.
- In such situation, the old foreign agent may not be able to forward those packets to the new foreign agent resulting in loss of packets.
- The traditional TCP the mechanism of detecting missing acknowledgements via time-outs and concluding packet loss due to congestion cannot distinguish between the above mentioned different causes of packet loss.
- Due to this, the traditional TCP may initiate slow start by assuming congestion when actually the loss of acknowledgement is due to some other reason.
- This is a fundamental design problem in traditional TCP.
- Thus, traditional TCP reacts to missing acknowledgements with slow start, which does not help in the case of transmission errors over wireless links and during handover.
- Due to this behavior there is a severe performance degradation of an unchanged TCP if used together with wireless links or mobile nodes.
- Now is it possible to change TCP completely ? The answer is no.
- We cannot change TCP completely just to support mobile users or wireless links because the installed base of computers using TCP is too large.
- Every enhancement to TCP, has to remain compatible with the standard TCP.
- Therefore, some classical solutions to these problems have been tried.

5.3 Classical TCP Improvements :

- The this section we will discuss some classical solutions in order to make standard TCP more compatible to the wireless mobile networks.

Some of such solutions are as follows :

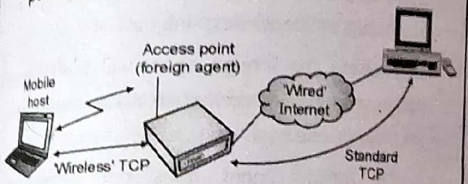
1. Indirect TCP
2. Snooping TCP and
3. Mobile TCP

5.3.1 Indirect TCP (I-TCP) :

The following two competing insights are responsible for the development of indirect TCP (I-TCP).

1. The first one is that TCP performs poorly in the wireless environment and
2. The second one it is not possible to change TCP within the fixed network.

The operation of I-TCP is based on segmenting a TCP connection into a fixed part and a wireless part as shown in Fig. 5.3.1.



(G-3072) **Fig. 5.3.1 : Indirect TCP segments a TCP connection into two parts**

- In Fig. 5.3.1, a mobile host is connected via a wireless link to an access point which is connected to the 'wired' internet where the correspondent node (CN) is also connected.
- The connection between the fixed desktop computer and the access point uses the standard (traditional) TCP.
- No computer in the Internet recognizes any changes to TCP.
- Thus, in this example network, instead of the mobile host, the access point terminates the standard TCP connection, and acts as a proxy.
- Thus the fixed host considers the access point as the mobile host whereas the mobile host will consider it as the fixed host.
- As shown in Fig. 5.3.1, a special TCP, adapted to wireless links, is used between the access point and the mobile host.
- However, changing TCP for the wireless link is not a requirement.
- Even for an unchanged TCP there is a benefit of much shorter round trip time, which can start retransmission much faster.
- The connection between mobile host and correspondent host is segmented at the foreign agent of mobile IP that is the access point as shown in Fig. 5.3.1.
- The foreign agent (access point) can control the mobility of the mobile host and can hand over the connection to the next foreign agent as well when the mobile host moves on.
- Note that, the correspondent host (desktop) in the fixed network absolutely does not notice the wireless link or the segmentation of the connection.
- The foreign agent acts as a proxy and relays all data in both directions.

Operation :

- When the correspondent host sends a packet, the foreign agent i.e. the access point acknowledges this packet and tries to forward the packet to the mobile host.
- On receiving the packet, the mobile host acknowledges the packet.
- However, only the foreign agent uses this acknowledgement.
- The correspondent node will not notice any loss of packet on the wireless link due to a transmission error.
- However, in such a situation, the foreign agent will try to retransmit this packet locally.
- When the mobile host sends a packet, the foreign agent acknowledges it and tries to forward the packet to the correspondent host.
- However, in case of loss of packet on the wireless link, the mobile host notices this loss much faster due to the lower round trip time and directly retransmits the packet.
- If the packet loss occurs in the wired network then, it is now handled by the foreign agent.

Advantages of I-TCP :

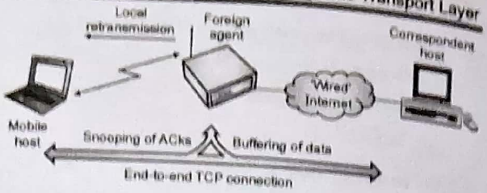
- Following are some of the major advantages of I-TCP :
  1. I-TCP does not require any changes in the standard TCP protocol which is used by the hosts in the fixed network.
  2. Due to the strict partitioning into two connections, the transmission errors such as lost packet on the wireless link, don't propagate into the fixed network.
  3. Due to partitioning into two connections it is possible to use a different transport layer protocol between the foreign agent and the mobile host.

Disadvantages of I-TCP :

- The idea of segmentation in I-TCP has the following disadvantages :
  1. Problems may arise due to the loss of the end-to-end semantics of TCP if the foreign agent partitioning the TCP connection crashes.
  2. There is an increase in the handover latency.
  3. The foreign agent must be a trusted entity because the TCP connections end at this point. If users apply end-to-end encryption, the foreign agent has to be integrated into all security mechanisms.

5.3.2 Snooping TCP :

- One of the drawbacks of I-TCP is that it segments the single TCP connection into two TCP connections which loses the original end-to-end TCP semantic.
- Snooping TCP is a TCP enhancement which works completely transparently and leaves the TCP end-to-end connection intact.
- The main function of the snooping TCP is that, it buffers data close to the mobile host in order to perform fast local retransmission in case of packet loss.
- This enhancement in the standard TCP should be carried out at the foreign agent in the Mobile IP context as shown in Fig. 5.3.2.



(G-3074) Fig.5.3.2 : Snooping TCP as extension of transparent TCP

Principle of operation :

- In the snooping TCP approach, the foreign agent will buffer all packets with destination mobile host and also will 'snoop' the packet flow in both directions to recognize acknowledgements.
- The buffering of packets going toward the mobile node is carried out in order to enable the foreign agent to perform a local retransmission if packet loss occurs on the wireless link.
- That means, the foreign agent will buffer every packet until the FA receives an acknowledgement from the mobile host.
- If the foreign agent does not receive an acknowledgement from the mobile host within a certain amount of time, then it means either the packet or its acknowledgement has been lost.
- If the foreign agent receives a duplicate ACK then it also indicates the loss of a packet.
- When a packet is lost in the wireless link, the foreign agent will retransmit the packet directly from the buffer.
- This retransmission is much faster as compared to that by the correspondent host.
- Therefore, the time out for acknowledgements can be much shorter, because it corresponds only to the delay of one hop plus processing time.
- In order to remain transparent, the foreign agent should not acknowledge data to the correspondent host.
- Because then, the correspondent host would believe that the mobile host had received the data and would violate the end-to-end semantic if the foreign agent fails.

- However, the foreign agent can avoid unnecessary retransmissions of data from the correspondent host by filtering the duplicate acknowledgements.
- Now if the foreign agent crashes, then the time-out of the correspondent host will work and will trigger a retransmission.
- The foreign agent may discard duplicates of packets already retransmitted locally and acknowledged by the mobile host.
- Due to this mechanism, the unnecessary traffic on the wireless link is avoided.

From mobile host to correspondent host :

- Data transfer from the mobile host to the correspondent host takes place as follows :
- In order to detect gaps in the sequence numbers of TCP, the foreign agent snoops into the packet stream.
- As soon as it (FA) detects a missing packet, it sends a negative acknowledgement (NACK) to the mobile host.
- On receiving a NACK the mobile host can retransmit the missing packet immediately.
- TCP does the reordering of packets automatically at the correspondent host.

Advantages of snooping TCP :

- The snooping TCP has the following advantages :
  1. The end-to-end TCP semantic is preserved.
  2. It is not necessary to change the correspondent host.
  3. It does not need a handover of state as soon as the mobile host moves to another foreign agent.
  4. It does not matter if the next foreign agent uses the enhancement or not.

Disadvantages of snooping TCP :

- Some of the important disadvantages of snooping TCP are as follows :
  1. Snooping TCP does not isolate the behaviour of the wireless link from the wired connection as well as I-TCP.
  2. Due to the use of negative acknowledgements between the foreign agent and the mobile host additional mechanisms are added on the mobile host.
  3. The snooping and buffering of data may be useless if certain encryption schemes are applied between the correspondent host and mobile host.

5.3.3 Mobile TCP :

- There are two main problems of wireless links and mobility :
  1. The first one is dropping packets due to a handover or higher bit error rates and
  2. The occurrence of lengthy and/or frequent disconnections.
- Very often the mobile users cannot connect at all.
- Let us understand what happens to standard TCP in the case of disconnection.
- In case of a disconnection, the TCP sender tries to retransmit data, controlled by a retransmission timer that doubles with each unsuccessful retransmission attempt.
- The maximum time is of one minute (the initial value depends on the round trip time).
- That means, in TCP, the sender tries to retransmit an unacknowledged packet every minute and will give up after 12 retransmissions.
- Now what happens if connectivity is back earlier than this ?
- The answer is, data is not successfully transmitted for a period of one minute.
- This is because, the retransmission time-out is still valid and the sender has to wait. The sender also goes into slow-start because it assumes congestion.
- Now what happens in I-TCP if the mobile is disconnected ?
- In I-TCP as the proxy (FA) has to buffer more and more data, more buffer is needed with increase in the period of disconnection.

- If a handover follows the disconnection, which is normally the case, even more state has to be transferred to the new proxy.
- The snooping TCP also suffers from the problem of frequent and long disconnections because then, the mobile will not be able to send ACKs and snooping cannot help.

Mobile TCP :

- The goals of M-TCP (mobile TCP) approach are same as those of I-TCP and snooping TCP.
- Its goals are to prevent the sender window from shrinking if bit errors or disconnection are responsible for loss of packets but congestion is not the cause.
- M-TCP wants to do the following things :
  1. Improve overall throughput,
  2. Reduce the delay,
  3. Maintain end-to-end semantics of TCP, and
  4. Provide a more efficient handover.
- In addition to this, M-TCP is well adapted to the problems arising from frequent and long disconnections.
- Similar to the I-TCP, the M-TCP splits the TCP connection into two parts i.e. the fixed wired connection and the mobile wireless connection.
- The connection between the standard host and **supervisory host (SH)** is the unmodified TCP.
- Whereas, that between the SH and MH (mobile host) is an optimized TCP.
- The supervisory host exchanges data between both parts similar to the proxy in I-TCP.
- In the M-TCP approach a relatively low bit error rate is assumed on the wireless link.
- Therefore, it does not perform the retransmission of data via the SH.
- That means, if a packet is lost on the wireless link, it has to be retransmitted by the original sender.
- This maintains the TCP end-to-end semantics.
- The SH will monitor all packets sent to the MH and ACKs returned from the MH.
- If no ACK is received for some time the SH assumes that the MH is disconnected.
- It stops retransmission from the sender by setting the sender's window size to 0.
- When the window size is set to 0, the sender goes into **persistent mode,** i.e., the mode in which the state of the sender will not change no matter how long the receiver is disconnected.
- This means that the sender will not try to retransmit data.
- As soon as the SH (either the old SH or a new SH) detects mobile host connectivity again, it reopens the sender window to the old value. So that the sender can continue sending at full speed.
- The M-TCP mechanism does not require changes to the sender's TCP.
- The wireless side in M-TCP makes use of an adapted TCP that can recover much faster from packet loss.
- This modified TCP does not use slow start.
- Therefore, M-TCP needs a **bandwidth manager** to implement fair bandwidth sharing over the wireless link.

Advantages of M-TCP :

- Some of the important advantages of M-TCP are as follows :
  1. It maintains the TCP end-to-end semantics. The SH does not send any ACK itself but forwards the ACKs from the MH.
  2. If the MH is disconnected, it simply shrinks the sender window to 0 and avoids useless retransmissions, slow starts or breaking connections.
  3. M-TCP does not buffer data in the SH as I-TCP does. Therefore, it is not necessary to forward buffers to a new SH. Lost packets will be automatically retransmitted to the new SH.

Disadvantages of M-TCP :

- Some of the important disadvantages of M-TCP are as follows :
  1. As the SH does not act as proxy as in I-TCP, the packet loss on the wireless link propagates to the sender. M-TCP assumes low bit error rates, which is not always a valid assumption.
  2. A modified TCP on the wireless link requires modifications to the MH protocol software as well as new network elements like the bandwidth manager.

5.3.4 Fast Retransmit / Fast Recovery :

- As discussed earlier, if a mobile host moves to a new foreign agent, it can cause packet loss or time out at mobile hosts or correspondent hosts.
- In such situations, TCP concludes that congestion has occurred and goes into slow start, even when there is no congestion.
- In section 5.2.3 we have seen the mechanisms of fast recovery / fast retransmit used by a host after receiving duplicate acknowledgements.
- With this mechanism it arrives at a conclusion that a packet loss is not due to the congestion.
- An idea was presented to artificially force the fast retransmit behavior on the mobile host and correspondent host.
- After registering with a new foreign agent using mobile IP, the mobile host starts sending duplicated acknowledgements to correspondent hosts.
- It has been proposed to send three duplicates because, this will force the correspondent host to go into fast retransmit mode and will not start slow start.
- This means that, the correspondent host will continue to send the packets with the same rate as it did before the mobile host moved to another foreign agent.
- This approach will additionally put the mobile host into fast retransmit, because of the possibility of the mobile host going into slow start, after moving to a new foreign agent.
- With this approach, the mobile host will retransmit all unacknowledged packets using the current congestion window size without going into slow start.

Advantages :

- The advantage of this approach are as follows :
  1. Simplicity.
  2. Performance improves with only minor changes in the mobile host's software.
  3. It not necessary to change the foreign agent or correspondent host.
  4. Forcing fast retransmission increases the efficiency.

Disadvantages :

- Some disadvantages of this approach are as follows :
  1. The insufficient isolation of packet losses.
  2. Due to forced fast retransmission, the efficiency increases, but retransmitted packets have to cross the entire network between correspondent host and mobile host.
  3. The correspondent host will start retransmission if the handover from one foreign agent to another takes a longer time.
  4. The approach only focuses on loss due to handover and does not consider the packet loss due to problems on the wireless link.
  5. More cooperation between the mobile IP and TCP layer is required in this approach that makes it harder to change one without influencing the other.

5.3.5 Transmission / Time-out Freezing :

- The approaches presented so far are able to handle short interruptions of the connection, due to handover or transmission errors on the wireless link.
- However some of them are designed to handle longer interruptions of transmission.
- The example of long interruption is the mobile hosts in a car driving into a tunnel, will lose their connection to a satellite.

- In this case, the mobile phone system will interrupt the connection.
- In this situation the TCP, even with all the enhancements discussed earlier, would disconnect after a time out.
- Very often, the MAC layer notices the connection problems, even before the TCP actually interrupts the connection.
- In addition, the MAC layer does not assume congestion, as TCP would and knows the real reason for the interruption.
- Therefore, the MAC layer can inform the TCP layer about an upcoming loss of connection or that the interruption is taking place due to congestion.
- Then TCP can now stop sending the packets and 'freezes' the current state of its congestion window and further timers.
- It is possible to inform both the mobile and correspondent host if the MAC layer notices the upcoming interruption early enough.
- With a fast interruption of the wireless link, it is necessary to have additional mechanisms in the access point to inform the correspondent host about the reason for interruption.
- Otherwise, the correspondent host will go into slow start by assuming congestion and will finally break the connection.
- The moment MAC layer detects connectivity again, it tells TCP to resume operation at the same point where it was forced to stop.
- For TCP as the time has not advanced, no timers expire.

Advantages :

- The advantage of this approach are as follows :
  1. It suggests a way of resuming TCP connections even after longer interruptions.
  2. This approach does not depend on any other TCP mechanism, hence, we can use it together with encrypted data.

Disadvantages :

- However, the disadvantages of this scheme are as follows :
  1. The software on the mobile host as well as the correspondent host needs to be changed.
  2. All mechanisms are dependent on the capability of the MAC layer to detect future interruptions.
  3. Freezing the state of TCP does may not be useful in case of some encryption schemes using time-dependent random numbers.
  4. These schemes need resynchronization after interruption.

5.3.6 Selective Retransmission :

- The use of selective retransmission is a very useful extension of TCP.
- TCP acknowledgements are cumulative. That means these acknowledgements will acknowledge the in-order receipt of packets up to a certain packet.
- Therefore, in the event of losing a single packet, the sender will have to retransmit everything starting from the lost packet (go-back-n retransmission).
- This is not acceptable as it wastes bandwidth for any network.
- Then, TCP can indirectly request a selective retransmission of packets using RFC 2018.
- For this, the receiver can acknowledge single packets, instead of acknowledging only trains of in-sequence packets.
- Therefore, the sender will know precisely which packet is lost and will retransmit it.

Advantages :

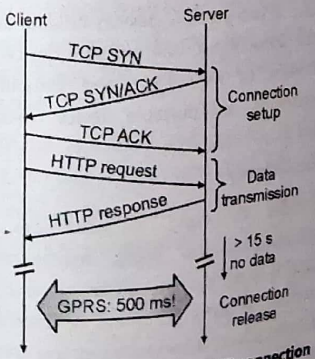
- The advantage of this approach is that the sender needs to retransmit only the lost packets.
- Therefore, the bandwidth requirement is reduced and efficiency increases. Hence this approach is very helpful in slow (low bandwidth) wireless links.

Disadvantage :

- A minor disadvantage of this approach is that more complex software on the receiver side is required.

5.3.7 Transaction-Oriented TCP :

- Assume that there is an application running on the mobile host that sends a short request to a server from time to time, and in response the server sends back a short message.
- Many applications of this type will use UDP and put the responsibility of reliability on a higher, application-oriented layer.
- However an application may use TCP if it requires reliable transport of the packets.
- For using TCP it is necessary to use several packets over the wireless link.
- First is connection establishment for which TCP uses a three-way handshake.
- At least one more packet is required for transmission of the request, and three more packets are required to close the connection via a three-way handshake.
- However, this overhead is minimal if connections have a lot of traffic or have a long duration.
- But if only one data packet is to be transmitted then TCP may need seven packets altogether.
- Refer Fig. 5.3.3 which shows an example for the overhead introduced by using TCP over GPRS on a web.



(G-3121) **Fig. 5.3.3 : Example TCP connection setup overhead**

- Web services are based on HTTP which requires a reliable transport system.
- Therefore, in the internet, TCP is used for this purpose.
- It is necessary to establish a TCP connection before a HTTP request can be transmitted.
- The connection establishment requires three messages.
- One-way delay of 500 ms is expected if GPRS is used as wide area transport system.
- Thus it requires far more than a second for the setup of a TCP connection which is a very long delay.
- This resulted in the development of a transaction-oriented TCP (T/TCP).
- In T/TCP it is possible to combine packets for connection establishment and connection release with user data packets.
- This will reduce the number of overhead packets down to two from seven and reduce the delay experienced in conventional TCP.

Advantage :

- The advantage of T/TCP is the reduction in the overhead to establish and release a connection, as compared to the standard TCP.

Disadvantages :

- The disadvantages of T/TCP are as follows :
  1. Since, T/TCP is not the original TCP anymore, some changes are required in the mobile host and all correspondent hosts.
  2. This approach does not hide mobility.
  3. T/TCP exhibits several security problems.

5.4 Support for Mobility :

- Data transfer from a sender to a single or many receivers is not enough.
- Only applications can make a communication network useful.
- In a mobile and wireless communication system, some additional components are required to use well-known applications from fixed networks.

- Examples of additional components are file systems, databases, security, accounting and billing mechanisms.
- Power consumption is an important issue in mobile and wireless communication system because mobile devices have limited energy resources.
- The maintenance of consistency is the main problem for distributed, loosely coupled file systems.

5.5 File Systems :

- This section explain some problems regarding file systems and presents some research projects.

Goal :

- The main goal of a file system is to support efficient, transparent, and consistent access to files. It does not matter where the client requesting files or the servers offering files are located.
- Efficiency of wireless systems is very important. The protocol overhead and updating operations should be kept at a minimum because of lower bandwidth.
- The problems of location-dependent views on a file system are addressed by the transparency.
- The file system must provide identical views on directories, file names, access rights etc. in order to support mobility independent of the current location.
- The consistency is the main problem in a file system.

Problems :

- General problems related to file systems are as follows :
  1. Limited resources on portable devices.
  2. Low bandwidth of the wireless access.
  3. Large caches in the end-system and performing updates through the wireless link are not reliable.
  4. Portable devices can be disconnected for a longer period.
  5. Hardware and software components of portable devices do not follow standard architectures of computer or operating systems.
  6. Mobile phones, PDAs, and other devices consists of their own OS, hardware, and application software.
  7. As compared to desktop systems or traditional file servers, portable devices are less reliable.
  8. In a mobile and wireless environment, the standard file systems (E.g. Network File System (NFS)) are inefficient and unusable.
  9. The consistency of the copy with the original data.
- Traditional file systems were not expecting disconnection, low bandwidth connections, and high latencies.
- The portable device can replicate files or single objects in order to support disconnected operation.
- This can be done in advance by pre-fetching or while data fetching (caching).

5.5.1 Consistency :

- The consistency of replicated objects (e.g. files, parts of files, parts of a data structure etc.) is the basic problem for distributed file systems that allow data replication for the performance purpose.
- In case of two portable devices holding the copies of the same object, one portable device changes the value of the object and without further mechanisms, one portable device reads an old value.

Strong consistency :

- In order to avoid inconsistencies, several traditional systems apply mechanisms that maintains a permanent consistent view of a file system for all users. Automatic updates similar to database systems achieves this strong consistency.
- After any change, a writer of an object locks the object, changes the object, and unlocks the object.

No other device can write the object if an object is locked. After a change, cached objects are invalidated.

Maintaining strong consistency is very expensive in terms of exchanging updates through the wireless link and it is sometimes impossible as well.

Consider a temporarily disconnected device with several objects in its cache. It is not possible to update the objects or invalidate them.

Other users cannot see the locking the cached objects.

One solution for this is to ban access to disconnected objects which would prohibit any real application based on the file system.

Weak consistency :

- A weak consistency model for file systems is used in the mobile systems.
- Weak consistency involves certain periods of inconsistency which have to be tolerated for the reason of performance.
- That means occasional inconsistencies have to be tolerated in a weak consistency.
- To ensure that the overall file system is consistent, conflict resolution strategies must be applied for reintegration.

Reintegration process :

- In this process, the objects from different users are merged that results in one consistent file system.
- A user can hold a copy of an object, disconnect from the network, change the object, and reconnect again. The changed object should be reintegrated.
- If an object has been changed by two users working with two copies, then a conflict can occur.
- During reintegration process, the file system can notice that both copies differ.
- The conflict resolution strategy decides which copy can be used or how to proceed.
- The file system can notice conflicts based on time stamps, version numbering, hash values, content comparison etc.

Example of reintegration :

- If a number of people are writing an article, with each person working on one section by using his or her own laptop.
- Reintegration is simple as long as each person stays within his or her section.
- Reintegration becomes difficult as soon as one person makes a copy of another section and starts making changes in it. In this case reintegration becomes content-dependent.

Different solutions for file systems :

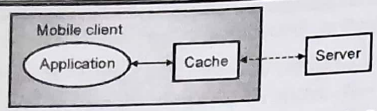
- Following examples show different solutions for file systems :
  1. Coda
  2. Little Work
  3. Ficus
  4. MIo-NFS
  5. Rover
- These solutions for file systems vary in :
  1. Granularity of caching and pre-fetching of files, directories, sub-trees, disk partitions.
  2. Location of mobility support (fixed network and/or mobile computer).
  3. Their conflict resolution strategies.
- In the following subsections we will discuss different solutions for file systems.

5.5.2 Coda :

- The Andrew File System (AFS) is the ancestor of many distributed file systems that can be used for mobile operation.
- **Coda** is the successor of AFS. It offers two different types of replications namely server replication and caching on clients.
- In Coda, disconnected clients work only on the cache, that means the applications use only cached replicated files.

Application, Cache and Server in Coda :

- The cache between an application and the server in Coda is as shown in Fig. 5.5.1.

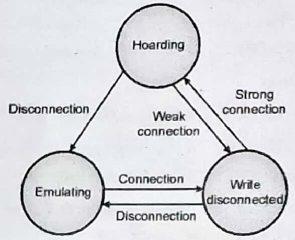


(G-3122) **Fig. 5.5.1 : Application, Cache and Server in Coda**

- Coda is a transparent extension of the cache manager of client.
- This architecture is valid for most of today's mobile systems that makes use of a cache.
- In order to provide all the required files for disconnected work, Coda offers extensive methods for pre-fetching of files while still connected called as hoarding.

States of a client in Coda :

- Fig. 5.5.2 illustrates the three states of a client in Coda.



(G-3123) **Fig. 5.5.2 : States of a client in Coda**

- Coda consist of following three states :
  1. Hoarding
  2. Emulating
  3. Write disconnected

1. Hoarding :

- If the client is connected to the server with a strong connection as shown in Fig. 5.5.2, hoarding transparently pre-fetches currently used files.
- This automatic data collection required for it is impossible for a standard user to know all currently used files.
- A user can be familiar with standard programs and application data but he or she does not know anything about the many additionally required small system files such as profiles, shared libraries, drivers, fonts etc.
- Coda must explicitly pre-fetch a list of files pre-determined by a user.
- A user may assign priorities to certain programs.
- Now, Coda makes a decision on the current cache content using the list and a least-recently-used (LRU) strategy.

2. Emulating :

- Emulating is the state where the applications work on the replicates as soon as the client is disconnected. This is shown in Fig. 5.5.2.
- In this state, Coda follows a positive approach and allows read and write access to all files.
- The system maintains a record of changed files, but it does not keep a history of changes for each file.
- The cache always has only one possibly changed replicate.
- Coda compares the replicates with the files on the server after reconnection.
- If Coda observes that two different users have changed a file then reintegration of this file fails.
- In this case, Coda saves the changed file as a copy on the server to allow manual reintegration.
- A very coarse grained optimistic approach of Coda is working on whole files.
- The success rate of Coda relies on the fact that files in UNIX are rarely written by more than one user.
- Only some files are changed, most files are just read.

3. Write disconnected :

- As shown in Fig. 5.5.2, if the client is connected to the server with a weak connection, Coda makes a decision whether it is useful to fetch a file through this connection or let the user wait until a better connection is available.
- That means, Coda models the user's patience and weighs it against the cost of fetching the file required by the user.
- When a strong connection between client and the server exists, the client performs hoarding.

- The client goes into emulating state and uses only the cached replicates if the connection breaks completely.
- If the client loses the strong connection with server and only a weak connection remains, the client does not perform hoarding operation.
- But it decides whether to fetch the file in case of a missing of cache considering user patience and file type.
- The weak connection is not used for reintegration of files.

Problems of Coda :

- Following are two problems of Coda :
  1. Application Specific Resolver (ASR) tool
  2. Definition of a conflict

Application Specific Resolver (ASR) tool :

- The tool Application Specific Resolver (ASR) was developed which automates conflict resolution after failure of reintegration.
- A problem with ASR tool is that they can only work after the fact. i.e. the tools have to reconstruct a history of changes based on the replicate as Coda does not record every single change in file.

Definition of a conflict :

- The definition of a conflict is another problem of Coda. Coda detects only write conflicts that means if two or more users change a file.
- Consider that there are two files f1 and f2.
- Both client uses values from files f1 and f2 to calculate something. One client stores the result in file f1 and other client stores the result in file f2.
- Coda will not detect any problem during the reintegration of the files.
- But the result is not correct based on the files. The execution order plays an important role in Coda.
- To resolve this problem, a simple optional transaction mechanism was introduced called as isolation-only transactions (IOT).
- The function of IOT is to allow grouping of certain operations and checks them for serial execution.

5.5.3 Little Work :

- Similar to Coda, the Little Work distributed file system is an extension of AFS (Andrew File System).
- Little Work distributed file system only needs changes to the client's cache manager.
- During reintegration, Little Work file system detects write conflicts.
- There are no specific tools in Little Work for reintegration and it offers no transaction service.

States of a client in Little Work :

- Little Work makes use of uses more client states to maintain consistency. It consists of the following states :
  1. Connected
  2. Partially connected
  3. Fetch only
  4. Disconnected

Connected :

- In this state, the operation of the client is normal. No special mechanisms from Little Work are needed.
- This connection mode requires a continuous high bandwidth as available in typical office environments using a WLAN.

Partially connected :

- In partially connected state, a client has only a lower bandwidth connection, still has the possibility to communicate continuously.
- The packet radio networks are examples of this type of network.
- These networks are charge based on the amount of traffic and not based on the duration of a connection.
- Partially connected client state allows use of cache consistency protocols similar to the normal client state, but with a delayed write to the server to lower the communication cost if the client changes the file again.
- This avoids the problems of consistency, even though no high bandwidth connection is available.

Fetch only :

- The client goes into the fetch only state available network available offers connections on demand.
- The examples of such type of networks are cellular networks such as GSM with costs per call.
- In this state, the client makes use of the replicates in the cache in an optimistic way. The client fetches files through the communication link if they are not available in the cache.
- This allows a user to access all files of the server. This state tries to minimize communication by working on replicates and reintegrate after reconnection by using a continuously high bandwidth link.

Disconnected :

- The client is disconnected without any network.
- Little Work terminates the connection if a cache does not occurs, otherwise replicates are used.

5.5.4 Ficus :

- Ficus distributed file system is not based on a client/server approach.
- Ficus systems allows the use of replicates, detects write conflicts, and solves conflicts on directories.
- Ficus makes use of so-called gossip protocols.
- In this system, a mobile computer do not need a direct connection to a server.
- It can spread updates through the network with the help of other mobile computers. It can spread updates until it reaches a fixed network and the server.
- Hence, step-by-step changes on files spread through the network.
- Ficus system tries to reduce the exchange of files that are valid only for a short time period e.g. temporary files.
- A serious issues in the gossip protocols are :
  1. How fast they propagate to the client that requires this information
  2. How much unnecessary traffic it causes to propagate information to clients that are not interested.

5.5.5 MIo-NFS :

- MIo-NFS stands for Mobile Integration of NFS. It is an extension of the Network File System (NFS).
- MIo-NFS system in contrast to many other systems, makes use of a pessimistic approach with tokens controlling access to files.
- In this system, only the token-holder for a specific file can change this file, thus MIo-NFS avoids write conflicts.
- MIo-NFS supports following three modes :
  1. Connected
  2. Loosely connected
  3. Disconnected

Connected :

- In this mode of MIo-NFS, the server handles all access to files.

Loosely connected :

- In this mode, clients makes use of local replicates, exchange tokens over the network, and update files through the network.

Disconnected :

- In this mode, the client makes use of only local replicates. If the client is token-holder then only writing allowed.

5.5.6 Rover :

- The Rover platform as compared to Coda makes use of another approach to support mobility.
- Rover platform develops new, mobility aware applications instead of adapting existing applications for mobile devices.
- Rover platform introduces two new components namely :
  1. Relocatable dynamic objects
  2. Queued remote procedure calls

Relocatable dynamic objects :

- Relocatable dynamic objects are the objects which can be dynamically loaded into a computer of client from a server (or vice-versa).
- Loading of relocatable dynamic objects to reduces the client-server communication.

- A trade-off among transferring objects and transferring only data for objects has to be found.
- It makes sense to migrate the object if a client requires an object quite frequently.
- Object migration for a single access generates too much overhead.

Queued remote procedure calls :

- Queued remote procedure calls allow for non-blocking RPCs even when a host is disconnected, queued remote procedure calls allow for non-blocking RPCs.
- The exchange between requests and responses are takes place as soon as a connection is available again.
- Conflict resolution is application specific and it is done in the server.
- None of the above discussed file system is in use every day, hence the focus of research has shifted towards the www.

5.6 World Wide Web (WWW) :

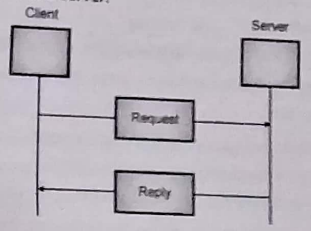
- In this section we will discuss some problems encountered by the web applications when used in a mobile and wireless environment.

5.6.1 HTTP (Hypertext Transfer Protocol) :

- The main function of HTTP is to access data on WWW.
- This protocol can access the data in various forms such as plaintext, hypertext, audio, video etc.
- The function of HTTP is equivalent to a combination of FTP and SMTP. It uses services of TCP. It uses only one TCP connection (port 80).
- There is no separate control connection like the one in FTP.
- Only the data transfer takes place between the client and server so there is only one connection and it is the data connection.
- The data transfer in HTTP is similar to SMTP. The format of the messages is controlled by MIME like headers.

HTTP Operation :

- The principle of HTTP is simple. A client sends a request. The server sends a response.
- The request and response messages carry data in the form of a letter with a MIME like format.
- Fig. 5.6.1 shows the HTTP transactions between client and server.



(G-657) **Fig. 5.6.1 : HTTP transaction**

- The client initializes the transaction by sending a request message and the server responds by sending a response.

The Web and HTTP :

- HTTP is the Web's application layer protocol. It is the heart of the Web. It has been defined in [RFC 1945] and [RFC 2616].
- HTTP is implemented in two programs :
  1. A client's program 2. A server's program.
- These programs are executed on different and systems and talk to each other by exchanging HTTP messages.
- HTTP defines how Web clients such as browsers request Web pages from Web servers and how servers transfer Web pages to clients.
- HTTP uses TCP as its under lying transport protocol (rather than using UDP). The HTTP client first initiates a TCP connection with the server.
- After establishing a connection, the browser and the server processes access TCP through their socket interface. TCP provides a reliable data transfer service to HTTP.
- That means each HTTP request message, transmitted by a client will eventually arrive intact at the server.

- Similarly each HTTP response message transmitted by the server will eventually arrive intact at the client, due to the reliable TCP connection.
- Due to this kind of layered architecture HTTP need not have to worry about the lost data or about the details of how TCP deals with the loss and retransmission of data. It is managed by TCP.

Statelessness :

- In HTTP, the server sends the files requested to the client without storing any state information about the client.
- So it may happen that the same client may ask the same information repeatedly to the server and the server would not even understand it.
- So it will keep resending those files. As the HTTP servers does not maintain any information about the state of client it is called as a stateless protocol.

Problems in HTTP :

- Some of the problems faced by HTTP in the wireless environments are as follows :
  1. Bandwidth and delay
  2. Caching
  3. POSTing

1. Bandwidth and delay :

- Note that HTTP was not designed to operate over a low bandwidth/high delay connections.
- It was designed to work in an environment with workstations running TCP/IP over wired networks with a large bandwidth of few Mbps.
- One problem with HTTP protocol is its large and redundant headers. Also many information fields in these headers are transferred repeatedly with each request because HTTP is stateless.
- These headers are readable for humans and transferred in plain ASCII format.
- In HTTP the content transferred by the server are uncompressed. If applications do not perform compression, then, the server does not perform it.
- HTTP uses TCP connections for each item on a web page such as icons, images etc. Therefore, there is a huge overhead coming with each item.
- A TCP connection is established with a three-way-handshake, data transmission, and reliable disconnection which may require seven PDUs exchanged between client and server.
- The problem is that, TCP has not been designed for this transaction like request/response scheme with only to exchange a small amount of data .
- Additional problems are caused due to the slow-start mechanism built into TCP.
- Due to this mechanism, TCP can be too cautious in the beginning of a transmission, and the transmission may be over before it can utilize the available bandwidth.
- That means, TCP never leaves the slow-start which causes unnecessary high delay.
- DNS look-up, necessary for many items on a web page, creates another problem to reduce bandwidth and increase the delay even further.

2. Caching :

- Caching is useful in many cases. Still content providers often disable caching.
- Many companies and content providers want the advertisements to be placed on web pages and need feedback.
- Companies cannot get realistic feedback if there is a cache between a server and a client,.
- In order to overcome this problem either caches need additional mechanisms or caching is disabled from using the **no-cache** keyword in the HTTP/1.0 header.
- Version 1.1 of HTTP provides more detailed caching mechanisms.
- Network providers need someone to follow this no-caching requirement from their customers.

- Due to no-caching, users will have to download the same content repeatedly from the server.
- Many present-day pages contain dynamic objects such as access counters, time, date, or other customized items that cannot be cached.
- The dynamic content changes over time or for each access.
- But, sometimes at least a part of a page is static and can be cached.
- Many companies generate customized pages on demand via CGI, ASP etc.
- Customization is saved in cookies which will prevent any caching because the names of links are also generated dynamically.
- Depending on the type of browser, client hardware, client location etc., the homepages of companies are created dynamically.
- A cache can store some static content, but it is generally not possible to merge the static content with the dynamic remainder of a page.
- Caching is quite often inhibited by the mobility because the ways in which web servers are accessed change over time due to changing access points.
- However if performed at entry points of mobile networks, caching may save some bandwidth and time.
- Caching is inhibited by many security mechanisms as well. Authentication is often between a client and a server, not between a client and its cache.
- Also note that the keys used for authentication have an associated time-out after which they are not valid.
- Therefore, caching the content for this type of secured transactions is useless.

3. POSTing :

- Additional problems are created while sending content from a client to a server if the client is currently disconnected.
- It is not possible to fulfill the **POST** request in a disconnected state.
- Therefore, a server needs to be simulated by accepting the posting via an additional process.
- This will cause additional problems.

5.6.2 Hypertext Markup Language (HTML) :

- **HTML** is used to describe the content of web pages in the world wide web.
- Irrespective of their version, they all share common properties.
- HTML has been designed to work with standard desktop computers connected to the internet with a fixed wire.
- These computers have some properties in common such as a higher performance (as compared to handheld devices), a color high-resolution display, mouse, sound system, and large hard disks.
- In comparison to these properties, what do standard handheld devices offer ?
- There is a restriction on the power consumption and form factor, these devices have small displays, a low resolution, low performance CPUs and limited user interfaces such as touch screens, soft keyboards, voice commands etc.
- The data rates offered in network connection of desktop computers are often high typically (few Mbps) and the round-trip delays are in the range of some ms.
- On the other hand, today's wireless connections have the data rates of few hundred kbps and round-trip delays in the range of some seconds.
- The web pages designed with the help of current HTML ignore these differences in end systems because these pages are designed for a good content and not for efficient transfer of content.
- That means HTML itself does not offer any way of optimizing pages for different clients or different transmission technologies.

- However, HTML is not the biggest problem when a user is accessing web pages from wireless handheld device.
- The web pages are often 'enriched' with special 'features' such as animated GIFs, ActiveX controls, Frames, Java Applets, multi-media content.
- Some of these applications use HTML, some not.
- The client's browser can interpret some of the applications directly, whereas some need a special **plug-in**.
- These additional content formats cause the following problems.
- The appropriate plug-ins are generally available only for the most common computer platforms.
- But they are not available for handheld devices, with each one having its own operating system.
- Even if a plug-in is available, the browser still has the problem of displaying a true-color video on a black and white display, or displaying a GIF with many 'clickable' areas etc.
- Many web pages use these GIFs for navigation purpose. The user just needs to click in the right area.
- But sometimes these GIFs cannot be displayed at all on the mobile handsets.
- In such condition, the approaches using content distillation or semantic compression might work with HTML.
- But they need many additional plug-ins and each need their own mechanism to translate them into a useful format for a wireless device.
- If such additional mechanisms are not used, then, large high-resolution pictures would be transferred to a low-resolution display mobile phone causing high costs.

5.6.3 Some Approaches That Might Help Wireless Access :

- Due to the problems with HTTP and HTML many different proprietary and standardized solutions or better partial solutions have been suggested.
- Some of these approaches are as follows :
  1. Image scaling
  2. Content transformation
  3. Content extraction / semantic compression
  4. Special languages and protocols
  5. Push technologies

1. Image scaling :

- If a page contains a true color, high-resolution picture, then it is possible to scale down this picture to fewer colors, lower resolution, or to just the title of the picture.
- Then the user can decide to download the picture separately.
- It is possible to offer clipping, zooming, or detail studies to users if they are interested in a part of the picture.

2. Content transformation :

- Many documents are only available in certain formats such as the PDF format.
- Then before transmitting such documents to a client without the appropriate PDF reader, it is necessary to use a special converter such as Fox, to translate this document into plain text.

3. Content extraction/semantic compression :

- In addition to transformation of the content, it is possible to extract headlines or keywords from a document and present them to a user.
- Then it is up to the user to download more information related to a certain headline or keyword.
- It is also possible to automatically generate an abstract from some given text.
- This semantic compression for an arbitrary text is very difficult but extracting headlines is comparatively simpler.

4. Special languages and protocols :

- Some other approaches try to replace HTML and HTTP with other languages and protocols that are better adapted to a wireless environment.

- Some examples of such protocols are : the handheld device transport protocol (HDTP) and the handheld device markup language (HDML).
- The ideas from these solutions have been integrated into a broader approach that is wireless application protocol (WAP) which is discussed later in this chapter.

5. Push technologies :

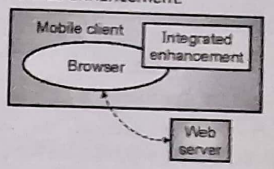
- In the push technologies, instead of client pulling content from a server, the server will push content to a client.
- The advantage of push technology is that it avoids the overhead of setting up connections for each item.
- However, this is only useful for some content such as news, weather information, road conditions etc. where users interaction is minimal.

5.6.4 System Architecture :

- WWW has the client / server based system architecture.
- Here the client which is a web browser running as an application on a computer, will request content from a server which is the web server running on another computer.
- With no enhancements, each click on a hyperlink will initiate transfer of the content that the link is pointing to.
- HTTP protocol is used by the browser for content transfer.
- HTML and many more (proprietary) formats are used to describe web pages.
- In the web client/server scenario, caching is a major topic.
- Caching is useful for wired computers because it reduces the delay of displaying previously accessed pages.
- However, in wireless environment, it is the only way of supporting (partially) disconnected web browsers.
- The wireless client to network connection especially on mobile can be often disrupted or have a bad quality.

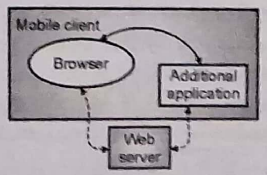
First enhancement :

- The first enhancement in the system architecture is to integrate caching into web browsers.
- It has now become a standard for all the modern browsers such as Microsoft.
- Refer Fig. 5.6.2 which shows a mobile client with a web browser that has an integrated caching mechanism as enhancement.



(G-3124) **Fig. 5.6.2 : Integrated browser enhancement**

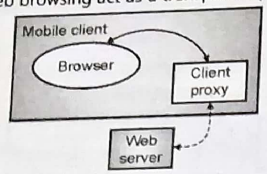
- This cache stores already transferred content up to a certain limit but does not perform automatic pre-fetching of pages.
- It is possible for the user to browse through the cached content (pages, pictures, multi-media objects etc.) even by going "offline".
- The caching strategies are generally very simple.
- The user can determine whether a check for updating is performed every time he or she accesses a page, only after he restarts the browser, or never (a page has to be refreshed manually).
- The architecture for an early approach to enhance web access for mobile clients has been shown in Fig. 5.6.3.



(G-3125) **Fig. 5.6.3 : Additional application supporting browsing**

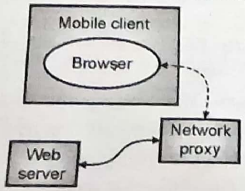
- For example the initial WebWhacker is a companion application for the browser which supports functions like pre-fetching of content, caching and disconnected service.

- However, problem with this approach is that it is not transparent for a browser because, now there are two different ways to access content.
- One way is directly to the web server and the other is one via the additional application.
- As shown in Fig. 5.6.4 the typical enhancements for web browsing act as a transparent proxy.



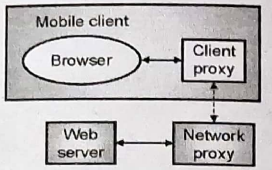
(G-3126) **Fig. 5.6.4 : Client proxy as browser support**

- Here, the browser accesses the web server through the client proxy as shown in Fig. 5.6.4.
- That means, the client proxy acts as server for the browser whereas it acts as client for the web server.
- According to many strategies, the client proxy can pre-fetch and cache content. The proxy serves the content, as soon as the client is disconnected.
- This scheme is followed by many approaches which is independent of the browser and therefore allows other developments.
- Some of the example strategies for pre-fetching are as given below :
  1. All pages pointed by the current pages,
  2. All pages including those pointed by the pre-fetched pages,
  3. Only pages but no pictures,
  4. All pages with the same keyword on the same server etc.
- As shown in Fig. 5.6.5, a network proxy can also support a mobile client on the network side.



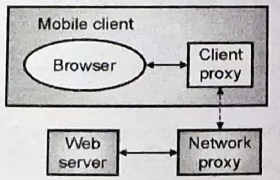
(G-3127) **Fig. 5.6.5 : Network proxy as browser support**

- This network proxy is capable of performing the adaptive content transformation such as semantic compression, headline extraction etc. or pre-fetch and cache content.
- Pre-fetching and caching can be very useful in a wireless environment which has a higher error probability.
- If the web access is split into a mobile and fixed part then it can improve overall system performance as shown in Fig. 5.6.6.



(G-3128) **Fig. 5.6.6 : Client and network proxy as browser support**

- This is similar to the enhancement achieved by I-TCP by splitting web access into a mobile and fixed part.
- The network proxy acts as any fixed browser for the web server with wired access.
- Therefore, even if the mobile client gets disconnected it does not influence the web server.
- Fig. 5.6.7 shows a system architecture which combines the benefits of client and network proxies.



(G-3129) **Fig. 5.6.7 : Client and network proxy with special transmission protocol**

- Client proxy and network proxy can now interact between them in a better manner to pre-fetch and cache of data.
- In this system, the client proxy may inform, the network proxy about behavior of a user and the network proxy can then pre-fetch pages based on this information.

- By going one step further we can implement a specialized network subsystem as shown in Fig. 5.6.7.
- The advantages of this system are same as those of the previous one but now, it is possible to further optimize the content transfer.
- Examples of this architecture are on line compression and replacement of protocols like HTTP and TCP, with protocols that are better adapted to the mobility and wireless access of the client.
- This system can also support web access over cellular telephone networks, that have low bandwidth and relatively high delay.
- The system replaces transport protocols and performs additional content transformation needed for mobile phones as well.
- The browser still uses HTTP to the client proxy.
- The client proxy uses a specialized transport service, the Mowgli data channel service, to the network proxy.
- We can use the standard protocols to the web servers.
- The exchange of messages between Client and network proxy takes place over Mowgli connections.
- This will avoid TCP's slow start and the one TCP connection per HTTP request requirement of HTTP/1.0.

5.7 Wireless Application Protocol (Version 1.x) :

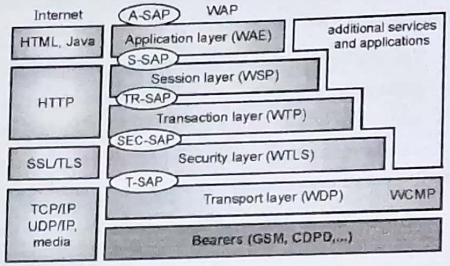
- The growth of the Internet, Internet applications, and mobile communications is responsible for many early proprietary solutions that provide Internet services for mobile, wireless devices.
- In order to avoid many incompatible solutions, e.g. special solutions for GSM, IS-136, or certain manufacturers, the wireless application protocol forum **(WAP Forum)** was founded in June 1997.
- It was renamed to Open wave; WAP Forum, 2000a.
- In summer 2002, the WAP forum, open mobile architecture forum and the SyncML initiative formed the open mobile alliance **(OMA, 2002) which can** cooperates with many other standardization bodies.
- In this section we will describes version 1.x of WAP which are still known as WAP Forum standards.

Objectives of WAP forum :

- The main objective of the WAP Forum and the OMA is to bring diverse internet content such as web pages and other data services to digital cellular phones as well as other wireless, mobile terminals like laptops.
- In addition to this, a protocol suite should make global wireless communication across different wireless network technologies, e.g., GSM, CDPD, UMTS etc. possible.
- The WAP forum embraces and extends the existing standards and technologies of the Internet and creates a framework for the development of contents and applications over a very wide range of wireless networks and wireless device types.
- All solutions must have the following features :
  1. **Interopearability**, i.e., to ensure that terminals and software from different vendors can communicate with networks from different providers;
  2. **Scalability**, i.e., to ensure that protocols and services scale up or down with customer needs and number of customers;
  3. **Efficiency**, i.e., to provide QoS suited to the characteristics of the wireless and mobile networks;
  4. **Reliability**, i.e., to provide a consistent and predictable platform for deploying services; and
  5. **Security**, i.e., to prevent the integrity of user data, protect devices and services from security problems.

5.7.1 Architecture of WAP :

- The WAP architecture is as shown in Fig. 5.7.1, which gives overview of its protocols and components.
- Fig. 5.7.1 also compares this architecture with the typical Internet architecture when using the world wide web.



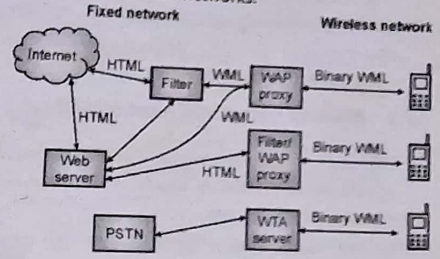
(G-3079) **Fig. 5.7.1 : Components and interface of the WAP 1x architecture**

- This comparison helps to understand the architecture of WAP Forum, 2000a.
- However, this comparison can be misleading as all components and protocols shown at the same layer are not comparable.
- The transmission of data is carried out on the basis of different **bearer services.**
- Note that, WAP does not specify bearer services, but it uses existing data services and integrates further services.
- Some example of such services 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.
- It also supports many other bearers, such as CDPD, IS-136, PHS.
- WAP forum architecture does not specify any special interface between the bearer service and the next higher layer, the transport layer with its wireless datagram protocol (WDP) and the additional wireless control message protocol (WCMP).
- In the WAP architecture the transport layer offers a bearer independent, consistent datagram-oriented service to the higher layers.
- It carries out communication transparently over one of the available bearer services.
- The higher layers use the transport layer service access point **(T-SAP)** which is the common interface, independent of the underlying network.
- WDP and WCMP are discussed in more detail later.
- The next higher layer which is the security layer along with its wireless transport layer security protocol **WTLS** offers its service at the security SAP (SEC-SAP).
- WTLS protocol is based on the transport layer security (TLS, formerly SSL, secure sockets layer) which we already know from the www.
- WTLS protocol has been optimized for the wireless networks with narrow-band channels.
- This protocol offers data integrity, privacy, authentication, and denial-of-service protection.
- The WAP transaction layer and its wireless transaction protocol **(WTP)** offer a lightweight transaction service at the transaction SAP **(TR-SAP).**
- This service provides reliable or unreliable requests and asynchronous transactions efficiently.
- The next higher layer is tightly coupled to this layer, if used for connection-oriented service.
- The session layer with the wireless session protocol (WSP) will offer two services at the session-SAP **(S-SAP)**, namely connection-oriented and connectionless if used directly on top of WDP.
- A special service for browsing the web (WSP/B) has been defined that offers the following features needed for wireless mobile access to the web.:
  1. HTTP/1.1 functionality.
  2. Long-lived session state.
  3. Session suspend and resume.
  4. Session migration.

- Finally job of the application layer and the wireless application environment **(WAE)** is to offer a framework for the integration of different www and mobile telephony applications.
- Some of the important issues here are :
  1. Scripting languages,
  2. Special mark up languages,
  3. Interfaces to telephony applications, and
  4. Many content formats adapted to the special requirements of small, handheld, wireless devices.
- Fig. 5.7.1 shows the overall WAP architecture and its relation to the traditional Internet architecture for www applications as well.
- The WAP transport layer together with the bearers is roughly comparable to the services offered by TCP or UDP over IP and different media in the Internet.
- In case of a bearer in the WAP architecture already offering IP services (e.g., GPRS, CDPD) the UDP is used as WDP.
- The WAP architecture has also adopted the TLS/SSL layer of the Internet with some changes required for optimization.
- The functionality of the session and transaction layer is roughly comparable with the role of HTTP in the web architecture.
- Also, the special formats and features that are optimized for the wireless scenario have been defined and telephony access has been added.
- WAP never forces all applications to use the whole protocol architecture. Instead, applications may use only a part of the architecture.
- That means, if an application does not want security but needs only the reliable transport of data, then it can directly use a service of the transaction layer.
- Similarly, simple applications can directly use WDP.

Integration of WAP components :

- Fig. 5.7.2 shows different possible scenarios for the integration of WAP components into existing wireless and fixed networks.



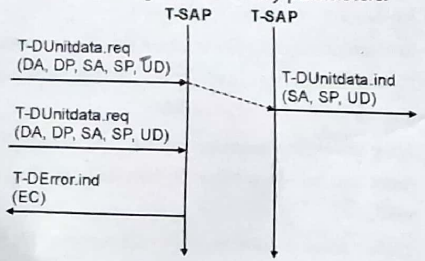
(G-3080) **Fig. 5.7.2 : Examples for the integration of WAP components**

- The left side of Fig. 5.7.2 shows different fixed networks, such as the traditional Internet and the public switched telephone network (PSTN).
- On its right side there is the WAP enabled wireless network.
- As we cannot change protocols and services of the fixed networks on the left side, we need to implement several new elements between these fixed networks and the WAP-enabled wireless, mobile devices in a wireless network on the right-hand side.
- The current www in the Internet offers web pages with the help of HTML and web servers.
- WAP defines a wireless markup language (WML) to enable the handheld devices to browse these web pages or additional pages.
- We can use special filters within the fixed network to translate HTML into WML, web servers can already provide pages in WML.
- Alternatively the gateways connected between the fixed and wireless network translate HTML into WML.
- These gateways filter pages and act as proxies for web access, as well, as explained in the following sections.
- In addition, for more efficient transmission, WML is converted into binary WML.

- Similarly, it is possible to implement a special gateway to access traditional telephony services via binary WML.
- This wireless telephony application (WTA) server translates, signaling of the telephone network (incoming call etc.) into WML events that are displayed at the handheld device.

5.7.2 Wireless Datagram Protocol (WDP) :

- The wireless datagram protocol **(WDP)** operates on top of many different bearer services that are capable of carrying data.
- In WAP Forum, 2000b at the T-SAP, the WDP offers a consistent datagram transport service which is independent of the underlying bearer.
- The adaptation needed in the transport layer, to offer this consistent service, can differ depending on the services of the bearer.
- The adaptation becomes smaller as the bearer service is closer to IP.
- If the bearer is already offering IP services, then UDP is used as WDP.
- WDP offers almost the same services as UDP.
- WDP offers source port numbers used for multiplexing and destination port numbers used for demultiplexing of data.
- Fig. 5.7.3 shows that the service primitive for sending a datagram is **TDUnitdata.req** with the destination address **(DA),** destination port **(DP),** Source address **(SA),** source port **(SP),** and user data **(UD)** being the mandatory parameters.



(G-3081) **Fig. 5.7.3 : WDP service primitives**

- Destination and source address are unique addresses for the receiver and sender respectively.
- These addresses could be MSISDNs (i.e., a telephone number), IP addresses, or any other unique identifiers.
- The **T-DUnitdata.ind** service primitive indicates that the data is received where destination address and port are only optional parameters.
- In the event of a higher layer requesting a service that the WDP cannot fulfill, there is an error.
- As shown in **Fig. 5.7.3**, this error is indicated with the **T-DError.ind** service primitive and an **error code (EC)** is returned to the higher layer indicating the reason for the error.
- It is not allowed for WDP to use this primitive for indicating problems with the bearer service.
- It can only use the primitive to indicate local problems, such as a user data size too large.
- In case of errors happening when WDP datagrams are sent from one WDP entity to another (e.g. the destination is unreachable, no application is listening to the specified destination port etc.), the **wireless control message protocol (WCMP)** provides error handling mechanisms for WDP.
- Therefore WCMP should be implemented together with WDP.
- The WCMP control messages resemble the Internet control message protocol (ICMP) for IPv4 and IPv6) messages.
- WCMP control messages can also be used for diagnostic and informational purposes.
- WDP nodes and gateways can use WCMP to report errors.
- However, we should not send the WCMP error messages as response to other WCMP error messages.
- In IP-based networks, ICMP will be used as WCMP.
- Typical WCMP messages are as follows :
  1. Destination unreachable (route, port, address unreachable),
  2. Parameter problem (errors in the packet header),
  3. Message too big,

  4. Reassembly failure, or
  5. Echo request/reply.
- An additional **WDP management entity** will support WDP and will provide information about changes in the environment, because this may influence the correct operation of WDP.
- The information such as the current configuration of the device, currently available bearer services, processing and memory resources etc. is important.
- However, design and implementation of this management component is considered to be outside the scope of WAP.
- If the bearer is already offering IP transmission, then WDP (i.e., UDP in this case) will rely on the segmentation (fragmentation in IP) and reassembly capabilities of the IP layer.
- Otherwise, WDP has to include these capabilities.

5.7.3 Wireless Transaction Layer Security (WTLS) :

- WTLS is based on a Transport Layer Security (TLS). It offers transport layer security between a WAP client and the WAP gateway/proxy.
- WTLS is required for the WAP to make sure the data integrity, privacy, authentication and protection from denial-of-service.
- It supports delivery of uncorrupted and unchanged data with encryption and it performs authentication and denial of service protection.
- WTLS is planned to use with the WAP transport protocols, which has the following features :
- **Data integrity :** WTLS makes sure that the data sent between the terminal and an application server is not damaged.
- **Confidentiality :** WTLS makes sure that the data sent between the terminal and an application server remains confidential.
- **Authentication :** WTLS makes sure the authenticity of the terminal and the application server.
- **Denial-of-service protection :** WTLS includes features, which will recognize and reject repeated data or data that has been not verified successfully.

5.7.4 Wireless Transaction Protocol (WTP) :

- The **wireless transaction protocol (WTP)** is placed on top of either WDP or, WTLS if security is required.
- WTP has been designed in such a way that it runs on very thin clients, such as mobile phones.

Advantages of WTP :

- WTP offers the following advantages to higher layers :
  1. Improvement in reliability over datagram services,
  2. Improved in efficiency over connection-oriented services, and
  3. It supports the transaction-oriented services such as web browsing.
- Here, we define a transaction as a request with its response, say for a web page.
- WTP can offer many features to the higher layers, on the basis of three classes of transaction service (class 0, class 1 and class 2).
- Class 0 is designed for unreliable message transfer without any result message.
- On the other hand, classes 1 and 2 provide reliable message transfer.
- Class 1 provides reliable message transfer without reliable result message, whereas, class 2 provides reliable message transfer with exactly one reliable result message (the typical request/response case).
- WTP uses duplicate removal, retransmission, acknowledgements and unique transaction identifiers to achieve reliability.
- Any of the WTP-classes does not require any connection set-up or tear-down phase to avoid unnecessary overhead on the communication link.

Features of WTP :

- WTP also has the following features :
  1. Asynchronous transactions,

2. Abort of transactions,
3. Concatenation of messages, and
4. Reporting success or failure of reliable messages

- To be consistent with the specification, here after, we will use the term **initiator** for a WTP entity initiating a transaction (client), and the term **responder** for the WTP entity responding to a transaction (server).

Service primitives :

- WTP offers the following three service primitives :
  1. **TR-Invoke** to initiate a new transaction,
  2. **TR-Result** to send back the result of a previously initiated transaction, and
  3. **TR-Abort** to abort an existing transaction.
- The WTPs exchanges following PDUs between two WTP entities for normal transactions :the **invoke PDU, ack PDU**, and **result PDU**.
- The service primitives, the PDUs, and the associated parameters vary with the classes of transaction service.
- A special feature of WTP is that the WTP entity can provide a **user acknowledgement** or, alternatively, an **automatic acknowledgement**.
- WTP user has to confirm every message received by a WTP entity if a user acknowledgement is required.
- A user acknowledgement provides a stronger version of a confirmed service.

5.7.5 Wireless Session Protocol (WSP) :

- The fifth layer of WAP is a wireless sessions layer (WSP).
- It is similar to HTTP/1.1 but some restrictions and extensions are made for the purpose of optimization.
- It corresponds to the sessions layer in the OSI model.
- For the connection-oriented and connectionless session services, the Wireless Session Protocol (WSP) provides an interface.
- The connection-oriented session service uses the transaction layer protocol. On the other hand, the connectionless session service uses a secure or non-secure datagram service.
- WSP is optimized for low-bandwidth networks with relatively long latency (delay). It can handle session interruptions and re-establish the sessions.
- WSP provides reliable, ordered exchange of data between client and server. Binary form of HTTP is the core of WSP design.

Features of WSP :

- Some of the important features of WSP are as given below :
  1. Session management
  2. Capability negotiation
  3. Content encoding

1. Session management :

- WSP can introduce long duration sessions that are established from a client to a server.
- WSP can also release the sessions in an orderly manner.
- WSP is capable of suspending and resuming a session that is important to mobile applications.
- If a mobile device is being switched off, then it would be useful if a user is able to continue operation at exactly the point where the device was switched off.
- Session lifetime does not depend on transport connection lifetime or continuous operation of a bearer network.

2. Capability negotiation :

- During the session establishment, clients and servers can agree upon a common level of protocol functionality.
- The parameters like maximum client SDU size, maximum outstanding requests, protocol options, and server SDU size can be negotiated.

3. Content encoding :

- WSP can also define the efficient binary encoding scheme for the content it is going to transfer.
- WSP offers content typing and composite objects as well.

5.7.6 Wireless Application Environment (WAE) :

- WAE is the topmost layer of WAP. It corresponds to the application layer in OSI model. In addition to higher cost, the other disadvantage of WAP is that it does not use HTML.
- Instead, the WAE layer uses a markup language called **WML** (wireless markup language). So a WAP device can access only those pages which are converted to WML.
- Therefore, it is necessary to increase the set of available pages. This is achieved by using an on-the-fly filter from HTML to WML.
- The goal of WAP is to achieve interoperable environment, which allows mobile operators and service providers to build applications that can reach wide variety of wireless platforms.
- WAP uses language such as WML and WML script. WML script can be used for validation of user input.
- The Wireless Application Environment, or WAE, provides an architecture for communication between wireless devices and Web servers.
- WAE follows an architecture that's more complicated than the WWW model because it needs to address the specific limitations of wireless devices.
- The cellular phone community is made up of a series of private networks. Because cellular phone providers maintain a partnership with a discrete network, each phone complies with the standards of only one network and there's limited synergy among the networks.
- WAE defines WAP as a public standard protocol. Wireless devices communicate via WAP in addition to the pre-existing protocols supplied by private networks.
- WAP is designed to handle slower processors and the specific constraints of the wireless network, such as limited bandwidth and high error rates, the markup languages available to wireless devices, such as WML and WML script, are scaled-down and conform to a format that requires less memory and processing power than HTML.
- In short, WAE consists of two parts: protocols (WAP, which includes WSP, WTP, WDP) and content (WML). Because a Web server only speaks HTTP, WAE uses a gateway to translate between WAP and HTTP.
- Each wireless device communicates with a designated gateway that makes calls to any number of Web servers.
- Wireless Application Environment (WAE) standards are most applicable to application developers.
- Based on the Internet standards the WAE provides vendor-neutral application architecture.
- The WAE specifications outline an application-programming model that supports browsing, scripting and extensions that allow cellular network operators to offer network services within WAP.
- Like the protocol stack specifications, the WAE standards are modified to suit the requirements of mobile devices and networks.
- The WAE defines user agents, services, and formats. User agents are simply applications that run inside a WAP-capable device such as a mobile phone.
- The standards support independent user agents to allow for expanded device functionality and to ensure that special services such as mobile network access are isolated from regular Internet services.
- The services that comprise the WAE include an extensible Markup Language (XML)-compliant Wireless Markup Language, a scripting language (WML Script) and supporting libraries, as well as telephony services provided by the Wireless Telephony Application libraries.

- Each class of information within the WAE is identified by a unique format.
- Encoding and decoding of content (e.g., WML) ensures that information sent between a user agent and the WAP gateway uses minimal bandwidth.

5.7.7 Wireless Markup Language (WML) :

- The **wireless markup language (WML)** (WAP Forum, 2000j) is based on the standard HTML.
- WML has been designed by taking several constraints of wireless handheld devices into consideration.
- First constraint is that, the wireless link always has a very limited capacity compared to that of a wire.
- The other constraints of the current handheld devices are as follows :
  1. Small displays,
  2. Limited user input facilities,
  3. Limited memory, and
  4. Only low performance computational resources.
- However, the gap between mobile and fixed devices regarding processing power is getting narrower.
- Today's CPUs used in PDAs have a performance close to that of desktop CPUs just a few years ago.
- WML is based on a deck and card metaphor. That means, a WML document is made up of multiple **cards** which can be grouped together to form 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 will navigate with the WML browser using a series of WML cards, review its contents, will enter requested data and makes choices.
- The WML browser would fetch decks as per requirement from origin servers.
- These decks can either be static files on the server or they can be generated dynamically.
- WML does not specify the way in which a WML browser has to interact with a user instead WML describes the intent of such interaction in an abstract manner.
- The user agent present on a handheld device needs to decide the best way to present all elements of a card.
- This presentation depends a lot on the capabilities of the device.

Features of WML :

- WML has the following basic features :

1. **Text and images :**

- WML gives, hints about way of presenting the text and images to a user. This is similar to the other mark-up languages.
- However, the exact presentation of data to a user depends on the user agent running on the handheld device.
- That means, WML only provides a set of mark-up elements, such as emphasis elements(bold, italic, etc.) for text.

2. **User interaction :**

- WML can support different elements for user input such as : text entry controls for text or password entry, option selections or controls for task invocation.
- Here also, the user agent has freedom of choosing how these inputs are implemented. They could be either bound to (physical keys, soft keys) or voice input.

3. **Navigation :**

- WML also offers a history mechanism with navigation with the help of browsing history, hyperlinks and other inter card navigation elements.

4. **Context management :**

- WML allows to save the state between different decks without interaction with the server, i.e., variable state can last longer than a single deck, and so it is possible to share the state across different decks.

- Over the narrow-band wireless channel, cards parameters can be defined by using this state without access to the server.

5.7.8 WML Script :

- The long form of WML script is Wireless Markup Language script. It is the scripting language of WML on the client's side.
- A scripting language is similar to a programming language but it is of lighter weight. WML script is a procedural programming language.
- It is dialect of JavaScript used for WML pages and it is an integral part of WAP (Wireless Access Protocol).
- WML script is very similar to JavaScript. It is used to carry out the tasks such as, user input validation, error message generation, creation of dialog boxes etc.
- Using WML script, it is possible for the wireless devices to carry out some of the processing and computation.
- This is important because it reduces the number of requests to and responses from the server.
- WML script is based on ECMA script i.e. European Computer Manufacturers Association. ECMA script is the standardized version of JavaScript.
- Therefore, the syntax of WML script is very similar to that of JavaScript but it is not fully compatible.
- WML script does not have objects or arrays, which are present in JavaScript.
- WML script has been optimized to work for low power devices and it is a compiled language.
- Various WML script operators are : Arithmetic operators, comparison operators, logical operators, Assignment operators and conditional operators.

5.7.9 Wireless Telephony Application (WTA) :

- For a handheld device user, browsing the web using the WML browser is just one application.
- The other applications could be : making phone calls and access all the features of the mobile phone network as with a traditional mobile phone.
- This is where we need to use the wireless telephony application **(WTA)**, the **WTA user agent** and the wireless telephony application interface **WTAI** come in.
- We can define WTA as a collection of telephony specific extensions for call and feature control mechanisms, which merges data networks and voice networks.
- The WTA framework performs the following functions :
  1. It integrates advanced telephony services using a consistent user interface (e.g., the WML browser) and
  2. It allows network operators to increase accessibility for various special services in their network.
- It is possible for a network operator to reach more end-devices using WTA.
- It is necessary for the WTA to enable third-party developers and network operators to create network-independent content that accesses the basic features of the bearer network.
- However, for security and stability reasons, most of the WTA functionality is reserved for the network operators.
- WTA extends the basic WAE application model in the following ways :

1. **Content push :**

- A WTA origin server can push content, such as WML decks or WML Script, to the client.
- Such a content push can take place without prior client request.
- The content enable the client to handle new network events that were not known before.

2. **Access to telephony functions :**

- The wireless telephony application interface (WTAI) provides multiple functions for handling telephony events like call accept, call setup, change of phone book entries etc.

3. Repository for event handlers :

- The repository is defined as a persistent storage on the client for content that are required to offer WTA services.
- Content can be either channels or resources. The resources are WML decks, WML Script objects, or WBMP pictures.
- Resources are either loaded using WSP or are pre-installed.
- A channel consists of references to resources and is associated with a lifetime.
- Within this lifetime, it is guaranteed that all resources that are pointed at by the channel are locally available in the repository.
- The repository essential because, it is necessary to react very quickly for time-critical events such as call accept.
- For such events it would take too long to load content from a server.

4. Security model :

- A security model is mandatory for WTA because many frauds happen with wrong phone numbers or faked services.
- WTA takes care that the client is connected only to trustworthy gateways.
- These gateways then check if the servers providing content are authorized to send this content to the client.
- It is not easy to define the term trustworthy in this context.
- In the beginning, the only trusted gateway would be the network operator's gateway and the network operator may decide the servers to be allowed to provide content.
- The WTAI specification (WAP Forum 2000m) has defined these libraries.
- These libraries allow the creation of telephony applications using the WTA user agent.
- We can use the library functions from WML decks or WML Script.

Library classes :

- The following three classes of libraries have been defined :
  1. Common network services
  2. Network specific services
  3. Public services

1. Common network services :

- This class contains libraries for services common to all mobile networks.
- The **call control** library contains, functions to set up, accept, and release calls.
- **Network text** library contains functions to send, read, and delete text messages.
- **Phonebook** allows for read, write and delete functions on the local phonebook entries.
- Finally, the library **miscellaneous** contains, a function that indicates the incoming data, e-mail, fax, or voice messages.

2. Network specific services :

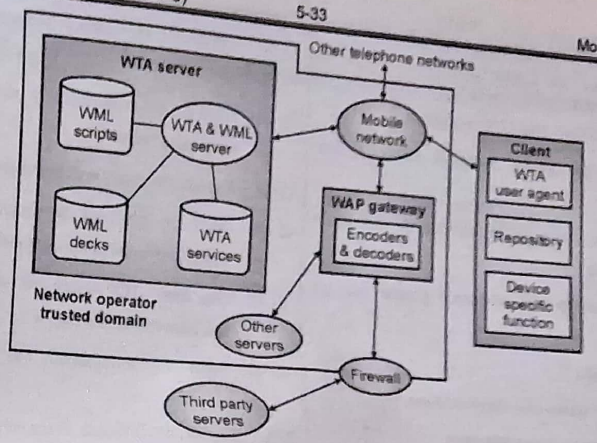
- Libraries in this class are dependent on the capabilities of the mobile network.
- In addition, this class may contain operator specific libraries.

3. Public services :

- Libraries in this class have publicly available functions, i.e., functions that are used by third-party providers, and not just by network operators.
- An example of such function is "make call" to set up a phone call.

5.7.10 WTA Logical Architecture :

- Fig. 5.7.4 shows the WTA logical architecture. All the components shown are not mandatory in this architecture; however, firewalls or other origin servers are useful.
- A minimal configuration of the logical architecture could contain a single server from the network operator serving all clients.



(G-3087) **Fig. 5.7.4 : WTA logical architecture**

- The **client** is connected with a **WTA server**, other telephone networks (e.g., fixed PSTN), and a **WAP gateway** via a mobile network.
- Fig. 5.7.4 does not show a WML user agent running on the client or on other user agents.
- The client can have voice and data connections over the mobile network.
- Other origin servers within the trusted domain can be connected via the WAP gateway.
- A firewall is used to connect third-party origin servers outside the trusted domain.
- One difference between WTA servers and other servers besides security is the tighter control of QoS.
- A network operator knows the latency, reliability, and capacity of its mobile network and can have more control over the behavior of the services.
- Other servers, that are probably located in the Internet, may not be able to give as good QoS guarantees as the network operator.

5.7.11 Advantages of WAP :

- Following are the advantages of WAP :
  1. It is an open standard.
  2. It is vendor independent.
  3. It is network standard independent.
  4. Fast speed technology.
  5. Can be implemented on multiple platforms.
  6. Most modern mobile telephone devices support WAP.

5.7.12 Disadvantages of WAP :

- Following are the disadvantages of WAP :
  1. Low speeds, very small user interface.
  2. Not very familiar to users.
  3. Business model is expensive.
  4. Forms are difficult to design.
  5. Third party is included.
  6. Poor security.

5.7.13 Applications of WAP :

- Following are the applications of WAP :
  1. Accessing the Internet from mobile devices such as E-mails.
  2. Playing games on mobile devices over wireless devices.
  3. Online banking via mobile phones.
  4. Weather forecasting.
  5. Flight information.
  6. Movie and cinema information.
  7. Traffic updates.

Review Questions

Q. 1 What is the main goal of a file system.

Q. 2 What are the general problems related to file systems.

Q. 3 Explain the terms strong consistency and weak consistency.

Q. 4 What is reintegration process ? Explain with an example.

Q. 5 What is Coda ?

Q. 6 Explain the states of a client in Coda.

Q. 7 Write a short note on Little work.

Q. 8 Write a short note on Ficus.

Q. 9 Write a short note on Little work.

Q. 10 Explain different problems faced by HTTP in the wireless environments.

Q. 11 What is the reaction of standard TCP in case of packet loss ?

Q. 12 Explain the congestion control taking place in wireless networks using traditional TCP.

Q. 13 Explain the slow start in wireless networks using traditional TCP.

Q. 14 What are the implications of using traditional TCP in wireless networks.

Q. 15 Explain the principle of operation of I-TCP.

Q. 16 What are the main drawbacks of the solution suggested in the previous question.

Q. 17 How does I-TCP isolate problems on the wireless link from the fixed network ?

Q. 18 State the advantages and disadvantages of I-TCP.

Q. 19 Explain the principle of snooping TCP.

Q. 20 State the advantages and disadvantages of snooping TCP.

Q. 21 State the advantages and disadvantages of M-TCP.

Q. 22 Explain the WAP architecture.

Q. 23 Write a note on WDP protocol.

Q. 24 Explain the WTP protocol.

Q. 25 Write a note on : WML.

□□□

Unit VI

Chapter 6

Mobile Platforms and Applications

Syllabus

Mobile device operating systems, Special constrains and requirements, Commercial mobile operating systems. **Software Development Kit :** iOS, Android, Blackberry, Windows Phone.

M-Commerce : Structure, Pros and Cons, **Mobile Payment System** : Security issues.

Chapter Contents

6.1 Operating Systems for Mobile Computing :

Definition :

- An Operating System (OS) is a program which acts as an interface between the system hardware and the user.
- A Mobile Operating System (Mobile OS) is an OS built exclusively for a mobile device, such as a smart phone, Personal Digital Assistant (PDA) or tablet.
- The smart phones are used to make phone calls, to make video conference calls, send multimedia messages, take pictures, play media files, browse World Wide Web (WWW), run remote applications, etc.
- The smart phone allows multiple tasks to be run on the device. Therefore, a powerful operating system is an important part of every smart phone.

6.2 Responsibilities of OSs in Mobile Devices :

- Two main responsibilities of operating systems in mobile devices are as follows :
  1. Managing resources
  2. Providing different interfaces

6.2.1 Managing Resources :

- Managing resources in a mobile device is an important task of the operating system which makes efficient utilization of the resources by performing multiple tasks.
- The operating system manages the resources such as processor, memory, files, and various types of attached devices such as camera, speaker, keyboard, and screen.
- The expectation from a mobile device is to run multiple applications at the same time and each application may require running multiple tasks. A task includes multiple threads.
- An example of such applications are voice communication, text messaging, e-mail, video play, music play, recording, web browsing, running remote applications, etc.
- An example of multitasking in a smart phone is as follows :
- A person listening to music can answer an incoming call, and an SMS can arrive at the same time and a person can look-up the SMS while the call is on.
- When various tasks compete to use the same set of resources, the OS acts like a traffic cop which makes sure that different tasks do not interfere with each other.

6.2.2 Providing Different Interfaces :

- The OS of a mobile device provides a highly interactive interface to the user of the device. It also acts as interface between other devices and networks.
- An important concern about an interface is to control data, and voice communications with the base station by using different types of protocols.
- An OS recognizes inputs from the keyboard, outputs to the display screen, and interfacing with peripheral devices such as other mobile devices, computers, printers, etc.

6.3 Mobile Operating System :

- The operating system provides a set of services to the application programs.

Layers of Operating System :

- The operating system is usually structured into a layers : Shell layer and Kernel layer.

1. Shell layer

- The function of the shell layer is to provide facilities to user, for interaction between used and the kernel.
- The shell programs are not memory resident part of an operating system.

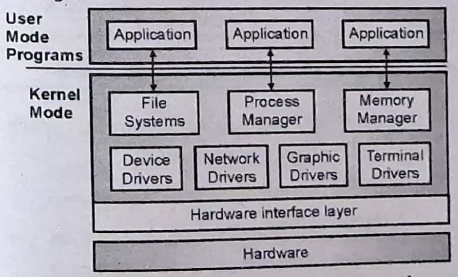
2. Kernel layer :

- A kernel is an important part of an OS and it manages system resources.
- It acts as a bridge between the software and hardware of the computer.

- The kernel layer executes in the supervisor mode. It can run privileged instructions that could not be run in the user mode.
- The kernel gets loaded first during booting process and it continues to remain in the main memory of the device.
- That means, paging does not apply to the kernel code and kernel data in a virtual memory system. Hence, the kernel is called as the **memory resident part** of an operating system.
- The kernel is responsible for interrupting servicing and management of processes, memory, and files.
- The kernel can be classified further into two categories :
  1. Monolithic Kernel
  2. Microkernel

Monolithic kernel :

- The traditional operating systems such as UNIX and Windows have a monolithic kernel design.
- The kernel in a monolithic OS design constitutes the entire operating system code, except the code for the shell.
- In this type of kernel approach, the entire operating system runs as a single program in kernel mode.
- The main motivation behind monolithic kernel design was the faith in the supervisor mode, where the operating system services can run more securely and efficiently.
- Fig. 6.3.1 shows architecture of monolithic kernel.



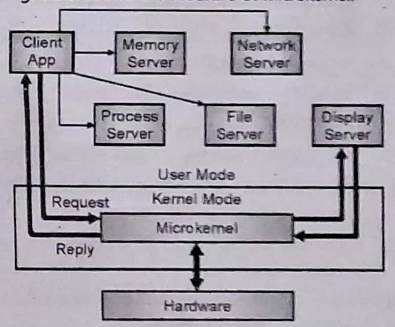
(G-3151) Fig. 6.3.1 : Monolithic kernel design of an operating system

Disadvantage of Monolithic kernel :

- The main problem with this design is that it makes the kernel huge and non-modular. It makes kernel code hard to tailor, maintain, extend and configure.
- Microkernel design approach has been proposed to overcome the problem of monolithic kernel design.

Microkernel :

- The microkernel design tries to minimize the size of the kernel code.
- In this design, a few important and the basic hardware-dependent functionalities are implemented in the kernel mode and all other functionalities are implemented in the user mode.
- The majority of the operating system services run as user level processes.
- Fig. 6.3.2 shows architecture of microkernel.



(G-3152) Fig. 6.3.2 : Microkernel design of an operating system

- The kernel code is very difficult to debug as compared to application programs because a bug in a kernel code can crash the system, thus crashing the debugger too.
- If some OS service crashes while being used by a user, it does not bring down the entire system.
- A microkernel operating system is expected to be more reliable than an equivalent monolithic kernel operating system.

- Most of the mobile operating systems are based on the microkernel design in order to minimize the size of kernel of a mobile operating system.

Advantage of Microkernel design :

- Following are the advantages of using microkernel design :
  1. The main advantage of microkernel approach is that it becomes easier to port, extend, and maintain the code of the operating system.
  2. Microkernel design is small and isolated, therefore it can function better.
  3. Microkernel is more secure as compared to monolithic kernel.
  4. The expansion of the system is more feasible, so it can be added to the system application without disturbing the Kernel.
  5. Microkernels are modular. The different modules can be replaced, reloaded, modified without even touching the Kernel.
  6. Microkernel interface helps to enforce a more modular system structure.
  7. New features can be added without recompiling.
  8. Microkernel system is flexible, so different strategies and APIs, implemented by different servers can exist in the system.

6.3.1 Differences between Monolithic Kernel and Microkernel :

Table 6.3.1 : Differences between Monolithic Kernel and Microkernel

Sr. No	Parameter	Monolithic kernel	Microkernel
1.	Address Space	User services and kernel services are kept in the same address space	User services and kernel services are kept in separate address space
2.	Design and Implementation	OS is easy to design and implement	OS is complex to design and implement
3.	Size	Larger than microkernel.	Smaller
4.	Functionality	Difficult to add new functionalities.	Easier to add new functionalities.
5.	Code	Less code required as compared to microkernel.	More code is required.
6.	Failure	Failure of one component leads to the failure of the entire system.	Failure of one component does not affect the working of micro kernel.
7.	Execution speed	High	Low
8.	Extend	Not easy to extend monolithic kernel	Easy to extend Microkernel
9.	Debugging	Difficult	Simple
10.	Maintenance	Extra time and resources are required for maintenance.	Easy to maintain
11.	Example	Microsoft Windows 95	Mac OS X

6.4 Special Constraints of Mobile OS :

- Following are the special constraints which influence the design of a mobile OS :
  1. Limited processing power
  2. Limited battery power
  3. Limited screen size
  4. Miniature keyboard
  5. Limited memory
  6. Limited and fluctuating bandwidth of the wireless medium

1. Limited processing power :

- A majority of the modern mobile devices consist of ARM (Advanced RISC Machine) based processors.
- ARM based processors are definitely energy efficient, powerful and cheaper as compared to the desktop or laptop processors, but these processors are considerably slower.
- Also, there is restriction on the sizes of the on-chip and off-chip memories.
- The operating system provides only a limited number of functionalities that are useful in the actual mobile operation in order to manage with the restricted processing power, storage and battery power.
- The mobile application development activity which needs to use memory-intensive utility programs (like editors and compilers) is carried out on a desktop or laptop.
- After the completion of simulation and testing of application it is cross-compiled and downloaded onto the mobile device.

2. Limited battery power :

- Mobile devices should be lightweight so as to increase their portability.
- Due to the strict limitations on the size and weight of mobile device, it has a small battery and it is not possible to recharge the device as and when required.
- Despite small battery, a mobile phone must support long talk time without need of frequent recharging.
- As a result, the mobile OS must be not only computationally efficient, but also it is expected to minimize the power consumption.
- Following techniques are used by OS to reduce power consumption of device :
  1. To keep the processor and display screen into sleep mode within a few seconds of inactivity.
  2. To vary the intensity of transmitted antennae power as per requirement.

3. Limited screen size :

- The screen size of a mobile handset needs to be small in order to make it portable which restricts the size of the display screen.
- To overcome the constraint of screen size and minimize user inconveniences, new innovative user interfaces should be supported by the mobile OS.

4. Miniature keyboard :

- Mobile devices are either provided with the small-sized display screen or a small keypad is designed to be used as a keyboard in a touch screen mode by using a stylus.
- The typing in the documents and entering the string commands is difficult with these arrangements.

5. Limited memory :

- In a mobile device there is less permanent and volatile storage as compared to that of a modern desktop or laptop computers.
- The OS should be as small as possible to cope with the limited memory of a mobile device.
- It should provide a rich set of functionalities to meet the user requirements and demands.
- The kernel size is very important figure of merit in a mobile OS.

6. Limited and fluctuating bandwidth of the wireless medium :

- The mobile OS needs to run complex protocols due to the inherent problems caused by mobility and the wireless medium.

- A wireless medium is vulnerable to atmospheric noise and causes high bit error rates.
- The bandwidth of a wireless channel can change randomly because of atmospheric noise, movement of some objects or the movements of the mobile handset itself.
- The bandwidth change can results in as short-term fades and there can be comparatively longer-term disconnections due to handoffs.
- In this situation, uninterrupted communication needs a special support for data caching, pre-fetching and integration.

6.5 Special Service Requirements :

- Some facilities and services are normally not expected to be supported by a traditional operating system but they are to be mandatorily supported by a mobile OS.
- Following are some special service requirements for mobile OS :
  1. Support for a variety of input mechanisms
  2. Support for specific communication protocols
  3. Extensive library support
  4. Compliance with open standards

1. Support for a variety of input mechanisms :

- In an inexpensive mobile device, a miniature keyboard forms the main user input mechanism whereas sophisticated mobile devices (smart phones) support the QUERTY keyboard.
- A lot of recent mobile devices support touch screen or even stylus-based input mechanisms along with the handwriting recognition capability.
- Different input mechanisms influence the intended use of a device as well as the specific customer segment for which it is positioned.
- These input mechanism issues dictate the choice and difficulty of the user interaction part of the OS to a large extent and the internal design of the OS to a smaller extent.
- A mobile OS must support a variety of input mechanisms in order to make it general and usable by mobile devices of different manufacturers.

2. Support for specific communication protocols :

- Improved communication support is required for mobile devices because they are always connected to the base station and various types of peripheral devices, computers and other mobile devices.
- The communication protocols used for communication with the base station are dependent on the generation of the communication technology (1G to 5G etc.) in which the mobile device is deployed.
- As a mobile should be usable across the existing technology spectrum, it becomes necessary to support two or more mobile generations simultaneously.
- TCP/IP and wireless LAN protocols must support the communicating with other devices and with computers.
- Even though mobile devices are equipped with USB and other types of ports, mobility constraints always make infrared or Bluetooth connections preferable for web browsing as well as communication with other personal devices like pen drive and headphones.
- It is required that the operating system must support various interfacing protocols and hardware interfaces.

3. Extensive library support :

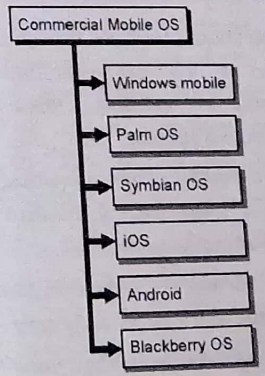
- The mobile OS requires extensive library support for the cost-effective development of third party applications.
- The library support must include the availability of programmer callable primitives for email, SMS, MMS, Bluetooth, multimedia, user interface primitives, and GSM/GPRS functionalities.

4. Compliance with open standards :

- It becomes easy for the third-party software developers to stick to an open standard for the development of innovative applications.
- If OS adhere to open standards then it becomes easy for the third party developer to reduce the development cost and time-to market by the mobile handset manufacturers.
- Smart phones come in different shapes and sizes with varying screen sizes and user input abilities.
- Hence the user interface and networking abilities of a mobile OS must be designed by keeping these diversities in mind.

6.6 Commercial Mobile Operating Systems :

- It becomes challenging to design a mobile OS with a set of core abilities that must be supported by mobile devices and it must provide a reliable programming environment across all smart phones that install the operating system.
- The mobile OS should facilitate third party development of application software.
- The mobile OS must allow manufacturers of different brands of mobile devices to build set of functionalities of their choice for the users.
- A few popular mobile OSs are as shown in Fig. 6.6.1.



(G-3153) **Fig. 6.6.1 : Commercial mobile operating systems**

6.6.1 Windows Mobile :

- Windows mobile is designed in a such way that it looks and feels very similar to the desktop version of Windows.
- Microsoft knows many users are familiar with the desktop version of Windows and they can easily operate Windows mobile.
- In addition to the core capabilities required by a mobile device, many third-party software applications are available in Windows mobile.
- The third-party software applications can be purchased through the Windows marketplace for mobiles.
- Windows marketplace is a website maintained by Microsoft.
- On the Windows marketplace, different application developers can submit their applications which can be downloaded by the subscribers.
- The developer gets 70 % of the fee received by hosting their applications.
- Windows mobile has recently launched the Windows phone 7.
- Windows phone 7 is a improved version of the Windows mobile OS.
- Windows phone 7 is not backward compatible with the Windows mobile operating system. That means a mobile application that runs on the Windows mobile may not run on the Windows phone OS.

Hardware specifications for Windows phone 7 :

- Microsoft has defined the following hardware specifications for Windows phone 7 device :
  1. It must support a screen resolution of 800 x 480 pixels.
  2. It should have an accelerometer and a compass in the device.
  3. It provides a touch screen interface with facilities for both command and text input.
  4. The OS is detected when a device is rotated from portrait to landscape orientation.

Features of the Windows mobile OS :

- Following are the important features of the Windows mobile OS :
  1. It provides a virtual memory management.
  2. It does not provide multitasking at present.
  3. It supports security by doing provision of a cryptographic library.
  4. Application development of Windows mobile OS is similar to that of the Win32 environment because many programmers are aware of Win 32-based application development.
  5. The Graphics/Window/Event manager (GWE) component in windows handles all input and output.
- An application in the background of windows mobile goes into hibernation and becomes active only when it comes to foreground.
- The expectation from Microsoft is it may support true multitasking in the future versions of the Windows Phone operating system.

6.6.2 Palm OS :

- Palm OS is also known as Garnet OS. In 1998, Palm Computing developed Palm OS for its highly successful PDA (Personal Digital Assistants) called Palm Pilot.
- Palm operating system was designed for simplicity and it provides a touch screen-based graphical user interface.
- Later on, Palm OS is upgraded which makes installation easy in several different mobile devices, such as smart phones of different makes, wrist watches, hand-held gaming consoles, bar code readers and GPS devices.

Features of the current Palm OS :

- The key features of the current Palm OS (Garnet) are as follows :
  1. A handwriting recognition-based system for user input is supported by palm OS.
  2. It supports Hot Sync technology for synchronization of data with desktop computers.
  3. It consists of simple memory management system. It does not separate the memory areas of applications from each other in order to keep the operating system small and fast. Thus, any misbehaving application can crash the system.
  4. It provides sound playback and recording abilities.
  5. It includes a very simple and basic security system in which a device can be locked by password.
  6. It is basically a single task operating system that means only one application can run at a time.
  7. Palm emulator supplied by Palm emulates the Palm hardware on a PC. Due to this Palm programs to be developed and debugged on a PC before being run on the Palm hardware.
  8. It supports interfaces like Serial port / USB, infrared, Bluetooth and Wi-Fi connections.
  9. With a proprietary format Palm stores calendar, address, task and note entries and that can be accessible by third-party applications.

6.6.3 Symbian OS :

- Symbian OS was developed through a association between a few well-known mobile device manufacturers including Nokia, Ericsson, Panasonic, and Samsung.
- The main objective of these manufacturers was to develop a single industry standard operating system (Hall and Anderson, 2009).
- Symbian OS was the undisputed leader in the smart phone OS market for many years.
- This OS was used in the handsets manufactured by Nokia, Ericsson, Panasonic, and Samsung.

- Symbian OS runs on ARM-based processor designs and it is a real time, multitasking, pre-emptive, 32-bit OS. Symbian OS is a microkernel-based operating system.
- When the applications are not directly dealing with an event, the CPU is switched into a low power mode.

Features of Symbian OS :

- A few important features of the Symbian OS are as follows :
  1. It is optimized for low-power and memory requirements. An object-oriented design paradigm is followed by applications, and the OS itself.
  2. The communication and networking protocols supported by Symbian OS includes TCP, UDP, PPP, DNS, FTP, WAP, etc. It supports Bluetooth, InfraRed and USB connectivity for personal area networking.
  3. It supports memory protection and pre-emptive multitasking scheduling.
  4. Programming in Symbian OS is event-based. When the applications are not directly dealing with an event then the CPU is switched into a low-power mode and this is achieved through a programming idiom called active objects.
  6. An Integrated Development Environment (IDE) toolkit Carbide is available for C++ application development on Symbian OS. It basically works as an Eclipse plug-in and it includes editor, compiler, emulator, libraries and header files required for Symbian OS development.

6.6.4 iOS :

- The iPhone was designed to replace Apple's iPod.
- Apple developed iOS as iPhone's operating system and it is originally known as iPhone OS, but later it renamed as iOS.
- iOS is derived from Mac OS. Later on, Mac OS was extended for use in other Apple devices such as the iPod touch, iPad and Apple TV.
- It is mainly written in C, C++, Objective-C, assembly language, and Swift. The first version of iOS was launched in 2007.
- It is a closed and proprietary OS, owned and controlled by Apple. It cannot be used by various mobile phone vendors on their systems.
- Apple does not allow iOS installation on third-party hardware. iOS grabbed the market attention because several innovative features are provided in iOS.
- iOS includes some user interactions with gestures such as swipe, tap, pinch, and reverse pinch etc.
- An innovative user interaction supported by the iOS includes internal accelerometers.
- Internal accelerometers are used by some applications to respond for shaking the device for the undo command, rotating the device in three dimensions to switch the display mode from portrait to landscape, etc.

Features of iOS :

- The following are the popular features of iOS :
  1. iPhone offers multitasking features.
  2. Apple's iCloud service offers Internet-based data storage.
  3. Notification center in iOS shows you all of your app alerts in one place.
  4. A three-axis accelerometer is built into every iOS device, providing acceleration readings in each of the three axes.
  5. To detect your location, the iPhone uses an inbuilt Assisted GPS (AGPS) chip.
  6. Apple supplies the Core Bluetooth framework, which includes classes for connecting to Bluetooth-enabled low-energy wireless technology.

7. The iOS apps can be used in both portrait and landscape modes.
8. Apple offers an online mapping service that can be utilized as the iOS default map system.

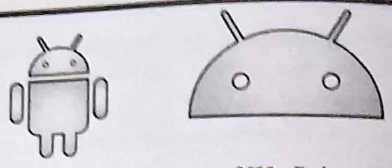
6.6.5 Android :

Definition :

- Android is a Linux-based operating system designed mainly for touch screen mobile devices such as smartphones and tablet computers
- Android system is developed by Google and later the company named **Open Handset Alliance (OHA)**.
- It is a Linux kernel-based system equipped with rich components that allows developers to create and run applications that can perform both basic and advanced functions.
- Mainly JAVA language is used to write the Android code even though other languages can be used.
- Android is an absolute set of software for other devices like tablets, smart watches, set-top boxes, smart TVs, notebooks, etc. The Google launched the first version of the Android platform in 2007.
- Since then, Google released a lot of android versions such as Apple Pie, Banana Bread, Cupcake, Donut, Éclair, Froyo, Gingerbread, Jellybeans, Kitkat, Lollipop, marshmallow, Nougat, Oreo, etc. with some extra functionalities and new features.
- Google released the first beta version of the Android Software Development Kit (SDK) in 2007 where as the first commercial version, Android 1.0, was released in September 2008.
- Google announced the next Android version, **4.1 Jelly Bean** on June 27, 2012.
- The android OS is an open-source operating system that means it is free and anyone can use it.

Android logo :

- *Fig. 6.6.2 shows the logos of Android Operating system.*



2008 – 2018 2019 – Today

(O-1464) **Fig. 6.6.2 : Android Logo**

- Android OS allows developer to build innovative applications and games with the use of java programming.

Goal of Android OS :

- The main goal of android OS is to create a successful real world product that improves the mobile experience for the end users.

Features of Android :

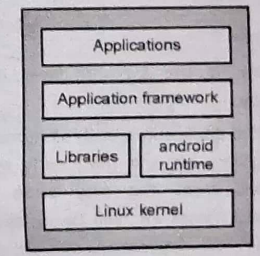
- Some of the important features of the Android system are as follows :

1. It is open-source and we can customize the OS based on our requirements.
2. A user can select lot of mobile applications with Android system.
3. Android offers many features such as whether details, opening screen, etc.
4. It provides support for messaging services (SMS and MMS), web browser, a lightweight relational database (SQLite), connectivity (GSM, CDMA, Bluetooth, Wi-Fi etc), media, handset layout etc.
5. Android supports connectivity for GSM, CDMA, Wi-Fi, Bluetooth, etc. for telephonic conversation or data transfer.
6. It contains multiple APIs that can support location-tracking services such as GPS.
7. With the use of file manager the user can manage all data storage related activities.
8. Android contains a wide range of media supports like AVI, MKV, FLV, MPEG4, etc. to play or record a variety of audio/video.

9. It supports various image formats like JPEG, PNG, GIF, BMP, MP3, etc.
10. It supports multimedia hardware control to perform playback or recording using a camera and microphone.
11. It provides support for virtual reality or 2D/3D Graphics.
12. Android OS basic screen provides a beautiful and intuitive user interface.
13. User can jump from one task to another and various applications can run simultaneously.
14. It supports the Google services like Gmail, Chrome, Google Search, Location Manager, Google Maps, Google Drive, Google Play Store, etc.
15. It supports various languages.
16. It supports development tools like android Studio, Eclipse IDE , Android Emulator.

Android architecture :

- Fig. 6.6.3 shows the architecture of an Android OS.

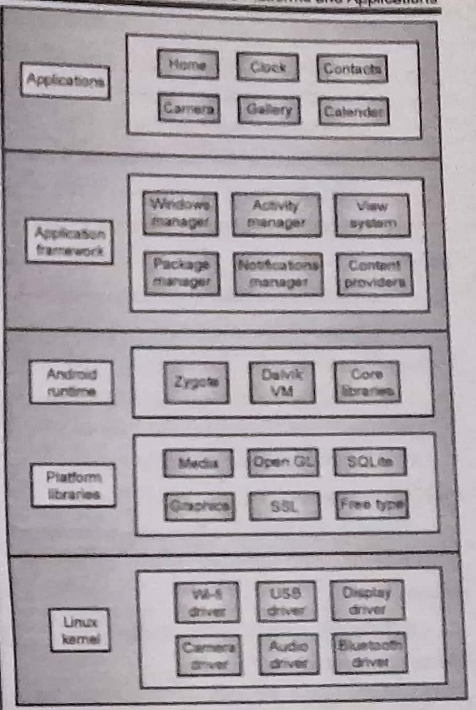


(O-1443) **Fig. 6.6.3 : Android Architecture**

- The main components of android architecture or Android software stack are as follows :

1. Linux Kernel
2. Platform / Native Libraries
3. Android Runtime
4. Application Framework
5. Applications

- Fig. 6.6.4 shows detailed architecture of an Android OS.



(O-1442) **Fig. 6.6.4 : Detail Android Architecture**

1. **The Linux Kernel :**
   - The android uses the powerful Linux kernel that supports a wide range of hardware drivers.
   - The Linux kernel is the heart of the android architecture.
   - It manages all the available drivers such as display drivers, camera drivers, Bluetooth drivers, audio drivers, memory drivers, etc. which are required during the runtime.
   - Linux kernel is very good at networking and it is not necessary to interface it to the peripheral hardware.
   - The Linux Kernel will provide an abstraction layer between the device hardware and the other components of android architecture.

- The Linux kernel does not interact directly with the user but it interacts with the shell and other programs as well as with the hardware devices on the system.
- It is responsible for management of memory, power, devices etc.

Functions / Features of Linux-Kernel :

Security :

- The Linux kernel handles the security between the application and the system.

Memory Management :

- It efficiently handles the memory management by providing the freedom to develop our applications.

Process Management :

- It manages the processes and whenever required it allocates resources to processes.

Network Stack :

- It handles the network communication.

Device management :

- It gives support for various drivers like audio, video, camera, USB, Wi-Fi, Bluetooth, Display etc.

2. Native Libraries :

- On the top of linux kernel there is a set of native libraries such as WebKit, Open GL, Free type, SQLite, Media etc.
- The Webkit library is responsible for browser support.
- SQLite is responsible for database, free type for font support, media for playing and recording audio and video formats.
- The responsibility of SSL (Secure Sockets Layer) is to establish an encrypted link between a web server and a web browser.

3. Android Runtime :

- Android runtime consists of the core libraries and Dalvik Virtual Machine (DVM) which runs the Android applications.
- DVM is like JVM but it is optimized for mobile devices to run faster and consume less memory.
- The Dalvik Virtual Machine is a kind of java virtual machine (JVM).
- It is specially designed and optimized for the mobile devices.
- The performance of DVM is faster as it consumes less memory.

4. Application Framework :

- Application framework is located on the top of native libraries and android runtime.
- Android / Application Framework consists of **Android API's** such as User Interface (UI), telephony, resource manager, location manager, data (content) providers, package managers, etc.
- Application Framework provides several important classes and interfaces which are used to create an Android application.
- The Android framework includes the following key services :
  1. Activity Manager
  2. Content Providers
  3. Resource Manager
  4. Notifications Manager
  5. View System

5. Applications :

- Applications is the topmost layer of android architecture.
- The installed applications such as home, contacts, camera, gallery etc and third party applications downloaded from the play store like chat applications, games etc. will be installed on this layer only. All applications are using android framework.
- Android framework uses native libraries and android runtime. Android runtime and libraries uses Linux kernel.

6.6.6 Blackberry OS :

- Blackberry OS is designed for Blackberry smart phones created by Research In Motion Limited (RIM).
- It is a proprietary operating system. Details of its architecture have not been published as it is a proprietary operating system.

- The very good email system that it deploys is easily noticed at the user level.
- Blackberry OS supports instant mailing. It maintains a high level of security through on-device hardware-based message encryption while mailing.
- The Blackberry OS runs on Blackberry variant phones such as Blackberry Bold, Curve, Pearl and Storm series.

Features of BlackBerry OS :

1. The BlackBerry OS is best known for its robust support for push Internet email.
2. This OS is designed in C++ programming language only.
3. It supports wireless communications using GSM, 3G, CDMA and Wi-Fi networks.
4. It provides multitasking dacility.
5. It provides email facility.
6. The blackberry handheld has all features of mobile phone along with walkie talkie and comes with comfortable headset.
7. Corporate data can be easily accessed.
8. It provides faster services.

Advantages BlackBerry OS :

1. Excellent Connectivity.
2. Amazing email client (can also sync with your enterprise server).
3. Most user friendly.
4. Easy typing due to QUERTY keyboard.
5. Very fast.

Disadvantages BlackBerry OS :

1. Battery life not good in some models.
2. Camera quality is not good.
3. Bad application support.

6.7 Comparison of Mobile Operating Systems :

- Table 6.7.1 gives the comparison of mobile operating systems.

Table 6.7.1 : Comparison of mobile operating systems

Sr. No.	Parameter	Android	IOS	Windows Phone	Blackberry OS	Symbian OS
1.	Symbol					
2.	Vendor	Open Handset Alliance, Google	Apple, Inc	Microsoft	Blackberry Ltd.	Accenture on the behalf of Nokia
3.	OS family	Linux	Darwin	Windows CE 7	QNX	RTOS
4.	License	Free, public, open source	Proprietary	Proprietary	Proprietary	Proprietary,
5.	Written in	C, C++, Java	C, C++, Objective C, Swift	C#, VB.NET, F#, C++, Jscript	C, C++, HTML 5, Java script, CSS, Action script, Java	C, C++, ME, Python, Ruby, Flash Lite
6.	Market Size	Very High	High	Medium	Low	Very low

Sr. No.	Parameter	Android	iOS	Windows Phone	Blackberry OS	Symbian OS
7.	Application Store	Google Play	iPhone App Store	Windows Phone Store	Blackberry App World	Nokia Ovi Store
8.	Cross Platforming	Supports cross platforming	Don't support cross platforming	Support cross platforming	Don't support cross platforming	Supports cross plat forming
9.	Future Prospect	Very High	High	Medium	Low	Low

6.8 Software Development Kit :

Definition :

- A software development kit (SDK or devkit) is a set of software development tools that allows the creation of applications for a certain software package, software framework, hardware platform, computer system, video game console, operating system, or similar development platform.
- An SDK contains an application programming interface (API which acts as a link between software applications and the platform they run on.
- It is possible to build APIs in many ways and includes helpful programming libraries and other tools. SDKs can be licensed by the software provider.

6.8.1 iOS SDK :

- The iOS SDK is a software development kit which helps developers to create native applications for Apple's iOS devices and platforms.
- The iOS SDK was formerly known as the iPhone SDK.
- The iPhone SDK provides tools for Apple's touchscreen interface and for its proprietary iOS operating system.
- iOS SDK is used to run Apple's iPhones as well as Apple's other mobile devices such as the iPad.
- With the help of the native iOS SDK you can communicate with the Client API.
- The iOS user interface is based on the concept of direct manipulation by using multi-touch gestures.
- In iOS, interface control elements consist of sliders, switches, and buttons and interaction with the OS contains gestures such as swipe, tap, pinch, and reverse pinch.
- Internal accelerometers are used by some applications to respond for shaking the device for the undo command, rotating the device in three dimensions to switch the display mode from portrait to landscape etc.

6.8.2 Android SDK :

- The Android SDK is a collection of software development tools and libraries used to develop Android applications.
- The Android SDK includes :
  1. Required libraries.
  2. Debugger.
  3. An emulator.
  4. Relevant documentation for the Android application program interfaces (APIs).
  5. Sample source code.
  6. Tutorials for the Android OS.
- When Google releases a new version of Android or an update, a corresponding SDK is also released.
- The Android SDK includes all the tools required to code programs from scratch and even test them.

- These tools provides a smooth flow of the development process from developing and debugging to packaging.
- The Android SDK is compatible with Windows, macOS and Linux.

6.8.3 BlackBerry SDK :

- The BlackBerry Dynamics SDK provides a powerful set of tools which allows you to focus on building useful productivity applications.
- The BlackBerry Dynamics SDK is used to develop applications for all major platforms that leverage valuable services, including secure communications, data exchange, presence, push, directory lookup, single sign-on authentication, and identity and access management.
- BlackBerry SDK allows application developers to integrate any BlackBerry service (i.e., BlackBerry UEM, BlackBerry Workspaces and BlackBerry Dynamics) into their applications through a Platform-as-a-Service model.
- The BlackBerry platform is known for its native support for corporate email which allows complete wireless activation and synchronization with Microsoft Exchange, Lotus Domino, or Novell GroupWise email, calendar, tasks, notes, and contacts, when it is used with BlackBerry Enterprise Server.

6.8.4 Windows Phone SDK :

- Windows SDK or Windows Software Development Kit developed by Microsoft is a collection of tools that allows developers to create software, frameworks, or applications for any organization.
- Windows Phone SDK contains :
  1. Documentation.
  2. Header files.
  3. Libraries.
  4. Samples and tools required to develop applications for windows phone.
- Windows SDK allows developers to include various functionality into their programs.
- It allows them to quickly and easily develop the typical features and components of the applications.
- The Windows Software Development Kits reduces the complexity of the integration process by simplifying basic operations like authorization signatures and SMS message interpretation through local platforms or languages.

6.9 M-Commerce (Mobile Commerce) :

- An important application of mobile computing is Mobile commerce (M-commerce).

Definition :

- Mobile commerce (m-commerce) carry out any activity related to buying and selling of product, services, or information by using the mobile hand-held devices such as hand-held computers, mobile phones or laptops.
- Mobile e-commerce is just one of the many subsets of electronic commerce. Mobile e-commerce can also be known as mobile commerce
- Over the last decade M-commerce has become extremely popular.
- M-commerce is very convenient to the buyers and sellers.

Examples of M-commerce :

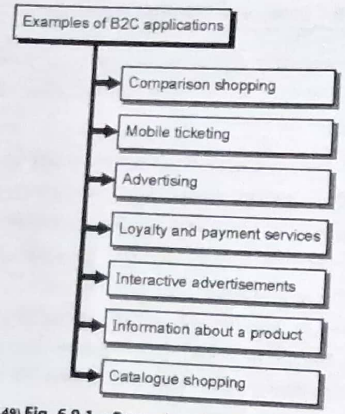
- Broken into three main categories (mobile shopping, mobile payments, and mobile banking), the highest growth areas for m-commerce are :
  1. In-app purchasing (buying clothing items via a retail app).
  2. Mobile banking
  3. Virtual marketplace apps like Amazon
  4. Digital wallets like Apple Pay, Android Pay, and Samsung Pay.
  5. Mobile ticketing

6.9.1 Applications of M-Commerce :

- M-commerce applications can be categorized into following two applications :
  - A. Business-to-Consumer (B2C) Applications
  - B. Business-to-Business (B2B) Applications

A. Business-to-Consumer (B2C) Applications :

- In Business-to-consumer (B2C) applications the products or services are sold by a business firm to a consumer.
- Fig. 6.9.1 shows few examples of B2C applications of M-commerce.



(G-3149) **Fig. 6.9.1. : Examples of B2C applications**

1. **Comparison shopping :**
- In this type of shopping, customers can compare pricing of a product at different stores and also the prices of the related products by using their mobile phones.

2. **Mobile ticketing :**
- With the help of mobile phones movie tickets can be purchased (called m-tickets) by using credit cards, debit cards, paytm, UPI etc.

After receiving the payment, a unique bar code is sent to the customers's mobile phone by an SMS.

The purchaser can enter into the movie hall by simply showing the bar code downloaded into the mobile device.

3. **Advertising :**
- A targeted advertising can be done by using the demographic data collected by the wireless service providers.
- The wireless service provider keeps track of the purchase history of customers by directing advertisements to mobile phones.

4. **Loyalty and payment services :**
- Mobile phones can replace the physical loyalty cards. After signing up for a supermarket loyalty scheme, a unique bar code is sent to a customer's mobile phone.
- The customer shows the bar code at the cash counter after shopping at the same supermarket, and collects points based on the total amount spent. The payments can be done by using mobile phones.

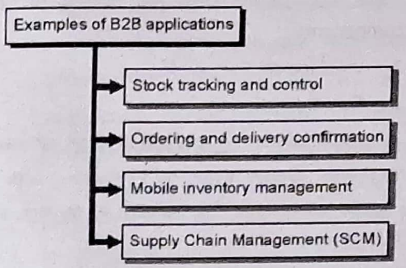
5. **Interactive advertisements :**
- In this application, consumers scans a bar code in an advertisement for a product appearing on a TV screen by using their mobile phones.
- The consumer scans the bar code and order the product by invoking an internet application.

6. **Information about a product :**
- Customer can get additional information about product through their mobile phones.

7. **Catalogue shopping :**
- A customer can place order for products listed in a catalogue by using their mobile phones.
- A customer may receive a catalogue by SMS from a catalogue shopping company.
- The consumer can buy products directly from the catalogue shopping company by scanning the bar codes.

B. Business-to-Business (B2B) Applications :
- In Business-to-Business (B2B) applications products or services are sold by a company to its dealers.

- E.g. A company of TV manufactures will normally sell its products through a dealer network instead of selling the product directly to the consumers.
- In this example, the manufacturer and the dealers are the B2B partners.
- Fig. 6.9.2 shows few examples of B2B applications of M-commerce.



(G-3150) **Fig. 6.9.2 : Examples of B2B applications**

1. **Stock tracking and control :**
- With the help of mobile phones, it is possible to keep track of the stock in a distributed inventory system and updates are sent to a central database.
- Mobile phone can scan bar codes or RFID tags on products and employees can update the stock in real time.

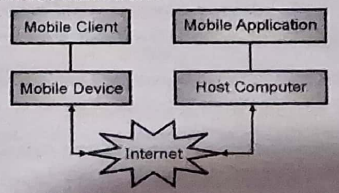
2. **Ordering and delivery confirmation :**
- In this application, dealers use mobile phones to order the products.
- The purchase orders can be sent to the supplier in a standard format.
- A dealer can automatically re-order the products and specify the quantity required through a simple application by scanning the bar code on a product.
- The bar code is scanned by using the camera of a mobile phone.
- The mobile phone keeps track of the consignment status during the transport and delivery process.
- A truck driver can confirm that a consignment has been delivered by reading the bar code on a packet using a mobile device.

3. **Mobile inventory management :**
- Mobile inventory management envisages a "rolling inventory" consisting of multiple trucks carrying large amounts of goods or products.
- If a store wants certain products then it locates the nearest truck and takes delivery of the required goods.
- Due to this, the cost and amount of inventory for both the producers and the retailers is reduced.
- It can considerably reduce the time required for delivery and can make just-in-time delivery of goods.

4. **Supply Chain Management (SCM) :**
- By using mobile device, information about the supply chain processes becomes available.
- A manager or anyone in the supply chain can check information about a product's state in the supply chain by scanning an RFID tag.
- To manage the business efficiently, this type of accurate information is helpful.

6.10 Structure of Mobile Commerce :

- A content provider in mobile commerce implements an application by providing two sets of programs namely client-side and server-side.
- The client side programs run on the micro browsers. Micro-browsers are installed on the mobile devices of user.
- The server-side programs reside on the host computer (servers). The server-side programs performs database access and computations.
- Fig. 6.10.1 shows the architecture of a mobile commerce framework.



(G-3154) **Fig. 6.10.1 : Architecture of a mobile commerce framework**

- An architecture of a mobile commerce framework consists of :
  1. Mobile devices
  2. Mobile middleware
  3. Network
  4. Host computers

1. Mobile devices :

- Hand-held devices are basically acts as the user interface to the mobile users.
- The mobile users specify their requests by using the suitable interface programs. Interface programs are then transmitted to the mobile commerce application on the Internet.
- The mobile commerce application displays the results in suitable formats.

2. Mobile middleware :

- The mobile middleware maps the Internet content to mobile phones seamlessly and transparently. It can sport various operating systems, markup languages, micro browsers and protocols.
- It also handles encryption and decryption of communication in order to provide secure transactions.

3. Network :

- Mobile commerce becomes possible available because of wireless networks.
- The request sent by user is delivered to the closest wireless access point in a wireless local area network environment or to a base station in a cellular network environment.
- In a mobile commerce system wired networks are optional.
- But the host computers (servers) are connected to wired networks such as the Internet.
- The user requests are routed to host computers by using transport and/or security mechanisms provided by wired networks.

4. Host computers :

- Host computers are basically servers. Host computers process and store all the information required for mobile commerce applications.
- Most application programs are hosted on host computers.
- These applications generally consist of three components :
  1. Web servers
  2. Database servers
  3. Application programs and support software
- The web servers helps in interacting with the mobile client and the database servers store information or data.
- The application program and support software is the middleware which implements the business logic of the mobile commerce application.

6.10.1 Features Required for Mobile Device to Enable Mobile Commerce :

- A mobile device should provide the following facilities to enable mobile commerce to be used widely :
  1. It must provide good internet connectivity.
  2. It must have ability to read the RFID tags.
  3. It should support services like MMS (Multimedia Message Service), SMS (Short Message Service).
  4. It should have ability to display rich content like images.
  5. It must have a good camera quality with auto focus.
  6. It should to communicate between the mobile device and the supporting network.
  7. It must scan bar codes.
  8. Mobile screen must be able to display the bar codes properly.
  9. It must interact with the Point-of-Sale (PoS) terminals. Point-of-Sale (PoS) means a checkout counter in a shop or supermarket.

6.11 Pros and Cons of M-Commerce :

Pros / Advantages of M-Commerce :

- The major advantages of using M-commerce are as follows :
  1. The advantages of M-commerce for the business organization include customer convenience, cost savings and new business opportunities.
  2. M-commerce for the customer's provides the flexibility of shopping at anytime, anywhere using just a lightweight device. The substantial time of customer is saved as compared to visiting several stores for identifying the right product at the lowest price.
  3. Highly personalized mobile devices are convenient to the customers. E.g. A repeat order for some items can be placed at the touch of a button.

Cons / Disadvantages of M-Commerce :

- The disadvantages of using M-commerce are as follows :
  1. Mobile devices do not provide graphics or processing power of a PC. Hence the users have to use small screen and keyboard and low resolution pictures and videos.
  2. Mobile devices with the small screens limit the complexity of applications. E.g. the menu choice and text typing ability are strictly restricted.
  3. Several types of limitations are imposed due to the underlying network. E.g. the available bandwidth is strictly restricted and international reach is prohibitively costly. Hence, ubiquity of M-commerce is difficult to achieve in practice.

6.12 Mobile Payment Systems :

Definition :

- A mobile payment (or M-payment) is a payment tool where a mobile device is used to initiate, authorize and confirm an exchange of financial value in return for products and services.
- Mobile payments are a natural development of E-payment schemes.
- Mobile devices in mobile payment system can be mobile phones, PDAs and any other device that connects to a mobile network for making payments.
- A mobile device is also used for online bill payment with access to account-based payment instruments like electronic funds transfer, Internet banking payments, direct debit and electronic bill presentment.
- The infrastructure required on the customer side is the main problem, which affects the establishment of the mobile payment procedure.
- If the customer is unable to handle the technology with ease, then a sophisticated technology can fail.
- The simple procedures based on simple message exchange through short messaging services (SMS) become more successful.
- Therefore, the important payment solutions at present and in the near future will be SMS-based.
- SMS-based payment solutions can easily be charged to the mobile phone bill of customers.
- Some other procedures can combine two or more solutions.
- The problems with M-payment schemes are security, privacy and guarding against frauds.
- There are many challenges for providing secure transactions ranges from physical theft of a mobile device which can be later used for fake payments.

6.12.1 Mobile Payment Schemes :

- Most popular types of M-payment schemes are as follows :
  1. Bank account based M-payment
  2. Credit card based M-payment
  3. Micropayment
- In all these types of schemes, a third party service provider such as bank, credit card company, or telecom company makes a payment on the behalf of customer.

1. Bank account based M-payment :

- The bank account of the customer in this scheme is linked to his / her mobile phone number.
- When the customer makes an M-payment transaction with a vendor, merchant or in a shopping complex, then the amount is debited from the customer's bank account and the value is credited to the vendor's account.
- **mChek** is a new payment scheme and it links a debit or credit card, or a bank account, to a mobile phone. It allows customer to make payments from the mobile phone.
- After registration, the user can pay phone bills, transfer talk time to a friend's account, book tickets for flights, movies, pay water bills, electric bills, etc. from the mobile phone.
- mChek has a tie-up with Airtel. It allows Airtel subscribers to download mChek application which provides simple graphic interface to use mChek.
- Airtel provides downloading as well as using mChek with free of cost.

2. Credit card based M-payment :

- In this scheme, the credit card number of customer is linked to the his/ her mobile number.
- When the customer makes an M-payment transaction with a vendor, merchant or in a shopping complex then the credit card is charged and the value is credited to the merchant's account.
- There is limitation on the Credit card based solutions being heavily dependent on the level of penetration of credit cards in a country.
- At present, the penetration level of credit cards is low but is expected to grow significantly in the coming years.

3. Micropayment :

- Micropayment scheme is proposed for small purchase payment such as payment from vending machines.
- The mobile device and vending machine directly communicate with each other by using a Bluetooth or wireless LAN connection in order to negotiate the payment and then the micropayment is carried out.
- A customer calls the number of a service provider where the per call charge and the cost of the vending item is equal.
- The micropayment is implemented through the cooperation of the mobile phone operator and a third party service provider.
- This micropayment scheme has been used for vending from Coca-Cola machines.

6.12.2 Security Issues :

- M-commerce is predicted to introduce new security and privacy risks beyond those currently found in E-commerce systems.
- It is difficult to trace the user of mobile devices because of roaming of the users.
- The mobile devices can frequently go on-line and off-line. Hence, security attacks are very difficult to trace.
- Another risk about the mobile devices is the risk of loss or theft.
- A stolen mobile device or device fallen into wrong hands can cause frauds which are difficult to track and prevent.
- A major problem with this is the lack of any satisfactory mechanism to authenticate a particular user.

Other security issues :

- Following are the security issues in mobile payment systems :
  1. Phishing scams
  2. Weak passwords
  3. Human error
  4. Lost or stolen devices
  5. Using Public Wifi

1. Phishing scams :

- Phishing scams leads to cyber attack. If you receive a suspected phishing text message then delete it immediately and do not click on any links.

2. Weak Passwords :

- Mobile device can be hacked due to weak passwords, or overused passwords.
- Don't use the same password for everything, and try and change them once a month.

3. Human Error :

- Human error or carelessness causes the security issues.
- When hackers planning for cyber attack, they rely on human error.
- Clicking on insecure links, opening emails containing security threats and accidentally download malware causes security issue to mobile device.
- Be vigilant with your passwords and about storing them.

4. Lost or Stolen Device :

- Lost or Stolen Device can cause risk. Now a day's Smartphone vendors continue to introduce protection technology that can prevent a hacker or thief from accessing your mobile wallet.
- They provide two-factor authentication to unlock your device which reduces the risk when device is stolen or lost.

5. Using Public Wifi :

- Attckers can hack your device by using public wifi.
- They create fake connections and then they intercept any data in transit, such as a bank transfer or online payment.
- One of the most secure form of protection against hacking is the use of a VPN, or Virtual Private Network.

6.12.3 Advantages of Mobile Payments :

- Following are the advantages of using mobile payments :
  1. Contactless transactions
  2. Instant Reponses
  3. No additional cost involved
  4. Safe to use
  5. 24×7 availability
  6. Very convenient
  7. Information Security

6.12.4 Disadvantages of Mobile Payments :

- Following are the disadvantages of using mobile payments :
  1. There are some security issues linked with mobile payments. All the transactions work with the servers, client, and wireless network. Where there is security glitch in either of one side it can harm the whole system.
  2. The users are dependent on the servers as well as the network service providers.
  3. Sometimes it becomes difficult for the users to adapt the application features and make a secured payment.
  4. The users should be educated or trained enough to make the payment without any issue.
  5. There are some features which are risky and less compatible to the current payment system.

6. The chip and mobile manufacturers coming up with the new functionality which forcers the clients and the network service providers to update the system and make it more secure.

Review Questions

Q. 1 Write a short note on : Operating Systems for Mobile Computing.

Q. 2 State and explain the responsibilities of OSs in Mobile Devices.

Q. 3 List and explain the layers of Operating System.

Q. 4 State the advantage of Microkernel design in OS.

Q. 5 Compare the Monolithic Kernel and Microkernel.

Q. 6 What are special Constraints of Mobile O/S.

Q. 7 What are special service requirements of mobile O/S ?

Q. 8 Enlist the important features of the Windows mobile OS.

Q. 9 Write a short note on : Palm OS.

Q. 10 Enlist the important features of the Symbian OS.

Q. 11 Write a short note on : iOS.

Q. 12 Explain Android OS.

Q. 13 With neat diagram explain the Android architecture.

Q. 14 Write a short note on : Blackberry OS.

Q. 15 What are the features of BlackBerry OS ?

Q. 16 What are the advantages and disadvantage BlackBerry OS ?

Q. 17 State and explain the types of mobile pay schemes.

Q. 18 Give comparison between different mob Operating Systems.

Q. 19 Define SDK. Explain iOS SDK.

Q. 20 Write a short note on : Android SDK.

Q. 21 Write a short note on : Windows Phone SDK.

Q. 22 Define and explain the concept of M-Commerce.

Q. 23 What is Business-to-consumer (B2C) ?

Q. 24 List the Examples of B2C applications.

Q. 25 List the Examples of B2B applications.

Q. 26 With neat diagram explain the structure of Mobile Commerce.

Q. 27 State the features required for mobile device to enable mobile commerce.

Q. 28 List the pros and cons of M-commerce.

Q. 29 Explain the security issues in mobile payment system.

Q. 30 State advantages and disadvantages of mobile payments.