GURUNANAK INSTITUTE OF TECHNOLOGY

157/F, Nilgunj Road, Panihati
Kolkata -700114
Website: www.gnit.ac.in
Email: info.gnit@jisgroup.org

Approved by A.I.C.T.E., New Delhi

Affiliated to MAKAUT, West Bengal

Online Course Ware (OCW)

Course: Mobile Computing

Course Level: Undergraduate

Credit: 3

Prepared by:

Dr. Suman Bhattacharya (CSE)

CS801A: Mobile Computing
Contact Hours: 3L
Credits: 3

Lecture: 35L

MODULE 1: Introduction [6L]

Evolution of different types of wireless communication devices; Effects of mobility of devices; Cellular mobile networks – mobility management (call setup, handoff, interoperability and internetworking), bandwidth management, energy management, security; Brief introduction about different generations of wireless communication technology – 1G, 2G, 3G, 4G, 5G.

1.1 Evolution of different types of wireless communication devices

Wireless communication plays a significant role in day to day life. Besides communication, wireless technology has become an integral part of our daily activities. The transmission of data or information from one place to another wirelessly is referred as wireless communication. This provides an exchange of data without any conductor through RF and radio signals. The information is transmitted across the devices over some meters to hundreds of kilometres through well-defined channels.

Fig.1.1 : Wireless Communication devices

Different types of signals are used in communication between the devices for wireless transmission of data. The following are the different electromagnetic signals are used depending on their wavelength and frequency.

- Radio Frequency Transmission
- Infrared Transmission
- Microwave Transmission
- Lightwave Transmission

Radio Frequency Transmission

Radio frequency is a form of electromagnetic transmission used in wireless communication. RF signals are easily generated, ranging 3kHz to 300GHz. These are used in wireless communication because of their property to penetrate through objects and travel long distances.

Radio communication depends on the wavelength, transmitter power, receiver quality, type, size and height of the antenna.

Drawbacks

- These are frequency dependent
- These have the relatively low bandwidth for data transmission.

Infrared Transmission

Infrared radiations are electromagnetic radiations with longer wavelengths than visible light. These are usually used for short-range communications. These signals do not pass through solid objects.

Examples like Television remote control, mobile data sharing.

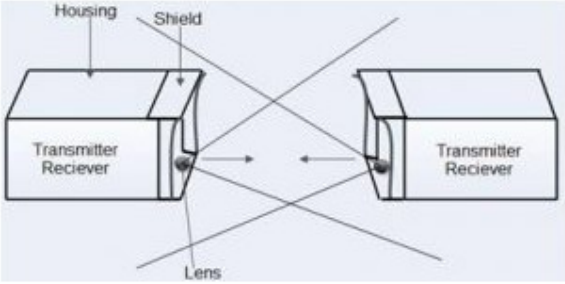


Fig. 1.2: Infrared Transmission

Microwave Transmission

Microwaves are the form of electromagnetic transmission used in wireless communication systems. The wavelength of microwave ranges from one meter to one millimetre. The frequency varies from 300MHz to 300GHz. These are widely used for long distance communications and are relatively less expensive.

Fig. 1.3: Microwave Transmission Node

Drawbacks

- The microwave does not pass through buildings.
- Bad weather affects the signal transmission.
- These are frequency dependent.

Lightwave Transmission

Light is an electromagnetic radiation with a wavelength ranging between infrared radiations and ultraviolet radiations. The wavelength ranges from 430 to 750THz. These are unguided optical signals such as laser and are unidirectional.

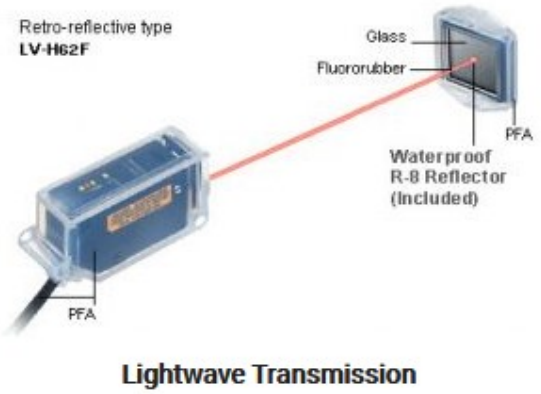


Fig. 1.4: Lightwave transmission

Drawbacks

- These signals cannot penetrate through rain and fog.
- The laser beam gets easily diverted by air.

Types of Wireless Communication Technologies

Wireless communication technology is categorized into different types depending on the distance of communication, the range of data and type of devices used. The following are the different types of wireless communication technologies.

- Radio and Television Broadcasting
- Radar Communication
- Satellite communication
- Cellular Communication
- Global Positioning System
- WiFi
- Bluetooth
- Radio Frequency Identification

Radio

Radio communication was one of the first wireless technology developed and it is still in use. The portable multi-channel radios allow the user to communicate over short distances whereas

citizen band and maritime radios provide communication services over long distances for truckers and sailors.

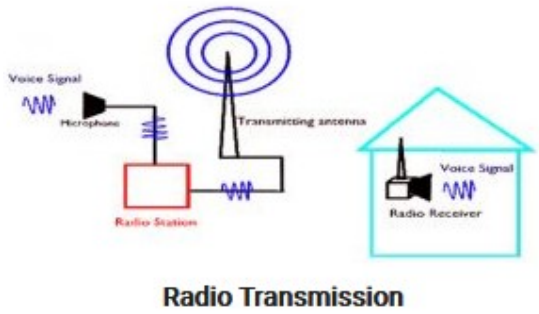


Fig. 1.5: Radio transmission

Mostly radio broadcasts sound through the air as radio waves. Radio has a transmitter which transmits the data in the form of radio signals to the receiver antenna.

To broadcast common programming stations are associated with the radio networks. The broadcast happens either in simulcast or syndication or both the forms. Radio broadcasting may be done via cable FM, and satellites over long distances at up to two megabits/Sec.

Cellular

A cellular network uses encrypted radio links, modulated to allow many users to communicate across the single frequency band. As the individual handsets lack significant broadcasting power, the system depends on a network of cellular towers which are capable of triangulating the source of any signal and handing reception duties off to the most suitable antenna.

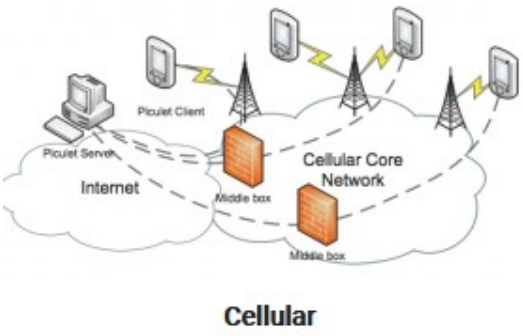


Fig. 1.6: Cellular communication

The data transmission over cellular networks is possible with modern 4G systems capable of speeds reaching that of wired DSL. Cellular companies charge their customers by a minute of their voice and by the kilobytes for data.

Satellite

Satellite communication is a wireless technology having significant importance across the globe. They have found widespread use in specialized situations.

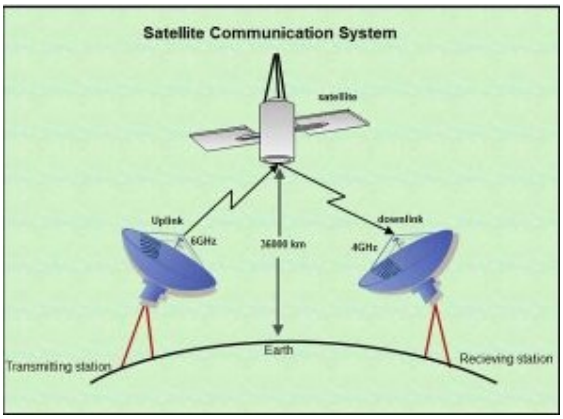


Fig. 1.7: Satellite communication

The devices using satellite technology to communicate directly with the orbiting satellite through radio signals.

This allows users to stay connected virtually from anywhere on the earth. Portable satellite phones and modems have powerful broadcast feature and reception hardware than the cellular devices due to the increased range.

The satellite communication consists of a space segment and a ground segment. When the signal is sent to the satellite through a device, the satellite amplifies the signal and sent it back to the receiver antenna which is located on the earth's surface. The ground segment consists of a transmitter, receiver and the space segment, which is the satellite itself.

Wi-Fi

Wi-Fi is a low-cost wireless communication technology. A WiFi setup consists of a wireless router which serves a communication hub, linking portable device with an internet connection.This network facilitates connection of many devices depending on the router configuration. These networks are limited in range due to the low power transmission, allowing the user to connect only in the close proximity.

Fig. 1.8 : Wifi communication

This network facilitates connection of many devices depending on the router configuration. These networks are limited in range due to the low power transmission, allowing the user to connect only in the close proximity.

Advantages

- Information can be transmitted quickly with a high speed and accuracy.
- The internet can be accessed from anywhere, at any time without any cables or wires.
- Emergency situations can be alerted through wireless communication.
- Wireless, no bunches of wire running out.
- Communication can reach where wiring is not feasible and costly.

Disadvantages

- An Unauthorized person can easily misuse the wireless signals which spread through the air.
- It is very important to secure the wireless network to protect information.
- High cost to set up the infrastructure.
- Wireless communication is influenced by physical constructions, climatic conditions and interference from other wireless devices.

Applications Wireless Communication

Wireless communication has wide applications.

- Space
- Military
- Telecommunications
- Wireless Power Transmission
- IoT
- Radar communication
- Artificial intelligence
- Fiber optics
- Intelligent Transport Systems

Therefore, this is all about Types of wireless communication and applications, these networks are one of the important technologies in the telecommunications market. WiFi, WiMax, Bluetooth, Femtocell, 3G and 4G are some of the most important standards of Wireless technology.

1.2: Cellular mobile networks

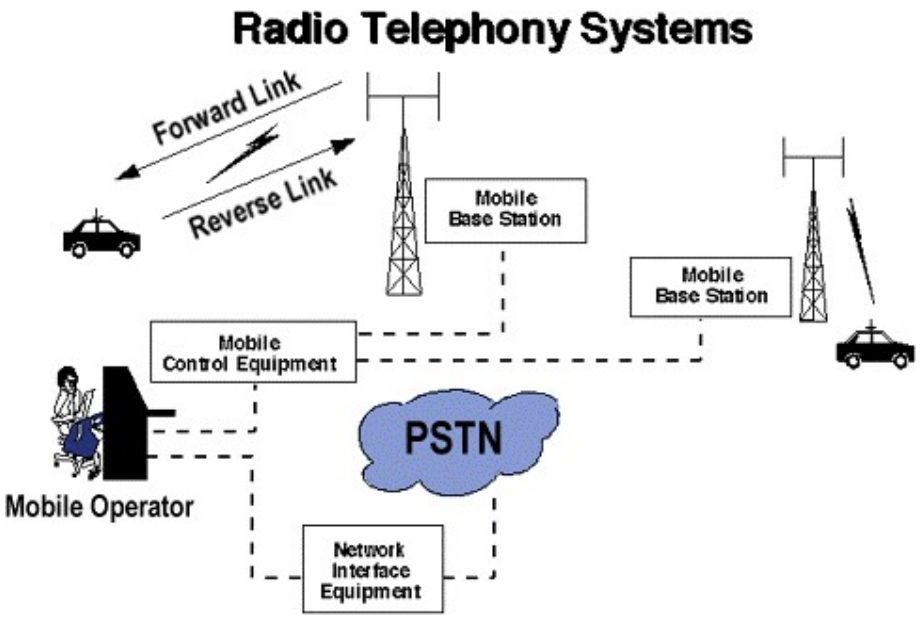
A cellular mobile communications system uses a large number of low-power wireless transmitters to create cells—the basic geographic service area of a wireless communications system. Variable power levels allow cells to be sized according to the subscriber density and demand within a particular region. As mobile users travel from cell to cell, their conversations are "handed off" between cells in order to maintain seamless service. Channels (frequencies) used in one cell can be reused in another cell some distance away. Cells can be added to accommodate growth, creating new cells in unserved areas or overlaying cells in existing areas.

1.2.1 Mobile Communications Principles

Each mobile uses a separate, temporary radio channel to talk to the cell site. The cell site talks to many mobiles at once, using one channel per mobile. Channels use a pair of frequencies for communication—one frequency, the forward link, for transmitting from the cell site, and one frequency, the reverse link, for the cell site to receive calls from the users. Radio energy dissipates over distance, so mobiles must stay near the base station to maintain communications. The basic structure of mobile networks include telephone systems and

radio services.
Where mobile radio service operates in a closed network and has no access to the telephone system, mobile telephone service allows interconnection to the telephone network (see Figure 1.9).

Figure 1.9: Basic Mobile Telephone Service Network



Early Mobile Telephone System Architecture

Traditional mobile service was structured similar to television broadcasting: One very powerful transmitter located at the highest spot in an area would broadcast in a radius of up to fifty kilometers. The cellular concept" structured the mobile telephone network in a different way. Instead of using one powerful transmitter, many low-power transmitters were placed throughout a coverage area. For example, by dividing a metropolitan region into one hundred different areas (cells) with low-power transmitters using twelve conversations (channels) each, the system capacity theoretically could be increased from twelve conversations— or voice channels using one powerful transmitter—to twelve hundred conversations (channels) using one hundred low-power transmitters. Figure 2 shows a metropolitan area configured as a traditional mobile telephone network with one high-power transmitter.

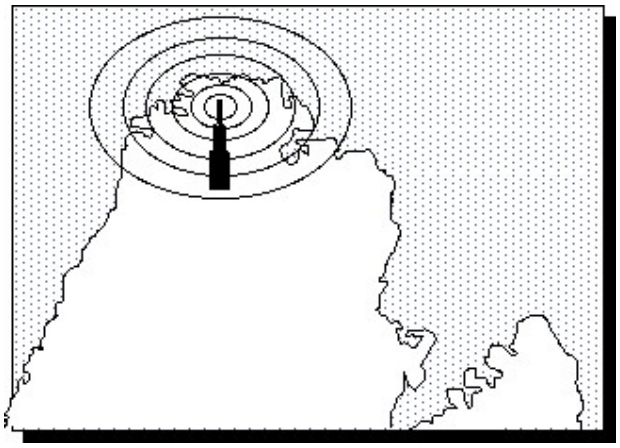


Figure 1.10: Early Mobile Telephone System Architecture

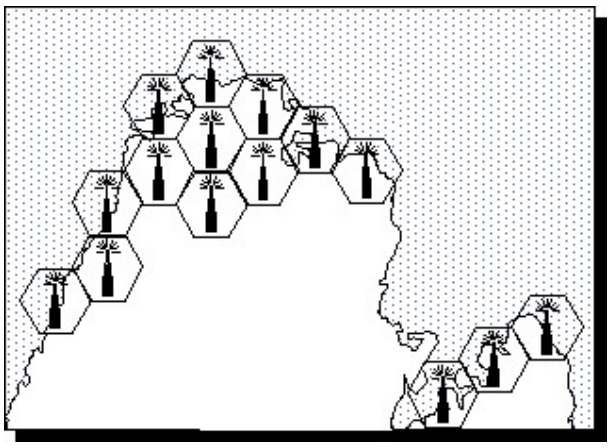
1.2.2 Mobile Telephone System Using the Cellular Concept

Interference problems caused by mobile units using the same channel in adjacent areas proved that all channels could not be reused in every cell. Areas had to be skipped before the same channel could be reused. Even though this affected the efficiency of the original concept, frequency reuse was still a viable solution to the problems of mobile telephony systems.

Engineers discovered that the interference effects were not due to the distance between areas, but to the ratio of the distance between areas to the transmitter power (radius) of the areas. By reducing the radius of an area by fifty percent, service providers could increase the number of potential customers in an area fourfold. Systems based on areas with a one-kilometer radius would have one hundred times more channels than systems with areas ten kilometers in radius. Speculation led to the conclusion that by reducing the radius of areas to a few hundred meters, millions of calls could be served.

The cellular concept employs variable low-power levels, which allows cells to be sized according to the subscriber density and demand of a given area. As the population grows, cells can be added to accommodate that growth. Frequencies used in one cell cluster can be reused in other cells. Conversations can be handed off from cell to cell to maintain constant phone service as the user moves between cells.

Figure 1.11: Mobile Telephone System Using a Cellular Architecture



The cellular radio equipment (base station) can communicate with mobiles as long as they are within range. Radio energy dissipates over distance, so the mobiles must be within the operating range of the base station. Like the early mobile radio system, the base station communicates with mobiles via a channel. The channel is made of two frequencies, one for transmitting to the base station and one to receive information from the base station.

1.2.3 Cellular System Architecture

Increases in demand and the poor quality of existing service led mobile service providers to research ways to improve the quality of service and to support more users in their systems. Because the amount of frequency spectrum available for mobile cellular use was limited, efficient use of the required frequencies was needed for mobile cellular coverage. In modern cellular telephony, rural and urban regions are divided into areas according to specific provisioning guidelines. Deployment parameters, such as amount of cell-splitting and cell sizes, are determined by engineers experienced in cellular system architecture.

Provisioning for each region is planned according to an engineering plan that includes cells, clusters, frequency reuse, and handovers.

Cells

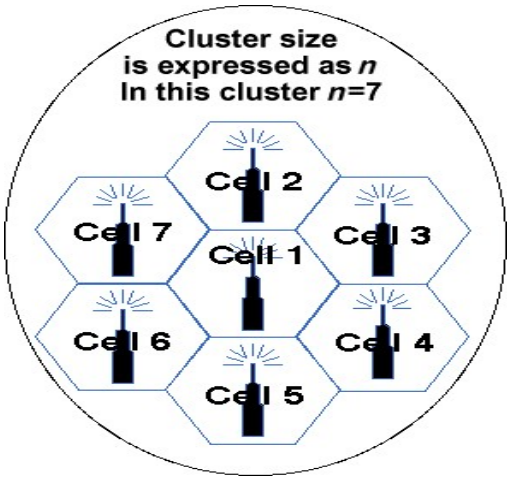
A cell is the basic geographic unit of a cellular system.

The term *cellular* comes from the honeycomb shape of the areas into which a coverage region is divided. Cells are base stations transmitting over small geographic areas that are represented as hexagons. Each cell size varies depending on the landscape. Because of constraints imposed by natural terrain and man-made structures, the true shape of cells is not a perfect hexagon.

Clusters

A cluster is a group of cells. No channels are reused within a cluster. Figure 4 illustrates a seven-cell cluster.

Figure 1.12: A Seven-Cell Cluster



Frequency Reuse

Because only a small number of radio channel frequencies were available for mobile systems, engineers had to find a way to reuse radio channels in order to carry more than one conversation at a time. The solution the industry adopted was called frequency planning or frequency reuse. Frequency reuse was implemented by restructuring the mobile telephone system architecture into the cellular concept.

The concept of frequency reuse is based on assigning to each cell a group of radio channels used within a small geographic area. Cells are assigned a group of channels that is completely different from neighboring cells. The coverage area of cells are called the footprint. This footprint is limited by a boundary so that the same group of channels can be used in different cells that are far enough away from each other so that their frequencies do not interfere.

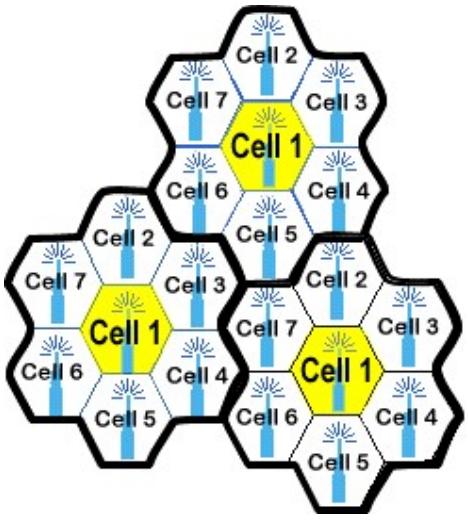


Figure 1.13: Frequency Reuse

Cells with the same number have the same set of frequencies. Here, because the number of available frequencies is 7, the frequency reuse factor is 1/7. That is, each cell is using 1/7 of available cellular channels.

Cell Splitting

Unfortunately, economic considerations made the concept of creating full systems with many small areas impractical. To overcome this difficulty, system operators developed the idea of cell splitting. As a service area becomes full of users, this approach is used to split a single area into smaller ones. In this way, urban centers can be split into as many areas as necessary in order to provide acceptable service levels in heavy-traffic regions, while larger, less expensive cells can be used to cover remote rural regions.

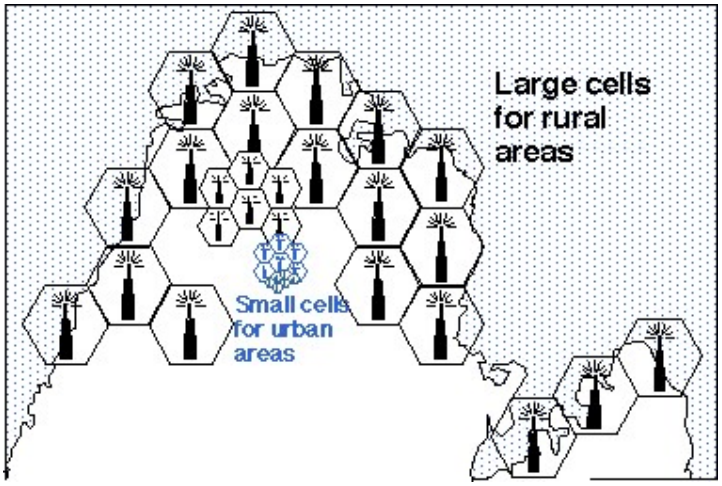
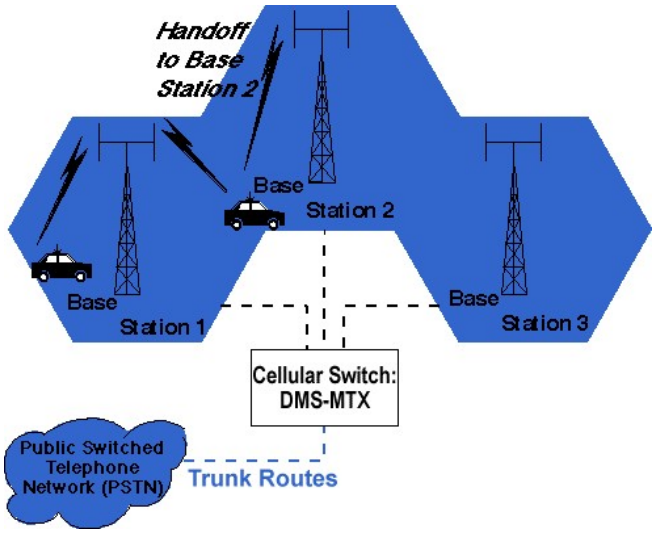


Figure 1.14: Cell Splitting

Handoff

The final obstacle in the development of the cellular network involved the problem created when a mobile subscriber traveled from one cell to another during a call. As adjacent areas do not use the same radio channels, a call must either be dropped or transferred from one radio channel to another when a user crosses the line between adjacent cells. Because dropping the call is unacceptable, the process of handoff was created. Handoff occurs when the mobile telephone network automatically transfers a call from radio channel to radio channel as a mobile crosses adjacent cells.

Figure 1.15: Handoff between Adjacent Cells



During a call, two parties are on one voice channel. When the mobile unit moves out of the coverage area of a given cell site, the reception becomes weak. At this point, the cell site in use requests a handoff. The system switches the call to a stronger-frequency channel in a new site without interrupting the call or alerting the user. The call continues as long as the user is talking, and the user does not notice the handoff at all.

1.3 Brief introduction about different generations of wireless communication technology – 1G, 2G, 3G, 4G, 5G

Classical 0G phones stood for the first generation of mobile phones i.e. Satellite phones developed for boats mainly. Networks such as Iridium, Global Star and Eutelsat were truly worldwide (although for physical reasons, think of a satellite as a fixed point above the equator, some Northern parts of Scandinavia aren't reachable), and everybody thought at that time that satellite phones would become mainstream products as soon as devices got smaller and cheaper. This vision proved wrong when the GSM concretely came to life in 1990-91 in Finland.
1G: Firstly, there were analog GSM systems that existed for a few years. And then discover the digital systems.

1G refers to the first generation of wireless mobile communication where analog signals were used to transmit data. It was introduced in the US in early 1980s and designed exclusively for voice communication. Some characteristics of 1G communication are −

- Speeds up to 2.4 kbps
- Poor voice quality
- Large phones with limited battery life
- No data security

2G: the second generation of mobile telecommunications still is the most widespread technology in the world; you have basically all heard of the GSM norm (GSM stands for Groupe Special Mobile in French, renamed in Global System for Mobility). The GSM operates in the 850Mhz. and 1900Mhz. bands in the US, & 900 Mhz. and 1.8 Mhz bands in the rest of the world (eg did you know Bluetooth stands in the 2.4Ghz. area, just like your…microwave!? But that is another story, not related to this article) and delivers data at the slow rate of 9.6 Kbytes/sec.

2G refers to the second generation of mobile telephony which used digital signals for the first time. It was launched in Finland in 1991 and used GSM technology. Some prominent characteristics of 2G communication are −

- Data speeds up to 64 kbps
- Text and multimedia messaging possible
- Better quality than 1G

When GPRS technology was introduced, it enabled web browsing, e-mail services and fast upload/download speeds. 2G with GPRS is also referred as 2.5G, a step short of next mobile generation.

2.5G: For that last reason (9.6 Kbytes/sec doesn't allow you to browse the Net or up/download an image), telecom operators came up with the GPRS (remember all the hype around the Wap) which could enable much faster communications (115Kbytes.sec). But the market decided it was still not enough compared to what they had at home.
2.75G: EDGE, which is a pretty recent standard, allows for downloading faster. Since mobile devices have become both a TV and music player, people needed to be able to watch streaming video and download mp3 files faster – that is precisely what EDGE allows for and that is for the good news. The bad news is that if EDGE rocks at downloading, it is protocol is asymmetrical hence making EDGE suck at uploading ie, broadcasting videos of yours for instance. Still an interesting achievement thanks to which data packets can effectively reach 180kbytes/sec. EDGE is now widely being used.
3G: also called UMTS (Universal Mobile Telecommunications Standard). Aimed at enabling long expected videoconferencing, although nobody seems to actually use it (do you know any?).

Its other name is 3GSM, which says literally that UMTS is 3 times better than GSM. One issue though: depending on the deployment level of the area you are in and your device, your phone will (have to be) handle(d) from the GSM network to the UMTS network, and conversely – making billing more complex to understand for the consumers. One of the major positive points of UMTS is its global roaming capabilities (roaming is the process that allows you, at a cost, to borrow bandwidth from a telecom provider that is not yours; you usually use roaming when calling from abroad).

3G of mobile telephony began with the start of the new millennium and offered major advancement over previous generations. Some of the characteristics of this generation are −

- Data speeds of 144 kbps to 2 Mbps
- High speed web browsing
- Running web based applications like video conferencing, multimedia e-mails, etc.
- Fast and easy transfer of audio and video files
- 3D gaming

Every coin has two sides. Here are some downsides of 3G technology −

- Expensive mobile phones
- High infrastructure costs like licensing fees and mobile towers
- Trained personnel required for infrastructure set up

The intermediate generation, 3.5G grouped together dissimilar mobile telephony and data technologies and paved way for the next generation of mobile communication.

3.5G or 3G+: HSDPA is theoretically 6 times faster than UMTS (up to 3.6 Mbytes/sec)! Practically speaking, this would mean downloading an mp3 file would take about 30 sec instead of something like 2 minutes.

4G: still a research lab standard, at least to my knowledge, that should combine the best of cell phone network technologies with WiMax wireless Internet, voice over IP and IPv6 (a post about the latter soon). Data rates are expected to reach 100 Mbytes/sec[4].

Keeping up the trend of a new mobile generation every decade, fourth generation (4G) of mobile communication was introduced in 2011. Its major characteristics are −

- Speeds of 100 Mbps to 1 Gbps
- Mobile web access
- High definition mobile TV
- Cloud computing
- IP telephony

5G : 5th generation mobile networks or 5th generation wireless systems is a name used in some research papers and projects to denote the next major phase of mobile telecommunication standards beyond the upcoming 4G standards (which is expected to be finalized between approximately 2011 and 2013)[5].

5G offers Peak per terminal throughput – 10 Gbps outdoors, spectral efficiency. – 5 bps/Hz/cell, areal reliability – 99.5%, round trip delay < 1 ms, seamless coexistence with other radios. These goals are significantly ahead of 4G performance. The new tools that can take us these goals may include Opportunistic OFDMA, 20-60 MHz channel bandwidth, cognitive and opportunistic channel structure, flexible ,variable reuse, cooperative methods, interference management, client relay, hierarchical modulation, distributed MIMO and accumulative methods.

COMPARISON OF 1G TO 5G TECHNOLOGIES

Table1: General comparison of 1G to 5G technologies.

Technology/Features	1G	2G/2.5G	3G	4G	5G
Start/ Development	1970/ 1984	1980/ 1999	1990/ 2002	2000/ 2010	2010/ 2015
Data Bandwidth	2 kbps	14.4-64kpbs	2 Mbps	2000 Mbps to1 Gbps for low mobility	1 Gbps and higher
Standards	AMPS	2G:TDMA, CDMS, GSM 2.5:GPRS, EDGE, 1xRTT	WCDMA, CDMA-2000	Single unified standard	Single Unified standard
Technology	Analog Cellular technology	Digital cellular technology	Broad bandwidth CDMA, IP technology	Unified IP and seamless combination of broadband, LAN/WAN/PAN and WLAN	Unified IP and Seamless combination of broadband, LAN/WAN/PAN /WLAN and wwww
Service	Mobile Telephony (voice)	2G: Digital voice, Short Messaging 2.5G: Higher capacity Packetized data	Integrated Higher Quality audio, video and data	Dynamic Information Access, Wearable devices	Dynamic Information Access, wearable device with IA capabilities
Multiplexing	FDMA	TDMA, CDMA	CDMA	CDMA	CDMA
Switching	Circuit	2G: Circuit 2.5G: Circuit for access network & air interface; packet for core network and data	Packet except circuit for air interface	All packet	All packet
Core Network	PSTN	PSTN	Packet network	Internet	Internet
Handoff	Horizontal	Horizontal	Horizontal	Horizontal and vertical	Horizontal and vertical

Module II: Mobile Data Communication [5L]

Mobile Data Communication, WLANs (Wireless LANs) IEEE 802.11 standard, Bluetooth technology, Bluetooth Protocols, Ad hoc networks initialization, leader election, location identification, communication protocols, energy and security.

2. Mobile Data Communication and Wireless LAN

This module presents several wireless local area network (WLAN) technologies. This constitutes a fast-growing market introducing the flexibility of wireless access into office, home, or production environments. WLANs are typically restricted in their diameter to buildings, a campus, single rooms etc. and are operated by individuals, not by large-scale network providers. The global goal of WLANs is to replace office cabling, to enable tetherless access to the internet and, to introduce a higher flexibility for ad-hoc communication in, e.g., group meetings. The following points illustrate some general advantages and disadvantages of WLANs compared to their wired counterparts.

Some advantages of WLANs are:

● Flexibility: Within radio coverage, nodes can communicate without further restriction. Radio waves can penetrate walls, senders and receivers can be placed anywhere (also non-visible, e.g., within devices, in walls etc.). Sometimes wiring is difficult if firewalls separate buildings (real

firewalls made out of, e.g., bricks, not routers set up as a firewall). Penetration of a firewall is only permitted at certain points to prevent fire from spreading too fast.

● Planning: Only wireless ad-hoc networks allow for communication without previous planning, any wired network needs wiring plans. As long as devices follow the same standard, they can communicate. For wired networks, additional cabling with the right plugs and probably interworking units (such as switches) have to be provided.

● Design: Wireless networks allow for the design of small, independent devices which can for example be put into a pocket. Cables not only restrict users but also designers of small PDAs, notepads etc. Wireless senders and receivers can be hidden in historic buildings, i.e., current networking technology can be introduced without being visible.

● Robustness: Wireless networks can survive disasters, e.g., earthquakes or users pulling a plug. If the wireless devices survive, people can still communicate. Networks requiring a wired infrastructure will usually break down completely.

● Cost: After providing wireless access to the infrastructure via an access point for the first user, adding additional users to a wireless network will not increase the cost. This is, important for e.g., lecture halls, hotel lobbies or gate areas in airports where the numbers using the network may vary significantly. Using a fixed network, each seat in a lecture hall should have a plug for the network although many of them might not be used permanently. Constant plugging and unplugging will sooner or later destroy the plugs. Wireless connections do not wear out.

But WLANs also have several disadvantages:

● Quality of service: WLANs typically offer lower quality than their wired counterparts. The main reasons for this are the lower bandwidth due to limitations in radio transmission (e.g., only 1–10 Mbit/s user data rate instead of 100–1,000 Mbit/s), higher error rates due to interference (e.g., 10–4 instead of 10–12 for fiber optics), and higher delay/delay variation due to extensive error correction and detection mechanisms.

● Proprietary solutions: Due to slow standardization procedures, many companies have come up with proprietary solutions offering standardized functionality plus many enhanced features (typically a higher bit rate using a patented coding technology or special inter-access point protocols). However, these additional features only work in a homogeneous environment, i.e., when adapters from the same vendors are used for all wireless nodes. At least most components today adhere to the basic standards IEEE 802.11b or (newer) 802.11a.

● Restrictions: All wireless products have to comply with national regulations. Several government and non-government institutions worldwide regulate the operation and restrict frequencies to minimize interference. Consequently, it takes a very long time to establish global solutions like, e.g., IMT-2000, which comprises many individual standards. WLANs are limited to low-power senders and certain license-free frequency bands, which are not the same worldwide.

● Safety and security: Using radio waves for data transmission might interfere with other high-tech equipment in, e.g., hospitals. Senders and receivers are operated by laymen and, radiation has to be low. Special precautions have to be taken to prevent safety hazards. The open radio interface makes eavesdropping much easier in WLANs than, e.g., in the case of fiber optics. All standards must offer (automatic) encryption, privacy mechanisms, support for anonymity etc.

Otherwise more and more wireless networks will be hacked into as is the case already (aka war driving: driving around looking for unsecured wireless networks; WarDriving, 2002).

Many different, and sometimes competing, design goals have to be taken into account for WLANs to ensure their commercial success:

● Global operation: WLAN products should sell in all countries so, national and international frequency regulations have to be considered. In contrast to the infrastructure of wireless WANs, LAN equipment may be carried from one country into another – the operation should still be legal in this case.

● Low power: Devices communicating via a WLAN are typically also wireless devices running on battery power. The LAN design should take this into account and implement special power-saving modes and power management functions. Wireless communication with devices plugged into a power outlet is only useful in some cases (e.g., no additional cabling should be necessary for the network in historic buildings or at trade shows). However, the future clearly lies in small handheld devices without any restricting wire.

● License-free operation: LAN operators do not want to apply for a special license to be able to use the product. The equipment must operate in a license-free band, such as the 2.4 GHz ISM band.

● Robust transmission technology: Compared to their wired counterparts, WLANs operate under difficult conditions. If they use radio transmission, many other electrical devices can interfere with them (vacuum cleaners, hairdryers, train engines etc.). WLAN transceivers cannot be adjusted for perfect transmission in a standard office or production environment. Antennas are typically omnidirectional, not directed. Senders and receivers may move.

● Simplified spontaneous cooperation: To be useful in practice, WLANs should not require complicated setup routines but should operate spontaneously after power-up. These LANs would not be useful for supporting, e.g., ad-hoc meetings.

● Easy to use: In contrast to huge and complex wireless WANs, wireless LANs are made for simple use. They should not require complex management, but rather work on a plug-and-play basis.

● Protection of investment: A lot of money has already been invested into wired LANs. The new WLANs should protect this investment by being interoperable with the existing networks. This means that simple bridging between the different LANs should be enough to interoperate, i.e., the wireless LANs should support the same data types and services that standard LANs support.

● Safety and security: Wireless LANs should be safe to operate, especially regarding low radiation if used, e.g., in hospitals. Users cannot keep safety distances to antennas. The equipment has to be safe for pacemakers, too. Users should not be able to read personal data during transmission, i.e., encryption mechanisms should be integrated. The networks should also take into account user privacy, i.e., it should not be possible to collect roaming profiles for tracking persons if they do not agree.

● Transparency for applications: Existing applications should continue to run over WLANs, the only difference being higher delay and lower bandwidth. The fact of wireless access and

mobility should be hidden if it is not relevant, but the network should also support location aware applications, e.g., by providing location information.

2.1 Infra red vs radio transmission

Today, two different basic transmission technologies can be used to set up WLANs. One technology is based on the transmission of infra red light (e.g., at 900 nm wavelength), the other one, which is much more popular, uses radio transmission in the GHz range (e.g., 2.4 GHz in the license-free ISM band). Both technologies can be used to set up ad-hoc connections for work groups, to connect, e.g., a desktop with a printer without a wire, or to support mobility within a small area. Infra red technology uses diffuse light reflected at walls, furniture etc. or directed light if a line-of-sight (LOS) exists between sender and receiver. Senders can be simple light emitting diodes (LEDs) or laser diodes. Photodiodes act as receivers. Details about infra red technology, such as modulation, channel impairments etc. can be found in Wesel (1998) and Santamaría (1994).

● The main advantages of infra red technology are its simple and extremely cheap senders and receivers which are integrated into nearly all mobile devices available today. PDAs, laptops, notebooks, mobile phones etc. have an infra red data association (IrDA) interface. Version 1.0 of this industry standard implements data rates of up to 115 kbit/s, while IrDA 1.1 defines higher data rates of 1.152 and 4 Mbit/s. No licenses are needed for infra red technology and shielding is very simple. Electrical devices do not interfere with infra red transmission.

● Disadvantages of infra red transmission are its low bandwidth compared to other LAN technologies. Typically, IrDA devices are internally connected to a serial port limiting transfer rates to 115 kbit/s. Even 4 Mbit/s is not a particularly high data rate. However, their main disadvantage is that infra red is quite easily shielded. Infra red transmission cannot penetrate walls or other obstacles. Typically, for good transmission quality and high data rates a LOS, i.e., direct connection, is needed.

Almost all networks described in this book use radio waves for data transmission, e.g., GSM at 900, 1,800, and 1,900 MHz, DECT at 1,880 MHz etc.

● Advantages of radio transmission include the long-term experiences made with radio transmission for wide area networks (e.g., microwave links) and mobile cellular phones. Radio transmission can cover larger areas and can penetrate (thinner) walls, furniture, plants etc. Additional coverage is gained by reflection. Radio typically does not need a LOS if the frequencies are not too high. Furthermore, current radio-based products offer much higher transmission rates (e.g., 54 Mbit/s) than infra red (directed laser links, which offer data rates well above 100 Mbit/s. These are not considered here as it is very difficult to use them with mobile devices).

● Again, the main advantage is also a big disadvantage of radio transmission. Shielding is not so simple. Radio transmission can interfere with other senders, or electrical devices can destroy data transmitted via radio. Additionally, radio transmission is only permitted in certain frequency bands. Very limited ranges of license-free bands are available worldwide and those that are available are not the same in all countries. However, a lot of harmonization is going on due to market pressure.

Of the three WLAN technologies presented in this chapter, only one (IEEE 802.11) standardized infra red transmission in addition to radio transmission. The other two (HIPERLAN and

Bluetooth) rely on radio. The main reason for this are the shielding problems of infra red. WLANs should, e.g., cover a whole floor of a building and not just the one room where LOSs exist. Future mobile devices may have to communicate while still in a pocket or a suitcase so cannot rely on infra red. The big advantage of radio transmission in everyday use is indeed the ability to penetrate certain materials and that a LOS is not required. Many users experience a lot of difficulties adjusting infra red ports of, e.g., mobile phones to the infra red port of their PDA. Using, e.g., Bluetooth is much simpler.

2.2 Infrastructure and ad-hoc networks

Many WLANs of today need an infrastructure network. Infrastructure networks not only provide access to other networks, but also include forwarding functions, medium access control etc. In these infrastructure-based wireless networks, communication typically takes place only between the wireless nodes and the access point (Fig. 1), but not directly between the wireless nodes. The access point does not just control medium access, but also acts as a bridge to other wireless or wired networks. Figure 1 shows three access points with their three wireless networks and a wired network. Several wireless networks may form one logical wireless network, so the access points together with the fixed network in between can connect several wireless networks to form a larger network beyond actual radio coverage.

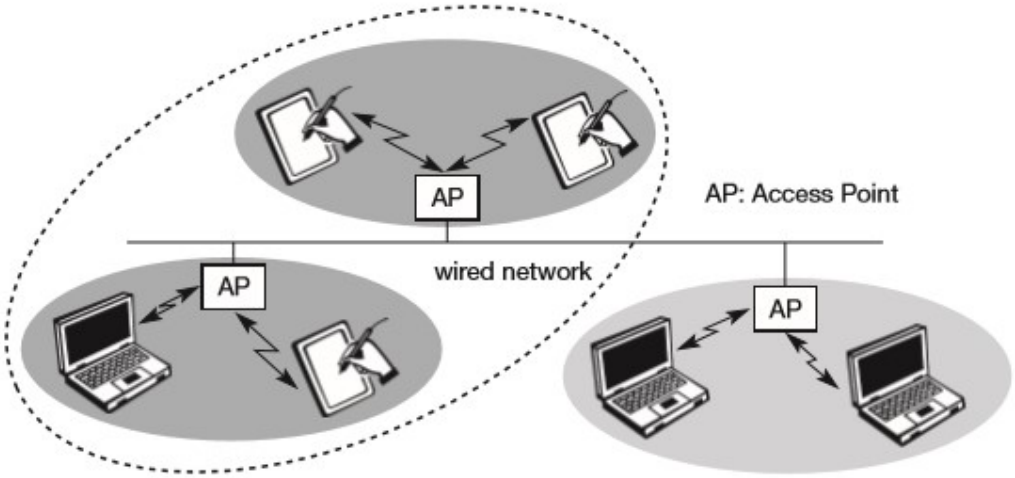


Fig . 1

Typically, the design of infrastructure-based wireless networks is simpler because most of the network functionality lies within the access point, whereas the wireless clients can remain quite simple. This structure is reminiscent of switched Ethernet or other star-based networks, where a central element (e.g., a switch) controls network flow. This type of network can use different access schemes with or without collision. Collisions may occur if medium access of the wireless nodes and the access point is not coordinated. However, if only the access point controls medium access, no collisions are possible. This setting may be useful for quality of service guarantees such as minimum bandwidth for certain nodes. The access point may poll the single wireless nodes to ensure the data rate.

Infrastructure-based networks lose some of the flexibility wireless networks can offer, e.g., they cannot be used for disaster relief in cases where no infrastructure is left. Typical cellular phone

networks are infrastructure-based networks for a wide area. Also satellite-based cellular phones have an infrastructure – the satellites . Infrastructure does not necessarily imply a wired fixed network.

Ad-hoc wireless networks, however, do not need any infrastructure to work. Each node can communicate directly with other nodes, so no access point controlling medium access is necessary. Figure 2 shows two ad-hoc networks with three nodes each. Nodes within an ad-hoc network can only communicate if they can reach each other physically, i.e., if they are within each other's radio range or if other nodes can forward the message. Nodes from the two networks shown in Figure 2 cannot, therefore, communicate with each other if they are not within the same radio range.

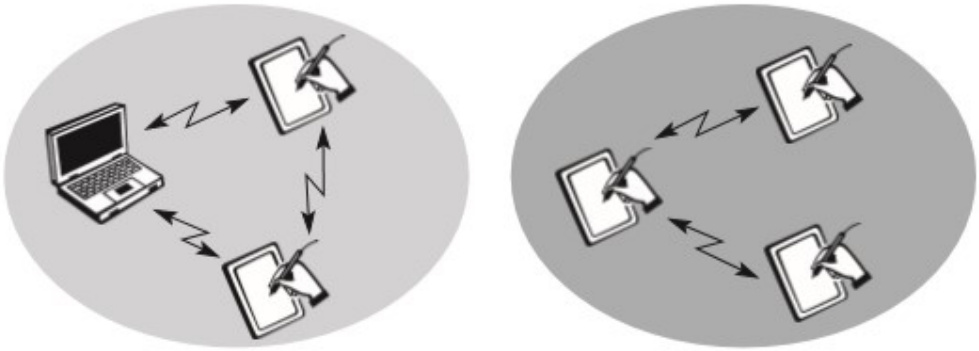


Fig. 2

In ad-hoc networks, the complexity of each node is higher because every node has to implement medium access mechanisms, mechanisms to handle hidden or exposed terminal problems, and perhaps priority mechanisms, to provide a certain quality of service. This type of wireless network exhibits the greatest possible flexibility as it is, for example, needed for unexpected meetings, quick replacements of infrastructure or communication scenarios far away from any infrastructure.

Clearly, the two basic variants of wireless networks (here especially WLANs), infrastructure-based and ad-hoc, do not always come in their pure form. There are networks that rely on access points and infrastructure for basic services (e.g., authentication of access, control of medium access for data with associated quality of service, management functions), but that also allow for direct communication between the wireless nodes. However, ad-hoc networks might only have selected nodes with the capabilities of forwarding data. Most of the nodes have to connect to such a special node first to transmit data if the receiver is out of their range. From the three WLANs presented, IEEE 802.11 and HiperLAN2 are typically infrastructure-based networks, which additionally support ad-hoc networking. However, many implementations only offer the basic infrastructure-based version. The third WLAN, Bluetooth, is a typical wireless ad-hoc network. Bluetooth focuses precisely on spontaneous ad-hoc meetings or on the simple connection of two or more devices without requiring the setup of an infrastructure.

2.3 IEEE 802.11

The IEEE standard 802.11 (IEEE, 1999) specifies the most famous family of WLANs in which many products are available. As the standard's number indicates, this standard belongs to the

group of 802.x LAN standards, e.g., 802.3 Ethernet or 802.5 Token Ring. This means that the standard specifies the physical and medium access layer adapted to the special requirements of wireless LANs, but offers the same interface as the others to higher layers to maintain interoperability.

The primary goal of the standard was the specification of a simple and robust WLAN which offers time-bounded and asynchronous services. The MAC layer should be able to operate with multiple physical layers, each of which exhibits a different medium sense and transmission characteristic. Candidates for physical layers were infra red and spread spectrum radio transmission techniques.

Additional features of the WLAN should include the support of power management to save battery power, the handling of hidden nodes, and the ability to operate worldwide. The 2.4 GHz ISM band, which is available in most countries around the world, was chosen for the original standard. Data rates envisaged for the standard were 1 Mbit/s mandatory and 2 Mbit/s optional.

The following sections will introduce the system and protocol architecture of the initial IEEE 802.11 and then discuss each layer, i.e., physical layer and medium access. After that, the complex and very important management functions of the standard are presented. Finally, this subsection presents the enhancements of the original standard for higher data rates, 802.11a (up to 54 Mbit/s at 5 GHz) and 802.11b (today the most successful with 11 Mbit/s) together with further developments for security support, harmonization, or other modulation schemes.

2.3.1 System architecture

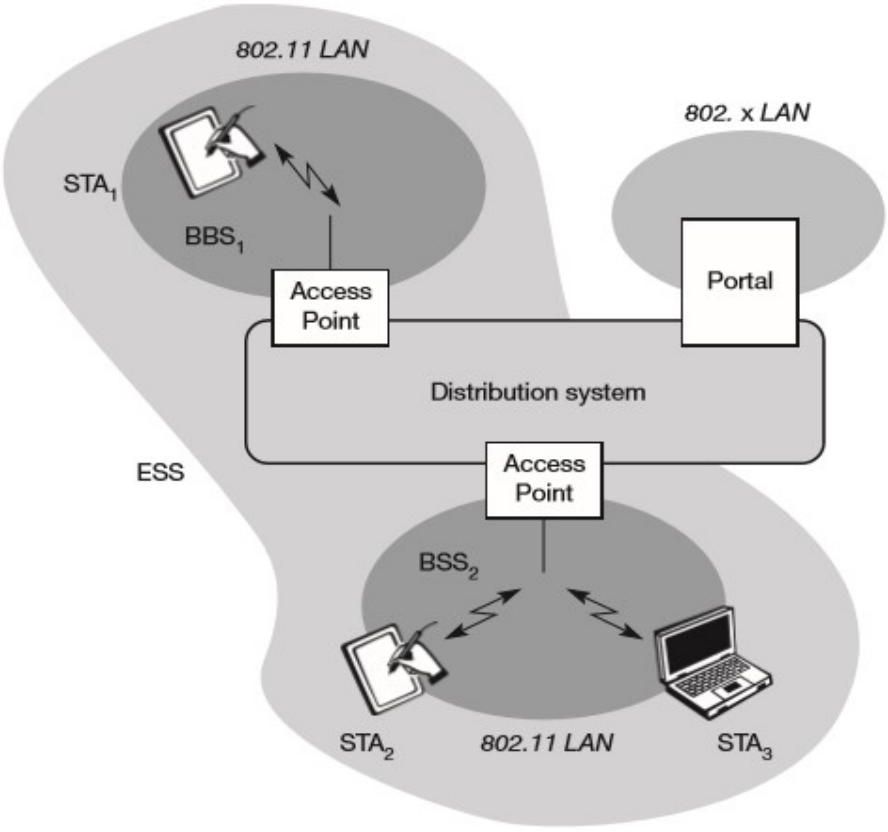


Fig. 3

Wireless networks can exhibit two different basic system architectures as shown in section 2: infrastructure-based or ad-hoc. Figure 3 shows the components of an infrastructure and a wireless part as specified for IEEE 802.11. Several nodes, called stations (STAi), are connected to access points (AP). Stations are terminals with access mechanisms to the wireless medium and radio contact to the AP. The stations and the AP which are within the same radio coverage form a basic service set (BSSi). The example shows two BSSs – BSS1 and BSS2 – which are connected via a distribution system. A distribution system connects several BSSs via the AP to form a single network and thereby extends the wireless coverage area. This network is now called an extended service set (ESS) and has its own identifier, the ESSID. The ESSID is the 'name' of a network and is used to separate different networks. Without knowing the ESSID (and assuming no hacking) it should not be possible to participate in the WLAN. The distribution system connects the wireless networks via the APs with a portal, which forms the interworking unit to other LANs.

The architecture of the distribution system is not specified further in IEEE 802.11. It could consist of bridged IEEE LANs, wireless links, or any other networks. However, distribution system services are defined in the standard (although, many products today cannot interoperate and needs the additional standard IEEE 802.11f to specify an inter access point protocol).

Stations can select an AP and associate with it. The APs support roaming (i.e., changing access points), the distribution system handles data transfer between the different APs. APs provide synchronization within a BSS, support power management, and can control medium access to support time-bounded service. These and further functions are explained in the following sections.

In addition to infrastructure-based networks, IEEE 802.11 allows the building of ad-hoc networks between stations, thus forming one or more independent BSSs (IBSS) as shown in Figure 4. In this case, an IBSS comprises a group of stations using the same radio frequency. Stations STA1, STA2, and STA3 are in IBSS1, STA4 and STA5 in IBSS2. This means for example that STA3 can communicate directly with STA2 but not with STA5. Several IBSSs can either be formed via the distance between the IBSSs (see Figure 7.4) or by using different carrier frequencies (then the IBSSs could overlap physically). IEEE 802.11 does not specify any special nodes that support routing, forwarding of data or exchange of topology information as, e.g., HIPERLAN 1 or Bluetooth .

2.3.2 Protocol architecture

As indicated by the standard number, IEEE 802.11 fits seamlessly into the other 802.x standards for wired LANs (see Halsall, 1996; IEEE, 1990). Figure 5 shows the most common scenario: an IEEE 802.11 wireless LAN connected to a switched IEEE 802.3 Ethernet via a bridge. Applications should not notice any difference apart from the lower bandwidth and perhaps higher access time from the wireless LAN. The WLAN behaves like a slow wired LAN. Consequently, the higher layers (application, TCP, IP) look the same for wireless nodes as for wired nodes. The upper part of the data link control layer, the logical link control (LLC), covers the differences of the medium access control layers needed for the different media. In many of today's networks, no explicit LLC layer is visible. Further details like Ethertype or sub-network access protocol (SNAP) and bridging technology are explained in, e.g., Perlman (1992).

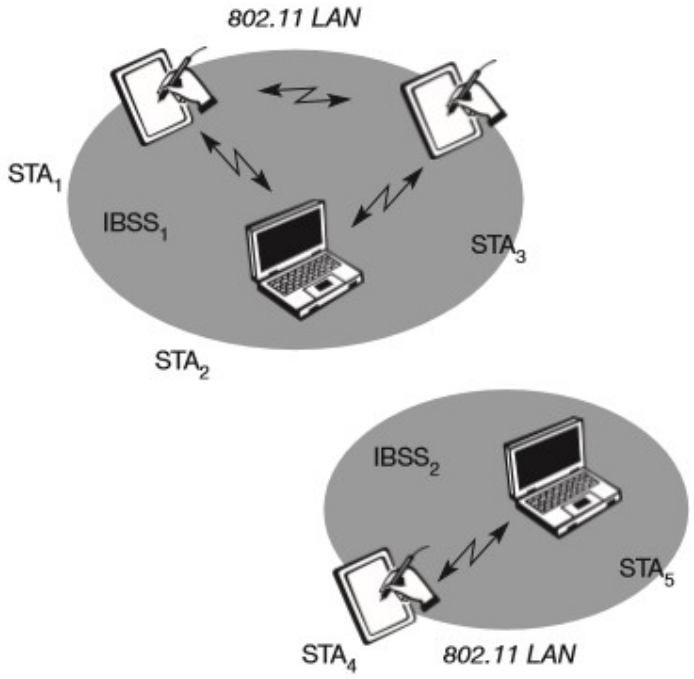


Fig. 4

The IEEE 802.11 standard only covers the physical layer PHY and medium access layer MAC like the other 802.x LANs do. The physical layer is subdivided into the physical layer convergence protocol (PLCP) and the physical medium dependent sublayer PMD (see Figure 6). The basic tasks of the MAC layer comprise medium access, fragmentation of user data, and encryption. The PLCP sublayer provides a carrier sense signal, called clear channel assessment (CCA), and provides a common PHY service access point (SAP) independent of the transmission technology. Finally, the PMD sublayer handles modulation and encoding/decoding of signals. The PHY layer (comprising PMD and PLCP) and the MAC layer will be explained in more detail in the following sections.

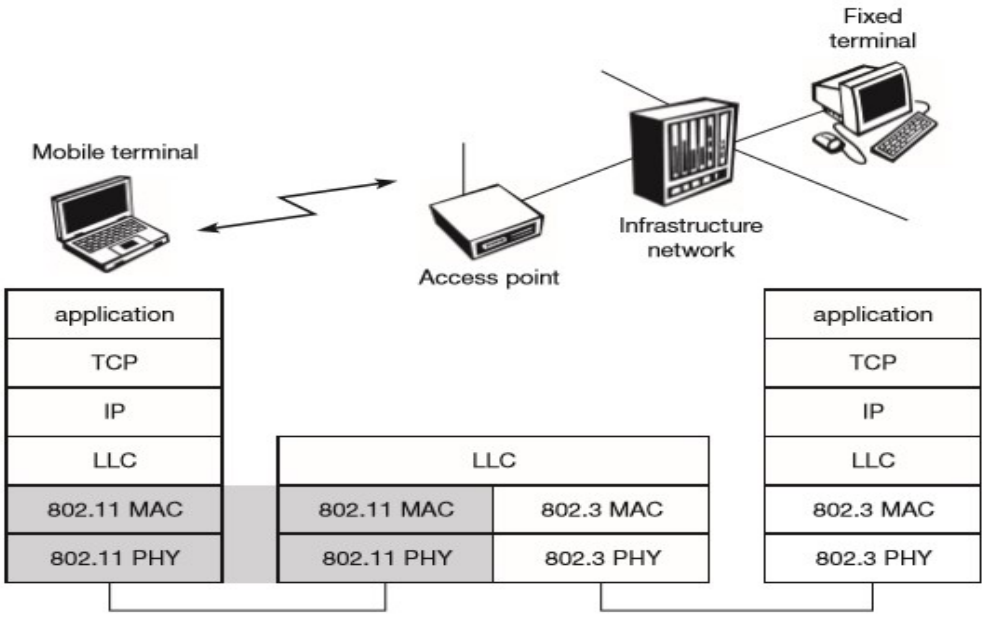


Fig. 5

Apart from the protocol sublayers, the standard specifies management layers and the station management. The MAC management supports the association and re-association of a station to an access point and roaming between different access points. It also controls authentication mechanisms, encryption, synchronization of a station with regard to an access point, and power management to save battery power. MAC management also maintains the MAC management information base (MIB).

The main tasks of the PHY management include channel tuning and PHY MIB maintenance. Finally, station management interacts with both management layers and is responsible for additional higher layer functions (e.g., control of bridging and interaction with the distribution system in the case of an access point).

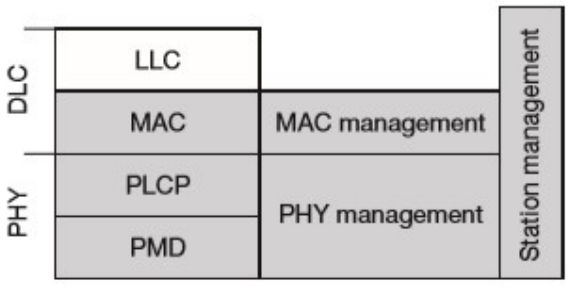


Fig. 6

2.3.3 Physical layer

IEEE 802.11 supports three different physical layers: one layer based on infra red and two layers based on radio transmission (primarily in the ISM band at 2.4 GHz, which is available worldwide). All PHY variants include the provision of the clear channel assessment signal (CCA). This is needed for the MAC mechanisms controlling medium access and indicates if the medium is currently idle. The transmission technology (which will be discussed later) determines exactly how this signal is obtained. The PHY layer offers a service access point (SAP) with 1 or 2 Mbit/s transfer rate to the MAC layer (basic version of the standard). The remainder of this section presents the three versions of a PHY layer defined in the standard.

2.3.3.1 Frequency hopping spread spectrum

Frequency hopping spread spectrum (FHSS) is a spread spectrum technique which allows for the coexistence of multiple networks in the same area by separating different networks using different hopping sequences. The original standard defines 79 hopping channels for North America and Europe, and 23 hopping channels for Japan (each with a bandwidth of 1 MHz in the 2.4 GHz ISM band). The selection of a particular channel is achieved by using a pseudo-random hopping pattern. National restrictions also determine further parameters, e.g., maximum transmit power is 1 W in the US, 100 mW EIRP (equivalent isotropic radiated power) in Europe and 10 mW/MHz in Japan. The standard specifies Gaussian shaped FSK (frequency shift keying), GFSK, as modulation for the FHSS PHY. For 1 Mbit/s a 2 level GFSK is used (i.e., 1 bit is mapped to one frequency, see chapter 2), a 4 level GFSK for 2 Mbit/s (i.e., 2 bits are mapped to one frequency). While sending and receiving at 1 Mbit/s is mandatory for all devices, operation at 2 Mbit/s is optional. This facilitated the production of low-cost devices for the lower rate only and more powerful devices for both transmission rates in the early days of 802.11.

Figure 7 shows a frame of the physical layer used with FHSS. The frame consists of two basic parts, the PLCP part (preamble and header) and the payload part. While the PLCP part is always transmitted at 1 Mbit/s, payload, i.e. MAC data, can use 1 or 2 Mbit/s. Additionally, MAC data is scrambled using the polynomial $s(z) = z7 + z4 + 1$ for DC blocking and whitening of the spectrum. The fields of the frame fulfill the following functions:

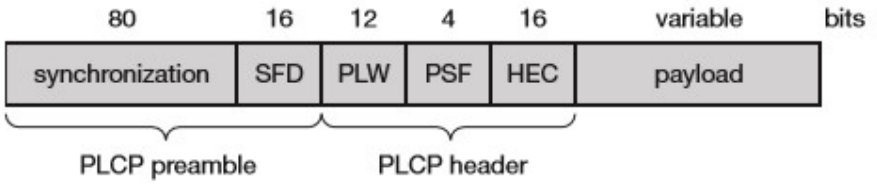


Fig. 7

Synchronization: The PLCP preamble starts with 80 bit synchronization, which is a 010101... bit pattern. This pattern is used for synchronization of potential receivers and signal detection by the CCA.

● Start frame delimiter (SFD): The following 16 bits indicate the start of the frame and provide frame synchronization. The SFD pattern is 0000110010111101.

● PLCP_PDU length word (PLW): This first field of the PLCP header indicates the length of the payload in bytes including the 32 bit CRC at the end of the payload. PLW can range between 0 and 4,095.

● PLCP signalling field (PSF): This 4 bit field indicates the data rate of the payload following. All bits set to zero (0000) indicates the lowest data rate of 1 Mbit/s. The granularity is 500 kbit/s, thus 2 Mbit/s is indicated by 0010 and the maximum is 8.5 Mbit/s (1111). This system obviously does not accommodate today's higher data rates.

● Header error check (HEC): Finally, the PLCP header is protected by a 16 bit checksum with the standard ITU-T generator polynomial G(x) = x16 + x12 + x5 + 1.

2.3.3.2 Direct sequence spread spectrum

Direct sequence spread spectrum (DSSS) is the alternative spread spectrum method separating by code and not by frequency. In the case of IEEE 802.11 DSSS, spreading is achieved using the 11-chip Barker sequence (+1, –1, +1, +1, –1, +1, +1, +1, –1, –1, –1). The key characteristics of this method are its robustness against interference and its insensitivity to multipath propagation (time delay spread). However, the implementation is more complex compared to FHSS. IEEE 802.11 DSSS PHY also uses the 2.4 GHz ISM band and offers both 1 and 2 Mbit/s data rates. The system uses differential binary phase shift keying (DBPSK) for 1 Mbit/s transmission and differential quadrature phase shift keying (DQPSK) for 2 Mbit/s as modulation schemes. Again, the maximum transmit power is 1 W in the US, 100 mW EIRP in Europe and 10 mW/MHz in Japan. The symbol rate is 1 MHz, resulting in a chipping rate of 11 MHz. All bits transmitted by the DSSS PHY are scrambled with the polynomial s(z) = z7 + z4 + 1 for DC blocking and whitening of the spectrum. Many of today's products offering 11 Mbit/s according to 802.11b are still backward compatible to these lower data rates.

Figure 8 shows a frame of the physical layer using DSSS. The frame consists of two basic parts, the PLCP part (preamble and header) and the payload part. While the PLCP part is always transmitted at 1 Mbit/s, payload, i.e., MAC data, can use 1 or 2 Mbit/s. The fields of the frame have the following functions:

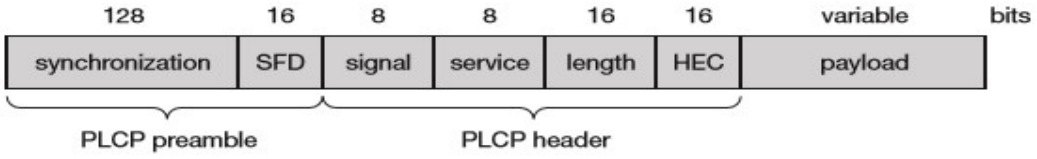


Fig. 8

Synchronization: The first 128 bits are not only used for synchronization, but also gain setting, energy detection (for the CCA), and frequency offset compensation. The synchronization field only consists of scrambled 1 bits.

● Start frame delimiter (SFD): This 16 bit field is used for synchronization at the beginning of a frame and consists of the pattern 1111001110100000.

● Signal: Originally, only two values have been defined for this field to indicate the data rate of the payload. The value 0x0A indicates 1 Mbit/s (and thus DBPSK), 0x14 indicates 2 Mbit/s (and

thus DQPSK). Other values have been reserved for future use, i.e., higher bit rates. Coding for higher data rates is explained in sections 7.3.6 and 7.3.7.

● Service: This field is reserved for future use; however, 0x00 indicates an IEEE 802.11 compliant frame.

● Length: 16 bits are used in this case for length indication of the payload in microseconds.

● Header error check (HEC): Signal, service, and length fields are protected by this checksum using the ITU-T CRC-16 standard polynomial.

2.3.3.3 Infra red

The PHY layer, which is based on infra red (IR) transmission, uses near visible light at 850–950 nm. Infra red light is not regulated apart from safety restrictions (using lasers instead of LEDs). The standard does not require a line-of-sight between sender and receiver, but should also work with diffuse light. This allows for point-to-multipoint communication. The maximum range is about 10 m if no sunlight or heat sources interfere with the transmission. Typically, such a network will only work in buildings, e.g., classrooms, meeting rooms etc. Frequency reuse is very simple – a wall is more than enough to shield one IR based IEEE 802.11 network from another. Today, no products are available that offer infra red communication based on 802.11. Proprietary products offer, e.g., up to 4 Mbit/s using diffuse infra red light. Alternatively, directed infra red communication based on IrDA can be used (IrDA, 2002).

2.3.4 Medium access control layer

The MAC layer has to fulfill several tasks. First of all, it has to control medium access, but it can also offer support for roaming, authentication, and power conservation. The basic services provided by the MAC layer are the mandatory asynchronous data service and an optional time-bounded service. While 802.11 only offers the asynchronous service in ad-hoc network mode, both service types can be offered using an infrastructure-based network together with the access point coordinating medium access. The asynchronous service supports broadcast and multi-cast packets, and packet exchange is based on a ‘best effort’ model, i.e., no delay bounds can be given for transmission. The following three basic access mechanisms have been defined for IEEE 802.11: the mandatory basic method based on a version of CSMA/CA, an optional method avoiding the hidden terminal problem, and finally a contention-free polling method for time-bounded service. The first two methods are also summarized as distributed coordination function (DCF), the third method is called point coordination function (PCF). DCF only offers asynchronous service, while PCF offers both asynchronous and time-bounded service but needs an access point to control medium access and to avoid contention. The MAC mechanisms are also called distributed foundation wireless medium access control (DFWMAC).

For all access methods, several parameters for controlling the waiting time before medium access are important. Figure 9 shows the three different parameters that define the priorities of medium access. The values of the parameters depend on the PHY and are defined in relation to a slot time. Slot time is derived from the medium propagation delay, transmitter delay, and other PHY dependent parameters. Slot time is 50 µs for FHSS and 20 µs for DSSS.

The medium, as shown, can be busy or idle (which is detected by the CCA). If the medium is busy this can be due to data frames or other control frames. During a contention phase several nodes try to access the medium.

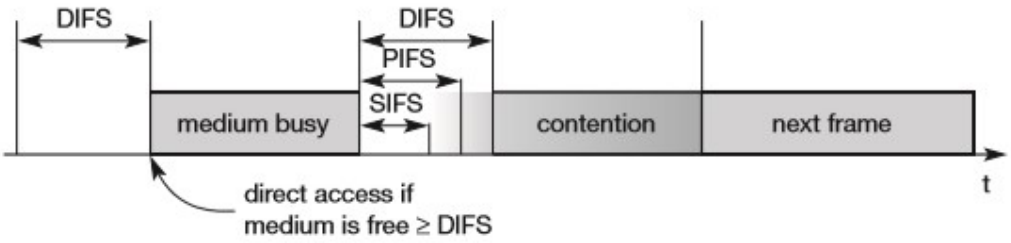


Fig. 9

Short inter-frame spacing (SIFS): The shortest waiting time for medium access (so the highest priority) is defined for short control messages, such as acknowledgements of data packets or polling responses. For DSSS SIFS is 10 µs and for FHSS it is 28 µs.

● PCF inter-frame spacing (PIFS): A waiting time between DIFS and SIFS (and thus a medium priority) is used for a time-bounded service. An access point polling other nodes only has to wait PIFS for medium access . PIFS is defined as SIFS plus one slot time.

● DCF inter-frame spacing (DIFS): This parameter denotes the longest waiting time and has the lowest priority for medium access. This waiting time is used for asynchronous data service within a contention period . DIFS is defined as SIFS plus two slot times.

2.3.4.1 Basic DFWMAC-DCF using CSMA/C

The mandatory access mechanism of IEEE 802.11 is based on carrier sense multiple access with collision avoidance (CSMA/CA), which is a random access scheme with carrier sense and collision avoidance through random backoff. The basic CSMA/CA mechanism is shown in Figure 10. If the medium is idle for at least the duration of DIFS (with the help of the CCA signal of the physical layer), a node can access the medium at once. This allows for short access delay under light load. But as more and more nodes try to access the medium, additional mechanisms are needed.

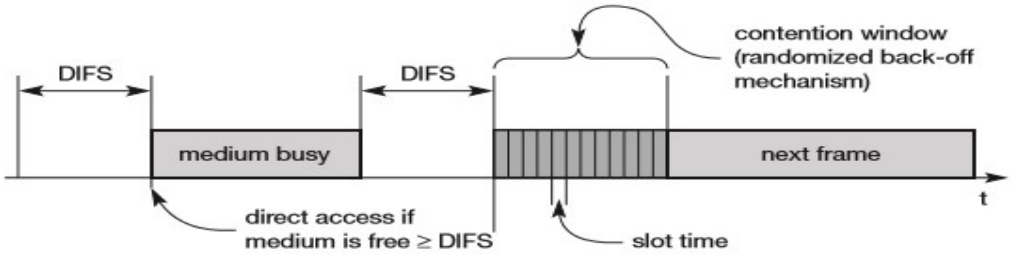


Fig. 10

If the medium is busy, nodes have to wait for the duration of DIFS, entering a contention phase afterwards. Each node now chooses a random backoff time within a contention window and delays medium access for this random amount of time. The node continues to sense the medium. As soon as a node senses the channel is busy, it has lost this cycle and has to wait for the next chance, i.e., until the medium is idle again for at least DIFS. But if the randomized additional waiting time for a node is over and the medium is still idle, the node can access the medium immediately (i.e., no other node has a shorter waiting time). The additional waiting time is

measured in multiples of the above-mentioned slots. This additional randomly distributed delay helps to avoid collisions – otherwise all stations would try to transmit data after waiting for the medium becoming idle again plus DIFS.

Obviously, the basic CSMA/CA mechanism is not fair. Independent of the overall time a node has already waited for transmission; each node has the same chances for transmitting data in the next cycle. To provide fairness, IEEE 802.11 adds a backoff timer. Again, each node selects a random waiting time within the range of the contention window. If a certain station does not get access to the medium in the first cycle, it stops its backoff timer, waits for the channel to be idle again for DIFS and starts the counter again. As soon as the counter expires, the node accesses the medium. This means that deferred stations do not choose a randomized backoff time again, but continue to count down. Stations that have waited longer have the advantage over stations that have just entered, in that they only have to wait for the remainder of their backoff timer from the previous cycle(s).

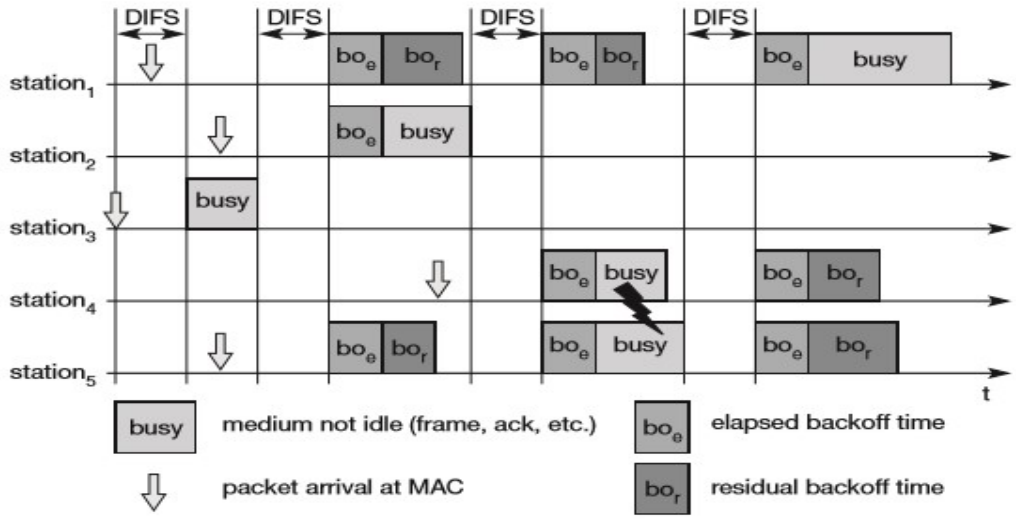


Fig. 11

Figure 11 explains the basic access mechanism of IEEE 802.11 for five stations trying to send a packet at the marked points in time. Station3 has the first request from a higher layer to send a packet (packet arrival at the MAC SAP). The station senses the medium, waits for DIFS and accesses the medium, i.e., sends the packet. Station1, station2, and station5 have to wait at least until the medium is idle for DIFS again after station3 has stopped sending. Now all three stations choose a backoff time within the contention window and start counting down their backoff timers.

Figure 11 shows the random backoff time of station1 as sum of boe (the elapsed backoff time) and bor (the residual backoff time). The same is shown for station5. Station2 has a total backoff time of only boe and gets access to the medium first. No residual backoff time for station2 is shown. The backoff timers of station1 and station5 stop, and the stations store their residual backoff times. While a new station has to choose its backoff time from the whole contention window, the two old stations have statistically smaller backoff values. The older values are on average lower than the new ones. Now station4 wants to send a packet as well, so after DIFS waiting time, three stations try to get access. It can now happen, as shown in the figure, that two

stations accidentally have the same backoff time, no matter whether remaining or newly chosen. This results in a collision on the medium as shown, i.e., the transmitted frames are destroyed. Station1 stores its residual backoff time again. In the last cycle shown station1 finally gets access to the medium, while station4 and station5 have to wait. A collision triggers a retransmission with a new random selection of the backoff time. Retransmissions are not privileged. Still, the access scheme has problems under heavy or light load. Depending on the size of the contention window (CW), the random values can either be too close together (causing too many collisions) or the values are too high (causing unnecessary delay). The system tries to adapt to the current number of stations trying to send.

The contention window starts with a size of, e.g., CWmin = 7. Each time a collision occurs, indicating a higher load on the medium, the contention window doubles up to a maximum of, e.g., CWmax = 255 (the window can take on the values 7, 15, 31, 63, 127, and 255). The larger the contention window is, the greater is the resolution power of the randomized scheme. It is less likely to choose the same random backoff time using a large CW. However, under a light load, a small CW ensures shorter access delays. This algorithm is also called exponential backoff and is already familiar from IEEE 802.3 CSMA/CD in a similar version.

While this process describes the complete access mechanism for broadcast frames, an additional feature is provided by the standard for unicast data transfer. Figure 12 shows a sender accessing the medium and sending its data. But now, the receiver answers directly with an acknowledgement (ACK). The receiver accesses the medium after waiting for duration of SIFS so no other station can access the medium in the meantime and cause a collision. The other stations have to wait for DIFS plus their backoff time. This acknowledgement ensures the correct reception (correct checksum CRC at the receiver) of a frame on the MAC layer, which is especially important in error-prone environments such as wireless connections. If no ACK is returned, the sender automatically retransmits the frame. But now the sender has to wait again and compete for the access right. There are no special rules for retransmissions. The number of retransmissions is limited, and final failure is reported to the higher layer.

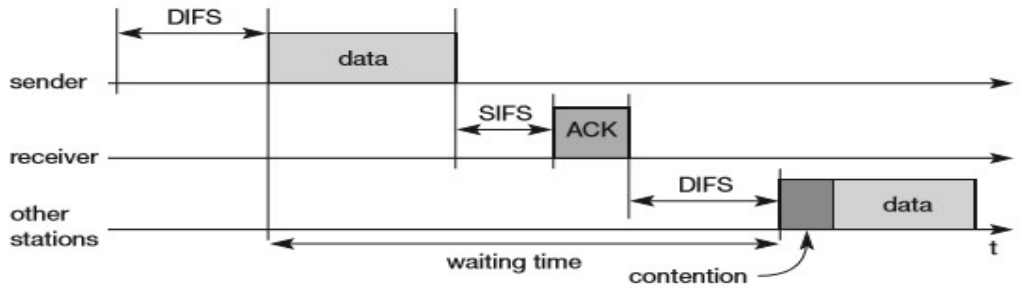


Fig. 12

2.3.4.2 DFWMAC-DCF with RTS/CTS extension

The problem of hidden terminals, a situation that can also occur in IEEE 802.11 networks. This problem occurs if one station can receive two others, but those stations cannot receive each other. The two stations may sense the channel is idle, send a frame, and cause a collision at the receiver in the middle. To deal with this problem, the standard defines an additional mechanism using two control packets, RTS and CTS. The use of the mechanism is optional; however, every 802.11 node has to implement the functions to react properly upon reception of RTS/CTS control packets.

Figure 13 illustrates the use of RTS and CTS. After waiting for DIFS (plus a random backoff time if the medium was busy), the sender can issue a request to send (RTS) control packet. The RTS packet thus is not given any higher priority compared to other data packets. The RTS packet includes the receiver of the data transmission to come and the duration of the whole data transmission. This duration specifies the time interval necessary to transmit the whole data frame and the acknowledgement related to it. Every node receiving this RTS now has to set its net allocation vector (NAV) in accordance with the duration field. The NAV then specifies the earliest point at which the station can try to access the medium again.

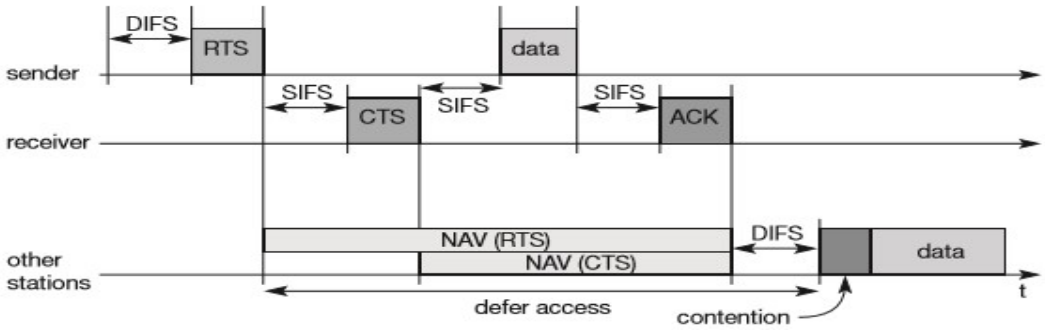


Fig. 13

If the receiver of the data transmission receives the RTS, it answers with a clear to send (CTS) message after waiting for SIFS. This CTS packet contains the duration field again and all stations receiving this packet from the receiver of the intended data transmission have to adjust their NAV. The latter set of receivers need not be the same as the first set receiving the RTS packet. Now all nodes within receiving distance around sender and receiver are informed that they have to wait more time before accessing the medium. Basically, this mechanism reserves the medium for one sender exclusively (this is why it is sometimes called a virtual reservation scheme).

Finally, the sender can send the data after SIFS. The receiver waits for SIFS after receiving the data packet and then acknowledges whether the transfer was correct. The transmission has now been completed, the NAV in each node marks the medium as free and the standard cycle can start again.

Within this scenario (i.e., using RTS and CTS to avoid the hidden terminal problem), collisions can only occur at the beginning while the RTS is sent. Two or more stations may start sending at the same time (RTS or other data packets). Using RTS/CTS can result in a non-negligible overhead causing a waste of bandwidth and higher delay. An RTS threshold can determine when to use the additional mechanism (basically at larger frame sizes) and when to disable it (short frames). Chhaya (1996) and Chhaya (1997) give an overview of the asynchronous services in 802.11 and discuss performance under different load scenarios. Wireless LANs have bit error rates in transmission that are typically several orders of magnitude higher than, e.g., fiber optics. The probability of an erroneous frame is much higher for wireless links assuming the same frame length. One way to decrease the error probability of frames is to use shorter frames. In this case, the bit error rate is the same, but now only short frames are destroyed and, the frame error rate decreases.

However, the mechanism of fragmenting a user data packet into several smaller parts should be transparent for a user. The MAC layer should have the possibility of adjusting the transmission frame size to the current error rate on the medium. The IEEE 802.11 standard specifies a fragmentation mode (see Figure 14). Again, a sender can send an RTS control packet to reserve the medium after a waiting time of DIFS. This RTS packet now includes the duration for the transmission of the first fragment and the corresponding acknowledgement. A certain set of nodes may receive this RTS and set their NAV according to the duration field. The receiver answers with a CTS, again including the duration of the transmission up to the acknowledgement. A (possibly different) set of receivers gets this CTS message and sets the NAV.

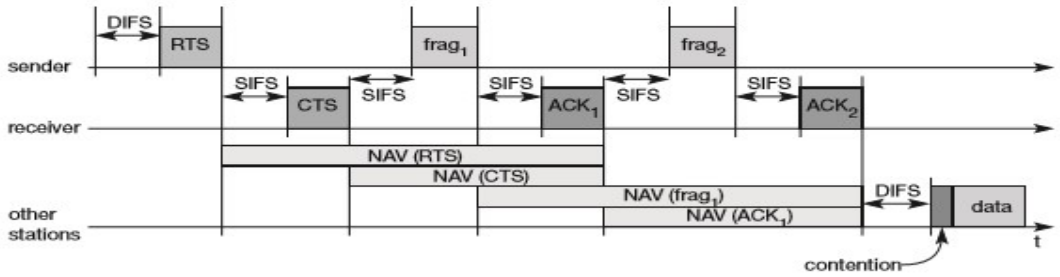


Fig. 14

As shown in Figure 13, the sender can now send the first data frame, frag1, after waiting only for SIFS. The new aspect of this fragmentation mode is that it includes another duration value in the frame frag1. This duration field reserves the medium for the duration of the transmission following, comprising the second fragment and its acknowledgement. Again, several nodes may receive this reservation and adjust their NAV. If all nodes are static and transmission conditions have not changed, then the set of nodes receiving the duration field in frag1 should be the same as the set that has received the initial reservation in the RTS control packet. However, due to the mobility of nodes and changes in the environment, this could also be a different set of nodes. The receiver of frag1 answers directly after SIFS with the acknowledgement packet ACK1 including the reservation for the next transmission as shown. Again, a fourth set of nodes may receive this reservation and adjust their NAV (which again could be the same as the second set of nodes that has received the reservation in the CTS frame).

If frag2 was not the last frame of this transmission, it would also include a new duration for the third consecutive transmission. (In the example shown, frag2 is the last fragment of this transmission so the sender does not reserve the medium any longer.) The receiver acknowledges this second fragment, not reserving the medium again. After ACK2, all nodes can compete for the medium again after having waited for DIFS.

2.3.4.3 DFWMAC-PCF with polling

The two access mechanisms presented so far cannot guarantee a maximum access delay or minimum transmission bandwidth. To provide a time-bounded service, the standard specifies a point coordination function (PCF) on top of the standard DCF mechanisms. Using PCF requires an access point that controls medium access and polls the single nodes. Ad-hoc networks cannot use this function so, provide no QoS but ‘best effort’ in IEEE 802.11 WLANs. The point co-ordinator in the access point splits the access time into super frame periods as shown in Figure

15. A super frame comprises a contentionfree period and a contention period. The contention period can be used for the two access mechanisms presented above. The figure also shows several wireless stations (all on the same line) and the stations' NAV (again on one line).

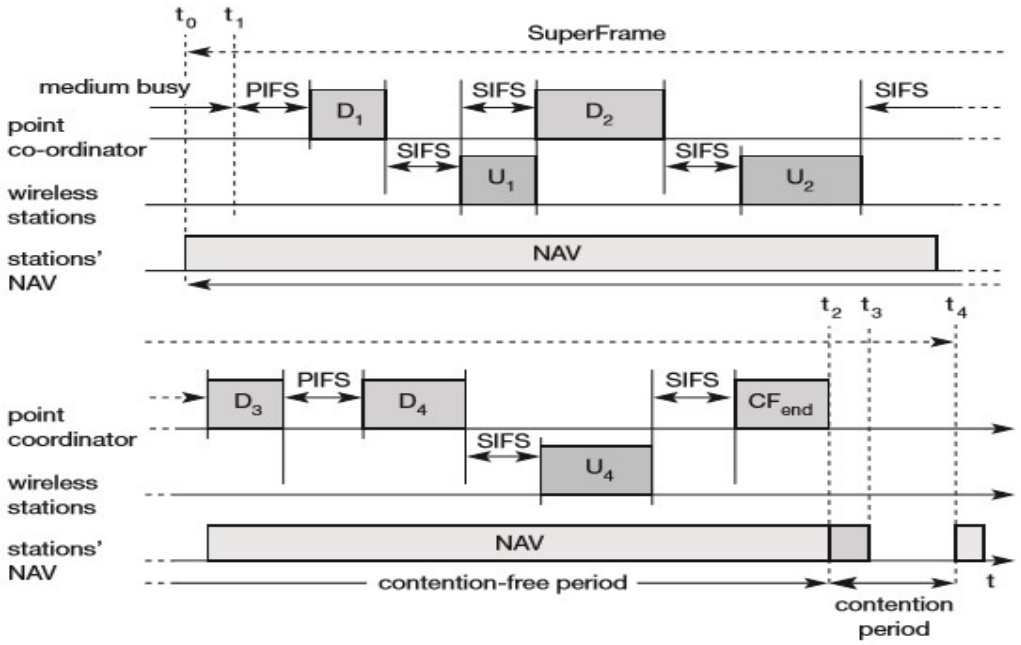


Fig. 15

At time t0 the contention-free period of the super frame should theoretically start, but another station is still transmitting data (i.e., the medium is busy). This means that PCF also defers to DCF, and the start of the super frame may be postponed. The only possibility of avoiding variations is not to have any contention period at all. After the medium has been idle until t1, the point coordinator has to wait for PIFS before accessing the medium. As PIFS is smaller than DIFS, no other station can start sending earlier. The point coordinator now sends data D1 downstream to the first wireless station. This station can answer at once after SIFS (see Figure 15). After waiting for SIFS again, the point coordinator can poll the second station by sending D2. This station may answer upstream to the coordinator with data U2. Polling continues with the third node. This time the node has nothing to answer and the point coordinator will not receive a packet after SIFS.

After waiting for PIFS, the coordinator can resume polling the stations. Finally, the point coordinator can issue an end marker (CFend), indicating that the contention period may start again. Using PCF automatically sets the NAV, preventing other stations from sending. In the example, the contention-free period planned initially would have been from t0 to t3. However, the point coordinator finished polling earlier, shifting the end of the contention-free period to t2. At t4, the cycle starts again with the next super frame.

The transmission properties of the whole wireless network are now determined by the polling behavior of the access point. If only PCF is used and polling is distributed evenly, the bandwidth is also distributed evenly among all polled nodes. This would resemble a static, centrally controlled time division multiple access (TDMA) system with time division duplex (TDD) transmission. This method comes with an overhead if nodes have nothing to send, but the access

point polls them permanently. Anastasi (1998) elaborates the example of voice transmission using 48 byte packets as payload. In this case, PCF introduces an overhead of 75 byte.

2.3.4.4 MAC frames

Figure 16 shows the basic structure of an IEEE 802.11 MAC data frame together with the content of the frame control field. The fields in the figure refer to the following:

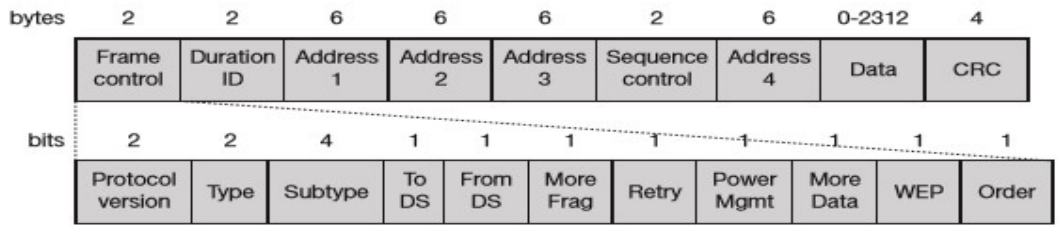


Fig. 16

● Frame control: The first 2 bytes serve several purposes. They contain several sub-fields as explained after the MAC frame.

● Duration/ID: If the field value is less than 32,768, the duration field contains the value indicating the period of time in which the medium is occupied (in µs). This field is used for setting the NAV for the virtual reservation mechanism using RTS/CTS and during fragmentation. Certain values above 32,768 are reserved for identifiers.

● Address 1 to 4: The four address fields contain standard IEEE 802 MAC addresses (48 bit each), as they are known from other 802.x LANs. The meaning of each address depends on the DS bits in the frame control field and is explained in more detail in a separate paragraph.

● Sequence control: Due to the acknowledgement mechanism frames may be duplicated. Therefore a sequence number is used to filter duplicates.

● Data: The MAC frame may contain arbitrary data (max. 2,312 byte), which is transferred transparently from a sender to the receiver(s).

● Checksum (CRC): Finally, a 32 bit checksum is used to protect the frame as it is common practice in all 802.x networks.

The frame control field shown in Figure 16 contains the following fields:

● Protocol version: This 2 bit field indicates the current protocol version and is fixed to 0 by now. If major revisions to the standard make it incompatible with the current version, this value will be increased.

● Type: The type field determines the function of a frame: management (=00), control (=01), or data (=10). The value 11 is reserved. Each type has several subtypes as indicated in the following field.

● Subtype: Example subtypes for management frames are: 0000 for association request, 1000 for beacon. RTS is a control frame with subtype 1011, CTS is coded as 1100. User data is transmitted as data frame with subtype 0000. All details can be found in IEEE, 1999.

● To DS/From DS: Explained in the following in more detail.

● More fragments: This field is set to 1 in all data or management frames that have another fragment of the current MSDU to follow.

● Retry: If the current frame is a retransmission of an earlier frame, this bit is set to 1. With the help of this bit it may be simpler for receivers to eliminate duplicate frames.

● Power management: This field indicates the mode of a station after successful transmission of a frame. Set to 1 the field indicates that the station goes into power-save mode. If the field is set to 0, the station stays active.

● More data: In general, this field is used to indicate a receiver that a sender has more data to send than the current frame. This can be used by an access point to indicate to a station in power-save mode that more packets are buffered. Or it can be used by a station to indicate to an access point after being polled that more polling is necessary as the station has more data ready to transmit.

● Wired equivalent privacy (WEP): This field indicates that the standard security mechanism of 802.11 is applied. However, due to many weaknesses found in the WEP algorithm higher layer security should be used to secure an 802.11 network (Borisov, 2001).

● Order: If this bit is set to 1 the received frames must be processed in strict order.

MAC frames can be transmitted between mobile stations; between mobile stations and an access point and between access points over a DS (see Figure 3). Two bits within the Frame Control field, ‘to DS’ and ‘from DS’, differentiate these cases and control the meaning of the four addresses used. Table 1 gives an overview of the four possible bit values of the DS bits and the associated interpretation of the four address fields.

to DS	from DS	Address 1	Address 2	Address 3	Address 4
0	0	DA	SA	BSSID	–
0	1	DA	BSSID	SA	–
1	0	BSSID	SA	DA	–
1	1	RA	TA	DA	SA

Table 1

Every station, access point or wireless node, filters on address 1. This address identifies the physical receiver(s) of the frame. Based on this address, a station can decide whether the frame is relevant or not. The second address, address 2, represents the physical transmitter of a frame. This information is important because this particular sender is also the recipient of the MAC layer acknowledgement. If a packet from a transmitter (address 2) is received by the receiver

with address 1, this receiver in turn acknowledges the data packet using address 2 as receiver address as shown in the ACK packet in Figure 17. The remaining two addresses, address 3 and address 4, are mainly necessary for the logical assignment of frames (logical sender, BSS identifier, logical receiver). If address 4 is not needed the field is omitted.

For addressing, the following four scenarios are possible:

● Ad-hoc network: If both DS bits are zero, the MAC frame constitutes a packet which is exchanged between two wireless nodes without a distribution system. DA indicates the destination address, SA the source address of the frame, which are identical to the physical receiver and sender addresses respectively. The third address identifies the basic service set (BSSID) (see Figure 4), the fourth address is unused.

● Infrastructure network, from AP: If only the 'from DS' bit is set, the frame physically originates from an access point. DA is the logical and physical receiver, the second address identifies the BSS, the third address specifies the logical sender, the source address of the MAC frame. This case is an example for a packet sent to the receiver via the access point.

● Infrastructure network, to AP: If a station sends a packet to another station via the access point, only the 'to DS' bit is set. Now the first address represents the physical receiver of the frame, the access point, via the BSS identifier. The second address is the logical and physical sender of the frame, while the third address indicates the logical receiver.

● Infrastructure network, within DS: For packets transmitted between two access points over the distribution system, both bits are set. The first receiver address (RA), represents the MAC address of the receiving access point. Similarly, the second address transmitter address (TA), identifies the sending access point within the distribution system. Now two more addresses are needed to identify the original destination DA of the frame and the original source of the frame SA. Without these additional addresses, some encapsulation mechanism would be necessary to transmit MAC frames over the distribution system transparently.

Figure 17 shows three control packets as examples for many special packets defined in the standard. The acknowledgement packet (ACK) is used to acknowledge the correct reception of a data frame as shown in Figure 12. The receiver address is directly copied from the address 2 field of the immediately previous frame. If no more fragments follow for a certain frame the duration field is set to 0. Otherwise the duration value of the previous frame (minus the time required to transmit the ACK minus SIFS) is stored in the duration field.

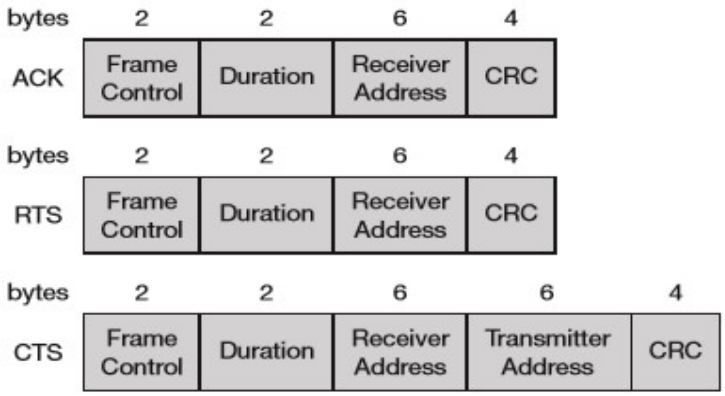


Fig. 17

For the MACA algorithm the RTS/CTS packets are needed. As Figure 13 shows, these packets have to reserve the medium to avoid collisions. Therefore, the request to send (RTS) packet contains the receiver address of the intended recipient of the following data transfer and the transmitter address of the station transmitting the RTS packet. The duration (in μs) comprises the time to send the CTS, data, and ACK plus three SIFS. The immediately following clear to send (CTS) frame copies the transmitter address from the RTS packet into its receiver address field. Additionally, it reads the duration field, subtracts the time to send the CTS and a SIFS and writes the result into its own duration field.

2.3.5 MAC management

MAC management plays a central role in an IEEE 802.11 station as it more or less controls all functions related to system integration, i.e., integration of a wireless station into a BSS, formation of an ESS, synchronization of stations etc. The following functional groups have been identified and will be discussed in more detail in the following sections:

● Synchronization: Functions to support finding a wireless LAN, synchronization of internal clocks, generation of beacon signals.

● Power management: Functions to control transmitter activity for power conservation, e.g., periodic sleep, buffering, without missing a frame.

● Roaming: Functions for joining a network (association), changing access points, scanning for access points.

● Management information base (MIB): All parameters representing the current state of a wireless station and an access point are stored within a MIB for internal and external access. A MIB can be accessed via standardized protocols such as the simple network management protocol (SNMP).

2.3.5.1 Synchronization

Each node of an 802.11 network maintains an internal clock. To synchronize the clocks of all nodes, IEEE 802.11 specifies a timing synchronization function (TSF). As we will see in the following section, synchronized clocks are needed for power management, but also for coordination of the PCF and for synchronization of the hopping sequence in an FHSS system. Using PCF, the local timer of a node can predict the start of a super frame, i.e., the contention free and contention period. FHSS physical layers need the same hopping sequences so that all nodes can communicate within a BSS. Within a BSS, timing is conveyed by the (quasi)periodic transmissions of a beacon frame. A beacon contains a timestamp and other management information used for power management and roaming (e.g., identification of the BSS). The timestamp is used by a node to adjust its local clock. The node is not required to hear every beacon to stay synchronized; however, from time to time internal clocks should be adjusted. The transmission of a beacon frame is not always periodic because the beacon frame is also deferred if the medium is busy.

Within infrastructure-based networks, the access point performs synchronization by transmitting the (quasi)periodic beacon signal, whereas all other wireless nodes adjust their local timer to the time stamp. This represents the simple case shown in Figure 18. The access point is not always

able to send its beacon B periodically if the medium is busy. However, the access point always tries to schedule transmissions according to the expected beacon interval (target beacon transmission time), i.e., beacon intervals are not shifted if one beacon is delayed. The timestamp of a beacon always reflects the real transmit time, not the scheduled time. For ad-hoc networks, the situation is slightly more complicated as they do not have an access point for beacon transmission. In this case, each node maintains its own synchronization timer and starts the transmission of a beacon frame after the beacon interval. Figure 19 shows an example where multiple stations try to send their beacon. However, the standard random backoff algorithm is also applied to the beacon frames so only one beacon wins. All other stations now adjust their internal clocks according to the received beacon and suppress their beacons for this cycle. If collision occurs, the beacon is lost. In this scenario, the beacon intervals can be shifted slightly because all clocks may vary as may the start of a beacon interval from a node's point of view. However, after successful synchronization all nodes again have the same consistent view.

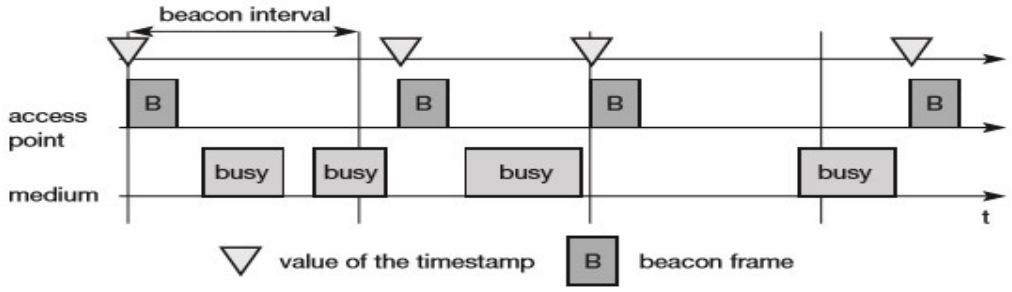


Fig. 18

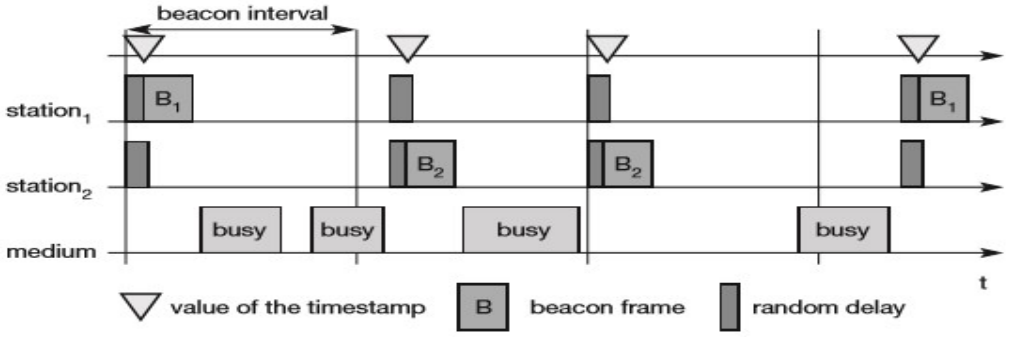


Fig. 19

2.3.5.2 Power management

Wireless devices are battery powered (unless a solar panel is used). Therefore, power-saving mechanisms are crucial for the commercial success of such devices. Standard LAN protocols assume that stations are always ready to receive data, although receivers are idle most of the time in lightly loaded networks. However, this permanent readiness of the receiving module is critical for battery life as the receiver current may be up to 100 mA (Woesner, 1998). The basic idea of IEEE 802.11 power management is to switch off the transceiver whenever it is not needed. For the sending device this is simple to achieve as the transfer is triggered by the device itself. However, since the power management of a receiver cannot know in advance when the transceiver has to be active for a specific packet, it has to 'wake up' the transceiver periodically.

Switching off the transceiver should be transparent to existing protocols and should be flexible enough to support different applications. However, throughput can be traded-off for battery life. Longer off-periods save battery life but reduce average throughput and vice versa.

The basic idea of power saving includes two states for a station: sleep and awake, and buffering of data in senders. If a sender intends to communicate with a power-saving station it has to buffer data if the station is asleep. The sleeping station on the other hand has to wake up periodically and stay awake for a certain time. During this time, all senders can announce the destinations of their buffered data frames. If a station detects that it is a destination of a buffered packet it has to stay awake until the transmission takes place. Waking up at the right moment requires the timing synchronization function (TSF). All stations have to wake up or be awake at the same time.

Power management in infrastructure-based networks is much simpler compared to ad-hoc networks. The access point buffers all frames destined for stations operating in power-save mode. With every beacon sent by the access point, a traffic indication map (TIM) is transmitted. The TIM contains a list of stations for which unicast data frames are buffered in the access point. The TSF assures that the sleeping stations will wake up periodically and listen to the beacon and TIM. If the TIM indicates a unicast frame buffered for the station, the station stays awake for transmission. For multi-cast/broadcast transmission, stations will always stay awake. Another reason for waking up is a frame which has to be transmitted from the station to the access point. A sleeping station still has the TSF timer running.

Figure 20 shows an example with an access point and one station. The state of the medium is indicated. Again, the access point transmits a beacon frame each beacon interval. This interval is now the same as the TIM interval. Additionally, the access point maintains a delivery traffic indication map (DTIM) interval for sending broadcast/multicast frames. The DTIM interval is always a multiple of the TIM interval.

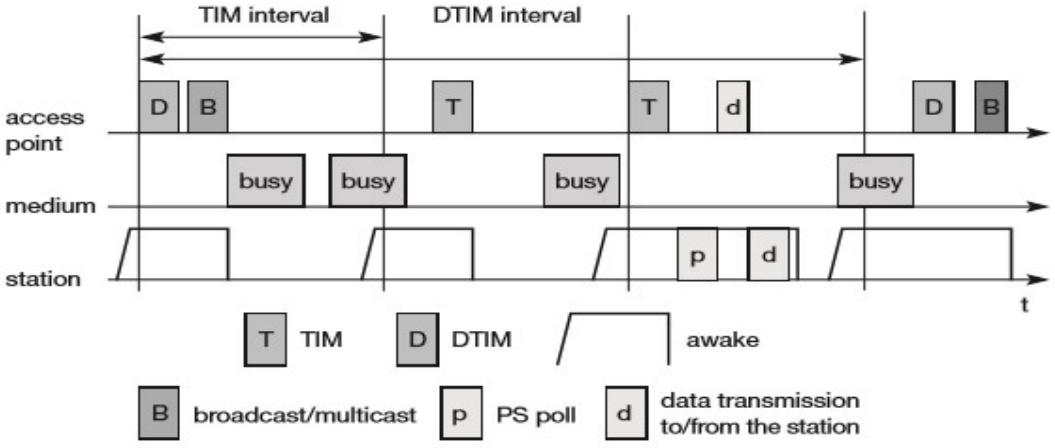


Fig. 20

All stations (in the example, only one is shown) wake up prior to an expected TIM or DTIM. In the first case, the access point has to transmit a broadcast frame and the station stays awake to receive it. After receiving the broadcast frame, the station returns to sleeping mode. The station wakes up again just before the next TIM transmission. This time the TIM is delayed due to a

busy medium so, the station stays awake. The access point has nothing to send and the station goes back to sleep. At the next TIM interval, the access point indicates that the station is the destination for a buffered frame. The station answers with a PS (power saving) poll and stays awake to receive data. The access point then transmits the data for the station, the station acknowledges the receipt and may also send some data (as shown in the example). This is acknowledged by the access point (acknowledgments are not shown in the figure). Afterwards, the station switches to sleep mode again.

Finally, the access point has more broadcast data to send at the next DTIM interval, which is again deferred by a busy medium. Depending on internal thresholds, a station may stay awake if the sleeping period would be too short. This mechanism clearly shows the trade-off between short delays in station access and saving battery power. The shorter the TIM interval, the shorter the delay, but the lower the power-saving effect.

In ad-hoc networks, power management is much more complicated than in infrastructure networks. In this case, there is no access point to buffer data in one location but each station needs the ability to buffer data if it wants to communicate with a power-saving station. All stations now announce a list of buffered frames during a period when they are all awake. Destinations are announced using ad-hoc traffic indication map (ATIMs) – the announcement period is called the ATIM window.

Figure 21 shows a simple ad-hoc network with two stations. Again, the beacon interval is determined by a distributed function (different stations may send the beacon). However, due to this synchronization, all stations within the ad-hoc network wake up at the same time. All stations stay awake for the ATIM interval as shown in the first two steps and go to sleep again if no frame is buffered for them. In the third step, station1 has data buffered for station2. This is indicated in an ATIM transmitted by station1. Station2 acknowledges this ATIM and stays awake for the transmission. After the ATIM window, station1 can transmit the data frame, and station2 acknowledges its receipt. In this case, the stations stay awake for the next beacon.

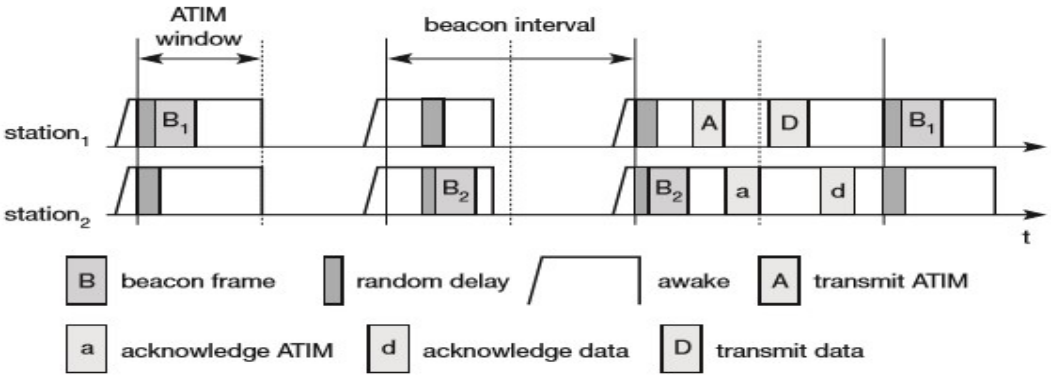


Fig. 21

One problem with this approach is that of scale. If many stations within an ad-hoc network operate in power-save mode, they may also want to transmit their ATIM within the ATIM window. More ATIM transmissions take place, more collisions happen and more stations are deferred. The access delay of large networks is difficult to predict. QoS guarantees can not be given under heavy load.

2.3.5.3 Roaming

Typically, wireless networks within buildings require more than just one access point to cover all rooms. Depending on the solidity and material of the walls, one access point has a transmission range of 10–20 m if transmission is to be of decent quality. Each storey of a building needs its own access point(s) as quite often walls are thinner than floors. If a user walks around with a wireless station, the station has to move from one access point to another to provide uninterrupted service. Moving between access points is called roaming. The term "handover" or "handoff" as used in the context of mobile or cellular phone systems would be more appropriate as it is simply a change of the active cell. However, for WLANs roaming is more common.

The steps for roaming between access points are:

● A station decides that the current link quality to its access point AP1 is too poor. The station then starts scanning for another access point.

● Scanning involves the active search for another BSS and can also be used for setting up a new BSS in case of ad-hoc networks. IEEE 802.11 specifies scanning on single or multiple channels (if available at the physical layer) and differentiates between passive scanning and active scanning. Passive scanning simply means listening into the medium to find other networks, i.e., receiving the beacon of another network issued by the synchronization function within an access point. Active scanning comprises sending a probe on each channel and waiting for a response. Beacon and probe responses contain the information necessary to join the new BSS.

● The station then selects the best access point for roaming based on, e.g., signal strength, and sends an association request to the selected access point AP2.

● The new access point AP2 answers with an association response. If the response is successful, the station has roamed to the new access point AP2. Otherwise, the station has to continue scanning for new access points.

● The access point accepting an association request indicates the new station in its BSS to the distribution system (DS). The DS then updates its database, which contains the current location of the wireless stations. This database is needed for forwarding frames between different BSSs, i.e. between the different access points controlling the BSSs, which combine to form an ESS (see Figure 3). Additionally, the DS can inform the old access point AP1 that the station is no longer within its BSS.

Unfortunately, many products implemented proprietary or incompatible versions of protocols that support roaming and inform the old access point about the change in the station's location. The standard IEEE 802.11f (Inter Access Point Protocol, IAPP) should provide a compatible solution for all vendors. This also includes load-balancing between access points and key generation for security algorithms based on IEEE 802.1x (IEEE, 2001).

2.3.6 802.11b

As standardization took some time, the capabilities of the physical layers also evolved. Soon after the first commercial 802.11 products came on the market some companies offered proprietary solutions with 11 Mbit/s. To avoid market segmentation, a common standard, IEEE 802.11b (IEEE 1999) soon followed and was added as supplement to the original standard

(Higher-speed physical layer extension in the 2.4 GHz band). This standard describes a new PHY layer and is by far the most successful version of IEEE 802.11 available today. Do not get confused about the fact that 802.11b hit the market before 802.11a. The standards are named according to the order in which the respective study groups have been established. As the name of the supplement implies, this standard only defines a new PHY layer. All the MAC schemes, management procedures etc. explained above are still used. Depending on the current interference and the distance between sender and receiver 802.11b systems offer 11, 5.5, 2, or 1 Mbit/s. Maximum user data rate is approx 6 Mbit/s. The lower data rates 1 and 2 Mbit/s use the 11-chip Barker sequence and DBPSK or DQPSK, respectively. The new data rates, 5.5 and 11 Mbit/s, use 8-chip complementary code keying (CCK) (see IEEE, 1999, or Pahlavan, 2002, for details).

The standard defines several packet formats for the physical layer. The mandatory format interoperates with the original versions of 802.11. The optional versions provide a more efficient data transfer due to shorter headers/different coding schemes and can coexist with other 802.11 versions. However, the standard states that control all frames shall be transmitted at one of the basic rates, so they will be understood by all stations in a BSS.

Figure 22 shows two packet formats standardized for 802.11b. The mandatory format is called long PLCP PPDU and is similar to the format illustrated in Figure 8. One difference is the rate encoded in the signal field this is encoded in multiples of 100 kbit/s. Thus, 0x0A represents 1 Mbit/s, 0x14 is used for 2 Mbit/s, 0x37 for 5.5 Mbit/s and 0x6E for 11 Mbit/s. Note that the preamble and the header are transmitted at 1 Mbit/s using DBPSK. The optional short PLCP PPDU format differs in several ways. The short synchronization field consists of 56 scrambled zeros instead of scrambled ones. The short start frame delimiter SFD consists of a mirrored bit pattern compared to the SFD of the long format: 0000 0101 1100 1111 is used for the short PLCP PDU instead of 1111 0011 1010 0000 for the long PLCP PPDU. Receivers that are unable to receive the short format will not detect the start of a frame (but will sense the medium is busy). Only the preamble is transmitted at 1 Mbit/s, DBPSK. The following header is already transmitted at 2 Mbit/s, DQPSK, which is also the lowest available data rate. As Figure 22 shows, the length of the overhead is only half for the short frames (96 µs instead of 192 µs). This is useful for, e.g., short, but timecritical, data transmissions.

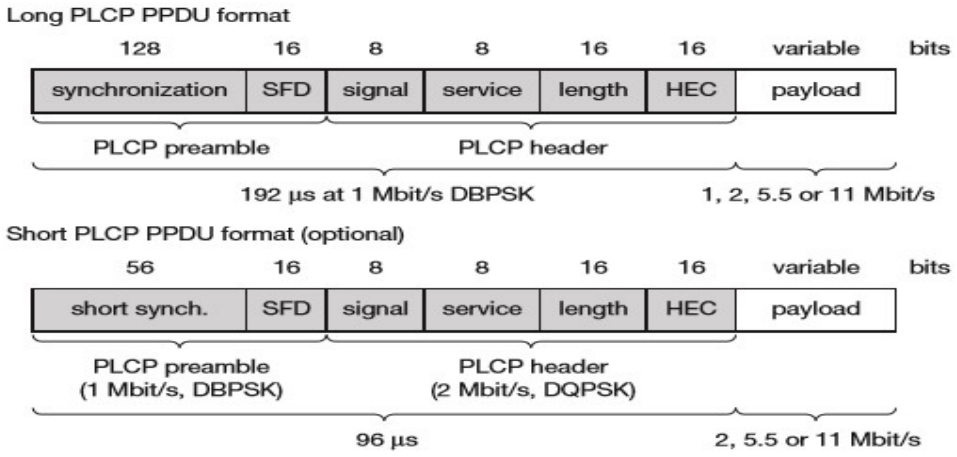


Fig. 22

As IEEE 802.11b is the most widespread version, some more information is given for practical usage. The standards operates (like the DSSS version of 802.11) on certain frequencies in the 2.4 GHz ISM band. These depend on national regulations. Altogether 14 channels have been defined as Table 2 shows. For each channel the center frequency is given. Depending on national restrictions 11 (US/Canada), 13 (Europe with some exceptions) or 14 channels (Japan) can be used.

Figure 23 illustrates the non-overlapping usage of channels for an IEEE 802.11b installation with minimal interference in the US/Canada and Europe. The spacing between the center frequencies should be at least 25 MHz (the occupied bandwidth of the main lobe of the signal is 22 MHz). This results in the channels 1, 6, and 11 for the US/Canada or 1, 7, 13 for Europe, respectively. It may be the case that, e.g., travellers from the US cannot use the additional channels (12 and 13) in Europe as their hardware is limited to 11 channels. Some European installations use channel 13 to minimize interference. Users can install overlapping cells for WLANs using the three non-overlapping channels to provide seamless coverage. This is similar to the cell planning for mobile phone systems.

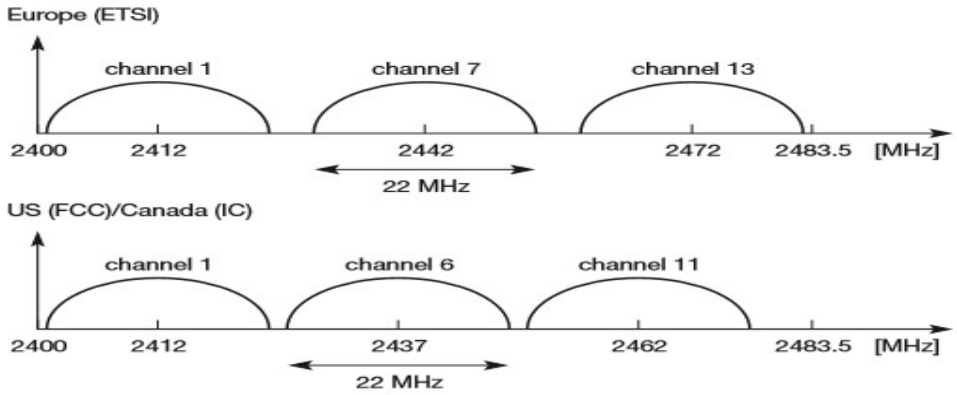


Fig. 23

Channel	Frequency [MHz]	US/Canada	Europe	Japan
1	2412	X	X	X
2	2417	X	X	X
3	2422	X	X	X
4	2427	X	X	X
5	2432	X	X	X
6	2437	X	X	X
7	2442	X	X	X
8	2447	X	X	X
9	2452	X	X	X
10	2457	X	X	X
11	2462	X	X	X
12	2467	–	X	X
13	2472	–	X	X
14	2484	–	–	X

Table 2

2.3.7 802.11a

Initially aimed at the US 5 GHz U-NII (Unlicensed National Information Infrastructure) bands IEEE 802.11a offers up to 54 Mbit/s using OFDM (IEEE, 1999). The first products were available in 2001 and can now be used (after some harmonization between IEEE and ETSI) in Europe. The FCC (US) regulations offer three different 100 MHz domains for the use of 802.11a, each with a different legal maximum power output: 5.15–5.25 GHz/50 mW, 5.25–5.35 GHz/250 mW, and 5.725–5.825 GHz/1 W. ETSI (Europe) defines different frequency bands for Europe: 5.15–5.35 GHz and 5.47–5.725 GHz and requires two additional mechanisms for operation: dynamic frequency selection (DFS) and transmit power control (TPC). (This is also the reason for introducing IEEE 802.11h) Maximum transmit power is 200 mW EIRP for the lower frequency band (indoor use) and 1 W EIRP for the higher frequency band (indoor and outdoor use). DFS and TPC are not necessary, if the transmit power stays below 50 mW EIRP and only 5.15–5.25 GHz are used. Japan allows operation in the frequency range 5.15–5.25 GHz and requires carrier sensing every 4 ms to minimize interference. Up to now, only 100 MHz are available 'worldwide' at 5.15–5.25 GHz.

The physical layer of IEEE 802.11a and the ETSI standard HiperLAN2 has been jointly developed, so both physical layers are almost identical. Most statements and explanations in the following, which are related to the transmission technology are also valid for HiperLAN2. However, HiperLAN2 differs in the MAC layer, the PHY layer packet formats, and the offered services (quality of service, real time etc.). It should be noted that most of the development for the physical layer for 802.11a was adopted from the HiperLAN2 standardization – but 802.11a products were available first and are already in widespread use. Again, IEEE 802.11a uses the same MAC layer as all 802.11 physical layers do and, in the following, only the lowest layer is explained in some detail. To be able to offer data rates up to 54 Mbit/s IEEE 802.11a uses many different technologies. The system uses 52 subcarriers (48 data + 4 pilot) that are modulated using BPSK, QPSK, 16-QAM, or 64-QAM. To mitigate transmission errors, FEC is applied using coding rates of 1/2, 2/3, or 3/4. Table 3 gives an overview of the standardized combinations of modulation and coding schemes together with the resulting data rates. To offer a data rate of 12 Mbit/s, 96 bits are coded into one OFDM symbol. These 96 bits are distributed over 48 subcarriers and 2 bits are modulated per sub-carrier using QPSK (2 bits per point in the constellation diagram). Using a coding rate of 1/2 only 48 data bits can be transmitted.

Data rate [Mbit/s]	Modulation	Coding rate	Coded bits per subcarrier	Coded bits per OFDM symbol	Data bits per OFDM symbol
6	BPSK	1/2	1	48	24
9	BPSK	3/4	1	48	36
12	QPSK	1/2	2	96	48
18	QPSK	3/4	2	96	72
24	16-QAM	1/2	4	192	96
36	16-QAM	3/4	4	192	144
48	64-QAM	2/3	6	288	192
54	64-QAM	3/4	6	288	216

Table 3

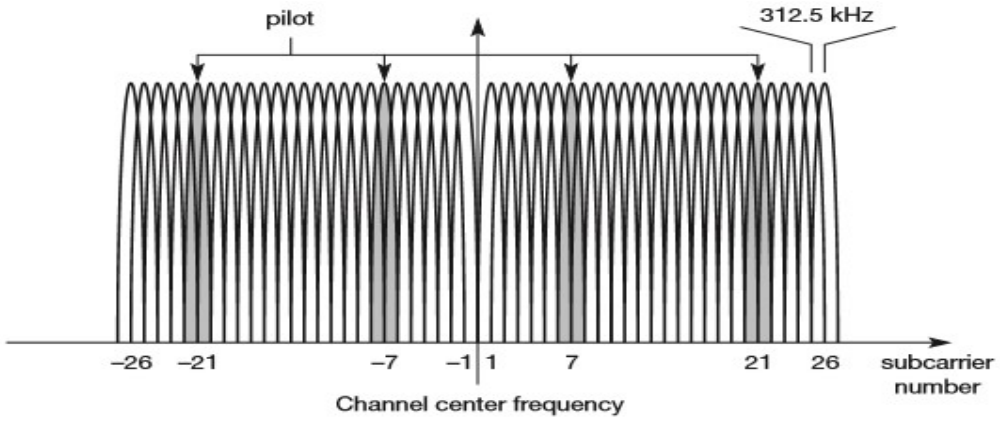


Fig. 24

Figure 24 shows the usage of OFDM in IEEE 802.11a. Remember, the basic idea of OFDM (or MCM in general) was the reduction of the symbol rate by distributing bits over numerous subcarriers. IEEE 802.11a uses a fixed symbol rate of 250,000 symbols per second independent of the data rate (0.8 µs guard interval for ISI mitigation plus 3.2 µs used for data results in a symbol duration of 4 µs). As Figure 24 shows, 52 subcarriers are equally spaced around a center frequency. (Center frequencies will be explained later). The spacing between the subcarriers is 312.5 kHz. 26 subcarriers are to the left of the center frequency and 26 are to the right. The center frequency itself is not used as subcarrier. Subcarriers with the numbers –21, –7, 7, and 21 are used for pilot signals to make the signal detection robust against frequency offsets.

Similar to 802.11b several operating channels have been standardized to minimize interference. Figure 25 shows the channel layout for the US U-NII bands. The center frequency of a channel is 5000 + 5*channel number [MHz]. This definition provides a unique numbering of channels with 5 MHz spacing starting from 5 GHz. Depending on national regulations, different sets of channels may be used. Eight channels have been defined for the lower two bands in the U-NII (36, 40, 44, 48, 52, 56, 60, and 64); four more are available in the high band (149, 153, 157, and 161). Using these channels allows for interference-free operation of overlapping 802.11a cells. Channel spacing is 20 MHz, the occupied bandwidth of 802.11a is 16.6 MHz. How is this related to the spacing of the sub-carriers? 20 MHz/64 equals 312.5 kHz. 802.11a uses 48 carriers for data, 4 for pilot signals, and 12 carriers are sometimes called virtual subcarriers. (Set to zero, they do not contribute to the data transmission but may be used for an implementation of OFDM with the help of FFT, see IEEE, 1999, or ETSI, 2001a, for more details). Multiplying 312.5 kHz by 52 subcarriers and adding the extra space for the center frequency results in approximately 16.6 MHz occupied bandwidth per channel (details of the transmit spectral power mask neglected, see ETSI, 2001a).

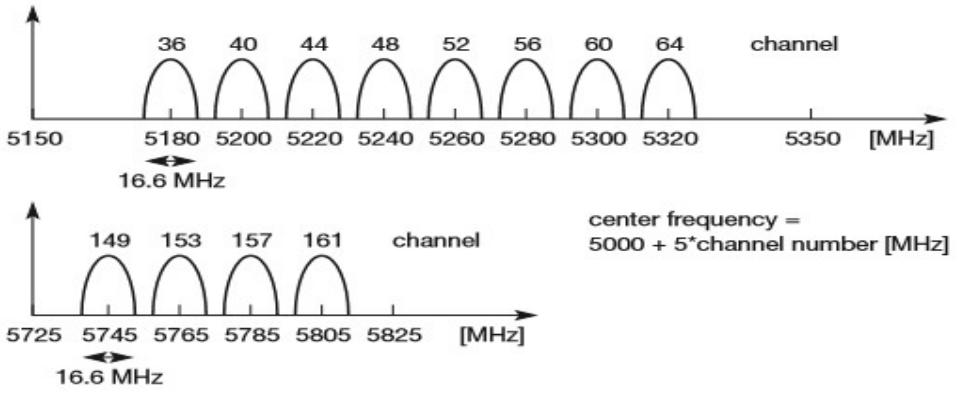


Fig. 25

Due to the nature of OFDM, the PDU on the physical layer of IEEE 802.11a looks quite different from 802.11b or the original 802.11 physical layers. Figure 26 shows the basic structure of an IEEE 802.11a PPDU.

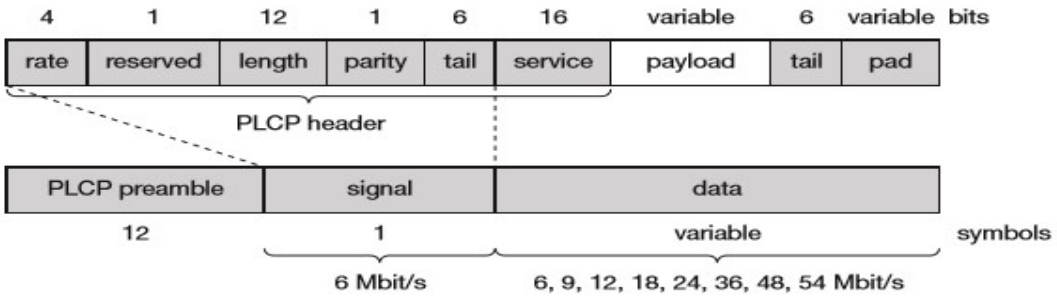


Fig. 26

● The PLCP preamble consists of 12 symbols and is used for frequency acquisition, channel estimation, and synchronization. The duration of the preamble is 16 µs.

● The following OFDM symbol, called signal, contains the following fields and is BPSK-modulated. The 4 bit rate field determines the data rate and the modulation of the rest of the packet (examples are 0x3 for 54 Mbit/s, 0x9 for 24 Mbit/s, or 0xF for 9 Mbit/s). The length field indicates the number of bytes in the payload field. The parity bit shall be an even parity for the first 16 bits of the signal field (rate, length and the reserved bit). Finally, the six tail bits are set to zero.

● The data field is sent with the rate determined in the rate field and contains a service field which is used to synchronize the descrambler of the receiver (the data stream is scrambled using the polynomial x7 + x4 + 1) and which contains bits for future use. The payload contains the MAC PDU (1-4095 byte). The tail bits are used to reset the encoder. Finally, the pad field ensures that the number of bits in the PDU maps to an integer number of OFDM symbols.

Compared to IEEE 802.11b working at 2.4 GHz IEEE 802.11a at 5 GHz offers much higher data rates. However, shading at 5 GHz is much more severe compared to 2.4 GHz and depending on the SNR, propagation conditions and the distance between sender and receiver, data rates may drop fast (e.g., 54 Mbit/s may be available only in an LOS or near LOS condition). Additionally,

the MAC layer of IEEE 802.11 adds overheads. User data rates are therefore much lower than the data rates listed above. Typical user rates in Mbit/s are (transmission rates in brackets) 5.3 (6), 18 (24), 24 (36), and 32 (54). The following section presents some additional developments in the context of 802.11, which also comprise a standard for higher data rates at 2.4 GHz that can benefit from the better propagation conditions at lower frequencies.

2.4 Bluetooth

The Bluetooth technology discussed here aims at so-called ad-hoc piconets, which are local area networks with a very limited coverage and without the need for an infrastructure. This is a different type of network is needed to connect different small devices in close proximity (about 10 m) without expensive wiring or the need for a wireless infrastructure (Bisdikian, 1998). The envisaged gross data rate is 1 Mbit/s, asynchronous (data) and synchronous (voice) services should be available. The necessary transceiver components should be cheap – the goal is about €5 per device. (In 2002, separate adapters are still at €50, however, the additional cost of the devices integrated in, e.g., PDAs, almost reached the target.) Many of today's devices offer an infra red data association (IrDA) interface with transmission rates of, e.g., 115 kbit/s or 4 Mbit/s. There are various problems with IrDA: its very limited range (typically 2 m for built-in interfaces), the need for a line-of-sight between the interfaces, and, it is usually limited to two participants, i.e., only point-to-point connections are supported. IrDA has no internet working functions, has no media access, or any other enhanced communication mechanisms. The big advantage of IrDA is its low cost, and it can be found in almost any mobile device (laptops, PDAs, mobile phones).

The history of Bluetooth starts in the tenth century, when Harald Gormsen, King of Denmark (son of Gorm), erected a rune stone in Jelling, Denmark, in memory of his parents. The stone has three sides with elaborate carvings. One side shows a picture of Christ, as Harald did not only unite Norway and Denmark, but also brought Christianity to Scandinavia. Harald had the common epithet of 'Blåtand', meaning that he had a rather dark complexion (not a blue tooth).

It took a thousand years before the Swedish IT-company Ericsson initiated some studies in 1994 around a so-called multi-communicator link (Haartsen, 1998). The project was renamed (because a friend of the designers liked the Vikings) and Bluetooth was born. In spring 1998 five companies (Ericsson, Intel, IBM, Nokia, Toshiba) founded the Bluetooth consortium with the goal of developing a single-chip, low-cost, radio-based wireless network technology. Many other companies and research institutions joined the special interest group around Bluetooth (2002), whose goal was the development of mobile phones, laptops, notebooks, headsets etc. including Bluetooth technology, by the end of 1999. In 1999, Ericsson erected a rune stone in Lund, Sweden, in memory of Harald Gormsen, called Blåtand, who gave his epithet for this new wireless communication technology. This new carving shows a man holding a laptop and a cellular phone, a picture which is quite often cited (of course there are no such things visible on the original stone, that's just a nice story!)

In 2001, the first products hit the mass market, and many mobile phones, laptops, PDAs, video cameras etc. are equipped with Bluetooth technology today. At the same time the Bluetooth development started, a study group within IEEE 802.11 discussed wireless personal area networks (WPAN) under the following five criteria:

- Market potential: How many applications, devices, vendors, customers are available for a certain technology?

● Compatibility: Compatibility with IEEE 802.

● Distinct identity: Originally, the study group did not want to establish a second 802.11 standard. However, topics such as, low cost, low power, or small form factor are not addressed in the 802.11 standard.

● Technical feasibility: Prototypes are necessary for further discussion, so the study group would not rely on paper work.

● Economic feasibility: Everything developed within this group should be cheaper than other solutions and allow for high-volume production.

Obviously, Bluetooth fulfills these criteria so the WPAN group cooperated with the Bluetooth consortium. IEEE founded its own group for WPANs, IEEE 802.15, in March 1999. This group should develop standards for wireless communications within a personal operating space (POS, IEEE, 2002c). A POS has been defined as a radius of 10 m around a person in which the person or devices of this person communicate with other devices.

2.4.1 User scenarios

Many different user scenarios can be imagined for wireless piconets or WPANs:

● **Connection of peripheral devices:** Today, most devices are connected to a desktop computer via wires (e.g., keyboard, mouse, joystick, headset, speakers). This type of connection has several disadvantages: each device has its own type of cable, different plugs are needed, wires block office space. In a wireless network, no wires are needed for data transmission. However, batteries now have to replace the power supply, as the wires not only transfer data but also supply the peripheral devices with power.

● **Support of ad-hoc networking:** Imagine several people coming together, discussing issues, exchanging data (schedules, sales figures etc.). For instance, students might join a lecture, with the teacher distributing data to their personal digital assistants (PDAs). Wireless networks can support this type of interaction; small devices might not have WLAN adapters following the IEEE 802.11 standard, but cheaper Bluetooth chips built in.

● **Bridging of networks:** Using wireless piconets, a mobile phone can be connected to a PDA or laptop in a simple way. Mobile phones will not have full WLAN adapters built in, but could have a Bluetooth chip. The mobile phone can then act as a bridge between the local piconet and, e.g., the global GSM network (see Figure 27). For instance, on arrival at an airport, a person's mobile phone could receive e-mail via GSM and forward it to the laptop which is still in a suitcase. Via a piconet, a file server could update local information stored on a laptop or PDA while the person is walking into the office.

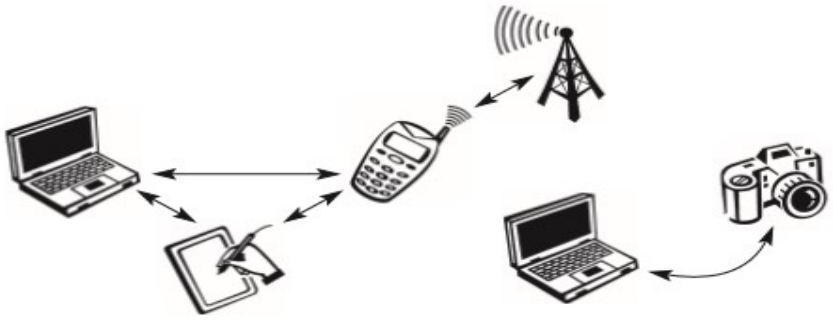


Fig. 27

When comparing Bluetooth with other WLAN technology we have to keep in mind that one of its goals was to provide local wireless access at very low cost. From a technical point of view, WLAN technologies like those above could also be used, however, WLAN adapters, e.g., for IEEE 802.11, have been designed for higher bandwidth and larger range and are more expensive and consume a lot more power.

2.4.2 Architecture

Like IEEE 802.11b, Bluetooth operates in the 2.4 GHz ISM band. However, MAC, physical layer and the offered services are completely different. After presenting the overall architecture of Bluetooth and its specialty, the piconets, the following sections explain all protocol layers and components in more detail.

2.4.2.1 Networking

To understand the networking of Bluetooth devices a quick introduction to its key features is necessary. Bluetooth operates on 79 channels in the 2.4 GHz band with 1 MHz carrier spacing. Each device performs frequency hopping with 1,600 hops/s in a pseudo random fashion. Bluetooth applies FHSS for interference mitigation (and FH-CDMA for separation of networks).

A very important term in the context of Bluetooth is a piconet. A piconet is a collection of Bluetooth devices which are synchronized to the same hopping sequence. Figure 28 shows a collection of devices with different roles. One device in the piconet can act as master (M), all other devices connected to the master must act as slaves (S). The master determines the hopping pattern in the piconet and the slaves have to synchronize to this pattern. Each piconet has a unique hopping pattern. If a device wants to participate it has to synchronize to this. Two additional types of devices are shown: parked devices (P) can not actively participate in the piconet (i.e., they do not have a connection), but are known and can be reactivated within some milliseconds (see section 7.5.5). Devices in stand-by (SB) do not participate in the piconet. Each piconet has exactly one master and up to seven simultaneous slaves. More than 200 devices can be parked. The reason for the upper limit of eight active devices, is the 3-bit address used in Bluetooth. If a parked device wants to communicate and there are already seven active slaves, one slave has to switch to park mode to allow the parked device to switch to active mode.

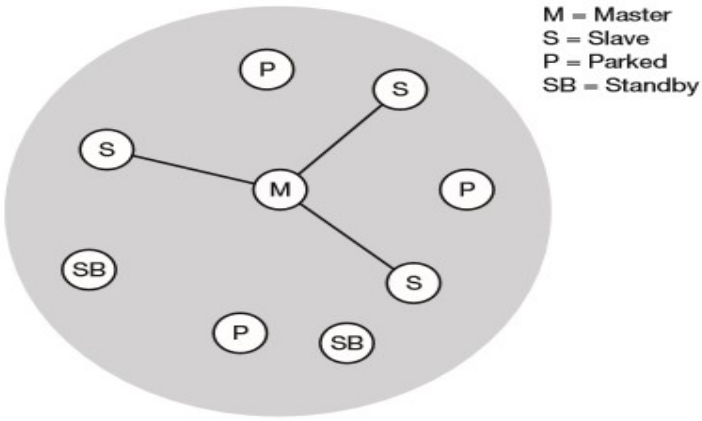


Fig. 28

Figure 29 gives an overview of the formation of a piconet. As all active devices have to use the same hopping sequence they must be synchronized. The first step involves a master sending its clock and device ID. All Bluetooth devices have the same networking capabilities, i.e., they can be master or slave. There is no distinction between terminals and base stations, any two or more devices can form a piconet. The unit establishing the piconet automatically becomes the master, all other devices will be slaves. The hopping pattern is determined by the device ID, a 48-bit worldwide unique identifier. The phase in the hopping pattern is determined by the master's clock. After adjusting the internal clock according to the master a device may participate in the piconet. All active devices are assigned a 3-bit active member address (AMA). All parked devices use an 8-bit parked member address (PMA). Devices in stand-by do not need an address.

All users within one piconet have the same hopping sequence and share the same 1 MHz channel. As more users join the piconet, the throughput per user drops quickly (a single piconet offers less than 1 Mbit/s gross data rate). (Only having one piconet available within the 80 MHz in total is not very efficient.) This led to the idea of forming groups of piconets called scatternet (see Figure 30). Only those units that really must exchange data share the same piconet, so that many piconets with overlapping coverage can exist simultaneously.

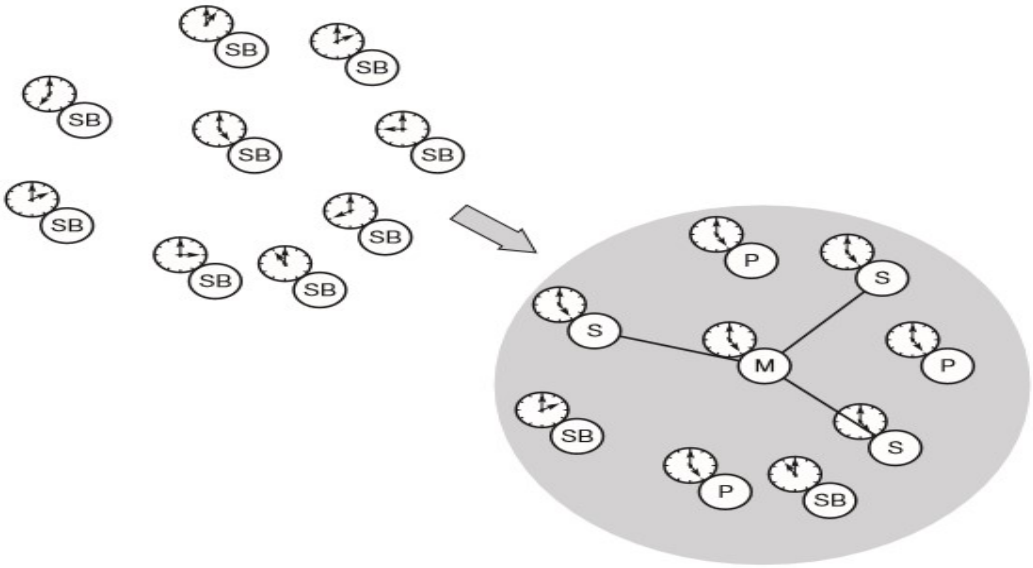


Fig. 29

In the example, the scatternet consists of two piconets, in which one device participates in two different piconets. Both piconets use a different hopping sequence, always determined by the master of the piconet. Bluetooth applies FH-CDMA for separation of piconets. In an average sense, all piconets can share the total of 80 MHz bandwidth available. Adding more piconets leads to a graceful performance degradation of a single piconet because more and more collisions may occur. A collision occurs if two or more piconets use the same carrier frequency at the same time. This will probably happen as the hopping sequences are not coordinated. If a device wants to participate in more than one piconet, it has to synchronize to the hopping sequence of the piconet it wants to take part in. If a device acts as slave in one piconet, it simply starts to synchronize with the hopping sequence of the piconet it wants to join. After synchronization, it acts as a slave in this piconet and no longer participates in its former piconet. To enable synchronization, a slave has to know the identity of the master that determines the hopping sequence of a piconet. Before leaving one piconet, a slave informs the current master that it will be unavailable for a certain amount of time. The remaining devices in the piconet continue to communicate as usual.

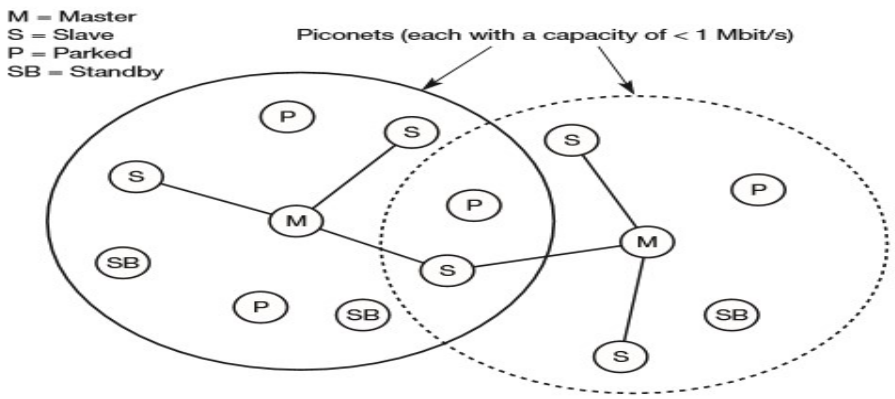


Fig. 30

A master can also leave its piconet and act as a slave in another piconet. It is clearly not possible for a master of one piconet to act as the master of another piconet as this would lead to identical behavior (both would have the same hopping sequence, which is determined by the master per definition). As soon as a master leaves a piconet, all traffic within this piconet is suspended until the master returns. Communication between different piconets takes place by devices jumping back and forth between theses nets. If this is done periodically, for instance, isochronous data streams can be forwarded from one piconet to another. However, scatternets are not yet supported by all devices.

2.4.2.2 Protocol stack

As Figure 31 shows, the Bluetooth specification already comprises many protocols and components. Starting as a simple idea, it now covers over 2,000 pages dealing with not only the Bluetooth protocols but many adaptation functions and enhancements. The Bluetooth protocol stack can be divided into a core specification (Bluetooth, 2001a), which describes the protocols from physical layer to the data link control together with management functions, and profile specifications (Bluetooth, 2001b). The latter describes many protocols and functions needed to adapt the wireless Bluetooth technology to legacy and new applications (see section 2.5.9). The core protocols of Bluetooth comprise the following elements:

● Radio: Specification of the air interface, i.e., frequencies, modulation, and transmit power .

● Baseband: Description of basic connection establishment, packet formats, timing, and basic QoS parameters.

● Link manager protocol: Link set-up and management between devices including security functions and parameter negotiation.

● Logical link control and adaptation protocol (L2CAP): Adaptation of higher layers to the baseband.

● Service discovery protocol: Device discovery in close proximity plus querying of service characteristics.

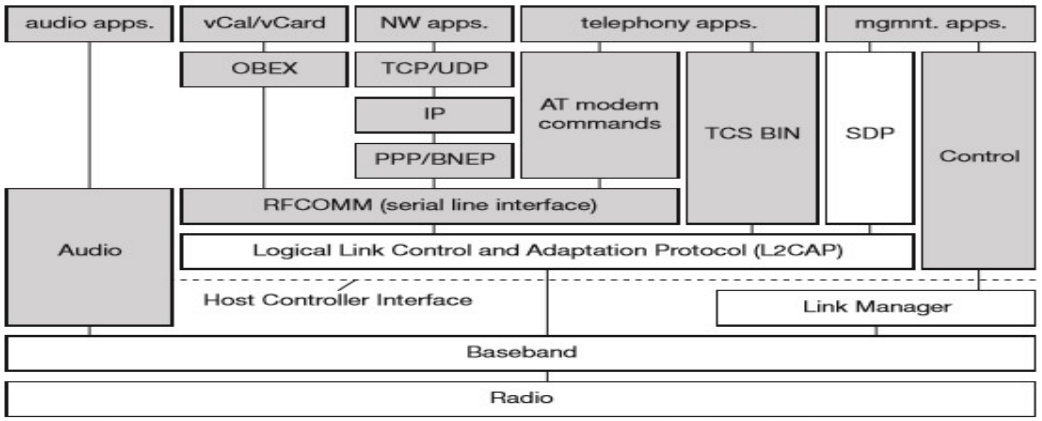


Fig. 31

On top of L2CAP is the cable replacement protocol RFCOMM that emulates a serial line interface following the EIA-232 (formerly RS-232) standards. This allows for a simple replacement of serial line cables and enables many legacy applications and protocols to run over Bluetooth. RFCOMM supports multiple serial ports over a single physical channel. The telephony control protocol specification – binary (TCS BIN) describes a bit-oriented protocol that defines call control signaling for the establishment of voice and data calls between Bluetooth devices. It also describes mobility and group management functions.

The host controller interface (HCI) between the baseband and L2CAP provides a command interface to the baseband controller and link manager, and access to the hardware status and control registers. The HCI can be seen as the hardware/software boundary.

Many protocols have been adopted in the Bluetooth standard. Classical Internet applications can still use the standard TCP/IP stack running over PPP or use the more efficient Bluetooth network encapsulation protocol (BNEP). Telephony applications can use the AT modem commands as if they were using a standard modem. Calendar and business card objects (vCalendar/vCard) can be exchanged using the object exchange protocol (OBEX) as common with IrDA interfaces. A real difference to other protocol stacks is the support of audio. Audio applications may directly use the baseband layer after encoding the audio signals.

2.4.3 Radio layer

The radio specification is a rather short document (less than ten pages) and only defines the carrier frequencies and output power. Several limitations had to be taken into account when Bluetooth's radio layer was designed. Bluetooth devices will be integrated into typical mobile devices and rely on battery power. This requires small, low power chips which can be built into handheld devices. Worldwide operation also requires a frequency which is available worldwide. The combined use for data and voice transmission has to be reflected in the design, i.e., Bluetooth has to support multi-media data.

Bluetooth uses the license-free frequency band at 2.4 GHz allowing for worldwide operation with some minor adaptations to national restrictions. A frequency-hopping/time-division duplex scheme is used for transmission, with a fast hopping rate of 1,600 hops per second. The time between two hops is called a slot, which is an interval of 625 µs. Each slot uses a different frequency. Bluetooth uses 79 hop carriers equally spaced with 1 MHz. After worldwide harmonization, Bluetooth devices can be used (almost) anywhere.

Bluetooth transceivers use Gaussian FSK for modulation and are available in three classes:

● Power class 1: Maximum power is 100 mW and minimum is 1 mW (typ. 100 m range without obstacles). Power control is mandatory.

● Power class 2: Maximum power is 2.5 mW, nominal power is 1 mW, and minimum power is 0.25 mW (typ. 10 m range without obstacles). Power control is optional.

● Power class 3: Maximum power is 1 mW.

2.4.4 Baseband layer

The functions of the baseband layer are quite complex as it not only performs frequency hopping for interference mitigation and medium access, but also defines physical links and many packet formats. Figure 32 shows several examples of frequency selection during data transmission. Remember that each device participating in a certain piconet hops at the same time to the same carrier frequency (fi in Figure 32). If, for example, the master sends data at fk, then a slave may answer at fk+1. This scenario shows another feature of Bluetooth. TDD is used for separation of the transmission directions. The upper part of Figure 32 shows so-called 1-slot packets as the data transmission uses one 625 µs slot. Within each slot the master or one out of seven slaves may transmit data in an alternating fashion. The control of medium access will be described later. Bluetooth also defines 3-slot and 5-slot packets for higher data rates (multi-slot packets). If a master or a slave sends a packet covering three or five slots, the radio transmitter remains on the same frequency. No frequency hopping is performed within packets. After transmitting the packet, the radio returns to the frequency required for its hopping sequence. The reason for this is quite simple: not every slave might receive a transmission (hidden terminal problem) and it can not react on a multi-slot transmission. Those slaves not involved in the transmission will continue with the hopping sequence. This behavior is important so that all devices can remain synchronized, because the piconet is uniquely defined by having the same hopping sequence with the same phase. Shifting the phase in one device would destroy the piconet.

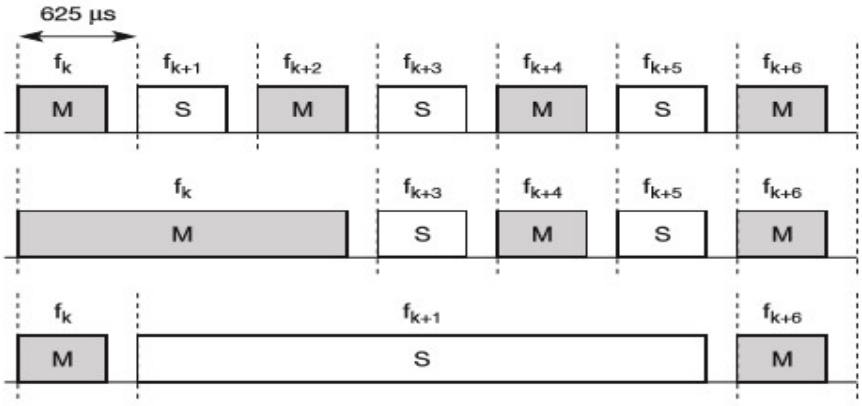


Fig. 32

Figure 33 shows the components of a Bluetooth packet at baseband layer. The packet typically consists of the following three fields:

● Access code: This first field of a packet is needed for timing synchronization and piconet identification (channel access code, CAC). It may represent special codes during paging (device access code, DAC) and inquiry. The access code consists of a 4 bit preamble, a synchronization field, and a trailer (if a packet header follows). The 64-bit synchronization field is derived from the lower 24 bit of an address (lower address part, LAP). If the access code is used for channel access (i.e., data transmission between a master and a slave or vice versa), the LAP is derived from the master's globally unique 48-bit address. In case of paging (DAC) the LAP of the paged device is used. If a Bluetooth device wants to discover other (arbitrary) devices in transmission range (general inquiry procedure) it uses a special reserved LAP. Special LAPs can be defined for inquiries of dedicated groups of devices.

● Packet header: This field contains typical layer 2 features: address, packet type, flow and error control, and checksum. The 3-bit active member address represents the active address of a slave.

Active addresses are temporarily assigned to a slave in a piconet. If a master sends data to a slave the address is interpreted as receiver address. If a slave sends data to the master the address represents the sender address. As only a master may communicate with a slave this scheme works well. Seven addresses may be used this way. The zero value is reserved for a broadcast from the master to all slaves. The 4-bit type field determines the type of the packet. Examples for packet types are given in Table 4. Packets may carry control, synchronous, or asynchronous data. A simple flow control mechanism for asynchronous traffic uses the 1-bit flow field. If a packet is received with flow=0 asynchronous data, transmission must stop. As soon as a packet with flow=1 is received, transmission may resume. If an acknowledgement of packets is required, Bluetooth sends this in the slot following the data (using its time division duplex scheme). A simple alternating bit protocol with a single bit sequence number SEQN and acknowledgement number ARQN can be used. An 8-bit header error check (HEC) is used to protect the packet header. The packet header is also protected by a one-third rate forward error correction (FEC) code because it contains valuable link information and should survive bit errors. Therefore, the 18-bit header requires 54 bits in the packet.

● Payload: Up to 343 bytes payload can be transferred. The structure of the payload field depends on the type of link and is explained in the following sections.

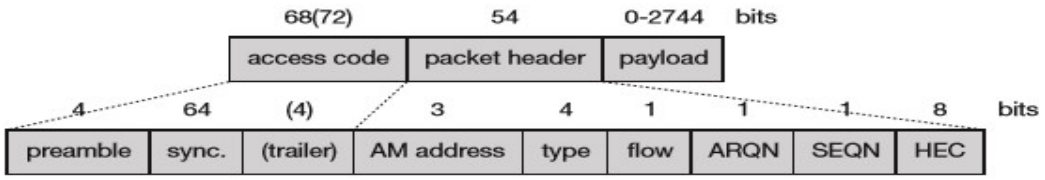


Fig. 33

Type	Payload header [byte]	User payload [byte]	FEC	CRC	Symmetric max. rate [kbit/s]	Asymmetric forward	Max. rate [kbit/s] reverse
DM1	1	0–17	2/3	yes	108.8	108.8	108.8
DH1	1	0–27	no	yes	172.8	172.8	172.8
DM3	2	0–121	2/3	yes	258.1	387.2	54.4
DH3	2	0–183	no	yes	390.4	585.6	86.4
DM5	2	0–224	2/3	yes	286.7	477.8	36.3
DH5	2	0–339	no	yes	433.9	723.2	57.6
AUX1	1	0–29	no	no	185.6	185.6	185.6
HV1	na	10	1/3	no	64.0	na	na
HV2	na	20	2/3	no	64.0	na	na
HV3	na	30	no	no	64.0	na	na
DV	1 D	10+ (0–9) D	2/3 D	yes D	64.0+ 57.6 D	na	na

Table 4

2.4.4.1 Physical links

Bluetooth offers two different types of links, a synchronous connection-oriented link and an asynchronous connectionless link:

● **Synchronous connection-oriented link (SCO):** Classical telephone (voice) connections require symmetrical, circuit-switched, point-to-point connections. For this type of link, the master reserves two consecutive slots (forward and return slots) at fixed intervals. A master can support up to three simultaneous SCO links to the same slave or to different slaves. A slave supports up to two links from different masters or up to three links from the same master. Using an SCO link, three different types of single-slot packets can be used (Figure 34). Each SCO link carries voice at 64 kbit/s, and no forward error correction (FEC), 2/3 FEC, or 1/3 FEC can be selected. The 1/3 FEC is as strong as the FEC for the packet header and triples the amount of data. Depending on the error rate of the channel, different FEC schemes can be applied. FEC always causes an overhead, but avoids retransmission of data with a higher probability. However, voice data over an SCO is never retransmitted. Instead, a very robust voice-encoding scheme, continuous variable slope delta (CVSD), is applied (Haartsen, 1998).

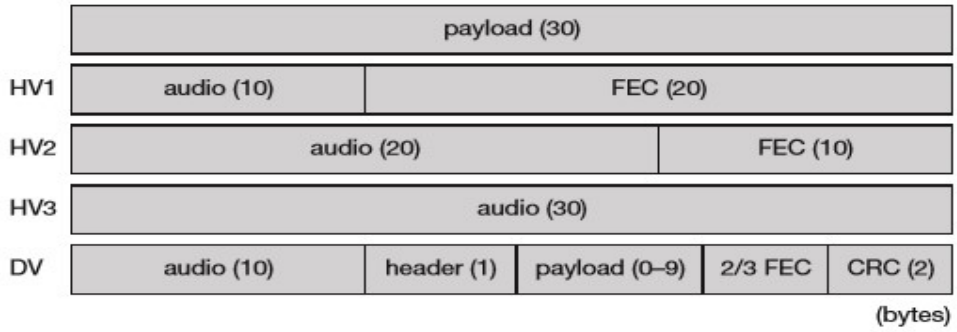


Fig. 34

● Asynchronous connectionless link (ACL): Typical data applications require symmetrical or asymmetrical (e.g., web traffic), packet-switched, point-to-multipoint transfer scenarios (including broadcast). Here the master uses a polling scheme. A slave may only answer if it has been addressed in the preceding slot. Only one ACL link can exist between a master and a slave. For ACLs carrying data, 1-slot, 3-slot or 5-slot packets can be used (Figure 35). Additionally, data can be protected using a 2/3 FEC scheme. This FEC protection helps in noisy environments with a high link error rate. However, the overhead introduced by FEC might be too high. Bluetooth therefore offers a fast automatic repeat request (ARQ) scheme for reliable transmission. The payload header (1 byte for 1-slot packets, 2 bytes for multi-slot packets) contains an identifier for a logical channel between L2CAP entities, a flow field for flow control at L2CAP level, and a length field indicating the number of bytes of data in the payload, excluding payload header and CRC. Payload is always CRC protected except for the AUX1 packet.

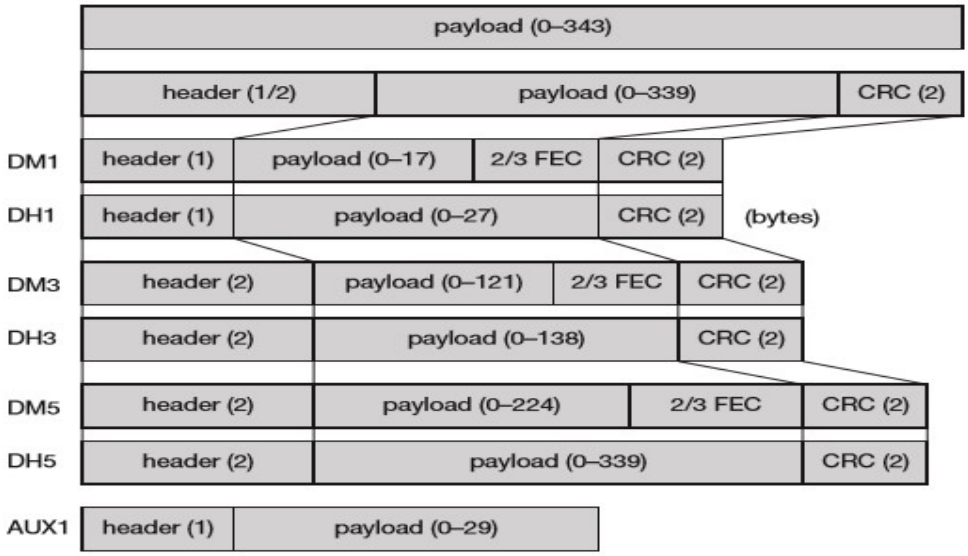


Fig. 35

Table 4 lists Bluetooth's ACL and SCO packets. Additionally, control packets are available for polling slaves, hopping synchronization, or acknowledgement. The ACL types DM1 (data medium rate) and DH1 (data high rate) use a single slot and a one byte header. DM3 and DH3 use three slots, DM5 and DH5 use five slots. Medium rates are always FEC protected, the high rates rely on CRC only for error detection. The highest available data rates for Bluetooth devices are 433.9 kbit/s (symmetric) or 723.3/57.6 kbit/s (asymmetric). High quality voice (HV) packets always use a single slot but differ with respect to the amount of redundancy for FEC. DV (data and voice) is a combined packet where CRC, FEC, and payload header are valid for the data part only.

Figure 36 shows an example transmission between a master and two slaves. The master always uses the even frequency slots, the odd slots are for the slaves. In this example every sixth slot is used for an SCO link between the master and slave 1. The ACL links use single or multiple slots providing asymmetric bandwidth for connectionless packet transmission. This example again shows the hopping sequence which is independent of the transmission of packets. The robustness of Bluetooth data transmissions is based on several technologies. FH-CDMA separates different piconets within a scatternet. FHSS mitigates interference from other devices operating in the 2.4 GHz ISM band. Additionally, FEC can be used to correct transmission errors. Bluetooth's 1/3 FEC simply sends three copies of each bit. The receiver then performs a majority decision: each received triple of bits is mapped into whichever bit is in majority. This simple scheme can correct all single bit errors in these triples. The 2/3 FEC encoding detects all double errors and can correct all single bit errors in a codeword.

ACL links can additionally be protected using an ARQ scheme and a checksum. Each packet can be acknowledged in the slot following the packet. If a packet is lost, a sender can retransmit it immediately in the next slot after the negative acknowledgement, so it is called a fast ARQ scheme. This scheme hardly exhibits any overheads in environments with low error rates, as only packets which are lost or destroyed have to be retransmitted. Retransmission is triggered by a negative acknowledgement or a time-out.

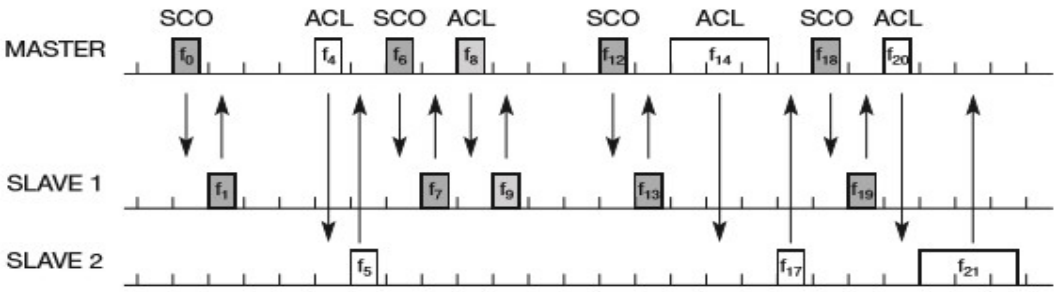


Fig. 36

2.4.5 Link manager protocol

The link manager protocol (LMP) manages various aspects of the radio link between a master and a slave and the current parameter setting of the devices. LMP enhances baseband functionality, but higher layers can still directly access the baseband. The following groups of functions are covered by the LMP:

● Authentication, pairing, and encryption: Although basic authentication is handled in the baseband, LMP has to control the exchange of random numbers and signed responses. The pairing service is needed to establish an initial trust relationship between two devices that have never communicated before. The result of pairing is a link key. This may be changed, accepted or rejected. LMP is not directly involved in the encryption process, but sets the encryption mode (no encryption, point-to-point, or broadcast), key size, and random speed.

● Synchronization: Precise synchronization is of major importance within a Bluetooth network. The clock offset is updated each time a packet is received from the master. Additionally, special synchronization packets can be received. Devices can also exchange timing information related to the time differences (slot boundaries) between two adjacent piconets.

● Capability negotiation: Not only the version of the LMP can be exchanged but also information about the supported features. Not all Bluetooth devices will support all features that are described in the standard, so devices have to agree the usage of, e.g., multi-slot packets, encryption, SCO links, voice encoding, park/sniff/hold mode (explained below), HV2/HV3 packets etc.

● Quality of service negotiation: Different parameters control the QoS of a Bluetooth device at these lower layers. The poll interval, i.e., the maximum time between transmissions from a master to a particular slave, controls the latency and transfer capacity. Depending on the quality of the channel, DM or DH packets may be used (i.e., 2/3 FEC protection or no protection). The number of repetitions for broadcast packets can be controlled. A master can also limit the number of slots available for slaves' answers to increase its own bandwidth.

● Power control: A Bluetooth device can measure the received signal strength. Depending on this signal level the device can direct the sender of the measured signal to increase or decrease its transmit power.

● Link supervision: LMP has to control the activity of a link, it may set up new SCO links, or it may declare the failure of a link.

● State and transmission mode change: Devices might switch the master/slave role, detach themselves from a connection, or change the operating mode. The available modes will be explained together with Figure 37.

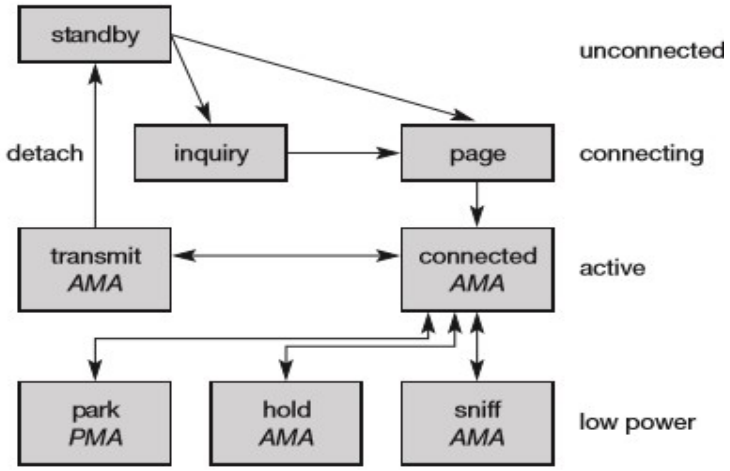


Fig. 37

With transmission power of up to 100 mW, Bluetooth devices can have a range of up to 100 m. Having this power and relying on batteries, a Bluetooth device cannot be in an active transmit mode all the time. Bluetooth defines several low-power states for a device. Figure 37 shows the major states of a Bluetooth device and typical transitions. Every device, which is currently not participating in a piconet (and not switched off), is in standby mode. This is a low-power mode where only the native clock is running. The next step towards the inquiry mode can happen in two different ways. Either a device wants to establish a piconet or a device just wants to listen to see if something is going on.

● A device wants to establish a piconet: A user of the device wants to scan for other devices in the radio range. The device starts the inquiry procedure by sending an inquiry access code (IAC) that is common to all Bluetooth devices. The IAC is broadcast over 32 so-called wake-up carriers in turn.

● Devices in standby that listen periodically: Devices in standby may enter the inquiry mode periodically to search for IAC messages on the wake-up carriers. As soon as a device detects an inquiry it returns a packet containing its device address and timing information required by the master to initiate a connection. From that moment on, the device acts as slave.

If the inquiry was successful, a device enters the page mode. The inquiry phase is not coordinated; inquiry messages and answers to these messages may collide, so it may take a while before the inquiry is successful. After a while (typically seconds but sometimes up to a minute) a Bluetooth device sees all the devices in its radio range. During the page state two different roles are defined. After finding all required devices the master is able to set up connections to each device, i.e., setting up a piconet. Depending on the device addresses received the master calculates special hopping sequences to contact each device individually.

The slaves answer and synchronize with the master's clock, i.e., start with the hopping sequence defined by the master. The master may continue to page more devices that will be added to the piconet. As soon as a device synchronizes to the hopping pattern of the piconet it also enters the connection state.

The connection state comprises the active state and the low power states park, sniff, and hold. In the active state the slave participates in the piconet by listening, transmitting, and receiving. ACL and SCO links can be used. A master periodically synchronizes with all slaves. All devices being active must have the 3-bit active member address (AMA). Within the active state devices either transmit data or are simply connected. A device can enter standby again, via a detach procedure.

To save battery power, a Bluetooth device can go into one of three low power states:

● Sniff state: The sniff state has the highest power consumption of the low power states. Here, the device listens to the piconet at a reduced rate (not on every other slot as is the case in the active state). The interval for listening into the medium can be programed and is application dependent. The master designates a reduced number of slots for transmission to slaves in sniff state. However, the device keeps its AMA.

● Hold state: The device does not release its AMA but stops ACL transmission. A slave may still exchange SCO packets. If there is no activity in the piconet, the slave may either reduce power consumption or participate in another piconet.

● Park state: In this state the device has the lowest duty cycle and the lowest power consumption. The device releases its AMA and receives a parked member address (PMA). The device is still a member of the piconet, but gives room for another device to become active (AMA is only 3 bit, PMA 8 bit). Parked devices are still FH synchronized and wake up at certain beacon intervals for re-synchronization. All PDUs sent to parked slaves are broadcast.

The effect of the low power states is shown in Table 5. This table shows the typical average power consumption of a Bluetooth device (BlueCore2, CSR, 2002). It is obvious that higher data rates also require more transmission power. The intervals in sniff mode also influence power consumption. Typical IEEE 802.11b products have an average current in the order of 200 mA while receiving, 300 mA while sending, and 20 mA in standby.

Operating mode	Average current [mA]
SCO, HV1	53
SCO, HV3, 1 s interval sniff mode	26
ACL, 723.2 kbit/s	53
ACL, 115.2 kbit/s	15.5
ACL, 38.4 kbit/s, 40 ms interval sniff mode	4
ACL, 38.4 kbit/s, 1.28 s interval sniff mode	0.5
Park mode, 1.28 s beacon interval	0.6
Standby (no RF activity)	0.047

Table 5

2.4.6 L2CAP

The logical link control and adaptation protocol (L2CAP) is a data link control protocol on top of the baseband layer offering logical channels between Bluetooth devices with QoS properties. L2CAP is available for ACLs only. Audio applications using SCOs have to use the baseband layer directly. L2CAP provides three different types of logical channels that are transported via the ACL between master and slave:

● Connectionless: These unidirectional channels are typically used for broadcasts from a master to its slave(s).

● Connection-oriented: Each channel of this type is bi-directional and supports QoS flow specifications for each direction. These flow specs follow RFC 1363 (Partridge, 1992) and define average/peak data rate, maximum burst size, latency, and jitter.

● Signaling: This third type of logical channel is used to exchanging signaling messages between L2CAP entities.

Each channel can be identified by its channel identifier (CID). Signaling channels always use a CID value of 1, a CID value of 2 is reserved for connectionless channels. For connection-oriented channels a unique CID (>= 64) is dynamically assigned at each end of the channel to identify the connection (CIDs 3 to 63 are reserved). Figure 38 gives an example for logical channels using the ACL link between master and slave. The master has a bi-directional signaling channel to each slave. The CID at each end is 1. Additionally, the master maintains a connectionless, unidirectional channel to both slaves. The CID at the slaves is 2, while the CID at the beginning of the connectionless channel is dynamically assigned. L2CAP provides mechanisms to add slaves to, and remove slaves from, such a multicast group. The master has one connection oriented channel to the left slave and two to the right slave. All CIDs for these channels are dynamically assigned (between 64 and 65535).

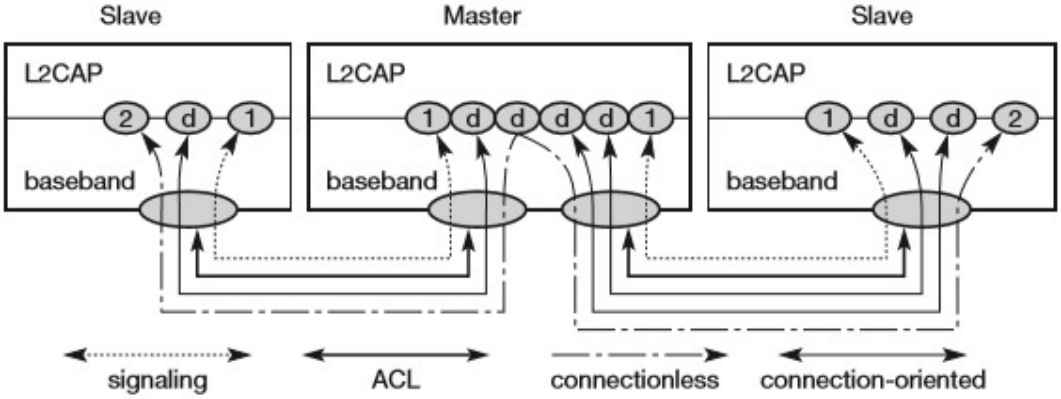


Fig. 38

Figure 39 shows the three packet types belonging to the three logical channel types. The length field indicates the length of the payload (plus PSM for connectionless PDUs). The CID has the multiplexing/demultiplexing function as explained above. For connectionless PDUs a protocol/service multiplexor (PSM) field is needed to identify the higher layer recipient for the

payload. For connection-oriented PDUs the CID already fulfills this function. Several PSM values have been defined, e.g., 1 (SDP), 3 (RFCOMM), 5 (TCS-BIN). Values above 4096 can be assigned dynamically. The payload of the signaling PDU contains one or more commands. Each command has its own code (e.g., for command reject, connection request, disconnection response etc.) and an ID that matches a request with its reply. The length field indicates the length of the data field for this command.

Besides protocol multiplexing, flow specification, and group management, the L2CAP layer also provides segmentation and reassembly functions. Depending on the baseband capabilities, large packets have to be chopped into smaller segments. DH5 links, for example, can carry a maximum of 339 bytes while the L2CAP layer accepts up to 64 kbyte.

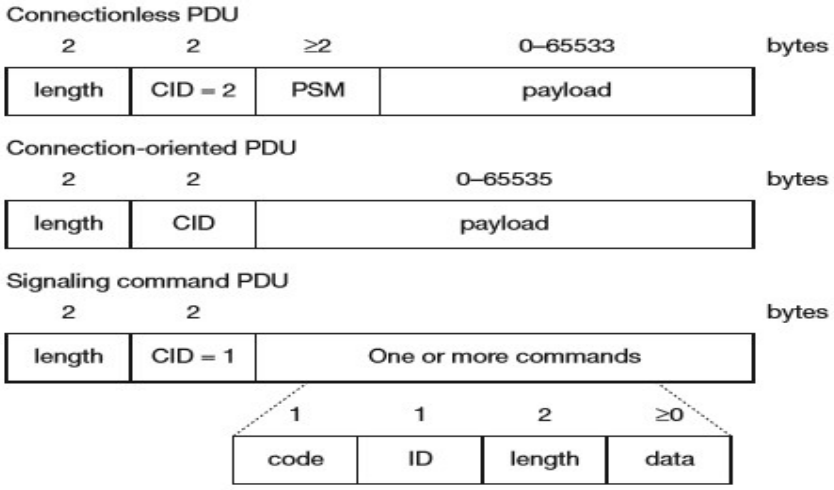


Fig. 39

2.4.7 Security

A radio interface is by nature easy to access. Bluetooth devices can transmit private data, e.g., schedules between a PDA and a mobile phone. A user clearly does not want another person to eavesdrop the data transfer. Just imagine a scenario where two Bluetooth enabled PDAs in suitcases ‘meet’ on the conveyor belt of an airport exchanging personal information! Bluetooth offers mechanisms for authentication and encryption on the MAC layer, which must be implemented in the same way within each device. The main security features offered by Bluetooth include a challeng-eresponse routine for authentication, a stream cipher for encryption, and a session key generation. Each connection may require a one-way, two-way, or no authentication using the challenge-response routine. All these schemes have to be implemented in silicon, and higher layers should offer stronger encryption if needed. The security features included in Bluetooth only help to set up a local domain of trust between devices.

The security algorithms use the public identity of a device, a secret private user key, and an internally generated random key as input parameters. For each transaction, a new random number is generated on the Bluetooth chip. Key management is left to higher layer software.

Figure 40 shows several steps in the security architecture of Bluetooth. The illustration is simplified and the interested reader is referred to Bluetooth (2001a) for further details. The first

step, called pairing, is necessary if two Bluetooth devices have never met before. To set up trust between the two devices a user can enter a secret PIN into both devices. This PIN can have a length of up to 16 byte. Unfortunately, most devices limit the length to four digits or, even worse, program the devices with the fixed PIN '0000' rendering the whole security concept of Bluetooth questionable at least. Based on the PIN, the device address, and random numbers, several keys can be computed which can be used as link key for authentication. Link keys are typically stored in a persistent storage. The authentication is a challenge-response process based on the link key, a random number generated by a verifier (the device that requests authentication), and the device address of the claimant (the device that is authenticated).

Based on the link key, values generated during the authentication, and again a random number an encryption key is generated during the encryption stage of the security architecture. This key has a maximum size of 128 bits and can be individually generated for each transmission. Based on the encryption key, the device address and the current clock a payload key is generated for ciphering user data. The payload key is a stream of pseudo-random bits. The ciphering process is a simple XOR of the user data and the payload key.

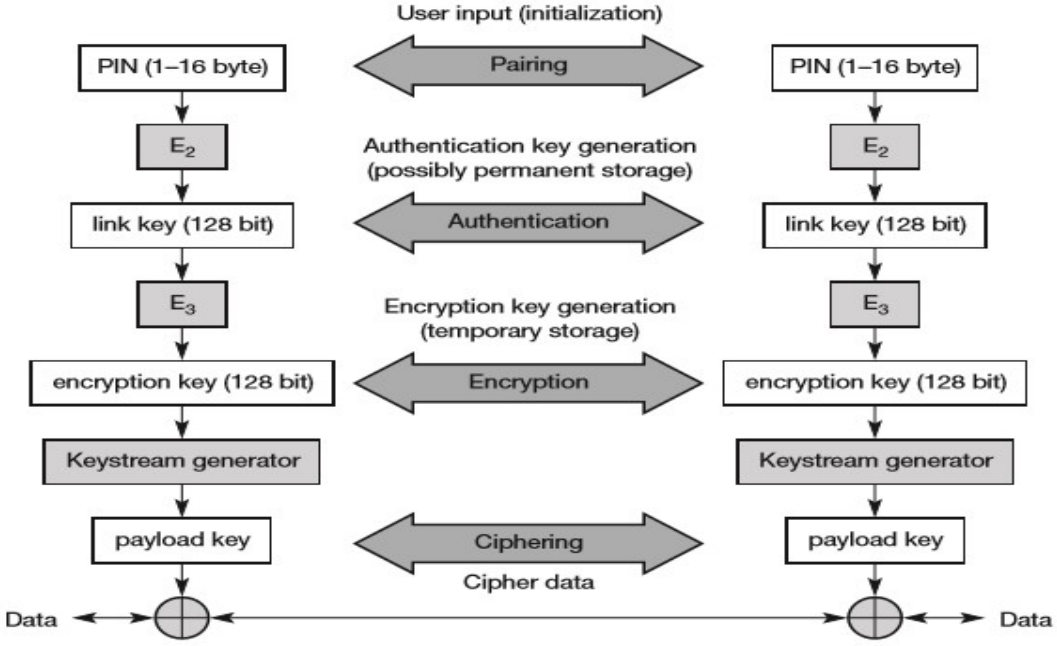


Fig. 40

Compared to WEP in 802.11, Bluetooth offers a lot more security. However, Bluetooth, too, has some weaknesses when it comes to real implementations. The PINs are quite often fixed. Some of the keys are permanently stored on the devices and the quality of the random number generators has not been specified. If Bluetooth devices are switched on they can be detected unless they operate in the non-discoverable mode (no answers to inquiry requests). Either a user can use all services as intended by the Bluetooth system, or the devices are hidden to protect privacy. Either roaming profiles can be established, or devices are hidden and, thus many services will not work. If a lot of people carry Bluetooth devices (mobile phones, PDAs etc.) this could give, e.g., department stores, a lot of information regarding consumer behavior.

2.4.8 SDP

Bluetooth devices should work together with other devices in unknown environments in an ad-hoc fashion. It is essential to know what devices, or more specifically what services, are available in radio proximity. To find new services, Bluetooth defined the service discovery protocol (SDP). SDP defines only the discovery of services, not their usage. Discovered services can be cached and gradual discovery is possible. Devices that want to offer a service have to instal an SDP server. For all other devices an SDP client is sufficient. All the information an SDP server has about a service is contained in a service record. This consists of a list of service attributes and is identified by a 32-bit service record handle. SDP does not inform clients of any added or removed services. There is no service access control or service brokerage. A service attribute consists of an attribute ID and an attribute value. The 16-bit attribute ID distinguishes each service attribute from other service attributes within a service record. The attribute ID also identifies the semantics of the associated attribute value. The attribute value can be an integer, a UUID (universally unique identifier), a string, a Boolean, a URL (uniform resource locator) etc. Table 6 gives some example attributes. The service handle as well as the ID list must be present. The ID list contains the UUIDs of the service classes in increasing generality (from the specific color postscript printer to printers in general). The protocol descriptor list comprises the protocols needed to access this service. Additionally, the URLs for service documentation, an icon for the service and a service name which can be displayed together with the icon are stored in the example service record.

Attribute name	Attribute ID	Attribute value type	Example
ServiceRecordHandle	0000	32-bit unsigned integer	1f3e4723
ServiceClassIDList	0001	Data element sequence (UUIDs)	ColorPostscriptPrinterService ClassID, PostscriptPrinterService ClassID, PrinterServiceClassID
ProtocolDescriptorList	0004	Data element sequence	((L2CAP, PSM=RFCOMM), (RFCOMM, CN=2), (PPP), (IP), (TCP), (IPP))
DocumentationURL	000A	URL	www.xy.zz/print/srvs.html
IconURL	000C	URL	www.xy.zz/print/ico.png
ServiceName	0100	String	Color Printer

Table 6

2.4.9 Profiles

Although Bluetooth started as a very simple architecture for spontaneous ad-hoc communication, many different protocols, components, extensions, and mechanisms have been developed over the last years. Application designers and vendors can implement similar, or even identical, services in many different ways using different components and protocols from the Bluetooth core standard. To provide compatibility among the devices offering the same services, Bluetooth specified many profiles in addition to the core protocols. Without the profiles too many parameters in Bluetooth would make interoperation between devices from different manufacturers almost impossible.

Profiles represent default solutions for a certain usage model. They use a selection of protocols and parameter set to form a basis for interoperability. Protocols can be seen as horizontal layers while profiles are vertical slices (as illustrated in Figure 41). The following basic profiles have been specified: generic access, service discovery, cordless telephony, intercom, serial port, headset, dialup networking, fax, LAN access, generic object exchange, object push, file transfer, and synchronization. Additional profiles are: advanced audio distribution, PAN, audio video remote control, basic printing, basic imaging, extended service discovery, generic audio video distribution, hands-free, and hardcopy cable replacement. Each profile selects a set of protocols. For example, the serial port profile needs RFCOMM, SDP, LMP, L2CAP. Baseband and radio are always required. The profile further defines all interoperability requirements, such as RS232 control signals for RFCOMM or configuration options for L2CAP (QoS, max. transmission unit).

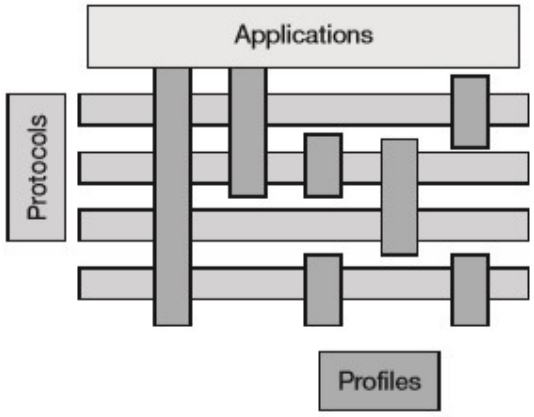


Fig. 41

Module III: Mobility Management in Cellular Networks [4L]

Call setup in PLMN (location update, paging), GPRS, Call setup in mobile IP networks; Handoff management; Mobility models- random walk, random waypoint, map-based, group-based.

3.1 History of Mobile Communication

Wireless communication was a magic to our ancestors but Marconi could initiate it with his wireless telegraph in 1895. Wireless Communication can be classified into three eras.

- Pioneer Era (Till 1920)
- Pre Cellular Era(1920-1979)
- Cellular Era (beyond 1979)

The first commercial mobile telephone system was launched by BELL in St. Louis, USA, in 1946. Few lucky customers got the services. Early mobile systems used single high power transmitters with analog Frequency Modulation techniques to give coverage up to about 50

miles and hence only limited customers could get the service due to this severe constraints of bandwidth.

Cellular Era

To overcome the constraints of bandwidth scarcity and to give coverage to larger sections, BELL lab introduced the principle of Cellular concept. By frequency reuse technique this method delivered better coverage, better utility of available frequency spectrum and reduced transmitter power. But the established calls are to be handed over between base stations while the phones are on move.

Even though the US based BELL lab introduced the cellular principle, the Nordic countries were the first to introduce cellular services for commercial use with the introduction of the Nordic Mobile Telephone (NMT) in 1981.

First Generation Systems

All these systems were analog systems, using FDMA technology. They are also known as First Generation (1G) systems. Different systems came into use based on the cellular principle. They are listed below.

Year	Mobile System
1981	Nordic Mobile Telephone(NMT)450
1982	American Mobile Phone System(AMPS)
1985	Total Access Communication System(TACS)
1986	Nordic Mobile Telephony(NMT)900

Disadvantages of 1G systems

- They were analog and hence are were not robust to interference.
- Different countries followed their own standards, which were incompatible.

To overcome the difficulties of 1G, digital technology was chosen by most of the countries and a new era, called 2G, started.

Advantages of 2G

- Improved Spectral Utilization achieved by using advanced modulation techniques.

- Lower bit rate voice coding enabled more users getting the services simultaneously.
- Reduction of overhead in signaling paved way for capacity enhancement.
- Good source and channel coding techniques make the signal more robust to Interference.
- New services like SMS were included.
- Improved efficiency of access and hand-off control were achieved.

Name of the Systems	Country
DAMPS-Digital Advanced Mobile Phone System	North America
GSM-Global System for Mobile communication	European Countries and International applications
JDC - Japanese Digital Cellular	Japan
CT-2 Cordless Telephone–2	UK
DECT-Digital European Cordless Telephone	European countries

History of GSM

GSM standard is a European standard, which has addressed many problems related to compatibility, especially with the development of digital radio technology.

Milestones of GSM

- 1982 - Confederation of European Post and Telegraph (CEPT) establishes Group Special Mobile.
- 1985 - Adoption of list of recommendation was decided to be generated by the group.
- 1986 - Different field tests were done for radio technique for the common air interface.
- 1987 - TDMA was chosen as the Access Standard. MoU was signed between 12 operators.
- 1988 - Validation of system was done.
- 1989 - Responsibility was taken up by European Telecommunication Standards Institute (ETSI).
- 1990 - First GSM specification was released.

- 1991 - First commercial GSM system was launched.

Frequency Range of GSM

GSM works on four different frequency ranges with FDMA-TDMA and FDD. They are as follows –

System	P-GSM (Primary)	E-GSM (Extended)	GSM 1800	GSM 1900
Freq Uplink	890-915MHz	880-915MHz	1710-1785Mhz	1850-1910MHz
Freq Downlink	935-960MHz	925-960MHz	1805-1880Mhz	1930-1990MHz

Cellular Concepts - Introduction

The immense potential of conventional telephone cannot be exploited to its maximum due to the limitation imposed by the connecting wires. But this restriction has been removed with the advent of the cellular radio.

Frequency Scarcity Problem

If we use dedicated RF loop for every subscriber, we need larger bandwidth to serve even a limited number of subsc in a single city.

Example

A single RF loop requires 50 kHz B/W; then for one lakh subscribers we need 1,00,000 x 50 kHz = 5 GHz.

To overcome this B/W problem, subscribers have to share the RF channels on need basis, instead of dedicated RF loops. This can be achieved by using multiple access methods FDMA, TDMA, or CDMA. Even then the number of RF channels required to serve the subscribers, works out to be impracticable.

Example

Consider a subs density of 30Sq.Km., Grade of service as 1%, Traffic offered per mobile sub as 30m E. Then number of RF channels required are −

Radius(km)	Area in Sq.km	Subs	RF Channels
1	3.14	100	8

3	28.03	900	38
10	314	10000	360

For 10,000 subs to allot 360 radio channels we need a B/Wof 360 × 50 KHz = 18 MHz. This is practically not feasible.

3.2 Cellular Approach

With limited frequency resource, cellular principle can serve thousands of subscribers at an affordable cost. In a cellular network, total area is subdivided into smaller areas called “cells”. Each cell can cover a limited number of mobile subscribers within its boundaries. Each cell can have a base station with a number of RF channels.

Frequencies used in a given cell area will be simultaneously reused at a different cell which is geographically separated. For example, a typical seven-cell pattern can be considered.

Cell Selection Criteria

The requirements that a cell must satisfy before a mobile station can receive service from it are −

It should be a cell of the selected PLMN. The mobile station checks whether the cell is part of the selected PLMN.

It should not be "barred". The PLMN operator may decide not to allow mobile stations to access certain cells. These cells may, for example only be used for handover traffic. Barred cell information is broadcast on the BCCH to instruct mobile stations not to access these cells.

The radio path loss between the mobile station and the selected BTS must be above a threshold set by the PLMN operator.

If no suitable cell is found then the MS enters a "limited service" state in which it can only make emergency calls.

Call to an Active Mobile Station

As an active mobile station (MS) moves in the coverage area of a public land mobile network (PLMN), it reports its movements so that it can be located as needed, using the update procedure locations. When a mobile services switching center (MSC) in the network needs to establish a call to a mobile station operating in its flow area, following things occur −

A page message its broadcast which contains the identification code of the MS. Not every Base Station Controller (BSC) in the network is requested to transmit the page message. The broadcast is limited to a cluster of radio cells that together form a location area. The last reported position of the MS identifies the location area to be used for the broadcast.

The MS monitors the page message transmitted by the radio cell in which it is located and, on detecting its own identification code, responds by transmitting a page response message to the Base Transceiver Station (BTS).

Communication is then established between the MSC and the MS via BTS that received the page response message.

Location Update

Case 1 − Location never updates.

If location never updates the implementation for location update, cost becomes zero. But we have to page every cell for locating the MS and this procedure will not be cost effective.

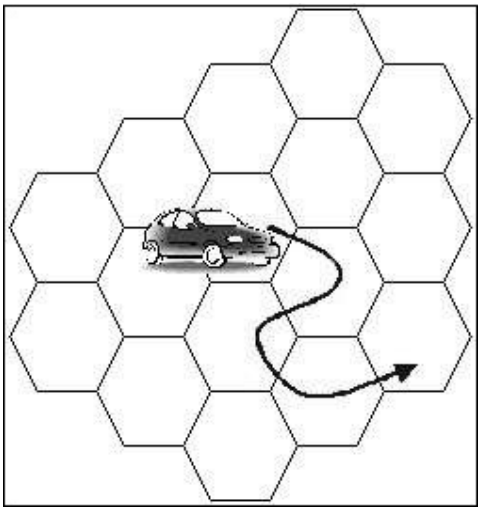


Fig. 3.1: cellular structure

Case 2 − Location update is implemented.

Location updates are taking place as per the requirements of the network, may be time or movement or distance based. This procedure involves high cost, but we have to page single cell or few cells only for locating the MS and this procedure will be cost effective.

Network Configuration

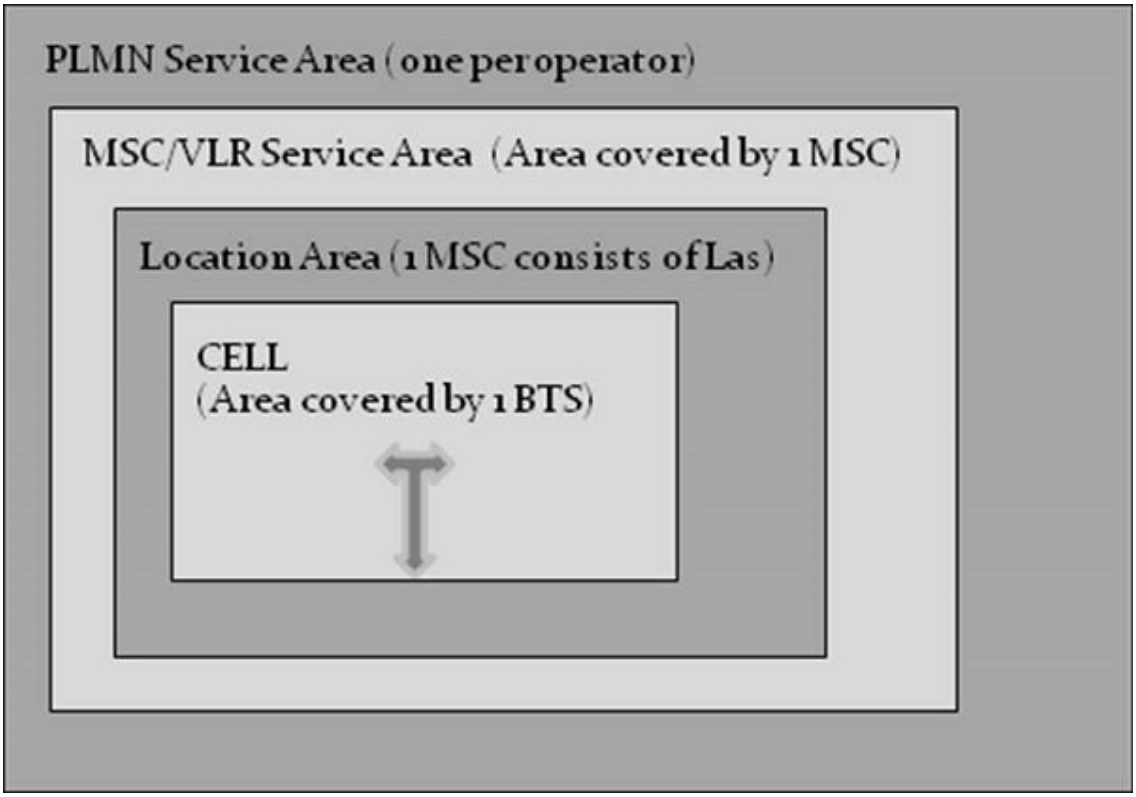


Figure 3.2: PLMN area

The configuration of a Public Land Mobile Network (PLMN) is designed so that active mobile station moving in the network area is still able to report its position. A network consists of different areas −

- PLMN area
- Location area
- MSC area
- PLMN Area

A PLMN area is the geographical area in which land mobile communication services are provided to the public by a particular PLMN operator. From any position within a PLMN area, the mobile user can set up calls to another user of the same network, or to a user of another network. The other network may be a fixed network, another GSM PLMN, or another type of PLMN. Users of the same PLMN or users of other networks can also call a mobile user who is active in the PLMN area. When there are several PLMN operators, the geographical areas covered by their networks may overlap. The extent of a PLMN area is normally limited by national borders.

Location Area

To eliminate the need for network-wide paging broadcasts, the PLMN needs to know the approximate positions of the MSs that are active within its coverage area. To enable the approximate positions of any MS to be represented by a single parameter, the total area covered by the network is divided into location areas. A Location Area (LA) is a group of one or more radio cells. This group fulfills the following requirements −

- BTSs in one location area may be controlled by one or more BSCs.
- BSCs those serve the same location area are always connected to the same MSC.
- Radio cells with BTSs controlled by a common BSC can lie in different location areas.

Location Area Identity

Every radio transmitter in the PLMN broadcast, via a control channel BCCH, a Location Area Identity (LAI), code to identify the location area that it serves. When an MS is not engaged in a call, it automatically scans the BCCH transmitted by the base stations in the locality and selects the channel that is delivering the strongest signal. The LAI code broadcast by the selected channel identifies the location area in which the MS is currently situated. This LAI code is stored in the Subscriber Identity Module (SIM) of the mobile equipment.

As the MS moves through the network area, the signal received from the selected control channel gradually diminishes in strength until it is no longer the strongest. At this point the MS re-tunes to the channel that has become dominant and examines the LAI code that it is broadcasting. If the received LAI code differs from that stored on the SIM, then the MS has entered another location area and initiates a location update procedure to report the change to the MSC. At the end of the procedure, the LAI code in the SIM is also updated.

Location Area Identity Format

It is a Location Area Identity (LAI) code to identify the location area in a PLMN. The LAI code has three components −

Mobile Country Code (MCC)

The MCC is a 3-digit code that uniquely identifies the country of domicile of the mobile subscriber (for example, India 404). It is assigned by the ITU-T.

Mobile Network Code (MNC)

The MNC is a 2-digit code (3-digit code for GSM-1900) that identifies the home GSM PLMN of the mobile subscriber. If more than one GSM PLMN exists in a country, a unique MNC is assigned to each of them. It is assigned by the government of each country. (For example Cell one, Chennai 64).

Location Area Code (LAC)

The LAC component identifies a location area within a PLMN; it has a fixed length of 2 octets and can be coded using hexadecimal representation. It is assigned by an operator.

MSC areas

An MSC area is a region of the network in which GSM operations are controlled by a single MSC. An MSC area consists of one more location areas. The boundary of an MSC area follows the external boundaries of the location areas on its periphery. Consequently, a location area never spans beyond the boundary of an MSC area.

VLR area

A VLR area is region of the network that is supervised by a single Visitor Location Register (VLR). In theory, a VLR area may consist of one more MSC areas. In practice, however the functions of the VLR are always integrated with those of the MSC so that the terms "VLR area" and "MSC area" have become synonymous.

Location Related Databases

Two databases are used by Location Management to store MS location related data.

- Visitor Location Register(VLR)
- Home Location Register(HLR)

Visitor Location Register

A VLR contains a data record for each of the MS that are currently operating in its area. Each record contains a set of subscriber identity codes, related subscription information, and a Location Area Identity (LAI) code. This information is used by the MSC when handling calls to or from an MS in the area. When an MS moves from one area to another, the responsibility for its supervision passes from one VLR to another. A new data record is created by the VLR that has adopted the MS, and the old record is deleted. Provided that aninter-working agreement exists between the network operators concerned, data transaction can cross both network and national boundaries.

Home Location Register

The HLR contains information relevant to mobile subscribers who are fee-paying customers of the organization that operates the PLMN.

The HLR stores two types of information −

Subscription Information

The subscription information includes the IMSI and directory number allocated to the subscriber, the type of services provided and any related restrictions.

Location Information

The location information includes the address of the VLR in the area where the subscribers MS is currently located and the address of the associated MSC.

The location information enables incoming calls to be routed to the MS. The absence of this information indicates that the MS is inactive and cannot be reached.

When an MS moves from one VLR area to another, the location information in the HLR is updated with the new entry for the MS, using subscription data copied from the HLR. Provided that an inter-working agreement exists between the network operators, concerned data transactions can move across both network and national boundaries.

Types of Identification Numbers

During the performance of the location update procedure and the processing of a mobile call different types of numbers are used −

- Mobile Station ISDN Number(MSISDN)
- Mobile Subscriber Roaming Number(MSRN)
- International Mobile Subscriber Identity(IMSI)
- Temporary Mobile Subscriber Identity(TMSI)
- Local Mobile Station Identity(LMSI)

Each number is stored in the HLR and/or VLR.

Mobile Station ISDN Number

The MSISDN is the directory number allocated to the mobile subscriber. It is dialed to make a telephone call to the mobile subscriber. The number consists of Country Code (CC) of the country in which the mobile station is registered (e.g. India 91), followed by national mobile number which consists of Network Destination Code (NDC) and Subscriber Number (SN). An NDC is allocated to each GSM PLMN.

The composition of the MSISDN is such that it can be used as a global title address in the Signaling Connection Control Part (SCCP) for routing message to the HLR of the mobile subscriber.

Mobile Station Roaming Number

The MSRN is the number required by the gateway MSC to route an incoming call to an MS that is not currently under the control of the gateway MSISDN. Using a mobile, terminated call is routed to the MSC gateway. Based on this, MSISDN gateway MSC requests for a MSRN to route the call to the current visited MSC International Mobile Subscriber Identity (IMSI).

An MS is identified by its IMSI. The IMSI is embedded in the SIM of the mobile equipment. It is provided by the MS anytime it accesses the network.

Mobile Country Code (MCC)

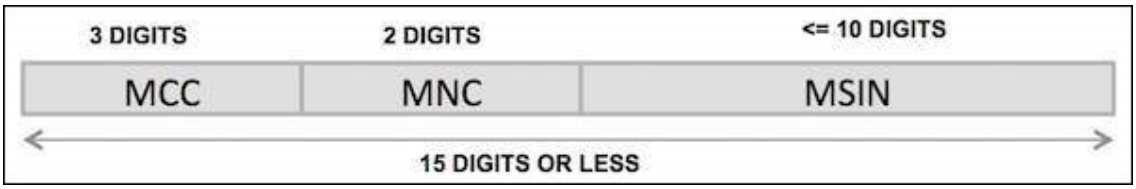
The MCC component of the IMSI is a 3-digit code that uniquely identifies the country of the domicile of the subscriber. It is assigned by the ITU-T.

Mobile Network Code (MNC)

The MNC component is a 2-digit code that identifies the home GSM PLMN of the mobile subscriber. It is assigned by the government of each country. For GSM-1900 a 3-digit MNC is used.

Mobile Subscriber Identification Number (MSIN)

The MSIN is a code that identifies the subscriber within a GSM PLMN. It is assigned by the operator.



Temporary Mobile Subscriber Identity (TMSI)

The TMSI is an identity alias which is used instead of the IMSI when possible. The use of a TMSI ensures that the true identity of the mobile subscriber remains confidential by eliminating the need to transfer a non ciphered IMSI code over a radio link.

A VLR allocates a unique TMSI code to each mobile subscriber that is operating in its area. This code which is only valid within the area supervised by the VLR is used to identify the subscriber, in messages to and from the MS. When a change of location area also involves a change of VLR area, a new TMSI code is allocated and communicated to the MS. The MS stores the TMSI on its SIM. The TMSI consists of four octets.

Location Update Scenario

In the following location update scenario, it is assumed that an MS enters a new location area that is under control of a different VLR (referred to as the "new VLR") than the one where the MS is currently registered (referred to as the "old VLR"). The following diagram shows the steps of the mobile location update scenario.

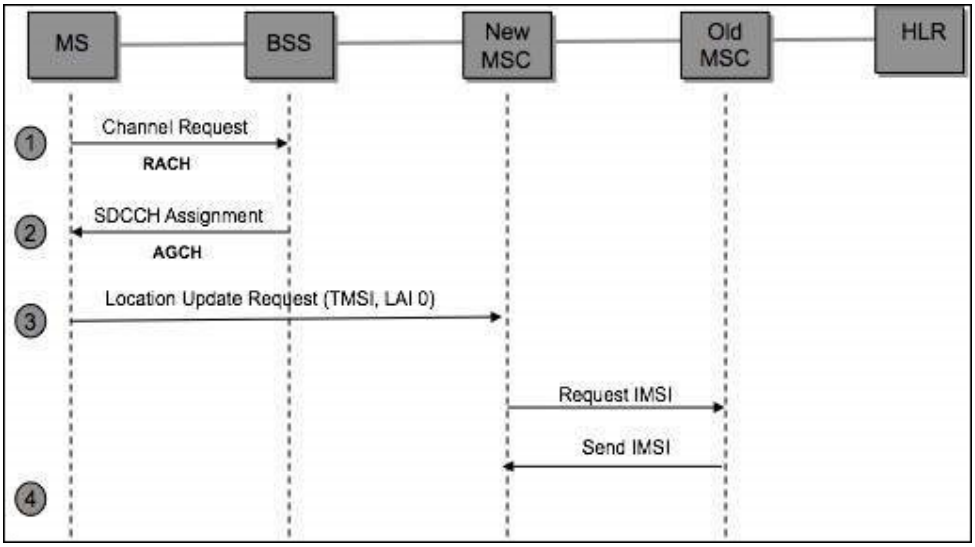


Figure 3.3: location update

The MS enters a new cell area, listens to the Location Area Identity (LAI) being transmitted on the broadcast channel (BCCH), and compares this LAI with the last LAI (stored in the SIM) representing the last area where the mobile was registered.

- The MS detects that it has entered a new Location Area and transmits a Channel Request message over the Random Access Channel (RACH).
- Once the BSS receives the Channel Request message, it allocates a Stand-alone Dedicated Control Channel (SDCCH) and forwards this channel assignment information to the MS over the Access Grant Channel (AGCH). It is over the SDCCH that the MS will communicate with the BSS and MSC.
- The MS transmits a location update request message to the BSS over the SDCCH. Included in this message are the MS Temporary Mobile Subscriber Identity (TMSI) and the old Location Area Subscriber (old LAI). The MS can identify itself either with its IMSI or TMSI. In this example, we will assume that the mobile provided a TMSI. The BSS forwards the location update request message to the MSC.
- The VLR analyses the LAI supplied in the message and determines that the TMSI received is associated with a different VLR (old VLR). In order to proceed with the registration the IMSI of the MS must be determined. The new VLR derives the identity of the old VLR by using the received LAI, supplied in the location update request message. It also requests the old VLR to supply the IMSI for a particular TMSI.
- Location Update Scenario-Update HLR/VLR is a point where we are ready to inform the HLR that the MS is under control of a new VLR and that the MS can be de-registered from the old VLR. The steps in update HLR/VLR phase are −

- The new VLR sends a message to the HLR informing it that the given IMSI has changed locations and can be reached by routing all incoming calls to the VLR address included in the message.
- The HLR requests the old VLR to remove the subscriber record associated with the given IMSI. The request is acknowledged.
- The HLR updates the new VLR with the subscriber data (mobiles subscribers' customer profile).

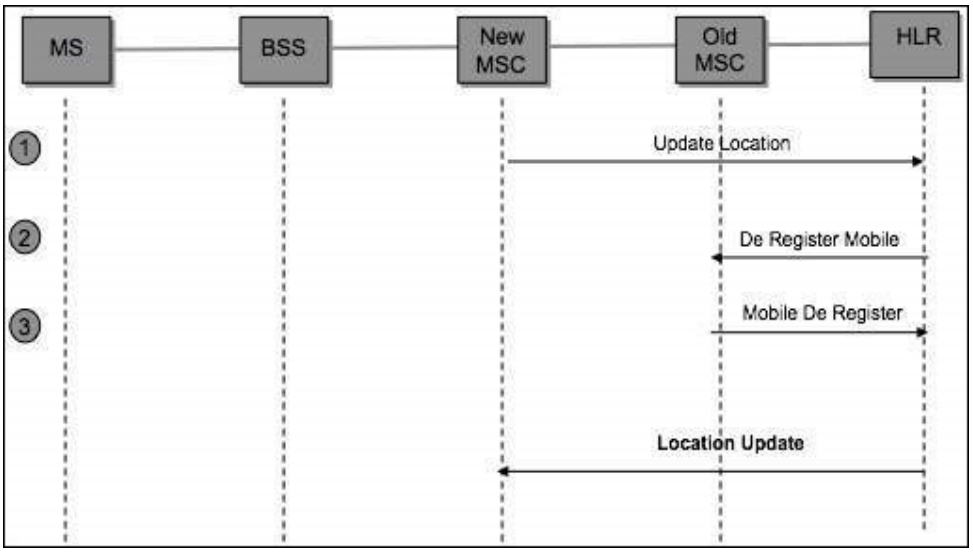


Fig. 3.4: Steps in TMSI Reallocation Phase

- The MSC forwards the location update accept message to the MS. This message includes the new TMSI.
- The MS retrieves the new TMSI value from the message and updates its SIM with this new value. The mobile then sends an update complete message back to the MSC.
- The MSC requests from the BSS, that the signaling connection be released between the MSC and the MS.
- The MSC releases its portion of the signaling connection when it receives the clear complete message from the BSS.
- The BSS sends a "radio resource" channel release message to the MS and then free up the Stand-alone Dedicated Control Channel (SDCCH) that was allocated previously. The BSS then informs the MSC that the signaling connection has been cleared.

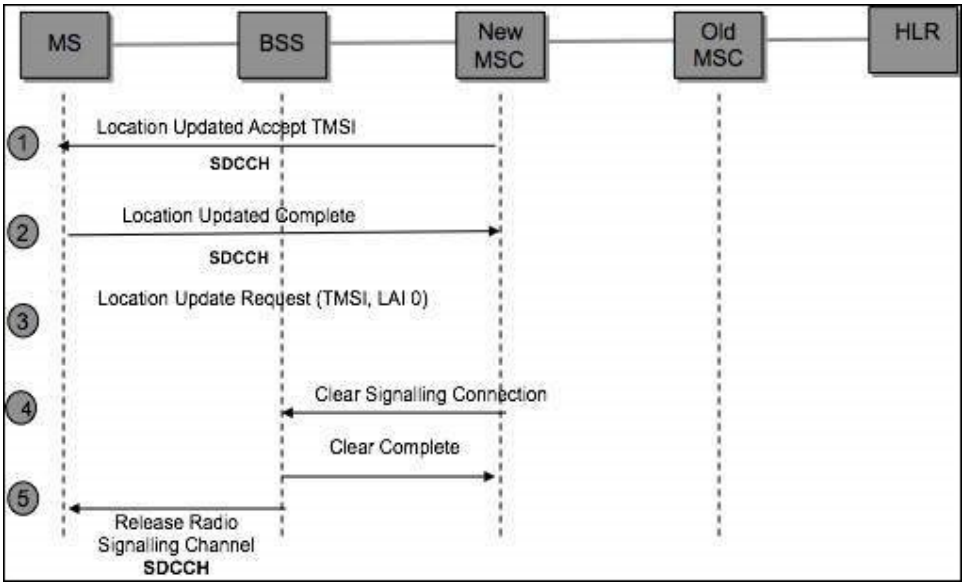


Fig. 3.5: Location Update Periodicity

Location Update automatically takes place when the MS changes its LA. A lot of location updates may be generated if a user crosses LA boundary frequently. If the MS remains in the same LA, Location Update may take place based on time/movement/distance, as defined by the network provider.

Hand Over

This is the process of automatically switching a call in progress from one traffic channel to another to neutralize the adverse effects of the user movements. Hand over process will be started only if the power control is not helpful anymore.

The Hand Over process is MAHO (Mobile Assisted Hand Over). It starts with the Down Link Measurements by the MS (Strength of the signal from BTS, Quality of the signal from BTS). MS can measure the signal strength of the 6 best neighboring BTS downlink (candidate list).

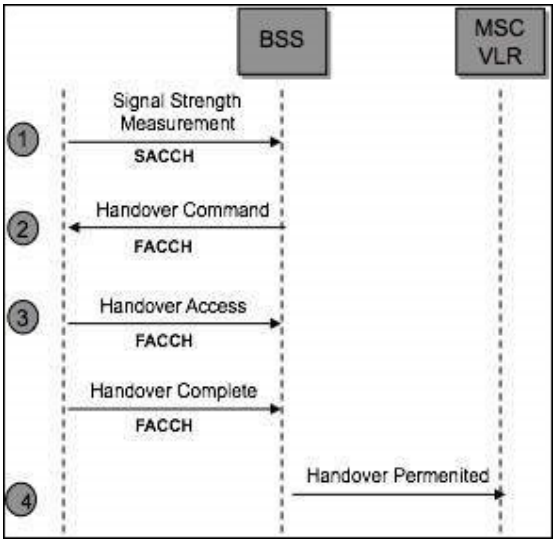


Fig. 3.6:Hand over

Hand Over Types

There are two types of Hand Over −

- Internal or Intra BSS Handover

  Intra-cell hand over

  Inter cell hand over

- External or Inter BSS Hand over

  Intra-MSC hand over

  Inter MSC hand over

  Internal handover is managed by the BSC and external handover by MSC.

The objectives of Hand Over are as follows −

- Maintain a good quality of speech.
- Minimize number of calls dropped.
- Maximize the amount of time the mobile station is in the best cell.
- Minimize the number of hand overs.

When will a Hand Over take place?

- Distance (propagation delay) between the MS and BTS becomes too big.
- If the received signal level is very low.
- If the received signal quality very low.

  Path loss situation for the mobile station to another cell is better.

The following new GPRS network adds the following elements to an existing GSM network.

- Packet Control Unit (PCU).
- Serving GPRS Support Node (SGSN) − the MSC of the GPRS network.
- Gateway GPRS Support Node (GGSN) − gateway to external networks.
- Border Gateway (BG) − a gateway to other PLMN.
- Intra-PLMN backbone − an IP based network inter-connecting all the GPRS elements.

General Packet Radio Service (GPRS)

- GPRS introduces packet data transmission to the mobile subscriber.

- GPRS is designed to work within the existing GSM infrastructure with additional packet switching nodes.
- This packet mode technique uses multi-slot technology together with support for all coding schemes (CS-1 to CS-4) to increase the data rates up to 160 kbit/s.
- The GPRS system uses the physical radio channels as defined for GSM. A physical channel used by GPRS is called a Packet Data Channel (PDCH).
- The PDCHs can either be allocated for GPRS (dedicated PDCH) or used by GPRS only if no circuit-switched connection requires them (on-demand). The operator can define 0-8 dedicated PDCHs per cell. The operator can specify where he wants his PDCHs to be located.
- The first dedicated PDCH in the cell is always a Master PDCH (MPDCH). The on-demand PDCHs can be pre-empted by incoming circuit switched calls in congestion situations in the cell.

Coding Scheme	Speed(kbit/s)
CS-1	8.0
CS-2	12.0
CS-3	14.4
CS-4	20.0

Serving GPRS Support Node (SGSN) Functions

The SGSN or Serving GPRS Support Node element of the GPRS network provides a number of takes focused on the IP elements of the overall system. It provides a variety of services to the mobiles −

- Packet routing and transfer
- Mobility management
- Authentication
- Attach/detach
- Logical link management
- Charging data

There is a location register within the SGSN and this stores the location information (e.g., current cell, current VLR). It also stores the user profiles (e.g., IMSI, packet addresses used) for all the GPRS users registered with the particular SGSN.

Gateway GPRS Support Node (GGSN) Functions

- The GGSN, Gateway GPRS Support Node is one of the most important entities within the GSM EDGE network architecture.
- The GGSN organizes the inter-working between the GPRS/EDGE network and external packet switched networks to which the mobiles may be connected. These may include both Internet and X.25 networks.
- The GGSN can be considered to be a combination of a gateway, router and firewall as it hides the internal network to the outside. In operation, when the GGSN receives data addressed to a specific user, it checks if the user is active, then forwards the data. In the opposite direction, packet data from the mobile is routed to the right destination network by the GGSN.

Upgradation of Equipment from GSM to GPRS

- **Mobile Station (MS)** − New Mobile Station is required to access GPRS services. These new terminals will be backward compatible with GSM for voice calls. Three types of handsets are available. Type-A: GPRS & Speech (simultaneously), Type-B: GPRS & Speech (Auto switch), Type-C: GPRS or Speech (manual switch).
- **BTS** − A software upgrade is required in the existing base transceiver site.
- **BSC** − Requires a software upgrade and the installation of new hardware called the packet control unit (PCU). PCU is responsible for handling the Medium Access Control (MAC) and Radio Link Control (RLC) layers of the radio interface and the BSSGP and Network Service layers of the Gb interface. There is one PCU per BSC. The Gb interface, carry GPRS/EGPRS traffic from the SGSN (Serving GPRS Support Node) to the PCU.
- **GPRS Support Nodes (GSNs)** − The deployment of GPRS requires the installation of new core network elements called the serving GPRS support node (SGSN) and gateway GPRS support node (GGSN).
- **Databases (HLR, VLR, etc.)** − All the databases involved in the network will require software upgrades to handle the new call models and functions introduced by GPRS.

Location Information - GSM Service Area Hierarchy

- **Cell** − Cell is the basic service area and one BTS covers one cell. Each cell is given a Cell Global Identity (CGI), a number that uniquely identifies the cell.
- **LA** − A group of cells form a Location Area. This is the area that is paged when a subscriber gets an incoming call. Each Location Area is assigned a Location Area Identity (LAI). Each Location Area is served by one or more BSCs.
- **MSC/VLR Service Area** − The area covered by one MSC is called the MSC/VLR service area.

- **PLMN** − The area covered by one network operator is called PLMN. A PLMN can contain one or more MSCs.
- **GSM Service Area** − The area in which a subscriber can access the network.

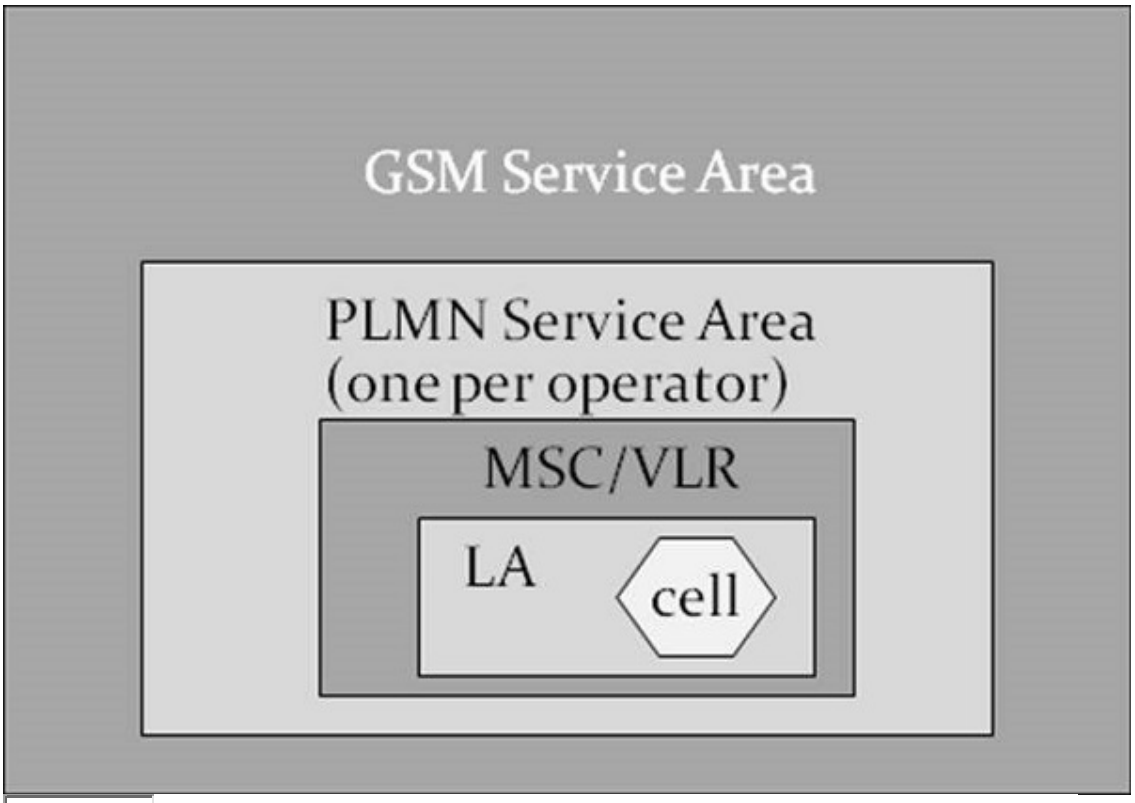


Fig. 3.7: GSM service area

3.3 Mobile IP:

Mobile IP (Internet Protocol) enables the transfer of information to and from mobile computers, such as laptops and wireless communications. The mobile computer can change its location to a foreign network and still access and communicate with and through the mobile computer's home network. The Solaris implementation of Mobile IP supports only IPv4.

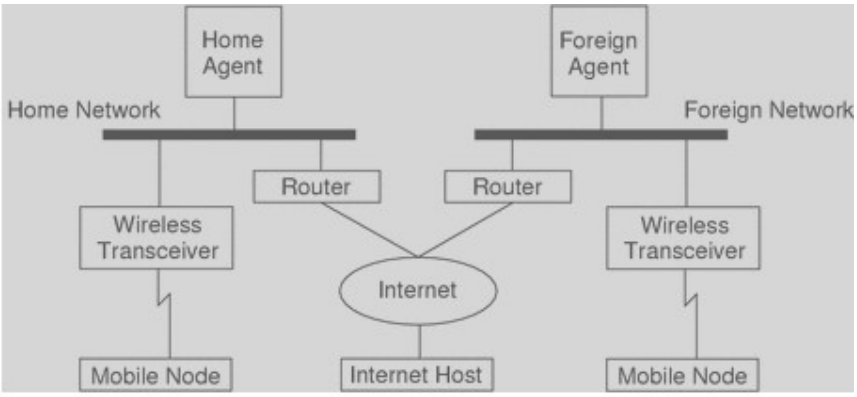
Introduction

Current versions of the Internet Protocol (IP) assume that the point at which a computer attaches to the Internet or a network is fixed and its IP address identifies the network to which it is attached. Datagrams are sent to a computer based on the location information contained in the IP address.

If a mobile computer, or **mobile node**, moves to a new network while keeping its IP address unchanged, its address does not reflect the new point of attachment. Consequently, existing routing protocols cannot route datagrams to the mobile node correctly. In this situation, you must reconfigure the mobile node with a different IP address representative of its new location, which is a cumbersome process. Thus, under the current Internet Protocol, if the mobile node moves without changing its address, it loses routing; but if it does change its address, it loses connections.

Mobile IP solves this problem by allowing the mobile node to use two IP addresses: a fixed **home address** and a **care-of address** that changes at each new point of attachment. Mobile IP enables a computer to roam freely on the Internet or an organization's network while still maintaining the same home address. Consequently, computing activities are not disrupted when the user changes the computer's point of attachment to the Internet or an organization's network. Instead, the network is updated with the new location of the mobile node. The following figure illustrates the general Mobile IP topology.

Figure 3.8 Mobile IP Topology



Using the previous illustration's Mobile IP topology, the following scenario shows how a datagram moves from one point to another within the Mobile IP framework.

1. The Internet host sends a datagram to the mobile node using the mobile node's home address (normal IP routing process).
2. If the mobile node is on its home network, the datagram is delivered through the normal IP process to the mobile node. Otherwise, the home agent picks up the datagram.
3. If the mobile node is on a foreign network, the home agent forwards the datagram to the foreign agent.
4. The foreign agent delivers the datagram to the mobile node.
5. Datagrams from the mobile node to the Internet host are sent using normal IP routing procedures. If the mobile node is on a foreign network, the packets are delivered to the foreign agent. The foreign agent forwards the datagram to the Internet host.

In the case of wireless communications, the illustrations depict the use of wireless transceivers to transmit the datagrams to the mobile node. Also, all datagrams between the Internet host and the mobile node use the mobile node's home address regardless of whether the mobile node is on a home or foreign network. The care-of address is used only for communication with mobility agents and is never seen by the Internet host.

Mobile IP Functional Entities

Mobile IP introduces the following new functional entities:

- **Mobile Node (MN)**–Host or router that changes its point of attachment from one network to another.
- **Home Agent (HA)**–Router on a mobile node's home network that intercepts datagrams destined for the mobile node, and delivers them through the care-of address. The home agent also maintains current location information for the mobile node.

- **Foreign Agent (FA)**–Router on a mobile node's visited network that provides routing services to the mobile node while the mobile node is registered.

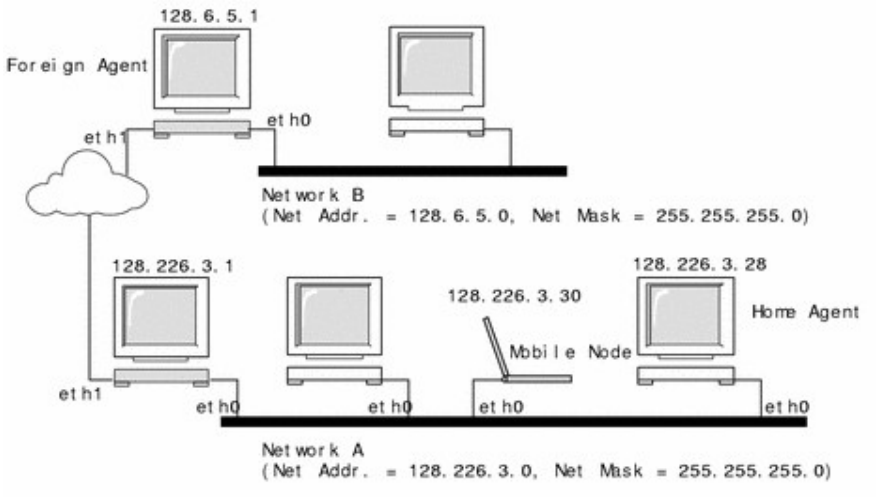
How Mobile IP Works

Mobile IP enables routing of IP datagrams to mobile nodes. The mobile node's home address always identifies the mobile node, regardless of its current point of attachment to the Internet or an organization's network. When away from home, a care-of address associates the mobile node with its home address by providing information about the mobile node's current point of attachment to the Internet or an organization's network. Mobile IP uses a registration mechanism to register the care-of address with a home agent.

The home agent redirects datagrams from the home network to the care-of address by constructing a new IP header that contains the mobile node's care-of address as the destination IP address. This new header then encapsulates the original IP datagram, causing the mobile node's home address to have no effect on the encapsulated datagram's routing until it arrives at the care-of address. This type of encapsulation is also called **tunneling**. After arriving at the care-of address, each datagram is de-encapsulated and then delivered to the mobile node.

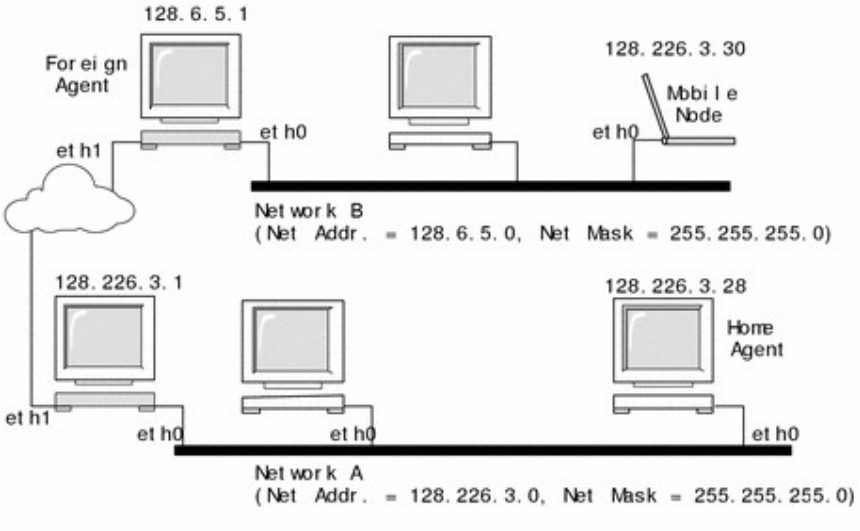
The following illustration shows a mobile node residing on its home network, Network A, before the mobile node moves to a foreign network, Network B. Both networks support Mobile IP. The mobile node is always associated with its home network by its permanent IP address, 128.226.3.30. Though Network A has a home agent, datagrams destined for the mobile node are delivered through the normal IP process.

Figure 3.9 Mobile Node Residing on Home Network



The following illustration shows the mobile node moving to a foreign network, Network B. Datagrams destined for the mobile node are intercepted by the home agent on the home network, Network A, encapsulated, and sent to the foreign agent on Network B. Upon receiving the encapsulated datagram, the foreign agent strips off the outer header and delivers the datagram to the mobile node visiting Network B.

Figure 3.10 Mobile Node Moving to a Foreign Network



The care-of address might belong to a foreign agent, or might be acquired by the mobile node through Dynamic Host Configuration Protocol (DHCP) or Point-to-Point Protocol (PPP). In the latter case, a mobile node is said to have a co-located care-of address.

The mobile node uses a special **registration** process to keep its home agent informed about its current location. Whenever a mobile node moves from its home network to a foreign network, or from one foreign network to another, it chooses a foreign agent on the new network and uses it to forward a registration message to its home agent.

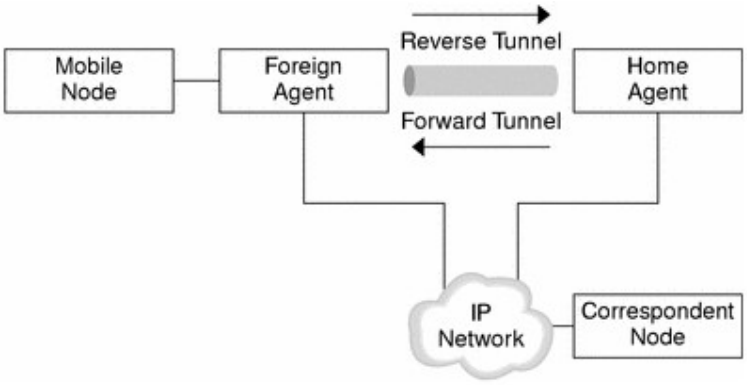
Mobility agents (home agents and foreign agents) advertise their presence using **agent advertisement** messages. A mobile node can optionally solicit an agent advertisement message from any locally attached mobility agents through an **agent solicitation** message. A mobile node receives these agent advertisements and determines whether they are on its home network or a foreign network.

When the mobile node detects that it is located on its home network, it operates without mobility services. If returning to its home network from being registered elsewhere, the mobile node **deregisters** with its home agent.

Mobile IP With Reverse Tunneling

The previous description of Mobile IP assumes that the routing within the Internet is independent of the data packet's source address. However, intermediate routers might check for a topologically correct source address. If an intermediate router does check, you should set up a reverse tunnel. By setting up a reverse tunnel from the mobile node's care-of address to the home agent, you ensure a topologically correct source address for the IP data packet. A mobile node can request a **reverse tunnel** between its foreign agent and its home agent when the mobile node registers. A reverse tunnel is a tunnel that starts at the mobile node's care-of address and terminates at the home agent. The following illustration shows the Mobile IP topology that uses a reverse tunnel.

Figure 3.11 Mobile IP With a Reverse Tunnel

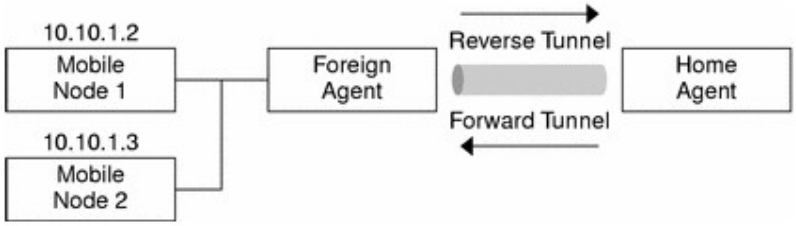


Limited Private Addresses Support

Mobile nodes that have private addresses which are not globally routable through the Internet require reverse tunnels. Solaris Mobile IP supports only privately addressed mobile nodes. See Overview of the Solaris Mobile IP Implementation for the functions that Solaris Mobile IP does not support.

Enterprises employ private addresses when external connectivity is not required. Private addresses are not routable through the Internet. When a mobile node has a private address, the mobile node can only communicate with a correspondent node through a reverse tunnel. The privately addressed correspondent node must belong to the same home agent's administrative domain. The following illustration shows a network topology with two privately addressed mobile nodes that use the same care-of address when registered to the same foreign agent.

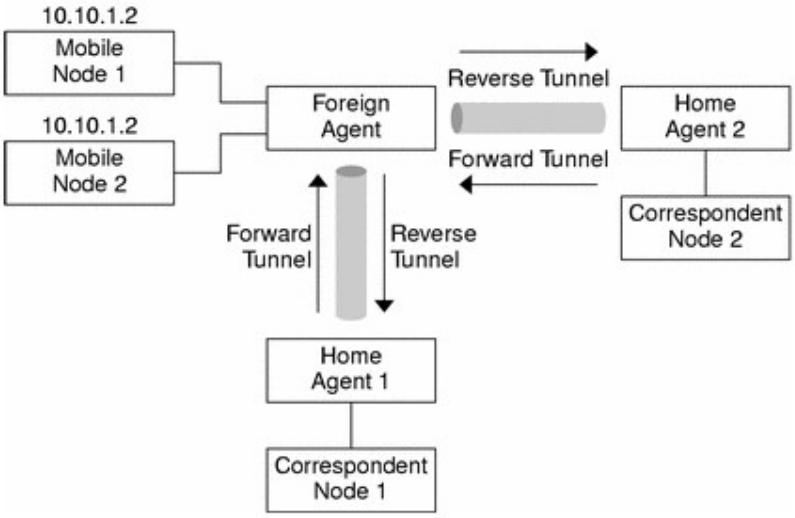
Figure 3.12 Privately Addressed Mobile Nodes Residing on the Same Foreign Network



Because both privately addressed mobile nodes belong to the same administrative domain, the home agent knows how to route data packets between the two mobile nodes. Also, the foreign agent's care-of address and the home agent's IP address must be globally routable addresses.

It is possible to have two privately addressed mobile nodes with the same IP address residing on the same foreign network. This situation is only possible when each mobile node has a different home agent. Also, this situation is only possible when each mobile node is on different advertising subnets of a single foreign agent. The following illustration shows a network topology that depicts this case.

Figure 3.13 Privately Addressed Mobile Nodes Residing on Different Foreign Networks



Because both privately addressed mobile nodes have the same IP address and because these mobile nodes belong to different home agent domains, the two nodes cannot communicate with each other. However, each node can communicate with nodes in its corresponding home agent's administrative domain through the reverse tunnel. For example, Mobile Node 2 can communicate with Correspondent Node 2 in the previous illustration.

Care-of Addresses

Mobile IP provides the following alternative modes for the acquisition of a care-of address:

- A foreign agent provides a **foreign agent care-of address** through its agent advertisement messages. In this case, the care-of address is an IP address of the foreign agent. The foreign agent is the endpoint of the tunnel and, on receiving tunneled datagrams, de-encapsulates them and delivers the inner datagram to the mobile node. In this mode, many mobile nodes can share the same care-of address. This sharing reduces demands on the IPv4 address space and can also save bandwidth, because the forwarded packets, from the foreign agent to the mobile node, are not encapsulated. Saving bandwidth is important on wireless links.
- A mobile node acquires a **co-located care-of address** as a local IP address through some external means, which the mobile node then associates with one of its own network interfaces. The address might be dynamically acquired as a temporary address by the mobile node, such as through DHCP. The address might also be owned by the mobile node as a long-term address for its use only while visiting some foreign network. When using a co-located care-of address, the mobile node serves as the endpoint of the tunnel and performs de-encapsulation of the datagrams tunneled to it.

Co-located care-of address enables a mobile node to function without a foreign agent, for example, in networks that have not yet deployed a foreign agent.

If a mobile node is using a co-located care-of address, the mobile node must be located on the link identified by the network prefix of this care-of address. Otherwise, datagrams destined to the care-of address are undeliverable.

Agent Discovery

A mobile node uses a method known as agent discovery to determine the following information:

- When the node has moved from one network to another
- Whether the network is the node's home or a foreign network
- What is the foreign agent care-of address offered by each foreign agent on that network

Mobility agents transmit **agent advertisements** to advertise their services on a network. In the absence of agent advertisements, a mobile node can solicit advertisements. This is known as **agent solicitation**.

Agent Advertisement

Mobile nodes use agent advertisements to determine their current point of attachment to the Internet or to an organization's network. An agent advertisement is an Internet Control Message Protocol (ICMP) router advertisement that has been extended to also carry a mobility agent advertisement extension.

A foreign agent can be too busy to serve additional mobile nodes. However, a foreign agent must continue to send agent advertisements. This way, mobile nodes that are already registered with it will know that they have not moved out of range of the foreign agent and that the foreign agent has not failed.

Also, a foreign agent that supports reverse tunnels must send it's advertisements with the reverse tunnel flag set on.

Agent Solicitation

Every mobile node should implement agent solicitation. The mobile node uses the same procedures, defaults, and constants for agent solicitation, as specified for ICMP router solicitation messages.

The rate at which a mobile node sends solicitations is limited by the mobile node. The mobile node can send three initial solicitations at a maximum rate of one per second while searching for an agent. After registering with an agent, the rate at which solicitations are sent is reduced, to limit the overhead on the local network.

Mobile IP Registration

When the mobile node receives an agent advertisement, the mobile node registers through the foreign agent, even when the mobile node might be able to acquire its own co-located care-of address. This feature enables sites to restrict access to mobility services. Through agent advertisements, mobile nodes detect when they have moved from one subnet to another.

Mobile IP registration provides a flexible mechanism for mobile nodes to communicate their current reachability information to their home agent. The registration process enables mobile nodes to perform the following tasks:

- Request forwarding services when visiting a foreign network
- Inform their home agent of their current care-of address
- Renew a registration that is due to expire
- Deregister when they return home
- Request a reverse tunnel

Registration messages exchange information between a mobile node, a foreign agent, and the home agent. Registration creates or modifies a mobility binding at the home agent, associating the mobile node's home address with its care-of address for the specified lifetime.

The registration process also enables mobile nodes to:

- Register with multiple foreign agents
- Deregister specific care-of addresses while retaining other mobility bindings
- Discover the address of a home agent if the mobile node is not configured with this information

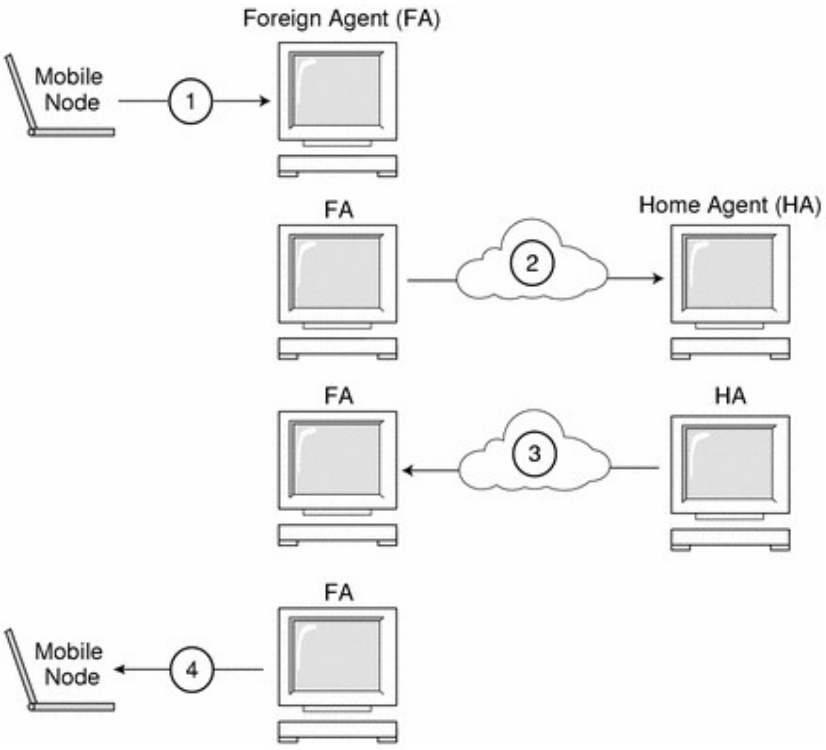
Mobile IP defines the following registration processes for a mobile node:

- If a mobile node is registering a foreign agent care-of address, the mobile node registers using that foreign agent.
- If a mobile node is using a co-located care-of address, and receives an agent advertisement from a foreign agent on the link on which it is using this care-of address, the mobile node registers using that foreign agent (or another foreign agent on this link).
- If a mobile node uses a co-located care-of address, the mobile node registers directly with its home agent.
- If a mobile node returns to its home network, the mobile node deregisters with its home agent.

These registration processes involve the exchange of registration requests and registration reply messages. When registering using a foreign agent, the registration process takes the following steps, which the subsequent illustration depicts:

1. The mobile node sends a registration request to the prospective foreign agent to begin the registration process.
2. The foreign agent processes the registration request and then relays it to the home agent.
3. The home agent sends a registration reply to the foreign agent to grant or deny the request.
4. The foreign agent processes the registration reply and then relays it to the mobile node to inform it of the disposition of its request.

Figure 3.14 Mobile IP Registration Process



When the mobile node registers directly with its home agent, the registration process requires only the following steps:

- The mobile node sends a deregistration request to the home agent.
- The home agent sends a registration reply to the mobile node, granting or denying the request.

Also, a reverse tunnel might be required by either the foreign agent or the home agent. If the foreign agent supports reverse tunneling, the mobile node uses the registration process to request a reverse tunnel. The mobile node does this by setting the reverse tunnel flag on in the mobile node's registration request.

Network Access Identifier (NAI)

AAA servers, in use within the Internet, provide authentication and authorization services for dial-up computers. These services are likely to be equally valuable for mobile nodes using Mobile IP when the nodes are attempting to connect to foreign domains with AAA servers. AAA servers identify clients by using the Network Access Identifier (NAI). A mobile node can identify itself by including the NAI in the Mobile IP registration request.

Since the NAI is typically used to identify the mobile node uniquely, the mobile node's home address is not always necessary to provide that function. Thus, it is possible for a mobile node to authenticate itself, and be authorized for connection to the foreign domain, without even having a home address. To request that a home address be assigned, a message containing the mobile node NAI extension can set the home address field to zero in the registration request.

Mobile IP Message Authentication

Each mobile node, foreign agent, and home agent supports a mobility security association between the various Mobile IP components, indexed by their security parameter index (SPI) and IP address. In the case of the mobile node, this address is its home address. Registration messages between a mobile node and its home agent are authenticated with the Mobile-home authentication extension. In addition to Mobile-home authentication, which is mandatory, you can use the optional Mobile-foreign agent and Home-foreign agent authentications.

Mobile Node Registration Request

A mobile node registers with its home agent using a **registration request** message so that its home agent can create or modify a mobility binding for that mobile node (for example, with a new lifetime). The foreign agent can relay the registration request to the home agent. However, if the mobile node is registering a co-located care-of address, then the mobile node can send the registration request directly to the home agent.

Registration Reply Message

A mobility agent returns a **registration reply** message to a mobile node that has sent a registration request message. If the mobile node is requesting service from a foreign agent, that foreign agent receives the reply from the home agent and subsequently relays it to the mobile node. The reply message contains the necessary codes to inform the mobile node about the status of its request, along with the lifetime granted by the home agent, which can be smaller than the original request. The registration reply can also contain a dynamic home address assignment.

Foreign Agent Considerations

The foreign agent plays a mostly passive role in Mobile IP registration. A foreign agent adds all registered mobile nodes to its visitor table. It relays registration requests between mobile nodes and home agents, and, when it provides the care-of address, de-encapsulates datagrams for delivery to the mobile node. It also sends periodic agent advertisement messages to advertise its presence.

If reverse tunnels are supported, the foreign agent establishes appropriate routes to reverse tunnel all the data packets from the mobile node for a correspondent node. A foreign agent that supports reverse tunnels advertises that the reverse tunnel is supported for registration. Given the local policy, the foreign agent can deny a registration request when the reverse tunnel flag is not set. Also, the foreign agent can only distinguish two different mobile nodes with the same IP address when the mobile nodes visit on two different advertising interfaces.

Home Agent Considerations

Home agents play an active role in the registration process. The home agent receives registration requests from the mobile node (perhaps relayed by a foreign agent), updates its record of the mobility bindings for this mobile node, and issues a suitable registration reply in response to each. The home agent also forwards packets to the mobile node when the mobile node is away from its home network.

A home agent might not have to have a physical subnet configured for mobile nodes. However, the home agent must recognize its mobile node's home address through the mipagent.conf file or some other mechanism when the home agent grants registration.

Dynamic Home Agent Discovery

In some cases, the mobile node might not know its home agent address when the mobile node attempts to register. If the mobile node does not know its home agent address, the mobile node can use dynamic home agent address resolution to learn the address of its home agent. In this case, the mobile node sets the home agent field of the registration request to the subnet-directed broadcast address of the mobile node's home network. Each home agent that receives a registration request with a broadcast destination address rejects the mobile node's registration by returning a rejection registration reply. By doing so, the mobile node can use the home agent's unicast IP address indicated in the rejection reply when the mobile node next attempts registration.

Routing Datagrams to and From Mobile Nodes

This section describes how mobile nodes, home agents, and foreign agents cooperate to route datagrams to and from mobile nodes that are connected to a foreign network.

Encapsulation Types

Home agents and foreign agents support tunneling datagrams using one of the available encapsulation methods (IP in IP Encapsulation, Minimal Encapsulation, or Generic Routing Encapsulation). Mobile nodes that use a co-located care-of address can receive tunneled datagrams using any encapsulation type.

3.4 Mobility models:

Random waypoint model

In mobility management, the **random waypoint model** is a random model for the movement of mobile users, and how their location, velocity and acceleration change over time. Mobility models are used for simulation purposes when new network protocols are evaluated. The random waypoint model was first proposed by Johnson and Maltz. It is one of the most popular mobility models to evaluate mobile ad hoc network (MANET) routing protocols, because of its simplicity and wide availability.

In random-based mobility simulation models, the mobile nodes move randomly and freely without restrictions. To be more specific, the destination, speed and direction are all chosen randomly and independently of other nodes. This kind of model has been used in many simulation studies.

The movement of nodes is governed in the following manner: Each node begins by pausing for a fixed number of seconds. The node then selects a random destination in the simulation area and a random speed between 0 (excluded) and some maximum speed. The node moves to this destination and again pauses for a fixed period before another random location and speed. This behaviour is repeated for the length of the simulation.

Random walk

A **random walk** is a mathematical object, known as a stochastic or random process, that describes a path that consists of a succession of random steps on some mathematical space such as the integers.

In this model, the nodes move randomly and freely without any restriction. In RW model, the destination, speed and direction all are chosen randomly and independently of other nodes. The RW Models produce memory-less mobility pattern because it does not keep records of previous

patterns formed by the speed and location values of mobile nodes. It has advantage that it does not need any memory space but nodes move randomly anywhere without having any particular destination to reach and without pausing at any location.

Random walk mobility model is based on the argument that entities naturally move around in unpredictable ways. In this model, every node moves towards a new randomly chosen location. A random direction and speed is assigned to each node from a predefined range and nodes of a network are independent from one another [3]. A new direction is again assigned from predefined ranges whenever any node reaches the destination location. In this model, the distributions of mobility parameters are a function of time. As the mobility parameters achieve the stable state of distributions, the simulation produces consistent results. This has kept its popularity to evaluate DTN protocols. However, the variations of direction and speed in this model are limited in a certain ranges that are defined beforehand. Besides, the RW Models produce memory-less mobility pattern because it cannot keep records of previous patterns formed by the locations and speed values of mobile nodes [6]. The current speed and direction of a node is independent of its past speed and direction as well. These characteristics can generate unrealistic movements such as sudden stops and sharp turns. A variant of this model, namely Gauss-Markov Mobility Model, has been introduced to resolve this discrepancy.

Map Based Mobility Model-

movement of nodes are constrained within a map.1)a)Random Map-Based Mobility Modals (RMBM)It is the simple random Map-Based Mobility Modal (MBM). It contain all features of random walk model. In this Model, nodes move to randomly determined directions on the map following the roads as defined by the map and also it has options to selectdifferent node groups that use only certain parts of the map. In this way, it can distinguish between cars and pedestrians so that the former do not drive on pedestrian paths or inside buildings. b)Shortest Path-Based Map Based Mobility Modal (SPBMM)This model adds the concept of finding shortest path in previous RMBM. This Model also initially places the nodes in random places on the map area. However, all nodes travel to a certain destination in the map and follow Dijkstra's shortest path algorithm to discover the shortest path to the destination. When nodes reach theirdestination, they wait for a while and select a new destination. In the map all the places usually have same probability to be chosen as the next destination, but the map can also contain Points of Interest (POIs).

Group Mobility Models:

Nodes' Movements Are Dependent Of One Another.

Community-based Mobility Model (CMM)

The CMM is the first and flexible model which is directly drawn from a social network. In CMM, nodes are grouped as friends who belong to the same community and non-friends who are with different community. At the beginning, the movement area is divided into some regions as a grid and each community is assigned into a cell of the grid. A link is established between all the friend and non-friend nodes in the network which will be used later to drive node movements. In this model, nodes move between the communities based on node attraction feature [25]. The drawback of this model is gregarious behaviour of nodes. In this case, when a node has decided to exit the community, all other nodes of the community follow this node. The variant that resolve this issue by considering both node and location attraction is home-cell community-based mobility model (HCMM). HCMM introduces the idea of foreign community to mention that some nodes have also social links with communities other than the home. In this model, each node is initially is assigned to a specific community as well as each node has social links with all the other members of its home community. Some special nodes also have social links with foreign communities other than the home community [26]. The probability of moving a node from its home community towards a given community is proportional to the number of ties with nodes of the destination community

Module IV: Bandwidth Management in Cellular Mobile networks [3L]
Mathematical formulation of the channel assignment problem (CAP); CAP and generalized graph coloring; Benchmark instances; Lower bound on bandwidth.

4.1 Introduction

During the past decades, we have been witnessing the revolution in telecommunications devices and their impact on our daily life. Indeed, mobile computing has emerged as an important topic of research. People now communicate via an increasing number of devices, many of which are mobile, and the number of mobile users keeps on increasing worldwide. To be able to satisfy users, wireless networks have been used to provide integrated services, but also to support the facilities of dynamically locating the mobile terminals and enabling efficient message routing among them. Therefore, an efficient allocation of channels for proper communication is important due to bandwidth limitation. Modern wireless networks are organised in geographical cells, each controlled by a base station (BS). The use of cells can increase the capacity of a wireless network, allowing more users to communicate simultaneously. The number of simultaneous calls a mobile wireless system can accommodate is essentially determined by the total spectral allocation for that system and the bandwidth required for transmitting signals used in handling a call. MacDonald (1979) introduced the cellular concept in which the radio coverage area of a base station is represented by a cell. The cellular networks involve a relatively simple architecture within which most of the communication aspects of wireless systems can be studied. A cellular network consists of a large number of wireless subscribers who have cellular phones that can be used in cars, in buildings, on the streets and almost anywhere. There is also a number of fixed base stations, arranged to provide coverage via wireless electromagnetic transmission of the cell phones. A regular hexagon is chosen to represent a cell because it covers a larger area with the same centre-to-vertex distance compared to a square or equilateral triangle. Consequently, fewer hexagonal cells need to be placed in a cellular structure to cover a given geographical area and these cells are grouped into clusters. The entire block of frequencies is completely allocated to each cluster and the cells in each cluster use different frequencies. In this way, the frequency in the bandwidth is reused. Two major fields of interest in cellular mobile networks have evolved, namely mobility management and bandwidth management. Mobility management consists of two basic components: location management and handoff management. Location management handles the tracking of mobile terminals and the channelling of incoming calls to the mobiles. Handoff management deals with providing continuity of a call with the required quality of service, even when the users move from the coverage area of one base station to that of another base station. With the ever-increasing number of mobile users and a pre-assigned communication bandwidth, the problem of efficiently using the radio spectrum for cellular mobile communication has become a critical research issue (Chen et al., 2002, 2003; Sarkar and Sivaranjan, 1998). Channel interference in the reuse of radio spectrum cells is the major factor which needs to be considered while solving the channel allocation problem (CAP). Neglecting other influencing factors, we assume that the channel interference is primarily a function of frequency and distance. A channel can simultaneously be used by multiple base stations if their mutual separation is more than the reuse distance, which is the minimum distance at which two signals of the same frequency do not interfere. In a cellular environment, the reuse distance is usually expressed in units of number of cells. There are three types of interference that exist in a cellular environment, as

described in Chapter 2. The task of assigning frequency channels to the cells that satisfies the frequency separation constraints with a view to avoiding channel interference and using as little bandwidth as possible is known as the channel allocation problem (CAP). In its most general form, the CAP is equivalent to the generalised graph-colouring problem (Metzger, 1970) which is a well-known NP-complete problem (Hale, 1980). Because of the nature of the CAP, much research has been carried out in order to develop timeefficient heuristic or approximate algorithms, which cannot guarantee optimal solutions, however research in this field dates from the early 1970s to the present. (Gamst and Rave, 1982) defined the general form of the channel allocation problem in an arbitrary inhomogeneous cellular radio network as an optimisation problem. Several methods have been used to solve the CAP problem.

These include:

• Graph colouring Method (Gamst and Rave, 1982; Sivaranjan et al., 1989; Yeung, 2000).

• Neural network Approach (Kunz, 1991; Funabiki and Takefuji, 1992; Kim et al., 1997; L´azaro and Girma, 2001).

• Genetic algorithm (Kim et al., 1996; Lai, 1996; Lima et al., 2002; Ghosh et al., 2003; Yen and Hanzo, 2004b).

• Evolutionary strategy (Creput et al., 2005; Sandalidis et al., 1998; Vidyarthi et al., 2005).

• Local search algorithm (Wang and Rushforth, 1996; Kendall and Mohamad, 2004c).

• Simulated annealing (Li and Wang, 2001; Wang and Rushforth, 1996; Santos et al., 2001).

• Tabu search (Castelino et al., 1996; Hao and Perrier, 1996; Hao et al., 1998).

• Q-learning approach (Nie and Haykin, 1999a,b).

The advent of the cellular concept was a major breakthrough in the development of wireless mobile communication. Mobile communication has improved our lives considerably and the most popular device at the moment is the cellular phone. The major advantages of the wireless link over the wired link are:

1. Faster speed of deployment;

2. Accessibility to difficult areas;

3. Low marginal cost and effort in adding or removing a subscriber compared to the cost required to install cables for wired access.

On the other hand, the main disadvantages are that wireless signals can be received anywhere within the coverage area, and since it is easier to gain unauthorised network access, security may be at risk. Besides, it is more difficult to transmit data at a high rate in a wireless channel than a wired channel since the wireless channel is more hostile. (MacDonald, 1979) introduced the cellular concept where the radio coverage area of a base station is represented by a cell. During the early part of the evolution of the cellular concept, the system designers recognised the concept of all cells having the same shape to be helpful in systematising the design and layout of the cellular system. In the Bell Laboratories paper (MacDonald, 1979), four possible geometrical shapes were discussed: the circle, the square, the equilateral triangle and the regular hexagon.

We use a regular hexagon to represent a cell. These cells are placed in a cellular structure covering a geographical area as shown in Figure 1. The cells in a geographical area are grouped into clusters. The entire block of frequencies is completely allocated to each cluster and the cells in each cluster use different frequencies. In this way, the limited block of frequency spectrum is reused. The concept of frequency is illustrated in Figure 1, where the cluster size is seven cells and a set of co-channel cells – that is, cells using the same frequency – is shown shaded. The co-channel reuse ratio is given as:

$$\frac{R_u}{R_b} = \sqrt{3N_c},$$

(1)

where Ru is the distance between the two closest co-channel cells, Rb is the radius of the cell and Nc is a positive integer representing the number of cells per cluster (Wong, 2003).

Each cell has a base station and a number of mobile terminals (e.g. for mobile phones, palmtops, laptops or other mobile devices). The base station is equipped with radio transmission and reception equipment. The mobile terminals within a cell communicate through wireless links with the base stations associated with the cell. Several base stations are connected to the base station controller (BSC) via microwave links or dedicated leased lines. The BSC contains logic for radio resource management of the base stations under its control. It is also responsible for transferring an ongoing call from one base station to another as a mobile user moves from cell to cell.

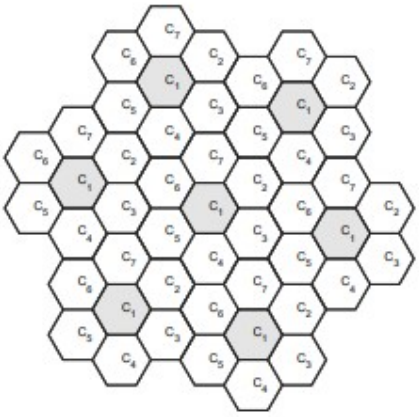


Fig. 1

Several BSCs are connected to a mobile switching centre (MSC), also known as mobile telephone switching office (MTSO). The MSC/MTSO is responsible for setting up and terminating calls to and from mobile subscribers. The MSC is connected to the backbone wire-line network such as the public switched telephone network (PSTN), integrated series digital network (ISDN) or any LAN-WAN based network. The MSC is also connected to a location database, which stores information about the location of each mobile terminal. The base station is responsible for the communication between the mobile terminal and the rest of the information network. It can communicate with mobiles as long as these are within its operating range. The range itself depends upon the transmission power of the base station. The number of simultaneous calls a mobile wireless system can accommodate is essentially determined by the

total spectral allocation for that system and the bandwidth required for transmitting signals in handling a call.

We shall now describe how the use of cells can increase the capacity of a wireless system, allowing more users to communicate simultaneously. The number of simultaneous calls a mobile wireless system can accommodate is essentially determined by the total spectral allocation for that system and the bandwidth required for transmitting signals used in handling a call. This is the same for other radio applications such as broadcast radio, AM or FM. If one considers the first-generation analogue mobile system in the US (the AMPS system), 25 MHz of spectrum is made available for each direction of transmission in the 800-900 MHz radio band. In making radio spectrum allocations for mobile radio communication, the Federal Communications Commission (FCC) in 1981 assigned the 824-849 MHz band for the uplink or reverse channel communication, mobile to base station; the 869-894 MHz was assigned to down-link or forward communication for BS to mobile. Frequency modulation (FM) was the modulation type adopted for these first-generation systems, with the 25 MHz band in each direction broken into signal channels that are 30 KHz wide, each accommodating one call. There are thus 832 analogue signal channels made available in each direction. The 832 analogue channels, or some multiple thereof using digital technology, are obviously insufficient to handle massive numbers of users mainly in urban/suburban areas. Dividing a region into a number of geographically distinct areas called cells and reusing the frequencies in these cells allows the number of communication channels to be increased. The idea of having cellular systems is to reuse channels in different cells, thus increasing the capacity. However, the same frequency assignments cannot be made in adjacent cells because of inter-channel interference. The assignments must be spaced far enough apart geographically to keep interference at tolerable levels. Channel reuse is thus not as efficient as might be expected, but the use of a large number of cells does provide an overall gain in system capacity - that is, the ability to handle simultaneous numbers of calls. In particular, if cells can be reduced in size, more of them can be added in a given geographical area, increasing the overall capacity. The recent trend is to use smaller cells (micro-cells).

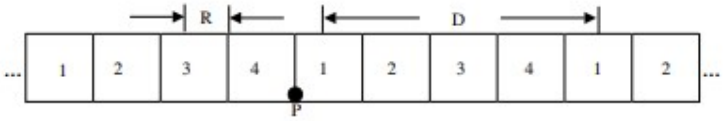


Fig. 2

To show how the introduction of cells has improved the system's capacity, we consider a onedimensional case with a first-generation analogue system as an example. Suppose that the overall band of 832 channels is first divided into four groups of 208 channels each. We label these groups as 1, 2, 3, 4 and their locations are shown in Figure 2. There are three separate cells with the same set of frequencies, so we can call this a four-cell reuse. So, given N cells in a system, 208N channels are made available compared to the original 832 possibilities when no cellular structure is introduced. With the number N of cells large enough, a significant increase in system capacity has been made possible over the original 832 channels. Had only two cells been required to separate cells using the same band of frequencies (three-cell reuse), a system with N cells would result in 277N usable channels, bringing an even larger improvement in capacity. The assignment strategy to be used depends on the tolerable interference, since spacing

same-band cells three cells apart results in less interference than spacing them into two cells apart.

4.2 The channel allocation problem

4.2.1 Channel allocation

A given radio spectrum (or bandwidth) can be divided into a set of disjoint or non-interfering radio channels. All such channels can be used simultaneously while still maintaining an acceptable received radio signal. Different methods exist to divide a given radio spectrum, these being frequency division (FD), time division (TD) and code division (CD). In frequency division, the spectrum is divided into disjoint frequency bands. In time division, the channel separation is achieved by dividing the usage of the channel into disjoint time periods called time slots. In code division, the channel separation is achieved by using different modulation codes. More elaborate techniques can be used to divide a radio spectrum into a set of disjoint channels based on a combination of the above techniques. For example, a combination of TD and FD can be used by dividing each frequency band of an FD scheme into time slots. The major driving factor in determining the number of channels with certain quality that can be used for a given wireless spectrum is the level of received signal quality that can be achieved in each channel.

In order to establish communication with a base station, a mobile terminal must first obtain a channel from the base station. A channel consists of a pair of frequencies: one frequency (the forward link/down link) for transmission from the base station to the mobile terminal and another frequency (the reverse link/up link) for transmission in the reverse direction. An allocated channel is released under two scenarios: the user completes the call or the mobile user moves to another cell before the call is completed. The capacity of a cellular system can be described in terms of the number of available channels, or the number of users the system can support. The total number of channels made available to a system depends on the allocated spectrum and the bandwidth of each channel. Due to limited availability of frequency spectrum and an increasing number of mobile users daily, the channels must be reused as much as possible in order to increase the system capacity. The assignment of channels to cells or mobiles is one of the fundamental resource management issues in a mobile communication system. The channel assignment problem was first introduced in Metzger (1970). The role of a channel assignment scheme is to allocate channels to cells or mobiles in such a way so as to minimise the probability that the incoming calls are blocked, ongoing calls are dropped, and the carrier-to-interference ratio of any call falls below a pre-specified threshold.

4.2.1.1 Interference

Radio transmission is such that the transmission in one channel causes interference with other channels. Such interference may degrade the signal quality and the quality of service. The potential types of radio interference to a call are:

1. Co-channel interference (c.c.c)- this interference is due to the allocation of the same channel to certain pair of cells close enough to cause interference (i.e. a pair of cells within the reuse distance); 2. Adjacent channel interference (a.c.c)- this interference is due to the allocation of adjacent channels (e.g fi and fi+1) to certain pairs of cells simultaneously; and

3. Co-site interferences (c.s.c)- this interference is due to the allocation of channels in the same cell that are not separated by some minimum spectral distance.

These constraints are known as Electromagnetic Compatibility Constraints and can be represented by a minimum channel separation between any pair of channels assigned to a pair of cells or a cell itself. If there are f channels to serve n cells in the system, the minimum channel separation required for an acceptable level of interference is described by a symmetric compatibility matrix C. Each element ci,j (i, j = 1, ..., n) represents the minimum separation required between channels assigned to cells i and j for an acceptable level of interference. (Gamst and Rave, 1982) defined the general form of the CAP in an arbitrary inhomogeneous cellular radio network. The reuse of channels in cellular systems is inevitable; on the other hand one has to take care of the co-channel interference, thus all channels may not be reused in every cell. However, the concept of a cellular system enables the discrete channels assigned to a specific cell to be reused in different cells separated by a distance sufficient to bring the value of co-channel interference to a tolerable level, thereby reusing each channel many times. The minimum distance required between the centres of two cells using the same channel to maintain the desired signal quality is known as the reuse distance (Ds). The cells with centre-to-centre distance of less than Ds belong to the same cluster within which no channels are allowed to be reused.

Channel allocation is one of many ways to reduce interference in a cellular network. Reduced interference leads to an increase in capacity and throughput of the system. Hence good channel allocation results in a more effective use of the frequency spectrum. Apart from reducing interference, channel allocation algorithms can also be used to adapt to traffic changes in a network, and together with reduced interference, the traffic that can be supported is higher. The task of channel allocation in a cellular system is to allocate c available channels to B base stations, where each base station i has a specific traffic demand that requires di channels. Hence, channel allocation is a permutation problem with a search space of

$$\frac{D!}{(D-c)!},$$

where D is the total traffic demand of all the B base stations. In other words D = ∑di , where di represents the number of frequencies assigned to cell i.

To illustrate the channel allocation problem, we consider a simple cellular network with three cells, each one with a base station, and let the set of base stations B = {b1, b2, b3} as shown in Figure 3.(Wong, 2003). The cellular system requires a channel separation of at least 2 for calls within the same cell and 0 for calls in neighbourhood cells. We therefore have the compatibility matrix as follows:

$$C = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} \tag{2}$$

Let us assume that the traffic demand D = [2 1 1] and the set of channels is given as {f1, f2, f3}. Define a solution A = [aij] where

$$a_{ij} = \begin{cases} 1, & \text{if channel } j \text{ is assigned to base station } i, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

A possible solution is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4}$$

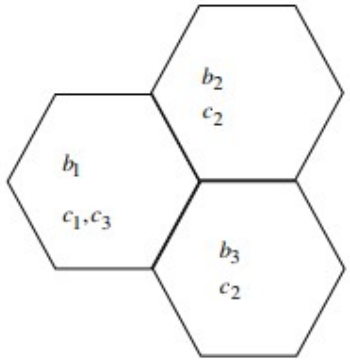


Fig. 3

4.3 Formulation of the CAP

The channel allocation problem in cellular networks has been modelled as an optimisation problem with binary solutions. The problem is characterised by a number of cells and a number of channels, n and c respectively. The solution to the problem is represented by a matrix A. Each element of the matrix is defined according to the equation (3).

A general form of matrix A is shown in Figure 4. Given the compatibility matrix C and the demand vector D, the aim of the channel allocation problem in the cellular network is to find a conflict-free assignment with the minimum number of frequencies.

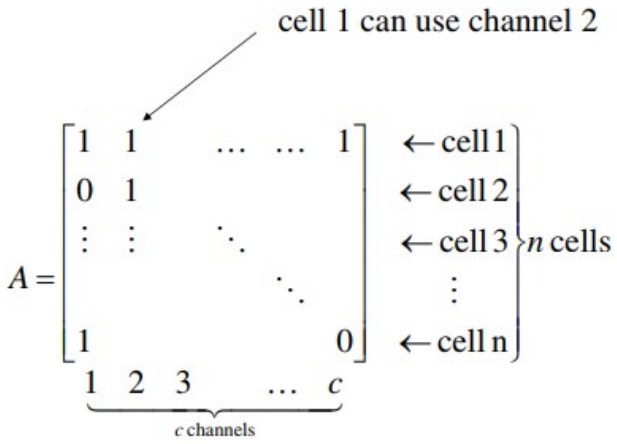


Fig. 4

4.4 Channel allocation schemes

Channel allocation schemes can be divided into a number of different categories depending on the comparison basis. For example, when channel algorithms are compared based on the manner in which co-channels are separated, they can be divided into three main categories:

1. Fixed channel allocation (FCA) (Cox and Reudink, 1972a),

2. Dynamic channel allocation (DCA) (Cox and Reudink, 1972b),

3. Hybrid channel allocation (HCA) (Kahwa and Georganas, 1978).

In FCA schemes, the area is partitioned into a number of cells and a number of channels assigned on the desired signal quality. Since the channels in FCA are static, it is not easy to adapt to changes in interference and traffic conditions. As mentioned earlier, in order to overcome those deficiencies of FCA schemes, DCA strategies have been introduced. In DCA, all channels are placed in a pool and they are assigned to new calls as needed, contingent upon a set of conditions (e.g. compatibility matrix) being satisfied. Unlike FCA, the base stations employing DCA do not own any particular channels and a channel is released when a call is completed. At the cost of higher complexity, DCA schemes provide flexibility and traffic adaptability. Although DCA performs better than FCA under high to moderate traffic, its performance is worse than that of FCA under conditions of heavy traffic changes (Katzela & Naghshineh, 1996). To take advantage of both strategies, HCA techniques have been designed. HCA schemes are a mixture of FCA and DCA techniques. In HCA strategies, channels are grouped into two sets; one set of channels is statistically assigned to a base station as in FCA and the second set is placed in a central pool, assigned in a DCA manner.

4.4.1 Fixed channel allocation

In the FCA strategy a set of nominal channels is permanently allocated to each cell for its exclusive use. A definite relationship is assumed between each channel and each cell, in accordance with co-channel reuse constraints. The total number of available channels in a system is divided into sets, and the minimum number of channel sets Nc required to serve the entire coverage area is related to the reuse distance as defined by equation (1).

In a simple FCA strategy, the same number of nominal channels is allocated to each cell. This uniform channel distribution is efficient if the traffic distribution of the system is also uniform. In this case, the overall blocking probability of the system is the same as the call blocking probability in a cell. Since cellular systems can be non-uniform with temporal and spatial fluctuations, a uniform allocation of channels to cells may result in high blocking in some cells, while others may have a sizeable number of redundant channels. This may result in poor channel utilisation. In non-uniform channel allocation the number of nominal channels allocated to each cell depends on the expected traffic profile in that cell. Therefore, heavily loaded cells are assigned more channels than lightly loaded ones. Channels are allocated to cells in such a way that the number of blocked calls of the entire system is minimised.

4.4.2 Dynamic channel allocation

Due to short-term temporal and spatial variations of traffic in cellular systems, FCA schemes are not able to attain high channel efficiency. To overcome this, DCA schemes have been studied, and in contrast to FCA, there is no fixed relationship between channels and cells in DCA. All

channels are kept in a central pool and are designed dynamically to radio cells as new calls arrive in the system. After a call is completed, its channel is returned to the central pool.

In DCA, a channel is eligible for use in any cell provided that signal interference constraints are satisfied. In general, more than one channel might be available in the central pool to be assigned to select a cell that requires a channel. Thus some strategy must be used to assign channels and the main idea in DCA schemes is to evaluate the cost of using each candidate. The one which gives the optimal solution is chosen. The cost function varies from DCA schemes. Some examples of cost functions (Tekinay and Jabbari, 1991):

• The future blocking probability in the vicinity of the cell;

• The usage frequency of the candidate channel;

• The reuse distance;

• Channel occupancy distribution under current traffic conditions;

• Radio channel measurements of individual mobile users; and

• The average blocking probability of the system.

Different DCA schemes exist, such as centralised DCA schemes and distributed DCA. These are discussed in more detail in Katzela & Naghshineh (1996).

4.5 Mathematical formulation of the CAP

4.5.1 Problem formulation – FCA

The basic model of the channel assignment problem can be represented as follows (Sivaranjan et al., 1989; Wang and Rushforth, 1996):

1. n: the number of cells in the network;

2. di : the number of frequencies required in cell i ($1 \leq i \leq n$) in order to satisfy channel demand;

3. C: compatibility matrix, C = (cij) denotes the frequency separation required between cell i and cell j; and

4. fiki : a radio channel is assigned to k th call in cell i.

The objective of the CAP is

$$\min_{i,k_i} f_{ik_i},$$

subject to

$$|f_{ik_i} - f_{jk_j}| \geq c_{ij},\ \forall\, i,\ j,\ k_i \neq k_j. \tag{5}$$

The channel allocation problem in the cellular network is finding a conflict-free assignment with the minimum number of total frequencies, where C, the compatibility matrix and D, the demand vector are given. In other words, one tries to find the minimum of

$$\max_{ik} f_{ik}.$$

4.5.2 Problem formulation – DCA

The DCA scheme which aims to reduce the effect of unbalanced loading due to evenly distributed traffic sources, assigns channels to cells on a call-to-call basis. Notation (Sivaranjan et al., 1990):

1. n: the number of cells in the system;

2. C: compatibility matrix, C = (c_{ij}) denotes the frequency separation required between cell i and cell j;

3. ni , 1 ≤ i ≤ n: the number of calls in progress in cell i;

4. pi , 1 ≤ i ≤ n: the probability that a new call arrives in cell i;

5. ρ: the total traffic in the system.

6. ρi = piρ, 1 ≤ i ≤ n: the traffic in cell i;

7. Nf : the number of (contiguous) frequency channels available. These channels are numbered 1 through Nf ; and

8. fiki , 1 ≤ i ≤ n, 1 ≤ ki ≤ mi : the frequency assigned to the k th call in the i th cell

Constraints: $|f_{iki} - f_{jlj}| \geq c_{ij}$ for all i, j, k_i , l_j except for i = j, $k_i = l_j$. Assumptions:

• Call arrivals in cell i are independent of all other arrivals and obey a Poisson distribution with parameter ρi ;

• Call holding times are exponentially distributed with mean call duration.

• There are no calls handed off between cells; and

• Blocked calls are cleared.

4.5.3 Problem formulation – HCA

The total number of channels available for service is divided into fixed and dynamic sets. The fixed set contains a number of nominal channels that are assigned to cells as in the FCA schemes and, in all cases, are to be preferred for use in their respective cells. The dynamic set is shared by all users in the system to increase flexibility. A request for a channel from the dynamic set is initiated only when the cell has exhausted using of all its channels from the fixed set.

4.6 Methods used to solve channel allocation problem

4.6.1 Graph colouring approach

Graph colouring is an assignment of colours to elements of a graph subject to certain constraints. It is a way of colouring the vertices of a graph such that no two adjacent vertices share the same colour. Graph colouring has many applications such as scheduling (Marx, 2004), sudoku, mobile

radio frequency assignment, pattern matching and register allocation (Chaitin, 1982), as well as theoretical challenges.

The problem discussed in the above can be modelled as a graph colouring problem in the following way. The vertices V of graph G consist of a set of calls to be assigned. In our case they are $call_{11}$, $call_{21}$, $call_{31}$ and $call_{41}$, $call_{42}$, $call_{43}$. An edge, $E \rightarrow e(call_{ij}, call_{kl})$ for a graph G can be defined as a constraint between $call_{ij}$ and $call_{kl}$, shown in Figure 5. The aim is to schedule the calls such that each 'clashed' call will be assigned a different channel. In the concept of graph colouring, the channel is represented by a colour and the objective is to minimise the number of colours used.

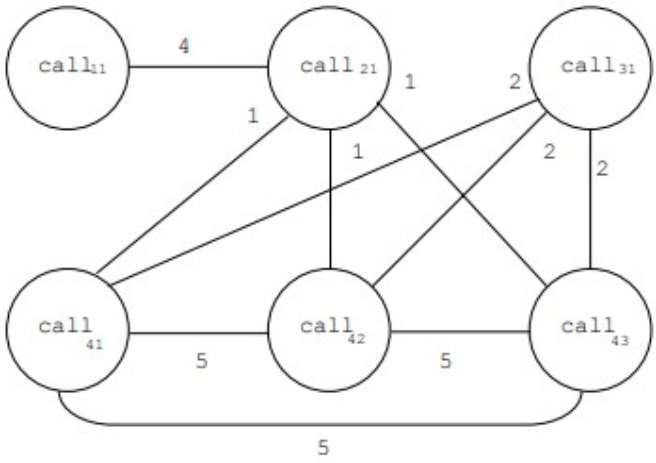


Fig. 5

In general, there are two objectives in solving the channel allocation problem:

• Given a traffic demand, base station number and compatibility matrix, we find the number of frequency channels without any interference constraints; in other words, we minimise the total bandwidth (span) of radio channels such that traffic demand and interference constraints are satisfied.

• Given a number of frequencies, a number of base stations, traffic demand and compatibility matrix, we minimise channel interferences such that demand constraints are satisfied.

The basic idea of the graph theoretic approach is to list the calls in some order and use either a frequency exhaustive assignment strategy (FEA) or a requirement exhaustive assignment strategy (REA). The FEA strategy starts with the first cell in the ordered list and each call is assigned the least possible frequency, consistent with previous assignment, that is, without violating the separation constraints. A pseudo-code of FEA is given in Algorithm 1.

Algorithm 1 Frequency exhaustive assignment strategy

1 - Loop 1: over all call i in the call list
2 - Loop 2: for each channel j within the lower bound
3 - If i is not assigned and j can be assigned to i
4 - Assign j to i
5 - Break Loop 2
6 - End If
7 - End Loop 2
8 - End Loop 1

In REA, an attempt is made to assign the first channel to all cells which have unsatisfied channel requirements, starting from the first cell.

Algorithm 2 Requirement exhaustive assignment strategy

1 - Loop 1: for each channel j within the lower bound
2 - Loop 2: for each call i in the call list
3 - If i is not assigned and j can be assigned to i
4 - Assign j to i
5 - Break Loop 2
6 - End If
7 - End Loop 2
8 - End Loop 1

Then the same procedure is followed to assign the remaining calls. The procedure is repeated until all the channel requirements are exhausted. We present a pseudo-code for the REA in Algorithm 2.

Metzger (1970) recognised that the classical node-colouring problem in graph theory is analogous to frequency assignment problems where only co-channel constraints are involved. A graph is a set of nodes partially or completely interconnected by lines. In colouring a graph, adjacent nodes may not receive the same colour. The objective is to find the minimum number of colours required to colour all the nodes. In the frequency assignment case, the nodes represent requirements and the lines connect pairs of requirements that cannot be assigned to the same channel, and the colours represent the channels.

Metzger (1970) employed decomposition procedures which are described as follows:

• A node of lowest degree from the graph along the lines connected to it is removed;

• Removal of a node of lowest degree in the reduced graph; and

• Procedure is repeated until the last node is removed from the graph.

The concept of the node colouring order technique by Metzger has shown to be an important technique in solving the CAP using the graph colouring approach. Another colouring procedure

in which the distribution of the number of colours used is made as even as possible is known as the uniform assignment strategy.

In Zoeller and Beall (1977), the authors have used three different assignment strategies: frequency exhaustive, requirement exhaustive and uniform assignment with three different order techniques, which are node colouring, node degree and random order. The degree of a cell i is defined as

$$deg_i = \sum_{j=1}^{n} d_i c_{ij}. \tag{6}$$

The node degree order arranges the cells in decreasing order of their degrees and the node colouring order is as follows: of the n cells, the cell with the least degree is placed at the last nth place list. This cell is eliminated from the system and the degrees of the remaining cells are recomputed. Now, the cell with the least degree is placed at the (n − 1)th position in the list, and eliminated from the system. This process is continued until the ordering is complete.

These different procedures were evaluated for the following cases: co-channel, adjacent and cosite cases. The results obtained by Zoeller and Beall (1977) show that about 35% more spectrum may have been committed to existing requirements than is really needed. This represents a significant under-utilisation of the spectrum. Upwards of 20-25% of the spectrum might be recovered in reassigning present users with the use of node-colouring order-based assignment procedure.

Gamst and Rave (1982) summarised four existing sequential algorithms. The first algorithm has four different versions by combining two different assignment strategies, namely the frequency exhaustive assignment and the requirement exhaustive assignment, and two different order strategies given by the node degree order and the node colouring order (Zoeller and Beall, 1977). The second algorithm repeatedly assigns frequencies according to the assignment difficulty of requirements (Box, 1978). The third algorithm uses the heuristic geometric principle of maximum overlap of denial areas. It states that a frequency should be assigned to a cell whose denial areas have the maximum overlap with the existing denial area of that frequency. The fourth algorithm is based on graph theory, where the clique number plays a key role. Sivaranjan et al. (1989) proposed an O(n 2) time sequential heuristic algorithm based on the first algorithm introduced by Gamst and Rave (1982). They applied their algorithm to several problems, where the values of total frequencies in solutions are shown without any actual assignment results. Later, Sengoku et al. (1991) formulated a channel offset system design using a graph theoretical concept in which the degree of co-channel interference between cells was represented.

Module V: Localization of Nodes in a Mobile Network [4L]

Different approaches, Indoor and outdoor localizations, LOS and NLOS signals, Outdoor localization techniques – triangulation (TOA-based, AOA- based), errors due to inaccuracies in coordinates of beacon nodes and in measurements.

Localization of wireless devices is a crucial requirement for many emerging applications such as environmental monitoring, intelligent transportation, home automation, health-care monitoring and social networking. For instance, in an environmental monitoring application such as forest fire detection or air pollution monitoring, the collected informa-tion is worthless without the location of nodes. Wireless nodes could be equipped with a GPS to acquire their locations, but

this is currently a costly solution in terms of energy and price. Thus, in the recent years, several localization algorithms that aim at obtaining nodes locations with a lower cost have been proposed. In this chapter we give a review of the state of the art research concerning localization in wireless networks. We first present the localization problem formulation, we then propose a taxonomy of the existing localization techniques and finally detail some representative localization algorithms.

5.1 Definitions and Problem Formulation

Some definitions are first needed in order to understand the localization problem.

Definition 1 (Unknown Nodes (U))**.** *A node* $u \in U \Longleftrightarrow$ *u is not aware of its own position. Unknown nodes are also referred to as normal nodes or blindfolded nodes.*

Definition 2 (Anchor nodes (A))**.** *A node* $a \in A \Longleftrightarrow$ *a is aware of its own posi-tion (through manual configuration or GPS). Anchor nodes are also called beacon nodes, landmarks or reference nodes.*

Definition 3 (Localization problem)**.** *Given a network with a set* N *of nodes, m an-chor nodes in* $A \subset N$ *with known positions* $\{X_1, ..., X_m\}$*, k unknown nodes in* $U \subset N$

$$\{ \qquad \} \qquad \{ \qquad \}$$

with unknown positions $X_{m+1}, ..., X_{m+k}$ *, estimate the positions* $X_{m+1}, ..., X_{m+k}$ *of the unknown nodes as close as possible to their true positions* $\{X_{m+1}, ..., X_{m+k}\}$.

5.2 Taxonomy of Localization Techniques

We provide in this section a taxonomy of the existing localization techniques. This taxonomy provides general guidelines for understanding the differences between the existing localization techniques.

5.2.1 Target vs. self-localization

Depending on their final goal and on their different fields of application, localization techniques can be categorized into two groups: target localization and self-localization.

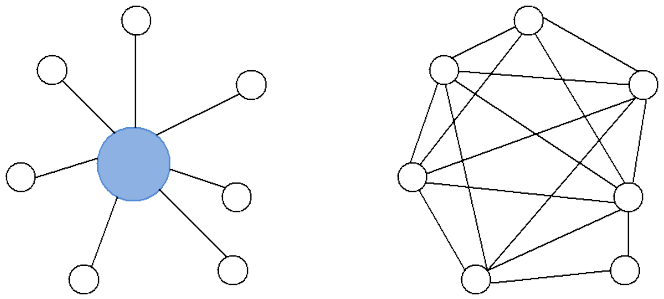
The objective of target localization is to determine the location of a target (e.g., hu-man, animal, vehicle, device). Target localization can be classified into two categories: active target localization and passive target localization. In active target localization the target actively emits a specific signal that can be received and analyzed by a reader [Savi 16]. Active target localization has a broad range of applications such as asset in-ventory and resource management. In passive target localization, the target does not actively participate in the localization process, it is rather just a reflecting/scattering ob-ject [Han 14]. Passive target localization is crucial for many applications such as crimes prevention and tracking, surveillance and medical patient monitoring.

In self-localization, unknown nodes determine their positions by themselves. Self-localization can be classified into two categories: active self-localization and passive self-localization. In passive self-localization, existing beacon signals are used by the unknown nodes to passively deduce their own positions [Hadd 16]. In active self-localization, un-known nodes actively inquire the location information from their surrounding environment [Reza 11].

Self-localization is necessary in many applications such as environmental mon-itoring applications where the measurement data are worthless without the location of the measuring node.

5.2.2 Centralized vs. distributed localization

Centralized localization algorithms are designed to run on a sufficiently powerful central base station [Tomi 15]. First the base station collects the environmental information from the different sensor nodes. Then, based on the collected information, it computes the po-sition of each sensor node and migrates them back to the respective nodes. Centralized algorithms eliminate the problem of nodes computational limitations but they introduce a large communication cost due to transporting data to and from the base station. Hence, centralized algorithms are only suitable for small networks. In contrast, distributed al-gorithms are designed to run on each node [Meye 16]. Unknown nodes positions are estimated based only on the inter-nodes communication. Due to the lack of global in-formation, distributed localization is usually less accurate than the centralized one but it considerably reduces the communication costs. Figure 5.1 illustrates the difference between the centralized and the distributed techniques.



CENTRALIZED LOCALIZATION DISTRIBUTED LOCALIZATION

Figure 5.1 – Centralized vs. distributed localization techniques

Range-based vs. range-free localization

Depending on their ranging assumption, localization techniques can be divided into range-based and range-free.

Range-based localization

Range-based algorithms use inter-nodes distances or angles to estimate the nodes posi-tions. They use special *measurements techniques* such as the time of arrival, the angle of arrival, and the received strength of a given transmitted *signal* to calculate the distance or angle separating two sensors.

Signal technologies: the choice of the signal technology used by sensor nodes for localization depends on the considered environment, application as well as the required precision, range and cost. These technologies include infrared (IR) [Seke 15], ultrasound [Filo 10], magnetic [Song 13], optical [Suh 16] and radio frequency signals. Radio technology is the most widely used technology for localization. Depending on the type of the used frequency range, the radio frequency signals can be classified into different groups: radio frequency identification (RFID) [DiGi 14], WIFI (IEEE 802.11) [Yang 15], zigbee (IEEE 802.15.4) [Chen 11], bluetooth (IEEE 802.15.1) [Gu 15], wide area cellular [Abu 16] and ultra-wideband (UWB) [Reyn 13]. Table 5.1 summarizes the different signal technologies used for localization.

Measurement techniques: There are three major measurement techniques to deter-mine the distance/angel between the nodes: the Time Of Arrival (TOA), the Angle Of Arrival (AOA) and the Received Signal Strength (RSS). In the TOA measurement tech-nique, the distance separating a receiver from a sender is calculated through multiplying the propagation time by the speed of the signal [Shen 12]. TOA based techniques re-quire a direct line-of-sight path between the transmitter and the receiver. The presence of obstacles in between them leads to later-arriving signals and hence inaccurate rang-ing estimations. Time synchronization between the transmitter and the receiver is also usually needed. There are however some existing works where TOA measurements are done without time synchronization [Chen 12a]. The AOA measurement technique typ-ically relays on the use of radio or microphone arrays to estimate the angel separating the receiver from the transmitter [Wang 15]. Systems based on the AOA measurement technique require specific hardware, they are thus expensive in terms of manufacturing cost, energy consumption and complexity. The accuracy of AOA based techniques also degrades as the distance between the transmitter and the measuring unit increases. The RSS measurement technique depends on the fact that the signal strength attenuates with distance [Yagh 14]. With the attenuation information, a receiving node is able to calculate its distance to the transmitting node. The RSS based techniques typically use radio signals. Indeed, the use of radio signals do not require any additional hardware since most of the radio communication devices come with built-in RSS indicator (RSSI) hardware.

Technology	Remarks
Infrared	• Widely used in indoor localization • Low cost and low power • Requires close proximity and line of sight between the transmitter and the receiver • Is difficult to read in the presence of sunlight

	• Typical range: up to 5m
Ultrasound	• Cannot penetrate through the walls • Affected by reflected signals and other noise sources (e.g., jangling metal objects) • Typical range: 3-10m
Magnetic	• Magnetic sensors are small and cheap • No line of sight requirement • Typical range: 1-3m
Optical	• Requires line of sight • Affected by many interference sources (e.g., light, weather) • Typical range: up to 5m
Radio frequency	• Can pass through buildings, human body and other obstructions • Affected by multipath 1) RFID: − Light and small tags that can be attached to people or equipments − Typical range: 1-10m 2) WIFI: − Uses the existing WLAN infrastructure for localization: lower cost − Typical range: 50-100m 3) Zigbee: − Low cost and low power − Typical range: 10-30m 4) Bluetooth: − Low cost and low power − Typical range: 10-15m 5) Cellular − Localizes a mobile within a cell coverage area

	− Typical range: 100-150m
UWB	• No line of sight requirement • Less multipath distortion than the other RF technologies • High penetration • Typical range: 10m

Table 5.1 – Signal technologies used for localization

late its distance to the transmitting node. The RSS based techniques typically use radio signals. Indeed, the use of radio signals do not require any additional hardware since most of the radio communication devices come with built-in RSS indicator (RSSI) hardware that directly provides the RSS measurements. Radio waves are nevertheless vulnerable to the environmental dynamics which may affect the accuracy of the distance estimations. Some works [Tomi 16] considered hybrid schemes combining two different measurement techniques in order to ameliorate the range estimations. Figure 1.2 illustrates the different range measurement techniques.

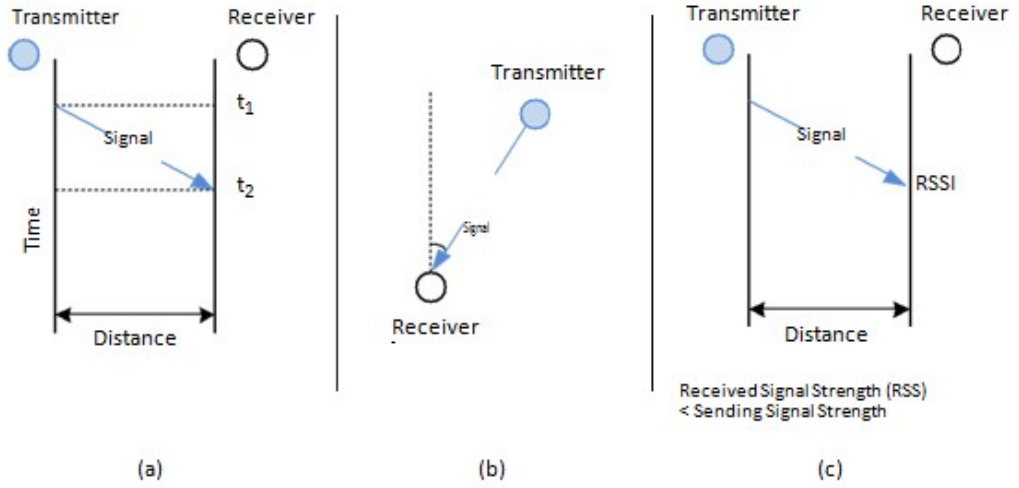


Figure 1.2 – Range based localization: (a) TOA (b) AOA (c) RSS

Figure 5.2: Range based localization

Range-free localization

Range-free localization algorithms make no assumption about the availability of inter-nodes distances or angles to estimate the locations [Zaid 16]. They instead rely on topol-ogy and connectivity information assuming an isotropic network where the hop count between nodes is proportional to their distance. Range-free algorithms provide promising solutions for the localization problem since they do not require extra hardware. However, because of the absence

of range information, the positions estimations obtained by these methods are usually less accurate than those obtained by the range-based methods.

5.2.4 Network-based vs. non-network-based localization

The network-based localization techniques use the already existing network infrastructure, such as WLAN, and consequently avoid the expensive and the time-consuming installation of the localization infrastructure [Wu 13]. The non-network-based localization techniques use dedicated infrastructure for positioning, such as sensor-based positioning systems [Suh 16]. The non-network-based localization techniques are more costly and less time

5.2.4 Network-based vs. non-network-based localization

The network-based localization techniques use the already existing network infrastructure, such as WLAN, and consequently avoid the expensive and the time-consuming installation of the localization infrastructure [Wu 13]. The non-network-based localization techniques use dedicated infrastructure for positioning, such as sensor-based positioning systems [Suh 16]. The non-network-based localization techniques are more costly and less time effective than the network-based localization techniques but offer, on the other hand, more control over the physical specifications and hence over the quality of the location estimations.

1.2.5 Outdoor vs. indoor localization

Indoor localization is more challenging than outdoor localization due to the complexity of the indoor environment. The various obstacles (e.g., walls, equipment), the mobility of people and the interference with other networks traffic degrade the accuracy of the positioning. Some works have tried to deal with the complexity of the indoor environment using the fingerprinting technique (also known as scene analysis) [Seet 12, He 16]. In this technique, an offline training phase is used in order to collect the signal features (fingerprints) for a particular indoor scene. The estimated location of a given node is calculated during the online phase based on these offline collected measurements. The drawback of the fingerprinting technique is that it requires a lot of pre-processing work and is ineffective in dynamic and changeable environments. A new training phase should be executed when there is any change in the environment.

1.2.6 Mobile vs. static nodes

Based on the mobility state of the anchor and normal nodes, existing localization algo-rithms can be classified into four groups: (1) static anchors and static normal nodes (2) static anchors and mobile normal nodes (3) mobile anchors and static normal nodes (4) mobile anchors and mobile normal nodes. The scenario of both static anchors and normal nodes is the most studied and hence the most mature localization scenario. In the second scenario (static anchors and mobile normal nodes), typically a small number of static anchors are mounted in discreet locations like ceilings or walls in order to track or help the unknown nodes estimate their coordinates. In the third scenario (mobile anchors and static normal nodes), a number of mobile anchors traverse the deployment region and periodically broadcast their coordinates. Unknown mobile nodes uses these location an-nouncements in order to infer their own location. In the last scenario (mobile anchors and mobile normal nodes), mobile anchors locations are periodically used in order to localize the set of mobile normal nodes.

When the unknown nodes are static (scenarios 1 and 3), the localization process can be executed only once (e.g., during initialization). However, when the unknown nodes are mobile (scenarios 2 and 4), the localization process must be frequently executed in order to determine the continuously changing positions of nodes.

We summarize in Figure 1.3 the different discussed categories and cite representative works in each category.

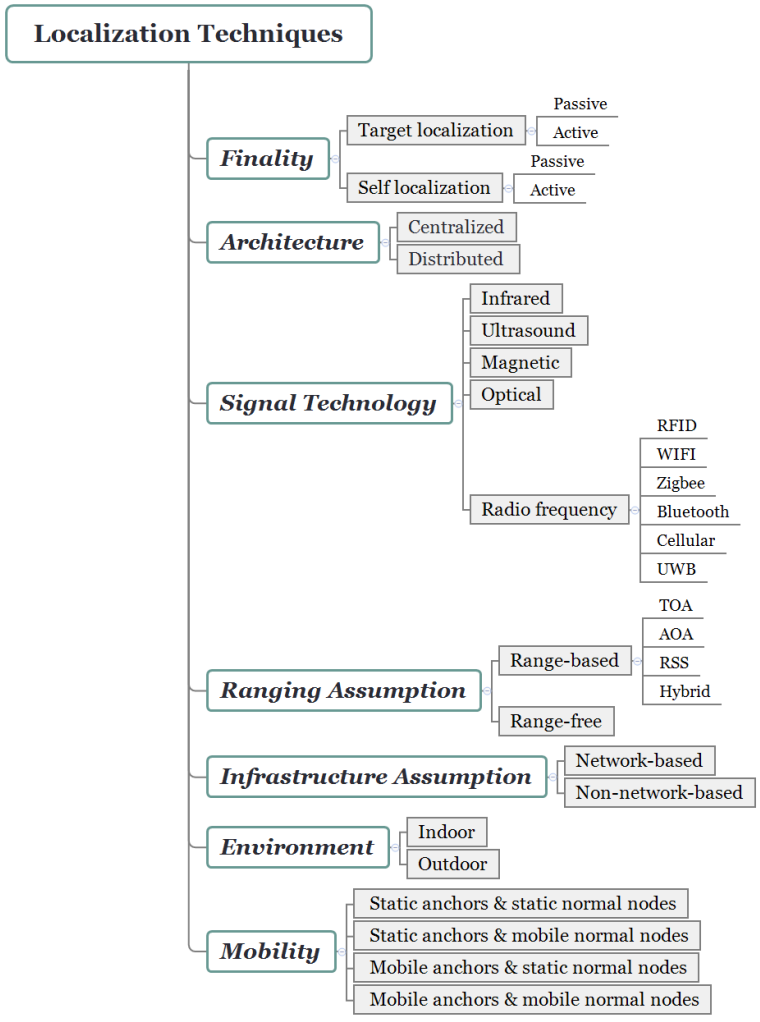


Figure 5.3 – Taxonomy of localization techniques

Solutions to Localization Problem

We review next a selected set of representative localization algorithms.

Dv-hop: Dv-hop [Nicu 01] is a classic range-free localization algorithm. It works as follows. First all anchor nodes broadcast their locations. The messages are propagated hop by hop to

reach all nodes in the network. Each node maintains a table containing all anchors locations and the least number of hops from each anchor. Once an anchor receives the coordinates of all the other anchors, it estimates the average distance per hop and broadcasts it. The average distance per hop d_i an anchor situated at (X_i, Y_i) computes is calculated as follow:

$$d_i = \frac{\sqrt{\sum_{i \neq j} (X_i - X_j)^2 + (Y_i - Y_j)^2}}{\sum_{i \neq j} h_{ij}}, \quad i = j, \ all\ landmarks\ j \tag{5.1}$$

Where h_{ij} denotes the minimum hop-count between anchors i and j. When receiving the average distance per hop (usually received from the closest anchor), a non-anchor node determines its distance from each anchor by multiplying the least number of hops to the anchor with the average distance per hop. Then it applies a multilateration (positioning using differences in distances) procedure to estimate its location. DV-Hop has the ad-vantage of involving only few beacon nodes. Its considerable disadvantage is that it fails in networks with irregular topologies, where the variance in actual hop distances is very large.

Module VI: Message Communication in Ad Hoc Networks [6L]

Collision avoidance mechanism (different schemes for a deterministic transmission schedule), collision resolution mechanism – successive partitioning approach; Time slot assignment based on location information, Point-to-point routing in ad hoc networks – proactive, reactive and hybrid approaches, different protocols - DSDV, DSR, AODV, TORA, ZRP

6.1 Collision avoidance & resolution mechanism

A mobile ad-hoc network (MANET) consists of mobile hosts equipped with wireless communication devices. The transmission of a mobile host is received by all hosts within its transmission range due to the broadcast nature of wireless communication and omni-directional antennae. If two wireless hosts are out of their transmission ranges in the ad hoc networks, other mobile hosts located between them can forward their messages, which effectively build connected networks among the mobile hosts in the deployed area. Due to the mobility of wireless hosts, each host needs to be equipped with the capability of an autonomous system, or a routing function without any statically established infrastructure or centralized administration. The mobile hosts can move arbitrarily and can be turned on or off without notifying other hosts. The mobility and autonomy introduces a dynamic topology of the networks not only because end-hosts are transient but also because intermediate hosts on a communication path are transient.

Characteristics

- Operating without a central coordinator
- Multi-hop radio relaying
- Frequent link breakage due to mobile nodes

- Constraint resources (bandwidth, computing power, battery lifetime, etc.)
- Instant deployment

Applications

- Military applications
- Collaborative computing
- Emergency rescue
- Mesh networks
- Wireless sensor networks
- Multi-hop cellular networks
- Wireless Community Network

Major Issues and Challenges

- Hidden terminal problem
- Exposed terminal problem
- Channel efficiency
- Access delay and fairness
- Differential service
- Realistic mobility modeling
- power-aware routing
- Constructing virtual backbone
- Distinguish contention, packet drop, and noise errors
- Security
- Efficient multicasting

In computer networks, there are many nodes and they have to transmit data packages over the same carrier. This carrier can be an optic cable in wired networks while it is a frequency in wireless networks. Owing to this networking principle, if two nodes in the same network attempt to send data packages to the communication line at the exact same time, a collision occurs. Collisions are important problems for networks because they violate data transmission and results in loss of information. When any collision occurs in the network, the communication stops; ultimately, data packages are dropped. Collisions always results in less throughput of the network, high network load, high delay and high data drop rate.

As mentioned previously, owing to collisions, network nodes face with loss of packet integrity. That means a proper communication cannot be established in the network. In seven layer OSI model, Media Access Control (MAC) layer is responsible for avoidance of package collision. MAC sub layer performs this task through avoidance protocols. These protocols play a critical role in preventing data collision; they aim to rule situations out which multiple nodes access to the network at the exact same time and to provide packet transmission to any node without any collision. There are some protocols that mostly used and are developed to prevent collisions in the networks such as CSMA, MACA and MACAW.

CSMA protocol

The popular Carrier Sense Multiple Access/Collision Detection (CSMA/CD) MAC method is used for wired network. In CSMA/CD, when a node wants to send over the network, first it sense the wire medium whether it's idle or busy. If it's idle, the node sends its data with sensing the medium continually. Otherwise, the node delays its transmission to avoid a collision with existing packets. While in the wireless networks, the signal strength is inversely proportional to the square distance from the transmitter node, thus nodes, which are out of transmitter's range, can't sense the transmitted signal causing problems as illustrated in fig. 6.1. There are three nodes A, B, C. Node B is within the range of each nodes A and B, node C is out of the range of

A. Node A wants to send to B, wherefore node A waits until the medium is idle, then A starts transmitting to B. Node C wants to send to node B while B is receiving data from A. But C can't sense the transmitted signal from A, thus C starts transmitting to B causing collision at node B. This problem is called "hidden- terminal problem".
Another problem is illustrated in fig.6.2, there are four nodes A, B, C, D. Nodes B and C are within the range of both nodes A and D, but D is out of A's range. While node B is transmitting data to node A, node C wants transmitting data to D. But C senses the transmitted signal from A, thus C delays its transmission to D. Even through a transmission from C does not interfere with the reception at node A, this case is called "exposed terminal problem".

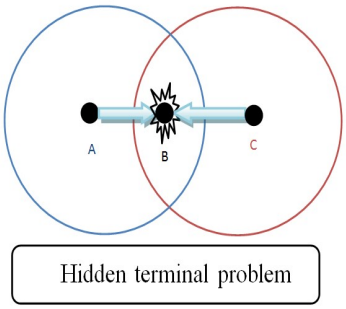


Fig. 1. Illustration of hidden terminal problem

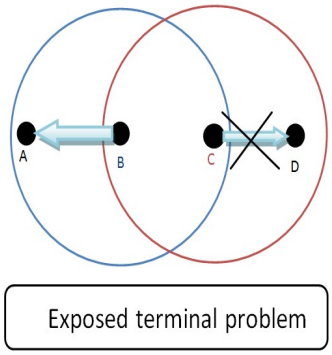


Fig. 2. Illustration of exposed terminal problem

MACA Protocol

The Multiple Access Collision Avoidance (MACA) protocol doesn't fully solve the problem of CSMA/CD. It uses two additional packets, Request To Send"RTS" and Clear To Send"CTS", to reduce the collision at receiver. These packets are shorter than data packets; however, they contain the length of the data frame that will follow. Let us conceder an example with four nodes A, B, C, D as shown in fig.6.3. The node A wants to send data to the node B, so A broadcasts a RTS packet then B replay to A by sending a CTS packet. At node C the CTS packet collides with a RTS packet sent from D, so C doesn't replay to the RTS from D. But A starts sending data to B after A receives CTS from B. The node D sends another RTS packet because it didn't receive a CTS from C and then C responses to D by sending back a CTS packet which may collide with data packet at B node if the data reception isn't complete. We notice there is no acknowledge for receiving the data, thus the retransmission is started by higher layer (transport layer). The wireless LANs use the MACA protocol, but they use additional control packets like

Acknowledgement packet (ACK) which is received by the sender from the receiver node after data reception is complete. Thus, the arrangement of transmitted packets is (RTS-CTS-DS - ACK).

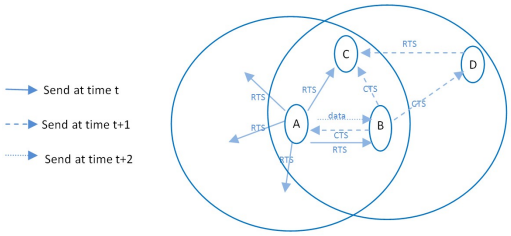


Fig. 3 : Illustration of RTSCTS mechanism

MACAW Protocol

Multiple Accesses with Collision Avoidance for Wireless (MACAW) is a widely used MAC sub layer protocol. MACAW is useful for mobile ad-hoc networks. It contains new collision avoidance mechanisms. By these mechanisms data transmission is completed in five steps. These five steps are Request-to-Send (RTS), Clear-to- Send (CTS), Data Sending (DS), data packages and Acknowledgement (ACK). RTS is a message, sent from data sender node to receiver node, notifies that a node attempts to transmit data to another node. CTS message is a respond for transmission request. If receiver node available for transmission then sends a CTS message. DS frame informs receiver node about the size of data package. After that, data transmission starts. When it completes properly, receiver node sends an ACK message to sender node. ACK notifies that data transmission completed successfully.
.

6.2 Time slot assignment based on location information

Position based routing protocol use additional information about the position of mobile nodes. It uses location services to determine the exact position of source node, neighbor node and the destination node. By the use of GPS or some kind of location services, it maintains the position information about the nodes and determines the exact co-ordinates of the nodes in any geographical direction, and thus leads in route discovery mechanism. It doesn't require the establishment and maintenance of the route, neither it has to update the routing table. It does all its activity by the use of location services and some kind of forwarding strategy, which is used in forwarding the packets from source node. The advantage of this protocol can supports the delivery of packets to all nodes in a given region of geographic this type of service is known as a geo-casting. It can distinguish three main forwarding the packets strategy for position based routing: Greedy forwarding, restricted directional flooding and hierarchical approach .

Performance of Position Based Routing Protocol

The following performance of position based routing strategy of the protocol can be according to their important design parameters are:

• Loop Free
• Distributed Operation
• Path Strategy

• Packet Forwarding
• Path Selection Metric
• Memory (State)
• Guaranteed Message Delivery
• Scalability
• Overhead
• Adaptive to Mobility

A. Location-aided Routing Protocol (LAR)

This protocol is based on the use of location information about the mobile nodes by using location services like GPS and many more to reduce the route discovery overhead, the two regions are defined i.e. Request zone and Expected zone. Request zone is the area in which the node forwards the route request only when the node is inside the zone. When the nodes does not belongs to request zone then it simply discards the message. Expected zone is the area in which there is the maximum probability of finding the destination nodes. Since the destination node is mobile, We can calculate its probabilistic position by assuming its average velocity multiplied by difference in time interval. We assume the expected zone to be circular with the radius v (t1-t0). Difference between at time t0 the location of destination node and at time t1 the location of destination node multiplied by its average velocity.

B. Distance Routing Effect Algorithm for Mobility Protocol (DREAM)

This protocol proposed which also used information from GPS systems for the communication by the node location. It is a part of proactive and reactive protocol where source node sends the data packet in the direction of the destination node by the selective flooding. Control packets are used to determine the node distance from source to destination which the direction is given by the distance line source to destination and the angle of alpha. It is based on the direction approach a recovery method is needed necessary when the destination node is not in the given direction. The performance of the basic techniques which can influence by the flooding, recovery procedure but it is not included in the specifications.

C. Adaptive Location-aided Mobile ad hoc Network Routing Protocol (ALARM)

This protocol uses feedback for adaption and location information for improvement the performance. It is a hybrid, adaptive to mobility protocol ‘which uses LAR and directed flooding. It introduces the number of hops to be flooded past the mobility hot spot by the flood horizon. It uses the link the duration of the feedback at each node to determine the appropriate forwarding method and it adapts the operation on the current network mobility conditions and it will increases the mobility of the packet overhead [6].

D. Greedy Perimeter Stateless Routing Protocol (GPSR)

This protocol uses the location of the node to selectively a forwarded the packets based on the distance. The node closest to the destination by forwarding is carried out on the basis by selecting the greedy approach. This process will continue until the destination is reached. This protocol uses two methods for data forwarding: greedy forwarding and perimeter forwarding. A node sends the packet to its neighbor nodes closed to its region of perimeter. In the route discovery the states are collected and cached in the nodes after the region of perimeter forwarding. For the study of mobility, we used a random waypoint model [5].

E. Grid or Geographic Location Service Protocol (GLS)

It is based a location service for the geographic locations. We can be simulated with the simple geographic routing and the GPSR. It breaks up the network area into a hierarchical forming of the system of squares a quad-tree, where each n-order squares contain four (n-1) order squares. It will make use of the location information and it can be a unique, permanent and random allocated node IPs, the local first order square that each node stores a table of all nodes. it use of the periodic broadcasts as the location which updates increase with the network size.

6.3 CLASSIFICATION OF ROUTING PROTOCOLS

Routing protocols[fig. 6.4] define a set of rules which governs the journey of message packets from source to destination in a network. In MANET, there are different types of routing protocols[fig. 6.5] each of them is applied according to the network circumstances.

Fig. 4 Classification of the routing protocols in MANET

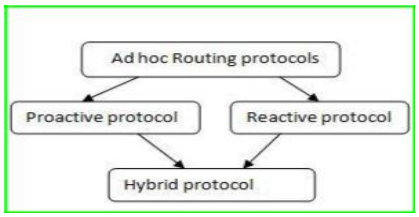
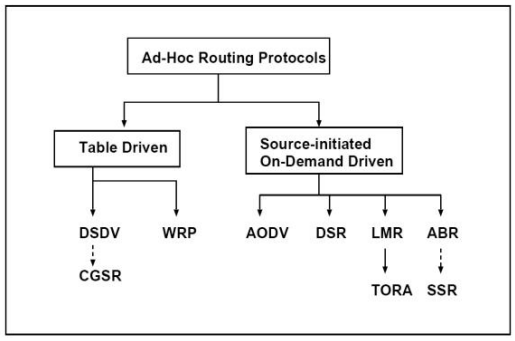


Fig. 5 the Family Tree



Proactive Routing Approaches

Proactive routing protocols are also called as table driven routing protocols. In this every node maintain routing table which contains information about the network topology even without requiring it. This feature although useful for datagram traffic, incurs substantial signalling traffic and power consumption. The routing tables are updated periodically whenever the network topology changes. Proactive protocols are not suitable for large networks as they need to maintain node entries for each and every node in the routing table of every node. These protocols maintain different number of routing tables varying from protocol to protocol. There are various well known proactive routing protocols. Example: DSDV, OLSR, WRP etc.

Reactive Routing Approaches

Reactive routing protocol is also known as on demand routing protocol. In this protocol route is discovered whenever it is needed Nodes initiate route discovery on demand basis. Source node

sees its route cache for the available route from source to destination if the route is not available then it initiates route discovery process. The on- demand routing protocols have two major components.

Route discovery: In this phase source node initiates route discovery on demand basis. Source nodes consults its route cache for the available route from source to destination otherwise if the route is not present it initiates route discovery. The source node, in the packet, includes the destination address of the node as well address of the intermediate nodes to the destination.

Route maintenance: Due to dynamic topology of the network cases of the route failure between the nodes arises due to link breakage etc, so route maintenance is done. Reactive protocols have acknowledgement mechanism due to which route maintenance is possible.

Reactive protocols add latency to the network due to the route discovery mechanism. Each intermediate node involved in the route discovery process adds latency. These protocols decrease the routing overhead but at the cost of increased latency in the network. Hence these protocols are suitable in the situations where low routing overhead is required. There are various well known reactive routing protocols present in MANET for example DSR, AODV, TORA and LMR.

Hybrid Routing **Approaches**

There is a trade-off between proactive and reactive protocols. Proactive protocols have large overhead and less latency while reactive protocols have less overhead and more latency. So a Hybrid protocol is presented to overcome the shortcomings of both proactive and reactive routing protocols. Hybrid routing protocol is combination of both proactive and reactive routing protocol. It uses the route discovery mechanism of reactive protocol and the table maintenance mechanism of proactive protocol so as to avoid latency and overhead problems in the network. Hybrid protocol is suitable for large networks where large numbers of nodes are present. In this large network is divided into set of zones where routing inside the zone is performed by using reactive approach and outside the zone routing is done using reactive approach. There are various popular hybrid routing protocols for MANET like ZRP, SHARP.

6.4 Proactive Routing Protocols

Dynamic Destination-Sequenced Distance-Vector Routing Protocol (DSDV)

DSDV is developed on the basis of Bellman–Ford routing algorithm with some modifications. In this routing protocol, each mobile node in the network keeps a routing table. Each of the routing table contains the list of all available destinations and the number of hops to each. Each table entry is tagged with a sequence number, which is originated by the destination node. Periodic transmissions of updates of the routing tables help maintaining the topology information of the network. If there is any new significant change for the routing information, the updates are transmitted immediately. So, the routing information updates might either be periodic or event driven. DSDV protocol requires each mobile node in the network to advertise its own routing table to its current neighbours. The advertisement is done either by broadcasting or by multicasting. By the advertisements, the neighbouring nodes can know about any change that has occurred in the network due to the movements of nodes. The routing updates could be sent in two ways: one is called a „„full dump‟‟‟ and another is „„incremental.‟‟‟ In case of full dump, the entire routing table is sent to the neighbours, where as in case of incremental upd ate, only the entries that require changes are sent.

Dynamic Source Routing (DSR)

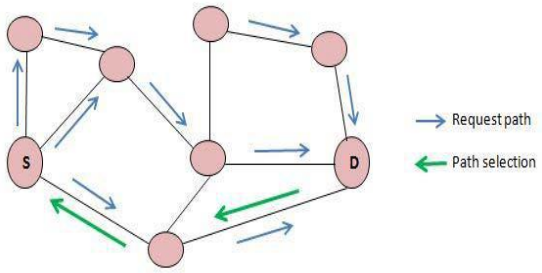


Fig. 6 DSR Routing Protocol

Dynamic Source Routing (DSR) is a reactive protocol based on the source route approach. In *Dynamic Source Routing (DSR)*, shown in Figure 6.6, the protocol is based on the link state algorithm in which source initiates route discovery on demand basis. The sender determines the route from source to destination and it includes the address of intermediate nodes to the route record in the packet. DSR was designed for multi hop networks for small Diameters. It is a beaconless protocol in which no HELLO messages are exchanged between nodes to notify them of their neighbours in the network.

Ad Hoc On-Demand Distance Vector Routing (AODV)

AODV is basically an improvement of DSDV. But, AODV is a reactive routing protocol instead of proactive. It minimizes the number of broadcasts by creating routes based on demand, which is not the case for DSDV. When any source node wants to send a packet to a destination, it broadcasts a route request (RREQ) packet. The neighboring nodes in turn broadcast the packet to their neighbors and the process continues until the packet reaches the destination. During the process of forwarding the route request, intermediate nodes record the address of the neighbor from which the first copy of the broadcast packet is received. This record is stored in their route tables, which helps for establishing a reverse path. If additional copies of the same RREQ are later received, these packets are discarded. The reply is sent using the reverse path. For route maintenance, when a source node moves, it can reinitiate a route discovery process. If any intermediate node moves within a particular route, the neighbor of the drifted node can detect the link failure and sends a link failure notification to its upstream neighbor. This process continues until the failure notification reaches the source node. Based on the received information, the source might decide to re-initiate the route discovery phase.

Zone Routing Protocol (ZRP)

ZRP is suitable for wide variety of MANETs, especially for the networks with large span and diverse mobility patterns. In this protocol, each node proactively maintains routes within a local region, which is termed as routing zone. Route creation is done using a query-reply mechanism. For creating different zones in the network, a node first has to know who its neighbours are. A neighbour is defined as a node with which direct communication can be established, and that is, within one hop transmission range of node.Neighbor discovery information is used as a basis for Intra-zone Routing Protocol (IARP). Rather than blind broadcasting, ZRP uses a query control mechanism to reduce route query traffic by directing query messages outward from the query source and away from covered routing zones. A covered node is a node which belongs to the routing zone of a node that has received a route query. During the forwarding of the query packet, a node identifies whether it is coming from its neighbour or not. If yes, then it marks all

of its known neighbouring nodes in its same zone as covered. The query is thus relayed till it reaches the destination. The destination in turn sends back a reply message via the reverse path and creates the route.

Module VII: Energy-efficient Communication [3L]

Energy efficiency at various layers - Physical layer, MAC layer, Network layer, Application layer, performance analysis in noisy channel environment.

Wireless data link layer network design issues

The context in this section is data link-level communication protocols for wireless networks that provide multimedia services to mobile users. As mentioned before, portable devices have severe constraints on the size, the energy consumption, and the communication bandwidth available, and are required to handle many classes of data transfer over a limited bandwidth wireless connection, including delay sensitive, real- time traffic such as speech and video. This combination of limited bandwidth, high error rates, and delay-sensitive data requires tight integration of all subsystems in the device, including aggressive optimisation of the protocols to suit the intended application. The protocols must be robust in the presence of errors; they must be able to differentiate between classes of data, giving each class the exact service it requires; and they must have an implementation suitable for low-power portable electronic devices.

The ISO/OSI network design model

Data communication protocols govern the way in which electronic systems exchange information by specifying a set of rules that, when followed, provide a consistent, repeatable, and well-understood data transfer service. In designing communication protocols and the systems that implement them, one would like to ensure that the protocol is correct and efficient. The ISO/OSI model is a design guide for how network software in general should be built. In this model, protocols are conceptually organised as a series of layers, each one built upon its predecessor. Most network architectures use some kind of layering model, although the specific layers may not be an exact match with the layers defined in the ISO/OSI model.

The rationale behind this layering approach is that it makes in principle possible to replace the implementation of a particular layer with another implementation, requiring only that each implementation provide a consistent interface that offers the same services and service access points to the upper layer. Thus, the goal of service abstraction is modularity and freedom to choose the implementation that is best suited for a particular environment. However, while this model provides an excellent starting point for conceptually partitioning a set of protocol services, it has two implicit assumptions that fail to hold in many practical contexts [78]. First, there is the assumption that cost of abstraction and separation is negligible compared to the gained modularity and flexibility. Second, there is the assumption that interchanging layers that provide the same logical services – for example, a wired physical layer and a wireless physical layer – provide equivalent service.

These assumptions are in general not valid for mobile systems and can impose severe limitations. For example, although the TCP specification contains no explicit reference to the characteristics of the lower layers, implicitly in the timeout and retransmission mechanisms there are the assumption that the error rate is low, and that lost packets occur due to network congestion. TCP has no way of distinguishing between a packet corrupted by bit errors in the wireless channel from packets that are lost due to congestion in the network. The applied measures result on a wireless channel in unnecessary increases in energy consumption and deterioration of QoS. This example attests the need to tailor protocols to the environment they operate in. Separating the design of the protocol from the context in which it exists leads to penalties in performance and energy consumption that are unacceptable for wireless,

multimedia applications.

The context of this section is mainly the data link layer. Data link protocols are usually divided into two main functional components: the *Logical Link Control* (LLC) and the *Medium Access Control* (MAC), that are responsible for providing a point-to-point packet transfer service to the network, and a means by which multiple users can share the same medium. The main task of the Data Link layer protocols on a wireless network is to provide access to the radio channel. Wireless link particularities, such as high error rate and scarce resources like bandwidth and energy, and the requirements to provide access for different connection classes with a variety of traffic characteristics and QoS requirements, makes this a non-trivial task. It requires a flexible, yet simple scheme that should be able to adjust itself to different operating conditions in order to satisfy all connections and overall requirements like efficient use of resources like energy and radio bandwidth. The protocols have to support traffic allocation according an agreed traffic contract of a connection, but must also be flexible enough to adapt to the dynamic environment and provide support for QoS renegotiations. It further has to provide error control and mobility related services.

Wireless link restrictions

The characteristics of the wireless channel the Data Link protocol has to deal with are basically high bit error rate (BER), limited bandwidth, broadcast transmission, high energy consumption and half duplex links.

Wireless networks have a much higher *error rate* than the normal wired networks. The errors that occur on the physical channel are caused by phenomena such as signal fading, transmission interference, user mobility and multi-path effects. Typically, the bit error rates observed may be as bad as 10^{-3} or 10^{-4}, which is far more worse than assumed by networks with wired connections. Additionally, the errors show a dynamic nature due to movement of the mobile. In indoor environments propagation mechanisms caused by the interactions between electromagnetic fields and various objects can increase error rates considerably. Especially in the outer regions of the radio cell, the low signal-to noise ratio (SNR) makes wireless link errors a norm rather than an exception in the system.

The *available bandwidth* on a wireless channel is usually much less than offered by wired networks. Consequently, an important design consideration in the design of a protocol, is the efficient use of the available bandwidth.

Closely related to this is the amount of *energy* that is needed to transmit or receive data. The required amount of energy is high, and typically depends on the distance that the radio signal has to propagate between sender and receiver. Since wireless networks for mobile systems will be used more widely and more intensively, the energy consumption that is required to communicate will take a large part of the available energy resources (batteries) of the mobile. So energy consumption will be another main design constraint for the wireless data link protocol of the mobile. In general, saving energy for the base station is not really an issue, as it is part of the fixed infrastructure and typically obtains energy from a mains outlet. However, since the current trend is to have ever smaller area cell sizes, and the complexity of the base station is increasing, this issue might become more important in the future mainly because of economical and thermal reasons.

By their nature wireless radio transmission is a *broadcast medium* to all receivers within the range of a transmitter. This characteristic gives rise to several problems in a wireless environment with multiple cells and mobiles. A mobile that is in reach of more than one base station and communicates with only one of them can cause errors on the communication in the neighbouring cell. Even if the mobiles are just in reach of one base station, interference between mobiles in different cells can also cause errors. Solutions on the physical layer are possible (colouring schemes with multiple frequencies, spread spectrum technologies, near field radio [72], etc.) but are out of the scope of our research. However, provisions for handoff when a mobile moves from one area cell to another, are important and have consequences for the design of a data link protocol.

A radio modem transceiver typically has one part dedicated to transmission, and the other part to reception. Consequently, the radio channel is generally used in *half duplex mode*. The only way to allow full duplex operation over the radio channel is to duplicate transceiver hardware and use two sub-bands in the frequency band, each of them being used for one-way transmission. Because such a solution is not economically viable, and also raises some technical problems, the data link protocol should be designed in such a way that connections in both directions are treated fairly.

Basic wireless networking functions

The challenge of a wireless data link protocol is to overcome the harsh reality of wireless transmission and to provide mobility and multimedia services. The data link layer of a wireless network has to provide assistance to several basic functions: *QoS management* when a connection is initiated or when the operating conditions have changed; *traffic and resource allocation* according to a traffic contract; *error control* to overcome the effect of errors on the wireless link, *flow control* to avoid buffer overflow and also to discard cells of which the maximum allowed delay is exceeded due to retransmissions; *security and privacy* for the mobile user, and *mobility features* to allow handover when a mobile moves to another area cell. In this section we will discuss these items briefly and describe the consequences for the data link layer.

QoS management

To support diverse traffic over a wireless channel, the notion of QoS of a connection is useful. Setting up a connection involves negotiation along a path from sender to receiver in order to reserve the required resources to fulfil the QoS needed. Due to the dynamic nature of wireless channels and the movement of the mobile the agreed QoS level in one or more contracts generally cannot be sustained for a longer period. These situations are not errors, but are modus operandi for mobile computers. Therefore, these situations must be handled efficiently, and QoS renegotiations will occur frequently. Multimedia applications can show a more dynamic range of acceptable performance parameters depending on the user's quality expectations, application usage modes, and application's tolerance to degradation.

Traffic and resource allocation

Each accepted connection has a certain traffic contract that describes the traffic type and required QoS parameters. A slot-scheduler is responsible to assign slots in a transmission frame according to the various traffic contracts. At the same time it must attain a high utilisation of the scarce radio bandwidth and minimise the energy consumption for the mobile.

Error control

Due to the high bit error rate (BER) that is typical for a wireless link, many packets can be corrupted during transmission. If this rate exceeds the allowable cell loss rate of a connection, an effective and efficient error control scheme must be implemented to handle such situations. At the radio physical level redundancy for detecting symbols reduces the bit error rate for the first time. However, it is usually inefficient to provide a very high degree of error correction, and some residual errors pass through. The residual channel characteristic is based on *erases*, i.e. missing packets in a stream. Erasures are easier to deal with than errors, since the exact location of the missing data is known. Then, integrated into the MAC layer (and possibly also into the higher layers), an error control scheme further enhances transmission quality by applying error correction and/or retransmission schemes.

Since different connections do not have the same requirements concerning cell loss rate and cell transfer delay, different error control schemes must be applied for different connection types [60]. The alternatives are Forward Error Correction (FEC), retransmission techniques like automatic repeat request (ARQ), or hybrid FEC/ARQ schemes. To reduce the overhead and energy involved the error control scheme can also be adapted to the current error condition of

the wireless connection. The error control mechanisms should trade off complexity, buffering requirements and energy requirements (taking into account the required energy for both computation and communication) for throughput and delay.

Flow control

A connection involves buffering at several places on the path between sender and receiver. Traffic type requirements concerning delay, and implementation restrictions on the buffer capacity generally limit the amount of buffer space available to a connection. Due to the dynamic character of wireless networks and user mobility, the stream of data might be hindered on the way from source to destination. Therefore, flow control mechanisms are needed to prevent buffer overflow, but also to discard packets that have exceeded the allowable transfer time. Depending on the service class and QoS of a connection a different flow control can be applied. For instance, in a video application it is useless to transmit images that are already outdated. It is more important to have the 'fresh' images. For such traffic the buffer is probably small, and when the connection is hindered somewhere, the oldest data will be discarded and the fresh data will be shifted into the fifo. Flow control can cover several hierarchical layers, but in the context of link access protocols we mainly deal with the buffering required directly at both sides of the wireless link.

Security and privacy

Since eavesdropping of the data bits is a real threat because they will be transmitted over the wireless air interface, security and privacy are important issues in wireless systems. These items are important on two levels: protection of the data on the wireless link, and end-to-end application security. The MAC layer is only capable to provide some basic protection of the data on the wireless link. Since it is hard to make this very secure, end- to-end security will be the most attractive and secure solution.

Mobility features

In a wireless environment the mobility of the mobile will enforce handover procedures when the mobile moves from one area cell to another. As the current trend is that the radius of an area cell decreases (because of the higher bandwidth density and lower energy requirements) handover situations will be encountered frequently.

The task of the link layer is to provide the higher layers of the mobile with information about which area cells are in range, and provide services to actually handle the handover. The radio link quality will be the first parameter to be taken into account for the handover initiation procedure. In the new area-cell a new connection has to be prepared and bandwidth reserved. When a mobile is being handovered to a new area cell, the connection will be dropped if there is insufficient bandwidth to support the connection. Since dropping connections is more undesirable than blocking new connection requests, some bandwidth can be reserved in neighbouring area cells in advance, before the mobile reaches that area cell. It is possible to provide a general pool of bandwidth that can be used for new connections. If it is possible to predict the movement of mobiles, then bandwidth can be saved since not in all neighbouring area cells bandwidth has to be reserved [75].

QoS renegotiation

In a wired network, QoS is usually guaranteed for the lifetime of a connection. In a wireless environment these guarantees are not realistic due to the movement of mobiles and the frequent occurrence of errors on the wireless link.

To prevent service interruptions in a proactive fashion, QoS renegotiations may be required to assure a lower, but deliverable level of service. The difficulty is to provide a mechanism with which QoS parameters of an active connection can be changed

dynamically.

QoS control is important during the handover procedure as the mobile moves into a cell and places demand on resources presently allocated to connections from other mobiles. If a mobile faces a significant drop in bandwidth availability as it moves from one cell to another, rather than dropping the connection, the QoS manager might be able to reallocate bandwidth among selected active connections in the new cell. The QoS manager of the new cell selects a set of connections, called *donors*, and changes their bandwidth reservations to attempt satisfactory service for all. To quickly process handover requests, the QoS manager can use cached bandwidth reserves. This cache can then be replenished after the QoS manager has obtained the required bandwidth from the donor connections.

When the mobile moves to a cell where the traffic on the wireless link is much higher, it is not just the current connection that needs renegotiations, even other connections of applications in that area cell may become subject to the QoS renegotiations to allow the new mobile access. The movement of the mobile also influences the (already poor) quality of wireless channels and can introduce dynamic changes in error rate. Especially an indoor environment with small rooms and corridors can cause interactions between the electromagnetic fields and various objects. These interactions can increase error rates considerably. To be able to guarantee an agreed QoS for – especially error sensitive – connections, error recovery techniques using error correcting codes or retransmission is required. In addition to this, before completely closing a connection on a faulty link, the link errors can be gracefully tolerated by renegotiations of QoS. Many multimedia applications can deal with varying bandwidth availability once provided with sufficient information about the operating conditions. For instance, video transmission schemes may adjust their resolution, their frame rates, and encoding mechanism to match the available bandwidth or deal with the current error conditions.

Energy-efficient wireless MAC design

The objective of an energy efficient MAC protocol design is to maximise the performance while minimising the energy consumption of the *mobile*. These requirements often conflict, and a trade-off has to be made.

Sources of unessential energy consumption

The focus of this work is on minimising the energy consumption of a mobile and in particular the wireless interface, the transceiver. Typically, the transceiver can be in five modes; in order of increasing energy consumption, these are off, sleep, idle, receive, and transmit. In transmit mode, the device is transmitting data; in receive mode, the receiver is receiving data; in idle mode, it is doing neither, but the transceiver is still powered and ready to receive or transmit; in sleep mode, the transceiver circuitry is powered down, except sometimes for a small amount of circuitry listening for incoming transmissions [50].

Several causes for unessential energy consumption exist. We will review in this section some of the most relevant sources of unessential energy consumption.

- First of all, most applications have low traffic needs, and hence the transceiver is *idling* most of the time. Measurements show that on typical applications like a web- browser or e-mail, the energy consumed while the interface is on and idle is more than the cost of actually receiving packets [54][74].
- Second, the typical *inactivity threshold*, which is the time before a transceiver will go in the off or standby state after a period of inactivity, causes the receiver to be in a too high energy consuming mode needlessly for a significant time.
- Third, in a typical wireless broadcast environment, the receiver has to be powered on at all times to be able to *receive messages* from the base station, resulting in significant

energy consumption. The receiver subsystem typically receives all packets and forwards only the packets destined for this mobile. Even in a scheme in which the base transmits a traffic schedule to a mobile, the mobile has to receive the traffic control information regularly to check for waiting downlink traffic. When the mobile is not synchronised with the base-station, then it might have to receive 'useless' data before it receives the traffic control.

- Fourth, significant time and energy is further spent by the mobile in switching from transmit to receive modes, and vice-versa. The *turnaround time* between these modes typically takes between 6 to 30 microseconds. The transition from sleep to transmit or receive generally takes even more time (e.g. 250 □s for WaveLAN). A protocol that assigns the channel per slot will cause significant overhead due to turnaround.

- Fifth, in broadcast networks *collisions* may occur (happens mainly at high load situations). This causes the data to become useless and the energy needed to transport that data to be lost.

- Sixth, the *overhead of a protocol* also influences the energy requirements due to the amount of 'useless' control data and the required computation for protocol handling. The overhead can be caused by long headers (e.g. for addressing, mobility control, etc), by long trailers (e.g. for error detection and correction), and by the number of required control messages (e.g. acknowledgements). In many protocols the overhead involved to receive or transmit an amount of data can be large, and may depend on the load of the network. In general, simple protocols need relatively less energy than complex protocols.

- Finally, the high *error rate* that is typical for wireless links is another source of energy consumption. First, when the data is not correctly received the energy that was needed to transport and process that data is wasted. Secondly, energy is used for error control mechanisms. On the data link layer level error correction is generally used to reduce the impact of errors on the wireless link. The residual errors occur as burst errors covering a period of up to a few hundred milliseconds. To overcome these errors retransmission techniques or error correction techniques are used. Furthermore, energy is consumed for the calculation and transfer of redundant data packets and an error detection code (e.g. a CRC). Finally, because in wireless communication the error rate and the channel's signal-to-noise ratio (SNR) vary widely over time and space, a fixed-point error control mechanism that is designed to be able to correct errors that rarely occur, wastes energy and bandwidth. If the application is error-resilient, trying to withstand all possible errors wastes even more energy in needless error control.

We define *energy efficiency* as the quotient between the intrinsic amount of energy needed to transfer a certain quantity of data and the actually used amount of energy (including all overheads). We will use this metric to quantify how well a MAC protocol behaves with respect to its energy consumption.

Main principles of energy-efficient MAC design

The above observations are just some of the possible sources of unessential energy consumption related to the medium access control protocol. We have no intention to provide a complete list. We can, however, deduce the following main principles that can be used to design a MAC protocol that is energy efficient for the mobile.

- *Avoid unsuccessful actions of the transceiver.* *(P1)*

Two main topics cause unsuccessful actions: collisions and errors.

Every time a *collision* occurs energy is wasted because the same transfer has to be repeated again after a backoff period. A protocol that does not suffer from collisions can have good throughput even under high load conditions. These protocols generally also have good energy consumption characteristics. However, if it requires the receiver to be turned on for long periods of time, the advantage diminishes.

A protocol, in which a base-station broadcasts traffic control for all mobiles in range with information about when a mobile is allowed to transmit or is supposed to receive data, reduces the occurrence of collisions significantly. Collisions can only occur when new requests

have to be made. New requests can be made per packet in a communication stream, per application of a mobile, or even per mobile. The trade- off between efficient use of resources and QoS determines the size to which a request applies. Note that this might waste bandwidth (but not energy) when slots are reserved for a request, but not used always. In such a reservation mechanism, energy consumption is further reduced because there is less need for a handshake to acknowledge the transfer.

Errors on the wireless link can be overcome by mechanisms like retransmissions or error correcting codes. Both mechanisms induce extra energy consumption. The error control mechanisms can be adapted to the current error condition in such a way that it minimises the energy consumption needed and still provides (just) enough fault tolerance for a certain connection. Due to the dynamic nature of wireless networks, *adaptive error control* can give significant gains in bandwidth and energy efficiency [23][82]. This avoids applying error control overhead to connections that do not need it, and allows the possibility to apply it selectively to match the required QoS and the conditions of the radio link. Note that this introduces a trade-off between communication and computation [36]. Section 5.5 goes into more detail on this issue. A different strategy to reduce the effect of errors is to avoid traffic during periods of bad error conditions. This, however, is not always possible for all traffic types as it influences the QoS.

Minimise the number of transitions. *(P2)*

Scheduling traffic into bursts in which a mobile can continuously transmit or receive data – possibly even bundled for different applications –, can reduce the number of transitions. Notice, however, that there is a trade-off with QoS parameters like delay and jitter. When the traffic is continuous and can be scheduled for a longer period ahead, then the mobile does not even have to listen to the traffic control since it knows when it can expect data or may transmit. The number of transitions needed can also be reduced by collecting multiple requests of multiple applications on a mobile, and by piggy-backing new requests on current data streams. Simple protocols can further reduce the required number of transitions due to the low amount of control messages needed.

Synchronise the mobile and the base station. *(P3)*

Synchronisation is beneficial for both uplink (mobile host to base station) and downlink (base station to mobile host) traffic. When the base-station and mobile are synchronised in time, the mobile can go in standby or off mode, and wake up just in time to communicate with the base-station. The energy consumption needed for downlink traffic can be reduced when the time that the receiver has to be on – just to listen whether the base-station has some data for the mobile – can be minimised. The premise is that the base has plenty of energy and can broadcast its beacon frequently. The application of a mobile with the least tolerable delay determines the frequency by which a mobile needs to turn its receiver on. If the wake-up call of the communication is implemented with a low-power low- performance radio, instead of the high-performance high-energy consuming radio, then the required energy can be reduced even more.

- *Migrate as much as possible work to the base-station*. *(P4)*

In a centralised wireless system architecture, the base-station that is connected to the fixed network and a mains outlet, can perform many tasks in lieu of the mobile. The calculation of a traffic control that adheres the QoS of all connections is an example of such task. At higher levels, the base-station can also perform tasks to process control information, or to manipulate user information that is being exchanged between the mobile device and a network-based server (see Chapter 2).

Note that these principles can reduce the energy consumption of the wireless interface. The energy consumption of the mobile system is much more complex and comprises many issues. The total achieved energy reduction is thus based on many trade-offs. For example, grouping traffic in multimedia video streams to minimise the number of transitions requires the data to be buffered in the client's memory. The required amount of energy needed for buffering reduces the effect of the energy savings principle in some sense.

There are many ways in which these principles can be implemented. We will consider an environment suitable for multimedia applications in which the MAC protocol also has other requirements like provisions for QoS of real-time traffic, and to provide a high throughput for bulk data. Due to the dynamic character of wireless multimedia systems and time-varying radio channel conditions, flexibility and adaptation play a crucial role in achieving an energy efficient design.

We have chosen to adopt *Asynchronous Transfer Mode* (ATM) mechanisms for the wireless network. We have no intention to build the full-blown B-ISDN ATM protocol stack, but merely adopt the small, fixed size packet and the QoS mechanisms. In the next section we give a short introduction to ATM and motivate why it is suitable for building an energy-efficient wireless network.

ATM

The challenge of designing a network that can cope with all different service types led to the development of the *Asynchronous Transfer Mode* (ATM). ATM is able to support different kind of connections with different QoS parameters. ATM technology provides deterministic or statistical guarantees with connection-oriented reservations. The original intent of ATM was to form a backbone network for high speed data transmission regardless of traffic type. Later, ATM has been found to be capable of more. Today, ATM scales well from backbone to the customer premises networks and is independent of the bit rate of the physical medium. By preserving the essential characteristics of ATM transmission, wireless ATM offers the promise of improved performance and QoS, not attainable by other wireless communication systems like cellular systems, cordless, or wireless LANs. In addition, wireless ATM access provides location independence that removes a major limiting factor in the use of computers and powerful telecom equipment over wired networks [58].

ATM transports data in small, fixed size (in B-ISDN ATM 53-byte) packets called *cells*. Having a fixed cell size allows for a simple implementation of ATM devices, and results in a more deterministic behaviour. Small cells have the benefit of a small scheduling granularity, and hence provide a good control over queuing delays. This also allows rapid switching that supports any mix of delay-sensitive traffic and bursty data traffic at varying bit rates. ATM carries cells across the network on connections known as *Virtual Circuits*. With a Virtual Circuit the flow of data is controlled at each stage in its path from source to destination. In ATM, the QoS requirements of Virtual Circuits are a key element as it relates to how cells for a Virtual Circuit are processed. The connection- oriented nature allows the user to specify certain QoS parameters for each connection. Network resources are reserved upon the acceptance of a Virtual Circuit, but they are consumed only when traffic is actually generated.

ATM service classes

The ATM service architecture uses procedures and parameters for traffic control and congestion control whose primary role is to protect the network and end-system to achieve network performance objectives. The design of these functions is also aimed at reducing network and end-system complexity while maximising network utilisation. The *ATM service categories* represent service building blocks and introduce the possibility for the user to select specific combinations of traffic and performance parameters. Most of the requirements that are specific to a given application may be resolved by choosing an appropriate ATM Adaptation Layer (AAL). However, given the presence of a heterogeneous traffic mix, and the need to adequately control the allocation of network resources for each traffic component, a much greater degree of flexibility, fairness and utilisation of the network can be achieved by providing a selectable set of capabilities within the ATM-layer itself.

The ATM forum has specified the following ATM Service Categories (ASC). ATM Service Category relates quality requirements for a given set of applications and traffic characteristics to network behaviour.

- *Constant Bit Rate* (CBR). A category based on constant (maximum) bandwidth

allocation. This category is used for connections that require constant amount of bandwidth continuously available during the connection lifetime. CBR is oriented to serve applications with stringent time delay and jitter requirements (like telephony), but is also suitable for any data transfer application which contains smooth enough traffic.

- *Variable Bit Rate* (VBR) for statistical (average) bandwidth allocation. This is further divided into real-time (rt-VBR) and non-real-time (nrt-VBR), depending on the QoS requirements. Rt-VBR is intended to model real-time applications with sources that transmit at a rate which varies in time (e.g. compressed images) and have strict delay constraints. Video-conference is a suitable application, in which the real-time constraint should guarantee a synchronisation of voice and image, and the network resources are efficiently utilised because of the varying bandwidth requirements due to compression. Nrt-VBR is for connections that carry variable bit rate traffic with no strict delay constraints, but with a required mean transfer delay and cell loss. Nrt-VBR can be used for data-transfer like response-time critical transaction processing (e.g. airline reservation, banking). The undetermined time constraints give the possibility to use large buffers.
- *Available Bit Rate* (ABR) where the amount of reserved resources varies in time, depending on network availability. The variation managed by the traffic control mechanisms is reported to the source via feedback traffic. Compliance to the variations from the feedback signal should guarantee a low cell loss ratio for the application. Generally, it is necessary to use large buffers to offer ABR service on the network due to the burst nature of the service. It has no guaranteed cell transfer delay, but just a minimum guaranteed bandwidth. This category provides an economical support to those applications that show vague requirements for throughput and delay and requires a low cell loss ratio. Applications are typically run over protocol stacks like TCP/IP, which can easily vary their emission as required by the ABR rate control policy.
- *Unspecified Bit Rate* (UBR) has no explicit resource allocation and does not specify bandwidth or QoS requirements. Losses and error recovery or congestion control mechanisms could be performed at higher layers, and not at lower network layers. UBR can provide a suitable solution for less demanding applications like data applications (e.g. background ftp) that are very tolerant to delay and cell loss. These services can take advantage of any spare bandwidth and will profit from the resultant reduced tariffs.

Admission control and policing

Setting up a virtual connection involves taking information on the required service class and QoS. Using this information the system negotiates along the path from source to destination in order to reserve the necessary resources. A traffic contract specifies the negotiated characteristics of a virtual connection at an ATM User Network Interface (UNI). Each QoS parameter consists of a value pair, one representing the low end, and the other the high end. This is called the tolerable range.

Once admitted, the system continually checks that the virtual connection sends data according to its allowance, known as *policing*. When the value of the delivered QoS parameter falls outside the tolerable range, the contract is be violated.

Functions related to the implementation of QoS in ATM networks are usage parameter control (UPC) and connection admission control (CAC). In essence, the UPC function (implemented at the network edge) ensures that the traffic generated over a connection conforms to the declared traffic parameters. Excess traffic may be dropped or carried on a best-effort basis. The CAC function is implemented by each switch in an ATM network to determine whether the QoS requirements of a connection can be satisfied with the available resources.

Wireless ATM

At the moment there are already wireless LANs and wireless systems offering data services and mobile data. These mobile systems offer low bit rate wireless data transmission with mobility and roaming possibility. Wireless LANs offer mobility only in restricted, smaller

areas of coverage without wide area roaming capabilities. The achieved bit rates are generally greater than with current mobile systems. The third generation mobile telecommunication systems, such as UMTS (Universal Mobile Telecommunication System) aim to achieve data services of up to 2 Mbit/s, which is a significant improvement over the second-generation mobile systems. However, the importance of speech service may overrun the 2 Mbit/s data service goals [58]. The third generation wireless networks will enable mobiles to carry integrated multimedia. Wireless ATM networks can be useful for these new generation wireless networks because of its ability to handle traffic of different classes and integrate them into one stream. A wireless ATM network consists generally of a cluster of base stations interconnected by a wired ATM network (see Figure 1).

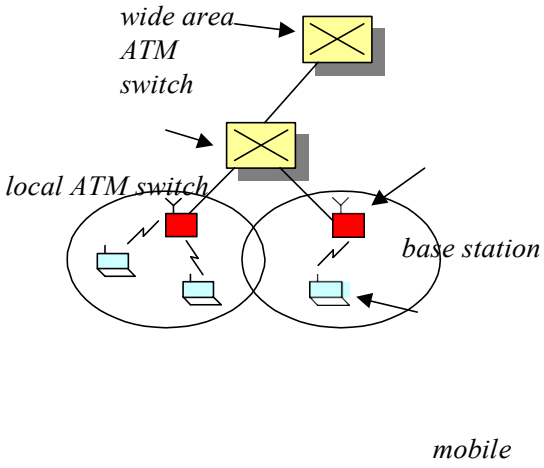


Figure 1: Wireless ATM architecture.

Originally, ATM was characterised by bandwidth on demand at megabits per second rates; it operates at very low bit error rate environments, supports packet switched transport, virtual circuit connections, and statistical sharing of the network resources among different connections.

Wireless networking is inherently unreliable and the bandwidth supported is usually lower than that of fixed networks. Various forms of interference on the wireless link result in high error rates, and thus introduces delay, jitter and an even lower effective bandwidth. Mobility of the user makes these problems even more dynamic and introduces the need for handover mechanisms when the user comes in reach of a different base-station.

The characterisation of ATM – that was designed for wired networks – seems rather contradictory with the operating conditions of wireless networks. Even with high redundancy introduced at several layers (i.e. physical, medium access control, transport and applications) the quality of service may not be guaranteed. Therefore, when adopting ATM in a wireless environment we need to adopt a more dynamic approach to resource usage. *Applications* must adapt their QoS requirements on the current operating environment. Explicit renegotiation of the QoS of a connection about the available resources between the application and the wireless system is essential in wireless ATM systems.

Since a connection typically involves both a fixed and a wireless part, the wireless link should support similar mechanisms as the fixed ATM network. Therefore it has to support all traffic types taking into account the characteristics of the wireless medium. The medium access protocol should be able to bridge the fixed and wireless world and provide ATM services transparently over a wireless link.

Wireless ATM is a topic on which many research activities are going on, e.g. Magic WAND [57], MEDIAN [18], NTT AWA [43]. Most projects aim to extend ATM to the mobile terminal. The main difference can be observed in air interface. No project explicitly

addresses reduction in power consumption as a major issue.

Energy-efficient error control

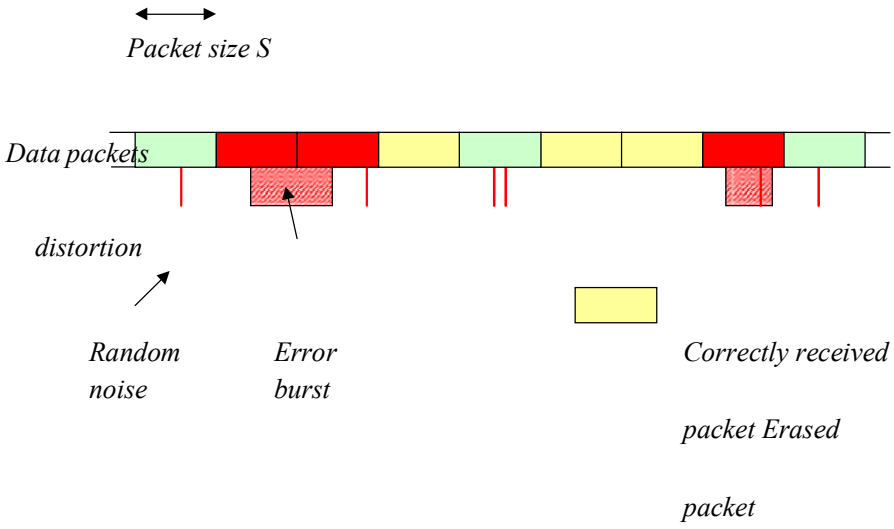
Since high error rates are inevitable to the wireless environment, *energy-efficient error- control* is an important issue for mobile computing systems. This includes energy spent in the physical radio transmission process, as well as energy spent in computation, such as signal processing and error control at the transmitter and the receiver.

Error-control mechanisms traditionally trades off complexity and buffering requirements for throughput and delay [46][48][15]. In our approach we apply energy consumption constraints to the error-control mechanisms in order to *enhance energy efficiency under various wireless channel conditions*. In a wireless environment these conditions not only vary dynamically because the physical conditions of a communication system can vary rapidly, but they can also vary because the user moves from an indoor office environment to a crowded city town. Not only the characteristics could have changed, it is even possible that a complete different infrastructure will be used [71]. The communication interface of the mobile must not only be able to adapt to these situations and provide the basic functionally, it must also do it energy efficient in all these situations. At the same time, the Quality of Service guarantees of the various connections should still be supported. In some cases it may be impossible to maintain the QoS guarantees originally promised to the application as the channel degrades, for example when the user moves into a radio shadow where the radio loses physical layer connectivity.

The error model

In any communication system, there have always been errors and the need to deal with them. Wireless networks have a much higher error rate than the normal wired networks. The errors that occur on the physical channel are caused by phenomena such as signal fading, transmission interference, and user mobility.

In characterising the wireless channel, there are two variables of importance. First, there is the Bit Error Rate (BER) – a function of Signal to Noise Ratio (SNR) at the receiver -, and second the burstiness of the errors on the channel. Figure 2 presents a graphical view of packets moving through this channel.



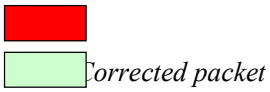


Figure 2: Error characteristics and packet erasures.

This leads to two basic classes of errors: packet erasures and bit corruption errors [21][83]. Error control is applied to handle these errors.

Note that when the bit errors are independent, the packet error rate (*PER*) is related to the size of the packet (*s*) and the bit error rate (*BER*) as

$$PER = 1 - (1 - BER)^s \qquad (1)$$

While this does not take into account the bursty nature of a wireless link, it gives an idea of the influence of the packet length on the error rate of a packet. Even one uncorrected bit error inside a packet will result in the loss of that packet. Each lost packet directly results in wasted energy consumption, wasted bandwidth, and in time spent. This loss might also result in the additional signalling overhead of an ARQ protocol [45]. Because of this, it is important to simultaneously adapt the error control mechanism when the packet size is maximised to minimise the number of transitions. In Section 5.7.2 we will analyse the effects of packet length and energy efficiency in more detail.

Error-control alternatives

There are a large variety of error-control strategies, each with its own advantages and disadvantages in terms of latency, throughput, and energy efficiency. Basically there are two methods of dealing with errors: retransmission (Automatic Repeat reQuest (ARQ) and Forward Error Correction (FEC). Hybrids of these two also exist. Within each category, there are numerous options. Computer communication generally implements a reliable data transfer using either methods or a combination of them at different levels in the communication protocol stack. Turning a poor reliability channel into one with moderate reliability is best done within the physical layer utilising signal space or binary coding techniques with soft decoding. FEC is mainly used at the data link layer to reduce the impact of errors in the wireless connection. In most cases, these codes provide less than perfect protection and some amount of residual errors pass through. The upper level protocol layers employ various block error detection and retransmission schemes (see e.g. [67][39]).

- With *FEC* redundancy bits are attached to a packet that allow the receiver to correct errors which may occur. In principle, FEC incurs a fixed overhead for every packet, irrespective of the channel conditions. This implies a reduction of the achievable data rate and causes additional delay. When the channel is good, we still pay this overhead. Areas of applications that can benefit in particular from error-correction mechanisms are *multicast applications* [74][65][61]. Even if the QoS requirement is not that demanding, insuring the QoS for all receiving applications is difficult with retransmission techniques since multiple receivers can experience losses on different packets. Individual repairs are not only extremely expensive, they also do not scale well to the number of receivers. Reducing the amount of feedback by the use of forward error correction, leads to a simple, scalable and energy-efficient protocol.

Several studies have shown that adaptive packet sizing and FEC can significantly increase the throughput of a wireless LAN, using relative simple adaptation policies (e.g. [21][24][60]). Note that, due to the burst errors, FEC block codes might require interleaving to spread the errors over the whole packet. However, burst error events on the indoor wireless channel caused by slow-moving interference may last for hundreds of milliseconds, rendering interleaving infeasible for time-critical (delay and jitter) applications [29].

- Using *ARQ*, feedback is propagated in the reverse direction to inform the sender of the status of packets sent. The use of ARQ results in an even more significant increase of delay and delay variations than FEC [66]. The retransmission requires additional

buffering at the transmitter and receiver. A large penalty is paid in waiting for and carrying out the retransmission of the packet. This can be unacceptable for systems where Quality of Service (QoS) provisioning is a major concern, e.g. in wireless ATM systems. These communications will include video, audio, images, and bulk data transfer, each with their own specific parameter settings regarding for example jitter, delay, reliability, and throughput [19]. Solutions to provide a predictable delay at the medium access control layer by reserving bandwidth for retransmission are possible [27], but spoil bandwidth.

ARQ schemes will perform well when the channel is good, since retransmissions will be rare, but perform poorly when channel conditions degrade since much effort is spent in retransmitting packets. Another often ignored side effect in ARQ schemes is that the round-trip-delay of a request-acknowledge can also cause the receiver to be waiting for the acknowledge with the receiver turned 'on', and thus wasting energy.

- *Hybrids* do not have to transmit with maximum FEC redundancy to deal with the worst possible channel. Under nominal channel conditions, the FEC will be sufficient, while under poor channel conditions ARQ will be used. Although more efficient than the pure categories, a hybrid system is still a rigid one since certain channel conditions are assumed.

- *Adaptive error control* allows the error-control strategy to vary as the channel conditions vary. The error control can be FEC, ARQ, or a hybrid. The wireless channel quality is a function of the distance of user from base station, local and average fading conditions, interference variations, and other factors. Furthermore, in packet data systems the bursty nature of data traffic also causes rapid changes in interference characteristics. In a wireless channel, link adaptations should occur frequently because of the rapid changes in signal and interference environment. In such a dynamic environment it is likely that any of the previous schemes is not optimal in terms of energy efficiency all the time. Adaptive error control seems likely a source of efficiency gain.

Adaptive error control can be added fairly easily to a MAC protocol and link layer protocols. First of all, the adaptive error-control techniques have to be present in the sender and receiver.

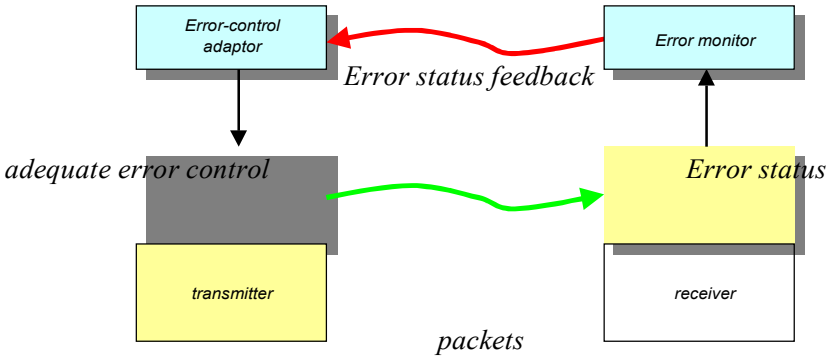


Figure 3: Feedback loop for adaptive error control.

Secondly, a *feedback loop* is required to allow the transmitter to adapt the error coding according to the error rate observed at the receiver. Normally, such information consists of parameters such as mean carrier-to-interference ratio (C/I) or signal-to-noise ratio (SNR), standard deviation of SNR channel impulse response characterisation, bit error statistics (mean and standard deviation), and packet error rate. The required feedback loop limits the responsiveness to the wireless link conditions. Additional information can be gathered with a technique that performs link adaptation in an implicit manner by purely relying on acknowledgement (ACK/NACK) information from the radio link layer.

Depending on the application, the adaptation might not need to be done frequently. If, for example, the application is an error-resilient compression algorithm that when channel distortion occurs, its effects will be a gradual degradation of video quality, then the best possible quality will be maintained at all BERs ([3][56][76][77]).

A more detailed comparison of the performance of ARQ and FEC techniques has been made by many researchers (e.g. [44],[66] and [85]), and is not part of our research.

The choice of energy-efficient error-control strategy is a strong function of QoS parameters, channel quality, and packet size [44]. Since different connections do not have the same requirements concerning e.g. cell loss rate and cell transfer delay, different error-control schemes must be applied for different connection types. The design goal of an error-control system is to find optimum output parameters for a given set of input parameters. Input parameters are e.g. channel BER or maximum delay. Examples of output parameters are FEC code rate and retransmission limit. The optimum might be defined as maximum throughput, minimum delay, or minimum energy consumption, depending on the service class (or QoS) of a connection. Real-time traffic will prefer minimum delay, while most traditional data services will prefer a maximum throughput solution. All solutions in a mobile environment should strive for minimal energy consumption.

Local versus end-to-end error-control

The networking community has explored a wide spectrum of solutions to deal with the wireless error environment. They range from local solutions that decrease the error rate observed by upper layer protocols or applications, to transport protocol modifications and proxies inside the network that modify the behaviour of the higher level protocols [23].

Addressing link errors near the site of their occurrence seems intuitively attractive for several reasons.

- It is most efficient that the error-correcting techniques to be tightly coupled to the transmission environment because they understand their particular characteristics [31].
- Entities on the link are likely to be able to respond more quickly to changes in the error environment, so that parameters such as FEC redundancy and packet length are varied with short time.
- Performing FEC on an end-to-end basis implies codes that deal with a variety of different loss and corruption mechanisms, even on one connection. In practice this implies that different codes have to be concatenated to deal with every possible circumstance, and the resulting multiple layers of redundancy would be carried by every link with a resultant traffic and energy consumption penalty [30]. End-to-end error control requires sufficient redundancy for the worst case link, resulting in a rate penalty on links with less impairment. Local error control requires only extra bandwidth where it is truly needed.
- Practically, deploying a new wireless link protocol on only those links that need it is easier than modifying code on all machines. Application-level proxies address this problem to some extend, but they are currently constrained to running end systems, whereas local error control can operate on exactly the links that require it [23].

Despite these attractions, trying to solve too much locally can lead to other problems. In the case of local error control for wireless links, there are at least three dangers [23].

- Local error control alters the characteristics of the network, which can confuse higher layer protocols. For example, local retransmission could result in packet reordering or in large fluctuations of the round-trip time, either of which could trigger TCP timeouts and retransmissions.
- Both local and end-to-end error control may respond to the same events, possibly resulting in undesirable interactions, causing inefficiencies and potentially even instability.
- End-to-end control has potentially better knowledge of the quality requirements of the connection. For example, a given data packet may bear information with a limited useful lifetime (e.g. multimedia video traffic), so error control that will cause the delay to exceed a certain value is

wasted effort. It might be better to drop a corrupt video packet, than to retransmit it, since retransmission may make the next packet late.

Given the significant advantages of local error control, we will pursue a local approach for the lower layers of the communication protocol stack. However, while we propose that the primary responsibility for error control fall to the local network, there is no reason to dogmatically preclude the involvement of higher level protocols. In particular, the application should be able to indicate to the local network the type of its traffic and the QoS expectations.

The lowest level solution to local error-control is by using hardware error-control techniques such as adaptive codecs and multi-rate modems. While these are attractive in terms of simplicity, they may leave a noticeable residual error rate. In addition, while they reduce the average error rate, they cannot typically differentiate between traffic of different connections. A MAC and link-layer approach that is able to apply error control on a per-traffic basis is an attractive alternative. These protocols, such as IEEE

802.11 [41], MASCARA [5], and E^{2}MaC [33], are □ or can be made □ traffic-aware (rather than protocol-aware) by tailoring the level of error control to the nature of the traffic (e.g. bounding retransmission for packets with a limited lifetime).

Energy-efficient wireless network design

This section describes the basic principles and mechanisms of the network interface architecture implemented in our research, and our energy efficient medium access control protocol for wireless links, called E^{2}MaC. The protocol and the architecture are targeted to a system in which quality of service (including the incurring energy consumption) plays a crucial role. The ability to integrate diverse functions of a system on the same chip provides the challenge and opportunity to do system architecture design and optimisations across diverse system layers and functions [73].

As mentioned before, two key requirements in mobile multimedia systems are:

- *Requirement 1*: the need to maintain quality of service in a mobile environment and,
- *Requirement 2*: the need to use limited battery resources available efficiently.

We have tackled these problem by making the system highly adaptive and by using energy saving techniques through all layers of the system. Adaptations to the dynamic nature of wireless networks are necessary to achieve an acceptable quality of service. It is not sufficient to adapt just one function, but it requires adaptation in several functions of the system, including radio, medium access protocols, error control, network protocols, codecs, and applications. Adaptation is also a key to enhancing battery life. Current research on several aspects of wireless networks (like error control, frame- length, access scheduling) indicate that continually adapting to the current condition of the wireless link have a big impact on the energy-efficiency of the system [13][16][36][45][41]. In our work these existing ideas and several new ideas have been combined into the design of adaptive energy efficient medium access protocols, communication protocol decomposition, and network interface architecture [37] using the previously mentioned principles P1, P2, P3, and P4.

System overview

The goals of low energy consumption and the required support for multiple traffic types lead to a system that is based on reservation and scheduling strategies. The wireless ATM network is composed of several base-stations that each handle a single radio cell[2] possibly covering several mobile stations. We consider an office environment in which the cells are small and have the size of one or several rooms. This not only saves energy because the transmitters can be low powered, it also provides a high aggregate bandwidth since it needs to be shared with only few mobiles. The backbone of the base- stations is a wired ATM network. In order to avoid a serious mismatch between the wired and wireless networks, the wireless network part should offer similar services as the wired network.

The general theme that influences many aspects of the design of the data link protocol is

adaptability and flexibility. This implies that for each connection a different set of parameters concerning scheduling, flow control and error control should be applied.

We do not intend to handle all aspects of a full-blown wireless ATM network that provides all possible services. We adapt some features of ATM because they can be used quite well for our purpose. To implement the full ATM stack would require a large investment in code and hardware. The QoS provisions of ATM fit quite well with the requirements of multimedia traffic. This provides much more possibilities for differentiating various media streams than an often used approach in QoS providing network systems with just two priority levels (real-time versus non-real-time) [17], or even multiple priority levels [33].

However, when adopting ATM in a wireless environment, we need a much more dynamic approach to resource usage. The small size packet structure and small header (in B-ISDN ATM 48 bytes data and 5 bytes header) allows for a simple implementation. Small cells have the benefit of a small scheduling granularity, and hence provide a good control over the quality of a connection. The fixed size also allows a simple implementation of a flexible buffering mechanism that can be adapted to the QoS of a connection. Also a flexible error control mechanism has advantage when these cells are adopted. When the base station is connected via a wired ATM network, then the required processing and adaptation can be minimal since they use the same cell structure and the same quality characteristics.

The system contains several QoS managers. Applications might need resources under control of several QoS managers. The QoS managers then need to communicate with each other via a wired network and wirelessly with applications on mobiles. The key to providing service quality will be the scheduling algorithm executed by the QoS manager that is typically located at the base-station. This QoS manager tries to find a (near) optimal 'schedule' that satisfies the wishes of all applications.

Each mobile can have multiple unidirectional connections with different Quality of Service requirements. Five service categories have been defined under ATM (see Section 5.4.1): constant bit rate (CBR), real time VBR, non-real time VBR, unspecified bit rate (UBR), and available bit rate (ABR). The scheduler gives priority to these categories in the same order as listed here possibly using different scheduling algorithms for each category.

The base-station receives transmission requests from the mobiles. The base-station controls access on the wireless channel based on these requests by dividing bandwidth into *transmission slots*. The key to providing QoS for these connections will be the scheduling algorithm that assigns the bandwidth. The premise is that the base-station has virtually no processing and energy limitations, and will perform actions in courtesy of the mobile. The main principles are (using the principles *P*1 to *P4* of Section 5.3): avoid unsuccessful actions by avoiding collisions and by providing provisions for adaptive error control, minimise the number of transitions by scheduling traffic in larger packets, synchronise the mobile and the base-station which allows the mobile to power-on precisely when needed, and migrate as much as possible work to the base-station.

The layers of the communication protocol are summarised in Figure 4. The column in the middle represents the layers used by the base-station; the columns on the left and right represents the layers used by the mobile.

base-station

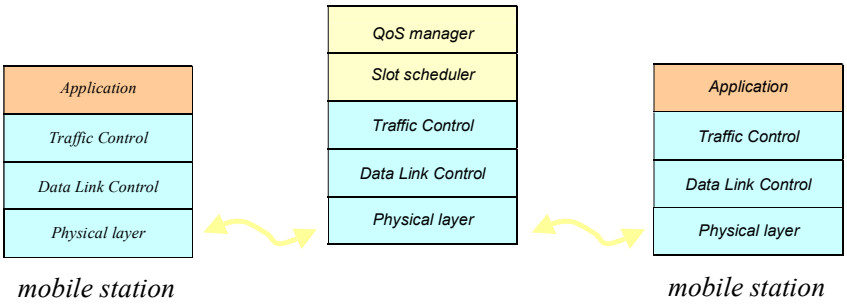


Figure 4: Protocol stack

The lower layers exist in both the mobile and the base station. The *Data link control* manages the data-transfer with the physical layer (using the E^2MaC protocol), and *Traffic control* performs error control and flow control. The base-station contains two additional layers: the *Slot Scheduler* that assigns slots within frames to connections, and the *QoS manager* that establishes, maintains and releases virtual connections.

The definition of the protocol in terms of multiple phases in a frame is similar to other protocols proposed earlier. The E^2MaC protocol goes beyond these protocols by having minimised the energy consumption of the mobile within the QoS requirements of a connection. The features of the protocol are support for multiple traffic types, per- connection flow control and error-control, provision of service quality to individual connections, and energy efficiency consideration.

E^2MaC protocol

In the E^2MaC protocol the scheduler of the base station is responsible for providing the required QoS for the connections on the wireless link and tries to minimise the amount of energy spend by the mobile. It uses the four main principles *P1* to *P4*. The protocol is able to provide near-optimal energy efficiency (i.e. energy is spent for the actual transfer only) for a mobile within the constraints of the QoS of all connections.

The protocol uses fixed-length frames of multiple slots. Each slot has a fixed size. A slot determines the time-frame in which data can be received or transmitted. The base-station and mobile are completely synchronised (the time unit is a slot), which allows the mobile to power-on precisely when needed. The base-station controls the traffic for all mobiles in range of the cell and broadcasts the schedule to the mobiles.

Module VIII: Secure Wireless Communication [4L]

Introduction-different types of attacks, internal attacks, external attacks; measures against attacks (authentication, intrusion detection, encryption); RC4 algorithm

Module VIII: Secure Wireless Communication [4L]

Introduction-different types of attacks, internal attacks, external attacks; measures against attacks (authentication, intrusion detection, encryption); RC4 algorithm

8.1 Collision avoidance & resolution mechanism

Introduction

Security is a critical issue in mobile radio applications both for the users and providers of such systems. Although the same may be said of all communications systems, mobile applications have special requirements and vulnerabilities, and are therefore of special concern. Wireless networks share many common characteristics with traditional wire-line networks such as public switch telephone/data

net-works, and therefore, many security issues with the wire-line networks also apply to the wireless environment. Wireless networks, while providing many benefits over their wired counterparts, including the elimination of cabling costs and increased user mobility, present some serious security concerns. Unlike wired networks, where the physical transmission medium can be secured, wireless networks use the air as a transmission medium. This allows easy access to transmit-ted data by potential eavesdroppers. The mobility of wireless net-works also introduces problems. The mobility of users, the transmission of signals through the open-air and the low power consumption of the mobile user bring to a wireless network a large num-ber of features distinctively different from those seen in a wire-line

8.2 Different types of attacks, internal attacks, external attacks

On the basis of source, attacks can be classified as external and internal attacks. External attacks are caused by the nodes which are not a part of the network. External attackers are the aims to cause congestion, propagate fake routing information or disturb nodes from providing services. Internal attacks are caused by the nodes which are a part of the network. Internal attacks, in which the adversary wants to gain the normal access to the network and participate the network activities, either by some malicious impersonation to get the access to the network as a new node, or by directly compromising a current node and using it as a basis to conduct its malicious behaviors.

Passive Attacks
Some important passive attacks are: Snooping Attacks, Eavesdropping Attacks, Traffic Analysis Attacks, and Traffic Monitoring Attacks.

Snooping Attacks
Snooping Attack is also known as masquerade or impersonation or spoofing Network attack. In this attack, a single malicious node attempts to take out the identity of other nodes' in the network by advertising false/fake routes. It then attempts to send packets over network with identity of other nodes making the destination believe that the packet is from original source.

Eavesdropping Attacks
The eavesdropping attacks are serious security threat to a wireless sensor network (WSN) since the eavesdropping attack is a prerequisite for other attacks.

Traffic Analysis Attacks
Traffic analysis is the process of intercepting and examining messages in order to deduce information from patterns in communication. It can be performed even when the messages are encrypted and cannot be decrypted. In general, the greater the number of messages observed, or even intercepted and stored, the more can be inferred from the traffic. Traffic analysis can be performed in the context of military intelligence or counter intelligence, and is a concern in computer security. In this type of attack, an attacker tries to sense the communication path between the sender and receiver. This way attacker found the amount of data which is travel between the route of sender and receiver. There is no alteration in data by the traffic analysis.

Monitoring Attacks
Monitoring is another passive attack in which attacker can see the confidential data but he cannot change the data or cannot modify the data

Active attack
Some important active attacks are: Blackmail, Denial of service attack Fabrication, Gray hole Attacks, Disclosure Attacks, Routing Attacks and Recourse Consumption Attacks.

Blackmail Attacks

A black mail attack is relevant against routing protocols that uses mechanisms for identification of malicious nodes and propagate messages that try to blacklist the offender.

Denial of service attacks
Denial of service attacks are aimed at complete disruption of routing information and therefore the whole operation of ad-hoc network.

Fabrication Attacks
The notation "fabrication" is used when referring to attacks performed by generating false routing messages. Such kind of attacks can be difficult to identify as they come as valid routing constructs, especially in the case of fabricated routing error messages, which claim that a neighbor can no longer be contacted.

Gray hole Attacks
A gray hole attack is a variation of the black hole attack, where the malicious node is not initially malicious, it turns malicious sometime later. In this attack, an attacker drops all data packets but it lets control messages to route through it.

Disclosure attacks
Disclosure attacks are aimed at acquiring system specific information about a website such as software distribution, version numbers, and patch levels. The acquired information might also contain the location of backup files or temporary files.

Routing Attacks
In Routing Attacks, attackers try to alter the routing information and data in the routing control packet. There are several types of routing attacks mounted on the routing protocol which are intended for disturbing the operation of the network

Resource Consumption Attack
In Resource Consumption Attack, a malicious node intentionally tries to consume or misuse of the resources (battery power, bandwidth, and computational power) of other nodes' exist in the network by requesting excessive route discovery (unnecessary route request control messages), very frequent generation of beacon packets, or by forwarding unnecessary packets (stale information) to that node.

On the basis of Behavior a passive attack attempts to retrieve valuable information by listening to traffic channel without proper authorization, but does not affect system resources and the normal functioning of the network. Passive attacks are very hard to detect because they do not involve any alteration of the data. An active attack attempts to change or destroy the system resources. It gains an authentication and tries to affect or disrupt the normal functioning of the network services by injecting or modifying arbitrary packets of the data being exchanged in the network. An active attack involves information interruption, modification, or fabrication.

On the basis of Nodes in these types of attacks, there are numerous nodes involved during the attack. These nodes can be physically existent or not existing at all.

Collaborative attacks

Collaborative attacks (CA) occur when more than one attacker or running process synchronize their actions to disturb a target network. Multiple attacks occur when a system is disturbed by more than one attacker, but not necessarily in collaboration. We have study different types of attacks and then provided the definition of collaborative attacks; we are now going to categorize these attacks into two different categories. First: Direct Collaborative Attacks and Second: Indirect Collaborative Attacks. Here, the attacker nodes are already in existence in the original network or a malicious node joins the

network or an internal node is compromised in the network. This kind of collaborative attacks can be referred to as direct collaborative attacks. A Black hole and Wormhole attack belongs to this category.

In the black hole attack, attacker uses the routing protocol to advertise itself as having the best path to the node whose packets it want to intercept. An attacker use the flooding based protocol for listing the request for a route from the initiator, then attacker create a reply message he has the shortest path to the receiver . As this message from the attacker reached to the initiator before the reply from the actual node, then initiator assume that it is the shortest path to the receiver. So that a fake route is create. Once the attacker has been able to insert himself between the communications node, then attacker may able to do anything with the packet which is send by the initiator for the receiver.

In a wormhole attack, an attacker receives packets at one point in the network, “tunnels” them to another point in the network, and then replays them into the network from that point. Routing can be disrupted when routing control message are tunneled. This tunnel between two colluding attacks is known as a wormhole. The attacker used different nonexistent nodes in order to fake other nodes to redirect data packets to malicious node. This kind of collaborative attacks can be referred to as indirect collaborative attacks. A Sybil and Routing table overflow attacks belongs to this category.

Sybil attack refers to the multiple copies of malicious nodes. It can be happen, if the malicious node shares its secret key with other malicious nodes. This way the number of malicious node is increased in the network and the probability of the attack is also increased. If we use the multipath routing, then the possibility of choosing a path in the network, those contain the malicious node will be increased. The malicious node makes routing services a target because it’s an important service in MANETs. There are two flavors to this routing attack. One is attack on routing protocol and another is attack on packet forwarding or delivery mechanism. The first is aimed at blocking the propagation of routing information to a node. The latter is aimed at disturbing the packet delivery against a predefined path.

8.3 Measures against attacks (authentication, intrusion detection, encryption)

Intrusion prevention

Security is an essential service for wired and wireless network communications. The success of MANET strongly depends on whether its security can be trusted. However, the characteristics of MANET pose both challenges and opportunities in achieving the security goals, such as confidentiality, authentication, integrity, availability, access control, and non-repudiation. The mobile hosts forming a MANET are normally mobile devices with limited physical protection and resources. Security modules, such as tokens and smart cards, can be used to protect against physical attacks. Cryptographic tools are widely used to provide powerful security services, such as confidentiality, authentication, integrity, and non-repudiation. Unfortunately, cryptography cannot guarantee availability; for example, it cannot prevent radio jamming. Meanwhile, strong cryptography often demands a heavy computation overhead and requires the auxiliary
complicated key distribution and trust management services, which mostly are restricted by the capabilities of physical devices (e.g. CPU or battery).

The characteristics and nature of MANET require the strict cooperation of participating mobile hosts. A number of security techniques have been invented and a list of security protocols has been proposed to enforce cooperation and prevent misbehavior, such as 802.11 WEP, IPsec, SEAD, SAODV, SRP, ARAN, SSL, and so on. However, none of those preventive approaches is perfect or capable to defend

against all attacks. A second line of defense called intrusion detection systems (IDS) is proposed and applied in MANET. IDS are some of the latest security tools in the battle against attacks. Distributed IDS were introduced in MANET to monitor either the misbehavior or selfishness of mobile hosts. Subsequent actions can be taken based on the information collected by IDS.

Security is the combination of processes, procedures, and systems used to ensure confidentiality, authentication, integrity, availability, access control, and nonrepudiation.

Confidentiality is to keep the information sent unreadable to unauthorized users or nodes. MANET uses an open medium, so usually all nodes within the direct transmission range can obtain the data. One way to keep information confidential is to encrypt the data, and another technique is to use directional antennas.
Authentication is to be able to identify a node or a user, and to be able to prevent impersonation. In wired networks and infrastructure-based wireless networks, it is possible to implement a central authority at a point such as a router, base station, or access point. But there is no central authority in MANET, and it is much more difficult to authenticate an entity.
Integrity is to be able to keep the message sent from being illegally altered or destroyed in the transmission. When the data is sent through the wireless medium, the data can be modified or deleted by malicious attackers. The malicious attackers can also resend it, which is called a replay attack.
Non-repudiation is related to a fact that if an entity sends a message, the entity cannot deny that the message was sent by it. By producing a signature for the message, the entity cannot later deny the message. In public key cryptography, a node A signs the message using its private key. All other nodes can verify the signed message by using A's public key, and A cannot deny that its signature is attached to the message.
Availability is to keep the network service or resources available to legitimate users. It ensures the survivability of the network despite malicious incidents.
Access control is to prevent unauthorized use of network services and system resources. Obviously, access control is tied to authentication attributes. In general, access control is the most commonly thought of service in both network communications and individual computer systems.

A variety of security mechanisms have been invented to counter malicious attacks. The conventional approaches such as authentication, access control, encryption, and digital signature provide a first line of defense. As a second line of defense, intrusion detection systems and cooperation enforcement mechanisms implemented in MANET can also help to defend against attacks or enforce cooperation, reducing selfish node behavior.

Preventive mechanism: The conventional authentication and encryption schemes are based on cryptography, which includes asymmetric and symmetric cryptography. Cryptographic primitives such as hash functions (message digests) can be used to enhance data integrity in transmission as well. Threshold cryptography can be used to hide data by dividing it into a number of shares. Digital signatures can be used to achieve data integrity and authentication services as well. It is also necessary to consider the physical safety of mobile devices, since the hosts are normally small devices, which are physically vulnerable. For example, a device could easily be stolen, lost, or damaged. In the battlefield they are at risk of being hijacked. The protection of the sensitive data on a physical device can be enforced by some security modules, such as tokens or a smart card that is accessible through PIN, passphrases, or biometrics. Although all of these cryptographic primitives combined can prevent most attacks in theory, in reality, due to the design, implementation, or selection of protocols and physical device restrictions, there are still a number of malicious attacks bypassing prevention mechanisms.

Reactive mechanism: An intrusion detection system is a second line of defense. There are widely used to detect misuse and anomalies. A misuse detection system attempts to define improper behavior based on the patterns of well-known attacks, but it lacks the ability to detect any attacks that were not considered during the creation of the patterns; Anomaly detection attempts to define normal or expected behavior statistically. It collects data from legitimate user behavior over a period of time,

and then statistical tests are applied to determine anomalous behavior with a high level of confidence. In practice, both approaches can be combined to be more effective against attacks.

Intrusion detection and prevention systems (IDPS)

Some systems may attempt to stop an intrusion attempt but this is neither required nor expected of a monitoring system. Intrusion detection and prevention systems (IDPS) are primarily focused on identifying possible incidents, logging information about them, and reporting attempts. In addition, organizations use IDPS for other purposes, such as identifying problems with security policies, documenting existing threats and deterring individuals from violating security policies. IDPS have become a necessary addition to the security infrastructure of nearly every organization.

IDPS typically record information related to observed events, notify security administrators of important observed events and produce reports. Many IDPS can also respond to a detected threat by attempting to prevent it from succeeding. They use several response techniques, which involve the IDPS stopping the attack itself, changing the security environment (e.g. reconfiguring a firewall) or changing the attack's content.

Intrusion prevention systems (**IPS**), also known as **intrusion detection and prevention systems** (**IDPS**), are network security appliances that monitor network or system activities for malicious activity. The main functions of intrusion prevention systems are to identify malicious activity, log information about this activity, report it and attempt to block or stop it.

Intrusion prevention systems are considered extensions of intrusion detection systems because they both monitor network traffic and/or system activities for malicious activity. The main differences are, unlike intrusion detection systems, intrusion prevention systems are placed in-line and are able to actively prevent or block intrusions that are detected. IPS can take such actions as sending an alarm, dropping detected malicious packets, resetting a connection or blocking traffic from the offending IP address. An IPS also can correct cyclic redundancy check (CRC) errors, defragment packet streams, mitigate TCP sequencing issues, and clean up unwanted transport and network layer options.

Intrusion prevention systems can be classified into four different type.

1. **Network-based intrusion prevention system (NIPS)**: monitors the entire network for suspicious traffic by analyzing protocol activity.
2. **Wireless intrusion prevention system (WIPS)**: monitor a wireless network for suspicious traffic by analyzing wireless networking protocols.
3. **Network behavior analysis (NBA)**: examines network traffic to identify threats that generate unusual traffic flows, such as distributed denial of service (DDoS) attacks, certain forms of malware and policy violations.
4. **Host-based intrusion prevention system (HIPS)**: an installed software package which monitors a single host for suspicious activity by analyzing events occurring within that host.

RC4 Encryption

RC4 is an encryption algorithm that was created by Ronald Rivest of RSA Security. It is used in WEP and WPA, which are encryption protocols commonly used on wireless routers. The workings of RC4 used to be a secret, but its code was leaked onto the internet in 1994.RC4 was originally very widely used due to its simplicity and speed. Typically 16 byte keys are used for strong encryption, but shorter key lengths are also widely used due to export restrictions. Overtime this code was shown to produce biased outputs towards certain sequences, mostly in first few bytes of the key stream generated. This led to a future version of the RC4 code that is more widely used today, called RC4-drop[n], in which the first n bytes of the key stream are dropped in order to get rid of this biased output. Some notable uses of RC4 are implemented in Microsoft Excel, Adobe's Acrobat 2.0 (1994), and Bit Torrent clients. To begin the process of RC4 encryption, you need a key, which is often user-defined and between 40-

bits and 256-bits.A 40-bit key represents a five character ASCII code that gets translated into its 40 character binary equivalent (for example, the ASCII key "pwd12" is equivalent to0111000001110111011001000011000100110010 in binary).The next part of RC4 is the key-scheduling algorithm (KSA), listed below

```
forifrom0to255
S[i] := i
Endfor
j := 0
for i from 0to255
        j := (j + S[i] + key[i mod keylength]) mod 256
        swap(S[i],S[j])end for

End for
```

KSA creates an array S that contains 256 entries with the digits 0 through 255, as in the tablebelow.012...ii+1...253254255Each of the 256 entries in S are then swapped with the j-th entry in S, which is computed to be j= [(j + S(i) + key[i mod key length]) mod 256],where j is the previous j value (which is initially zero).S[i]is the value of the current entry in S.key[i mod key length]is either a zero or a one. For example, if we are at the52th entry in S and the key length was 40-bit, then 52 mod 40 = 12. The 13th element (because numbering for arrays begins at zero)in the binary version of "pwd12" is 0.For example, consider the first iteration of KSA with key "pwd12". Then, since i = 0, i mod 256 = 0. So, the element at the index 0 of the key is p, and its ascii value is 112. So, the new j is computed as j = [(0 + 0 + 112) mod 256] = 112.So, swapping the i-th and the j-th elements, we obtain the following array after the first iteration:11212...1110113114...255After the terminus of KSA, we obtain the seemingly random array:[101, 124, 172, 10, 166, 26, 46, 91, 2, 137, 39, 243, 253, 25, 3, 30, 47, 238, 196, 38, 94, 149, 15,32, 248, 51, 158, 150, 106, 183, 67, 219, 95, 177, 138, 152, 13, 188, 118, 108, 207, 151, 41, 142,236, 103, 55, 72, 20, 244, 216, 14, 168, 90, 4, 42, 153, 64, 250, 129, 97, 225, 87, 199, 204, 100,16, 249,191, 82, 43, 131, 24, 169, 69, 54, 96, 77, 255, 84, 1, 143, 242, 123, 21, 93, 61, 102, 224, 107,109, 79, 80, 23, 229, 6, 156, 181, 105, 159, 33, 141, 18, 104, 9, 56, 233, 178, 127, 111, 135, 206,202, 128, 31, 71, 211, 222, 45, 66, 163, 189, 167, 201, 232, 17, 251, 198, 170, 155, 115, 57, 228,98, 190, 76, 59, 239, 37, 147, 180, 240, 197, 200, 19, 0, 213, 99, 125, 44, 195, 164, 176, 121,220, 212, 86, 186, 34, 214, 230, 254, 40, 203, 194, 231, 162, 226, 187, 116, 208, 22, 68, 88, 192, 140,
205, 234, 119, 83, 136, 63, 12, 112, 217, 154, 184, 81, 70, 35, 174, 78, 241, 179, 210, 215, 49,144, 130, 48, 133, 7, 209, 92, 73, 193, 28, 75, 117, 223, 50, 113, 114, 148, 173, 29, 53, 160, 8,139, 246, 65, 252, 161, 221, 185, 27, 36, 11, 110, 237, 165, 5, 182, 145, 171, 120, 157, 134, 175,122, 58, 235, 52, 62, 126, 85, 60, 132, 74, 245, 227, 218, 89, 247, 146]The next part of RC4 is the pseudo-random generation algorithm (PRGA). The PRGA is below: i := 0j := 0whileGeneratingOutput:i := (i + 1) mod 256j := (j + S[i]) mod 256swap(S[i],S[j])output S[(S[i] + S[j]) mod 256]end while In PRGA, we begin with the array S that was swapped in the KSA. In PRGA, an element in S(at index i) is swapped with another element in S(at index j). Then, the next element in the encrypted text is the element of S at the index calculated by (S[i] + S[j] mod 256). At each iteration, i is recalculated as (i + 1) mod 256, and j is recalculated as (j + S[i]) mod 256. The number of iterations performed is the length of the key, and every value of S is swapped at least once beyond 256 iterations (due to the fact that i and j are calculated by some numbernmod256). The result of this is the code. Following up with the previous example, let us examine the first iteration of PRGA. Since i, j =0, i becomes 1 and j becomes (0 + S[1]) mod 256. Since S[1] = 124 (see the resulting S from KSA), j becomes 124. Then, the elements of S at 1 and 124 are swapped, yielding the following new array:[101,232, 172, 10, 166, 26, 46, 91, 2, 137, 39, 243, 253, 25, 3, 30, 47, 238, 196, 38, 94, 149, 15,32, 248, 51, 158, 150, 106, 183, 67, 219, 95, 177, 138, 152, 13, 188, 118, 108, 207, 151, 41, 142,236, 103, 55, 72, 20, 244, 216, 14, 168,90, 4, 42, 153, 64, 250, 129, 97, 225, 87, 199, 204, 100,16, 249, 191, 82, 43, 131, 24, 169, 69, 54,

96, 77, 255, 84, 1, 143, 242, 123, 21, 93, 61, 102, 224,107, 109, 79, 80, 23, 229, 6, 156, 181, 105, 159,33, 141, 18, 104, 9, 56, 233, 178, 127, 111, 135,
206, 202, 128, 31, 71, 211, 222, 45, 66, 163, 189, 167, 201,124, 17, 251, 198, 170, 155, 115, 57,228, 98, 190, 76, 59, 239, 37, 147, 180, 240, 197, 200, 19, 0, 213, 99, 125, 44, 195, 164, 176,121, 220, 212, 86, 186, 34, 214, 230, 254, 40, 203, 194, 231, 162, 226, 187, 116, 208, 22, 68, 88,192, 140, 205, 234, 119, 83, 136, 63, 12, 112, 217, 154, 184, 81, 70, 35, 174, 78, 241, 179, 210,215, 49, 144, 130, 48, 133, 7, 209, 92, 73, 193, 28, 75, 117, 223, 50, 113, 114, 148, 173, 29, 53,160, 8, 139, 246, 65, 252, 161, 221, 185, 27, 36, 11, 110, 237, 165, 5, 182, 145, 171, 120, 157,134, 175, 122, 58, 235, 52, 62, 126, 85, 60, 132, 74, 245, 227, 218, 89, 247, 146].Then, we add the two numbers that we have just swapped: 232+124 = 356, mod by 256 to get100. then output the 100th element of S, which is 33. We then look at the i-th index of the input string, "Math 310 Proves!", which gives us 'M'. We take the Unicode code of that character, which is 77, and perform a bitwise AND with 33 to get 108. Finally, get the character value for the Unicode code 108, which is 'l', whose HEX representation is then "6C". This is shown in the beginning of the encrypted output string in HEX:"6CA86FE3CBC33C162595C3E78B9C97BC"In the case of a wireless router, the key will be required to generate the code. If the key is incorrect, then the encrypted outputs of the codes will not match and the user will not be allowed to connect to the router.

Text books:

1) K. Sinha, S.Ghosh and B. P. Sinha, Wireless Networks and Mobile Computing. CRC Press : New York, 2015.
2) J. Schiller, Mobile Communication, Pearson
3) Yi-Bing Lin & Imrich Chlamtac, Wireless and Mobile Networks Architectures, John Wiley & Sons, 2001
4) Raj Pandya, Mobile and Personal Communication systems and services, Prentice Hall of India, 2001
5) 5. XiangYang Li, Wireless Adhoc and Sensor Networks, Cambridge University Press.

Recommended books:

1) Research articles published on secure wireless communication (authentication, mitigation of DoS, DDoS, eavesdropping) published in leading journals.
2) Mark Ciampa, Guide to Designing and Implementing wireless LANs, Thomson learning, Vikas Publishing House, 2001.
3) P.Stavronlakis, Third Generation Mobile Telecommunication systems, Springer Publishers.