MOBILE COMPUTING NOTES
BY
VIKRAM NARAYANDAS

UNIT I

- **INTRODUCTION TO MOBILE COMMUNICATIONS AND COMPUTING**

There are two different kinds of mobility: **user mobility** and **device portability**. User mobility refers to a user who has access to the same or similar telecommunication services at different places, i.e., the user can be mobile, and the services will follow him or her.

With device portability, the communication device moves (with or without a user). Many mechanisms in the network and inside the device have to make sure that communication is still possible while the device is moving. A typical example for systems supporting device portability is the mobile phone system, where the system itself hands the device from one radio transmitter (also called a base station) to the next if the signal becomes too weak.

With regard to devices, the term wireless is used. This only describes the way of accessing a network or other communication partners, i.e., without a wire. The wire is replaced by the transmission of electromagnetic waves through 'the air' (although wireless transmission does not need any medium).

A communication device can thus exhibit one of the following characteristics:

- **Fixed and wired**: This configuration describes the typical desktop computer in an office. Neither weight nor power consumption of the devices allow for mobile usage. The devices use fixed networks for performance reasons.
- **Mobile and wired**: Many of today's laptops fall into this category; users carry the laptop from one hotel to the next, reconnecting to the company's network via the telephone network and a modem.
- **Fixed and wireless**: This mode is used for installing networks, e.g., in historical buildings to avoid damage by installing wires, or at trade shows to ensure fast network setup. Another example is bridging the last mile to a customer by a new operator that has no wired infrastructure and does not want to lease lines from a competitor.
- **Mobile and wireless**: This is the most interesting case. No cable restricts the user, who can roam between different wireless networks. Most technologies discussed in this book deal with this type of device and the networks supporting them. Today's most successful example for this category is GSM with more than 800 million users.

Applications

1. **Vehicles**

Today's cars already comprise some, but tomorrow's cars will comprise many wireless communication systems and mobility aware applications. Music, news, road conditions, weather reports, and other broadcast information are received via digital audio broadcasting (DAB) with 1.5 Mbit/s. For personal communication, a universal mobile telecommunications system (UMTS) phone might be available offering voice and data connectivity with 384 Kbit/s. For remote areas, satellite communication can be used, while the current position of the car is determined via the global positioning system (GPS). Cars driving in the same area build a local ad-hoc network for the fast exchange of information in emergency situations or to help each other keep a safe distance. In case of an accident, not only will the airbag be triggered, but the police and ambulance service will be informed via an emergency call to a service provider. Cars with this technology are already available. In the future, cars will also inform other cars about accidents via the ad-hoc network to

help them slow down in time, even before a driver can recognize an accident. Buses, trucks, and trains are already transmitting maintenance and logistic information to their home base, which helps to improve organization (fleet management), and saves time and money.

2. Emergencies

Just imagine the possibilities of an ambulance with a high-quality wireless connection to a hospital. Vital information about injured persons can be sent to the hospital from the scene of the accident. All the necessary steps for this particular type of accident can be prepared and specialists can be consulted for an early diagnosis. Wireless networks are the only means of communication in the case of natural disasters such as hurricanes or earthquakes. In the worst cases, only decentralized, wireless ad-hoc networks survive. The breakdown of all cabling not only implies the failure of the standard wired telephone system, but also the crash of all mobile phone systems requiring base stations!

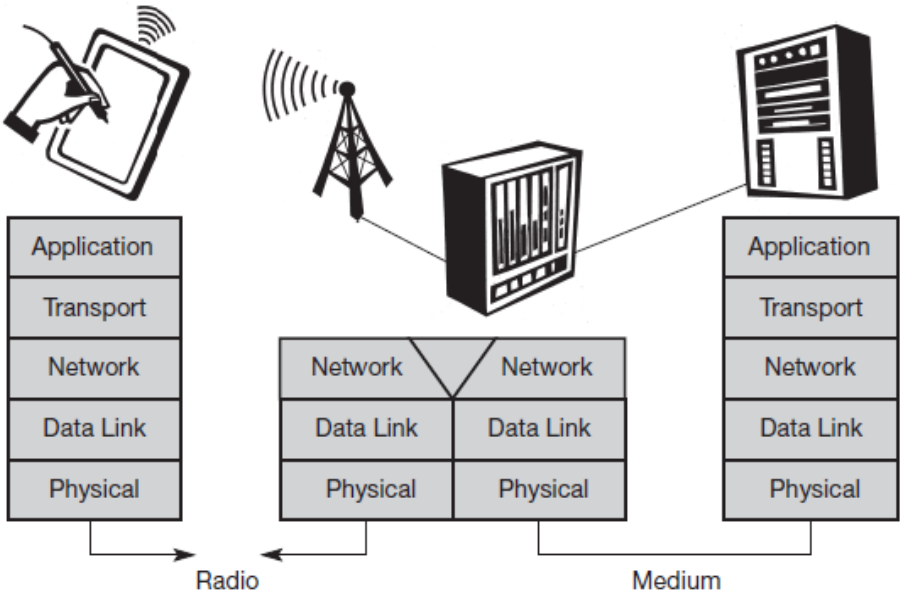
3. Business

A travelling salesman today needs instant access to the company's database: to ensure that files on his or her laptop reflect the current situation, to enable the company to keep track of all activities of their travelling employees, to keep databases consistent etc. With wireless access, the laptop can be turned into a true mobile office, but efficient and powerful synchronization mechanisms are needed to ensure data consistency. At home, the laptop connects via a WLAN or LAN and DSL to the Internet. Leaving home requires a handover to another technology, e.g., to an enhanced version of GSM, as soon as the WLAN coverage ends. Due to interference and other factors discussed in chapter 2, data rates drop while cruising at higher speed. Gas stations may offer WLAN hot spots as well as gas. Trains already offer support for wireless connectivity. Several more handovers to different technologies might be necessary before reaching the office. No matter when and where, mobile communications should always offer as good connectivity as possible to the internet, the company's intranet, or the telephone network.

4. Replacement of wired networks

In some cases, wireless networks can also be used to replace wired networks, e.g., remote sensors, for tradeshows, or in historic buildings. Due to economic reasons, it is often impossible to wire remote sensors for weather forecasts, earthquake detection, or to provide environmental information. Wireless connections, e.g., via satellite, can help in this situation. Tradeshows need a highly dynamic infrastructure, but cabling takes a long time and frequently proves to be too inflexible. Many computer fairs use WLANs as a replacement for cabling. Other cases for wireless networks are computers, sensors, or information displays in historical buildings, where excess cabling may destroy valuable walls or floors. Wireless access points in a corner of the room can represent a solution.

A simplified reference model



● **Physical layer**: This is the lowest layer in a communication system and is responsible for the conversion of a stream of bits into signals that can be transmitted on the sender side. The physical layer of the receiver then transforms the signals back into a bit stream. For wireless communication, the physical layer is responsible for frequency selection, generation of the carrier frequency, signal detection (although heavy interference may disturb the signal), modulation of data onto a carrier frequency and (depending on the transmission scheme) encryption.

● **Data link layer**: The main tasks of this layer include accessing the medium, multiplexing of different data streams, correction of transmission errors, and synchronization (i.e., detection of a data frame). Altogether, the data link layer is responsible for a reliable point-to point connection between two devices or a point-to-multipoint connection between one sender and several receivers.

● **Network layer**: This third layer is responsible for routing packets through a network or establishing a connection between two entities over many other intermediate systems. Important topics are addressing, routing, device location, and handover between different networks.

● **Transport layer**: This layer is used in the reference model to establish an end-to-end connection. Topics like quality of service, flow and congestion control are relevant, especially if the transport protocols known from the Internet, TCP and UDP, are to be used over a wireless link.

● **Application layer**: Finally, the applications (complemented by additional layers that can support applications) are situated on top of all transmission oriented layers. Topics of interest in this context are service location, support for multimedia applications, adaptive applications that can handle the large variations in transmission characteristics, and wireless access to the World Wide Web using a portable device. Very demanding applications are video (high data rate) and interactive gaming (low jitter, low latency).

Frequencies for radio transmission

Radio transmission can take place using many different frequency bands. Each frequency band exhibits certain advantages and disadvantages The figure shows frequencies starting at 300 Hz and going up to over 300 THz. Directly coupled to the frequency is the wavelength λ via the equation: $\lambda = c/f$, where $c \cong 3 \cdot 108$ m/s (the speed of light in vacuum) and f the frequency. For traditional wired networks, frequencies of up to several hundred kHz are used for distances up to some km with twisted pair copper wires, while frequencies of several hundred MHz are used with coaxial cable (new coding schemes work with several hundred MHz even with twisted pair copper wires over distances of some 100 m). Fiber optics are used for frequency ranges of several hundred THz, but here one typically refers to the wavelength which is, e.g., 1500 nm, 1350 nm etc. (infra red).

Radio transmission starts at several kHz, the very low frequency (VLF) range. These are very long waves. Waves in the low frequency (LF) range are used by submarines, because they can penetrate water and can follow the earth's surface. Some radio stations still use these frequencies, e.g., between 148.5 kHz and 283.5 kHz in Germany. The medium frequency (MF) and high frequency (HF) ranges are typical for transmission of hundreds of radio stations either as amplitude modulation (AM) between 520 kHz and 1605.5 kHz, as short wave (SW) between 5.9 MHz and 26.1 MHz, or as frequency modulation (FM) between 87.5 MHz and 108 MHz. The frequencies limiting these ranges are typically fixed by national regulation and, vary from country to country. Short waves are typically used for (amateur) radio transmission around the world, enabled by reflection at the ionosphere. Transmit power is up to 500 kW – which is quite high compared to the 1 W of a mobile phone. digital audio broadcasting (DAB) takes place as well (223–230 MHz and 1452–1472 MHz) and digital TV is planned or currently being installed (470– 862 MHz), reusing some of the old frequencies for analog TV. UHF is also used for mobile phones with analog technology (450–465 MHz), the digital GSM (890–960 MHz, 1710–1880 MHz), digital cordless telephones following the DECT standard (1880–1900 MHz), 3G cellular systems following the UMTS standard (1900–1980 MHz, 2020–2025 MHz, 2110–2190 MHz) and many more. VHF and especially UHF allow for small antennas and relatively reliable connections for mobile telephony. Super high frequencies (SHF) are typically used for directed microwave links (approx. 2–40 GHz) and fixed satellite services in the C-band (4 and 6 GHz), Ku-band (11 and 14 GHz), or Ka-band (19 and 29 GHz). Some systems are planned in the extremely high frequency (EHF) range which comes close to infra red. All radio frequencies are regulated to avoid interference, e.g., the German regulation covers 9 kHz–275 GHz. The next step into higher frequencies involves optical transmission, which is not only used for fiber optical links but also for wireless communications. Infra red (IR) transmission is used for directed links, e.g., to connect different buildings via laser links. The most widespread IR technology, infra red data association (IrDA), uses wavelengths of approximately 850–900 nm to connect laptops, PDAs etc. Finally, visible light has been used for wireless transmission for thousands of years. While light is not very reliable due to interference, but it is nevertheless useful due to built-in human receivers.

Data and Signals

To be transmitted, data must be transformed to electromagnetic signals
Both data and the signals that represent them can be either analog or digital in form.

Analog and Digital Data

Data can be analog or digital. The term analog data refers to information that is continuous;digital data refers to information that has discrete states

Analog data, such as the sounds made by a human voice, take on continuous values.When someonespeaks, an analog wave is created in the air. This can be captured by amicrophone and converted to an analog signal or sampled and converted to a digitalsignal.Digital data take on discrete values. For example, data are stored in computer memory in the form of Os and 1s. They can be converted to a digital signal or modulated into an analog signal for transmission across a medium.

Signals can be analog or digital. Analog signals can have an infinite number of values in a range; digital signals can have only a limited number of values

Antennas

As the name wireless already indicates, this communication mode involves 'getting rid' of wires and transmitting signals through space without guidance. We do not need any 'medium' (such as an ether) for the transport of electromagnetic waves.

Antennas couple electromagnetic energy to and from space to and from a wire or coaxial cable (or any other appropriate conductor). A theoretical reference antenna is the isotropic radiator, a point in space radiating equal power in all directions, i.e., all points with equal power are located on a sphere with the antenna as its center. The radiation pattern is symmetric in all directions two dimensional

Real antennas

exhibit directive effects, i.e., the intensity of radiation is not the same in all directions from the antenna. The simplest real antenna is a thin, center-fed dipole, also called Hertzian dipole, The dipole consists of two collinear conductors of equal length, separated by a small feeding gap. The length of the dipole is not arbitrary, but, for example, half the wavelength λ of the signal to transmit results in a very efficient radiation of the energy. If mounted on the roof of a car, the length of $\lambda/4$ is efficient. This is also known as Marconi antenna. A $\lambda/2$ dipole has a uniform or omni-directional radiation pattern in one plane .

Signal propagation

Like wired networks, wireless communication networks also have senders and receivers of signals. However, in connection with signal propagation, these two networks exhibit considerable differences. In wireless networks, the signal has no wire to determine the direction of propagation, whereas signals in wired networks only travel along the wire (which can be twisted pair copper wires, a coax cable, but also a fiber etc.). As long as the wire is not interrupted or damaged, it typically exhibits the same characteristics at each point. One can precisely determine the behavior of a signal travelling along this wire, e.g., received power depending on the length. For wireless

transmission, this predictable behavior is only valid in a vacuum, i.e., without matter between the sender and the receiver.

Transmission range: Within a certain radius of the sender transmission is possible, i.e., a receiver receives the signals with an error rate low enough to be able to communicate and can also act as sender.

- Detection range: Within a second radius, detection of the transmission is possible, i.e., the transmitted power is large enough to differ from background noise. However, the error rate is too high to establish communication.
- Interference range: Within a third even larger radius, the sender may interfere with other transmission by adding to the background noise. A receiver will not be able to detect the signals, but the signals may disturb other signals.

MULTIPLEXING

Bandwidth utilization is the wise use of available bandwidth to achieve specific goals.Efficiency can be achieved by multiplexing; privacy and antijamming can be achieved by spreading.

Multiplexing is the set of techniques that allows the simultaneous transmission of multiple signals across a single data link. As data and telecommunications use increases, so does traffic. We can accommodate this increase by continuing to add individual links each time a new channel is needed; or we can install higher-bandwidth links and use each to carry multiple signals.

There are three basic multiplexing techniques: frequency-division multiplexing, wavelength-division multiplexing, and time-division multiplexing. The first two are techniques designed for analog signals, the third, for digital signals.

FDM is an analog multiplexing technique that combines analog signals.

WDM is an analog multiplexing technique to combine optical signals

TDM is a digital multiplexing technique for combining several low-rate channels into one high-rate one.

In synchronous TDM, the data rate of the link is *n* times faster, and the unit duration is *n* times shorter.

Space division multiplexing

For wireless communication, multiplexing can be carried out in four dimensions: space, time, frequency, and code. In this field, the task of multiplexing is to assign space, time, frequency, and code to each communication channel with a minimum of interference and a maximum of medium utilization. The term communication channel here only refers to an association of sender(s) and receiver(s) who want to exchange data six channels ki and introduces a three dimensional coordinate system. This system shows the dimensions of code c, time t and frequency f. For this first type of multiplexing, space division multiplexing (SDM) The channels k1 to k3 can be mapped onto the three ‘spaces’ s1 to s3 which clearly separate the channels and prevent the interference

ranges from overlapping. The space between the interference ranges is sometimes called guard space. Such a guard space is needed in all four multiplexing schemes presented.

Frequency division multiplexing

Frequency division multiplexing (FDM) describes schemes to subdivide the frequency dimension into several non-overlapping frequency bands
. Each channel ki is now allotted its own frequency band as indicated. Senders using a certain frequency band can use this band continuously. Again, guard spaces are needed to avoid frequency band overlapping (also called adjacent channel interference). This scheme is used for radio stations within the same region, where each radio station has its own frequency. This very simple multiplexing scheme does not need complex coordination between sender and receiver: the receiver only has to tune in to the specific sender.

Time division multiplexing

A more flexible multiplexing scheme for typical mobile communications is time division multiplexing (TDM). Here a channel ki is given the whole bandwidth for a certain amount of time, i.e., all senders use the same frequency but at different points in time .Again, guard spaces, which now represent time gaps, have to separate the different periods when the senders use the medium. In our highway example, this would refer to the gap between two cars. If two transmissions overlap in time, this is called co-channel interference. (In the highway example, interference between two cars results in an accident.) To avoid this type of interference, precise synchronization between different senders is necessary. This is clearly a disadvantage, as all senders need precise clocks or, alternatively, a way has to be found to distribute a synchronization signal to all senders. For a receiver tuning in to a sender this does not just involve adjusting the frequency, but involves listening at exactly the right point in time. However, this scheme is quite flexible as one can assign more sending time to senders with a heavy load and less to those with a light load. requency and time division multiplexing can be combined, i.e., a channel ki can use a certain frequency band for a certain amount of time Now guard spaces are needed both in the time and in the frequency dimension. This scheme is more robust against frequency selective interference, i.e., interference in a certain small frequency band. A channel may use this band only for a short period of time. Additionally, this scheme provides some (weak) protection against tapping, as in this case the sequence of frequencies a sender uses has to be known to listen in to a channel. The mobile phone standard GSM uses this combination of frequency and time division multiplexing for transmission between a mobile phone and a so-called base station

Code division multiplexing

While SDM and FDM are well known from the early days of radio transmission and TDM is used in connection with many applications, code division multiplexing (CDM) is a relatively new scheme in commercial communication systems. First used in military applications due to its inherent security features (together with spread spectrum techniques, it now features in many civil wireless transmission scenarios thanks to the availability of cheap processing power channels ki use the same frequency at the same time for transmission. Separation is now achieved by assigning each channel its own 'code', guard spaces are realized by using codes with the necessary 'distance' in code space, e.g., orthogonal codes. The main advantage of CDM for wireless transmission is that it gives good protection against interference and tapping. Different codes have to be assigned, but code space is huge compared to the frequency space. Assigning individual codes to each sender does not usually cause problems. The main disadvantage of this scheme is the relatively high complexity of the receiver. A receiver has to know the code and must separate the channel with user data from the background noise composed of other signals and environmental noise. Additionally, a receiver must be precisely synchronized with the transmitter to apply the decoding correctly. The voice example also gives a hint to another problem of CDM receivers. All signals should reach a receiver with almost equal strength, otherwise some signals could drain others. If some people close to a receiver talk very loudly the language does not matter. The receiver cannot listen to any other person. To apply CDM, precise power control is required.

SPREAD SPECTRUM

In spread spectrum, we also combine signals from different sources to fit into a larger bandwidth, but our goals are somewhat different. Spread spectrum is designed to be used in wireless applications (LANs and WANs). In these types of applications, we have some concerns that outweigh bandwidth efficiency. In wireless applications, all stations use air (or a vacuum) as the medium for communicatiospread spectrum techniques add redundancy; they spread the original spectrum needed for each station.

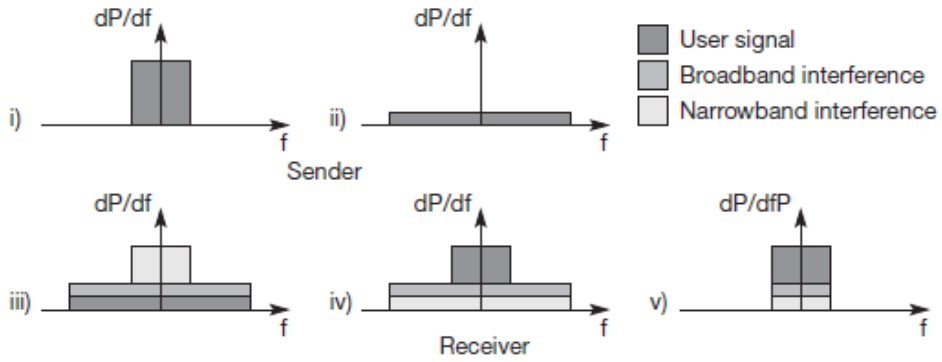


Figure 2.32 Spread spectrum: spreading and despreading

There are two techniques to spread the bandwidth: frequency hopping spread spectrum (FHSS) and direct sequence spread spectrum (DSSS).

Frequency Hopping Spread Spectrum (FHSS)

The frequency hopping spread spectrum (FHSS) technique uses *M* different carrier frequencies that are modulated by the source signal. At one moment, the signal modulates one carrier frequency; at the next moment, the signal modulates another carrier frequency. Although the modulation is done using one carrier frequency at a time, *M* frequencies are used in the long run.

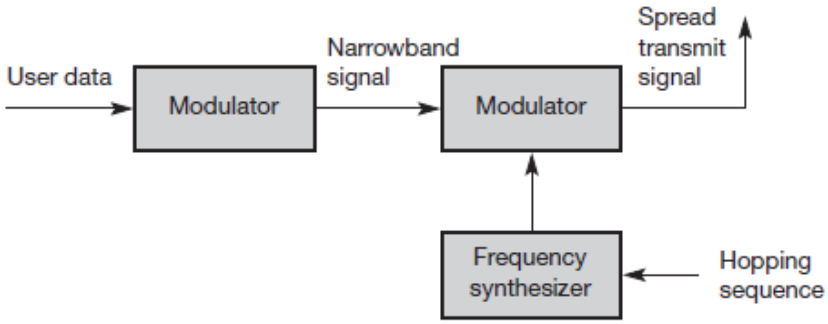


Figure 2.39
FHSS transmitter

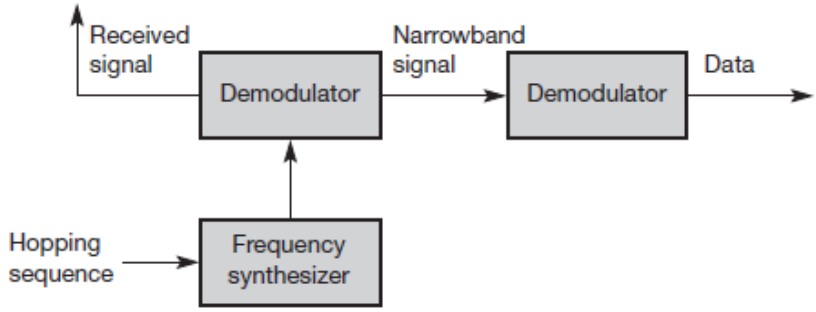


Figure 2.40
FHSS receiver

A pseudorandom code generator, called pseudorandom noise (PN), creates a k-bit pattern for every hopping period *Th*• The frequency table uses the pattern to find the frequency to be used for this hopping period and passes it to the frequency synthesizer. The frequency synthesizer creates a carrier signal of that frequency, and the source signal modulates the carrier signal.

Direct Sequence Spread Spectrum

The direct sequence spread spectrum (nSSS) technique also expands the bandwidth of the original signal, but the process is different. In DSSS, we replace each data bit with 11 bits using a spreading code. In other words, each bit is assigned a code of 11 bits, called chips, where the chip rate is 11 times that of the data bit. spreading code is 11 chips having the pattern 10110111000 (in this case). If the original signal rate is *N,* the rate of the spread signal is *lIN.* This means that the required

bandwidth for the spread signal is 11 times larger than the bandwidth of the original signal. The spread signal can provide privacy if the intruder does not know the code. It can also provide immunity against interference if each station uses a different code.

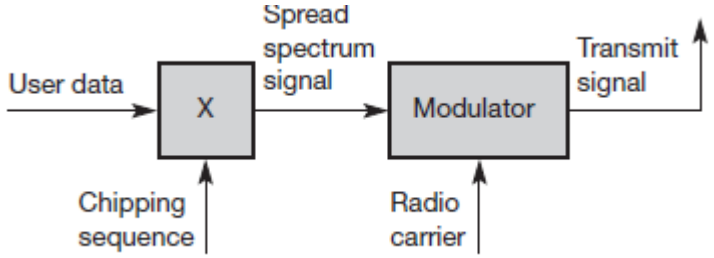
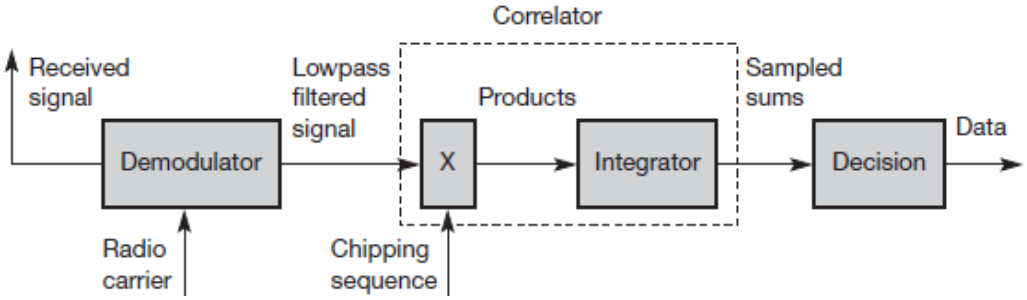


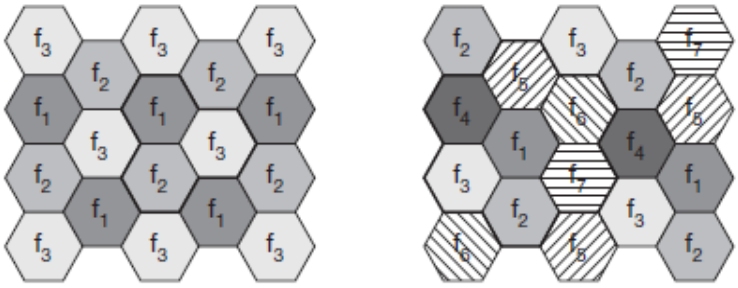
Figure 2.37
DSSS receiver



Cellular systems

Cellular systems for mobile communications implement SDM. Each transmitter,typically called a base station, covers a certain area, a cell. Cell radii can varyfrom tens of meters in buildings, and hundreds of meters in cities, up to tens ofkilometers in the countryside. The shape of cells are never perfect circles orhexagons but depend on the environment (buildings,mountains, valleys etc.), on weather conditions, and sometimes even on systemload. Typical systems using this approach are mobile telecommunication systems.

Figure 2.41
Cellular system with three and seven cell clusters



Advantages of cellular systems with small cells are the following:

● **Higher capacity**: Implementing SDM allows frequency reuse. If one transmitter is far away from another, i.e., outside the interference range, it can reuse the same frequencies. As most mobile phone systems assign frequencies to certain users (or certain hopping patterns), this frequency is

blocked for other users. But frequencies are a scarce resource and, the number of concurrent users per cell is very limited. Huge cells do not allow for more users. On the contrary, they are limited to less possible users per km2. This is also the reason for using very small cells in cities where many more people use mobile phones.

● **Less transmission power**: While power aspects are not a big problem for base stations, they are indeed problematic for mobile stations. A receiver far away from a base station would need much more transmit power than the current few Watts. But energy is a serious problem for mobile handheld devices.

● **Local interference only**: Having long distances between sender and receiver results in even more interference problems. With small cells, mobile stations and base stations only have to deal with 'local' interference.

● **Robustness**: Cellular systems are decentralized and so, more robust against the failure of single components. If one antenna fails, this only influences communication within a small area.

MEDIUM ACCESS CONTROL

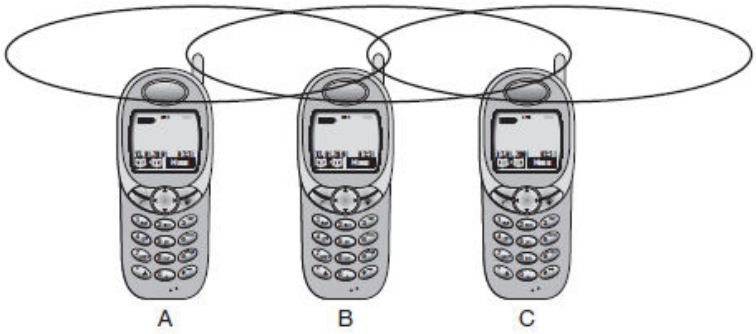
Motivation for a specialized MAC

The main question in connection with MAC in the wireless is whether it is possible to use elaborated MAC schemes from wired networks, for example, CSMA/CD as used in the original specification of IEEE 802.3 networks.

Consider carrier sense multiple access with collision detection, (CSMA/CD) which works as follows. A sender senses the medium (a wire or coaxial cable) to see if it is free. If the medium is busy, the sender waits until it is free. If the medium is free, the sender starts transmitting data and continues to listen into the medium. If the sender detects a collision while sending, it stops at once and sends a jamming signal.

This scheme fails in wireless networks. The strength of a signal decreases proportionally to the square of the distance to the sender. Obstacles attenuate the signal even further. The sender may now apply carrier sense and detect an idle medium. The sender starts sending – but a collision happens at the receiver due to a second sender. The same can happen to the collision detection. The sender detects no collision and assumes that the data has been transmitted without errors, but a collision might actually have destroyed the data at the receiver. Collision detection is very difficult in wireless scenarios as the transmission power in the area of the transmitting antenna is several magnitudes higher than the receiving power. So, this very common MAC scheme from wired network fails in a wireless scenario.

Hidden and exposed terminals

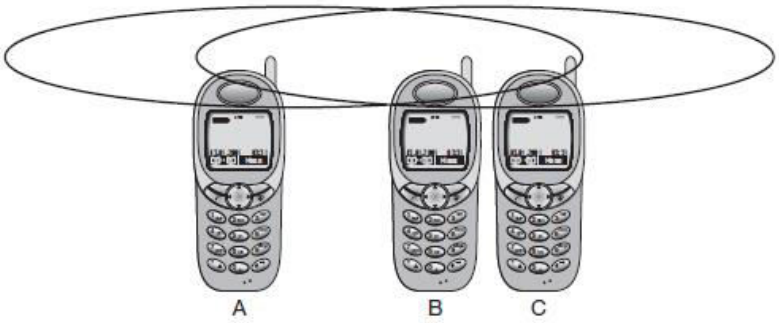


Consider the scenario with three mobile phones as shown in Figure. The transmission range of A reaches B, but not C (the detection range does not reach C either). The transmission range of C reaches B, but not A. Finally, the transmission range of B reaches A and C, i.e., A cannot detect C and vice versa.

A starts sending to B, C does not receive this transmission. C also wants to send something to B and senses the medium. The medium appears to be free, the carrier sense fails. C also starts sending causing a collision at B. But A cannot detect this collision at B and continues with its transmission. A is **hidden** for C and vice versa.

While hidden terminals may cause collisions, the next effect only causes unnecessary delay. Now consider the situation that B sends something to A and C wants to transmit data to some other mobile phone outside the interference ranges of A and B. C senses the carrier and detects that the carrier is busy (B's signal). C postpones its transmission until it detects the medium as being idle again. But as A is outside the interference range of C, waiting is not necessary. Causing a 'collision' at B does not matter because the collision is too weak to propagate to A. In this situation, C is **exposed** to B.

Near and far terminals



Consider the situation as shown in Figure. A and B are both sending with the same transmission power. As the signal strength decreases proportionally to the square of the distance, B's signal drowns out A's signal. As a result, C cannot receive A's transmission.

Now think of C as being an arbiter for sending rights (e.g., C acts as a base station coordinating media access). In this case, terminal B would already drown out terminal A on the physical layer. C in return would have no chance of applying a fair scheme as it would only hear B.

The **near/far effect** is a severe problem of wireless networks using CDM. All signals should arrive at the receiver with more or less the same strength. Otherwise a person standing closer to somebody could always speak louder than a person further away. Even if the senders were separated by code, the closest one would simply drown out the others. Precise power control is needed to receive all senders with the same strength at a receiver

SDMA

Space Division Multiple Access (SDMA) is used for allocating a separated space to users in wireless networks. A typical application involves assigning an optimal base station to a mobile phone user. The mobile phone may receive several base stations with different quality. The basis for the SDMA algorithm is formed by cells and sectorized antennas which constitute the infrastructure implementing space division multiplexing (SDM).

For wireless communication, multiplexing can be carried out in four dimensions: space, time, frequency, and code. In this field, the task of multiplexing is to assign space, time, frequency, and code to each communication channel with a minimum of interference and a maximum of medium utilization.

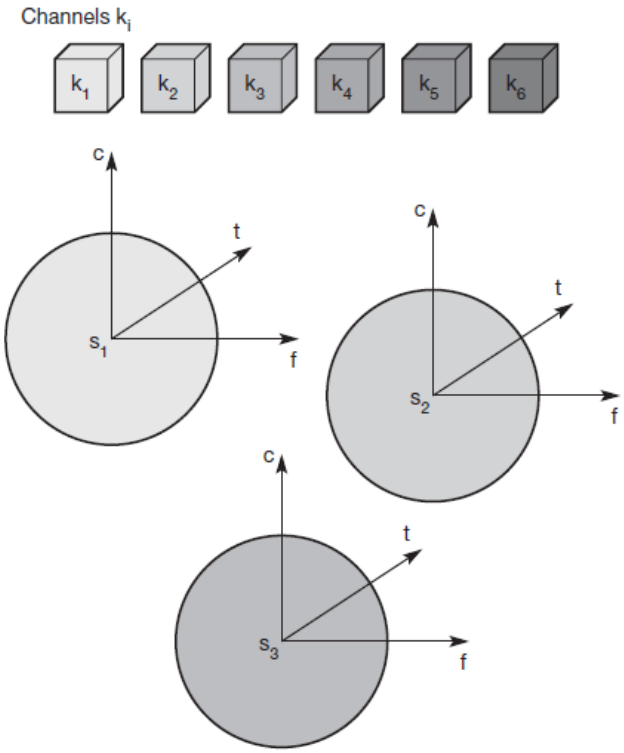


Figure shows six channels ki and introduces a three dimensional coordinate system. This system shows the dimensions of code c, time t and frequency f. For this first type of multiplexing,

space division multiplexing (SDM), the (three dimensional) space s_i is also shown. Here space is represented via circles indicating the interference range as introduced in Figure. The channels k1 to k3 can be mapped onto the three 'spaces' s1 to s3 which clearly separate the channels and prevent the interference ranges from overlapping. The space between the interference ranges is sometimes called **guard space**. Such a guard space is needed in all four multiplexing schemes presented.

For the remaining channels (k4 to k6) three additional spaces would be needed.

FDMA

Frequency division multiple access (FDMA) comprises all algorithms allocating frequencies to transmission channels according to the **frequency division multiplexing (FDM)**. Allocation can either be fixed (as for radio stations or the general planning and regulation of frequencies) or dynamic (i.e., demand driven).

Channels can be assigned to the same frequency at all times, i.e., pure FDMA, or change frequencies according to a certain pattern, i.e., FDMA combined with TDMA. However, this scheme also has disadvantages. While radio stations broadcast 24 hours a day, mobile

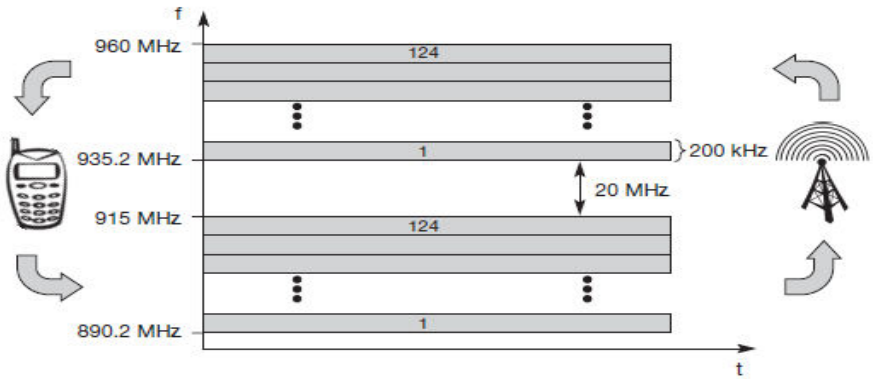


Figure 3.3
Frequency division multiplexing for multiple access and duplex

communication typically takes place for only a few minutes at a time. Assigning a separate frequency for each possible communication scenario would be a tremendous waste of (scarce) frequency resources. Additionally, the fixed assignment of a frequency to a sender makes the scheme very inflexible and limits the number of senders.

TDMA

Compared to FDMA, **time division multiple access (TDMA)** offers a much more flexible scheme, which comprises all technologies that allocate certain time slots for communication, i.e., controlling **TDM**. Now tuning in to a certain frequency is not necessary, i.e., the receiver can stay at the same frequency the whole time. Using only one frequency, and thus very simple receivers and transmitters, many different algorithms exist to control medium access. Listening to different frequencies at the same time is quite difficult, but listening to many channels separated in time at the same frequency is simple. Now synchronization between sender and receiver has to be achieved

in the time domain. Again this can be done by using a fixed pattern similar to FDMA techniques, i.e., allocating a certain time slot for a channel, or by using a dynamic allocation scheme.

It is too static, too inflexible for data communication. In this case, connectionless, demand-oriented TDMA schemes can be used.

Classical Aloha

TDMA comprises all mechanisms controlling medium access according to TDM. But what happens if TDM is applied without controlling access? This is exactly what the classical **Aloha** scheme does, a scheme which was invented at the University of Hawaii and was used in the ALOHANET for wireless connection of several stations. Aloha neither coordinates medium access nor does it resolve contention on the MAC layer. Instead, each station can access the medium at any time. This is a random access scheme, without a central arbiter controlling access and without coordination among the stations. If two or more stations access the medium at the same time, a **collision** occurs and the transmitted data is destroyed. Resolving this problem is left to higher layers (e.g., retransmission of data).

The simple Aloha works fine for a light load and does not require any complicated access mechanisms.

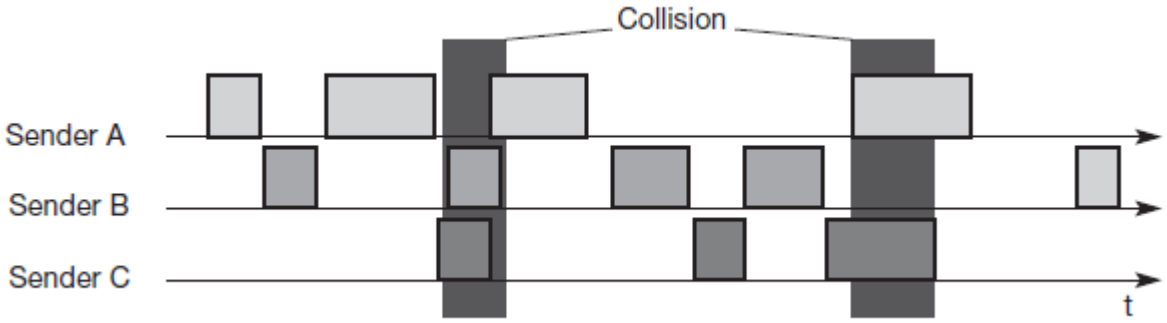


Fig: Classical Aloha multiple access

Slotted Aloha

The first refinement of the classical Aloha scheme is provided by the introduction of time slots (slotted Aloha). In this case, all senders have to be synchronized transmission can only start at the beginning of a time slot as shown in Figure. Still, access is not coordinated.

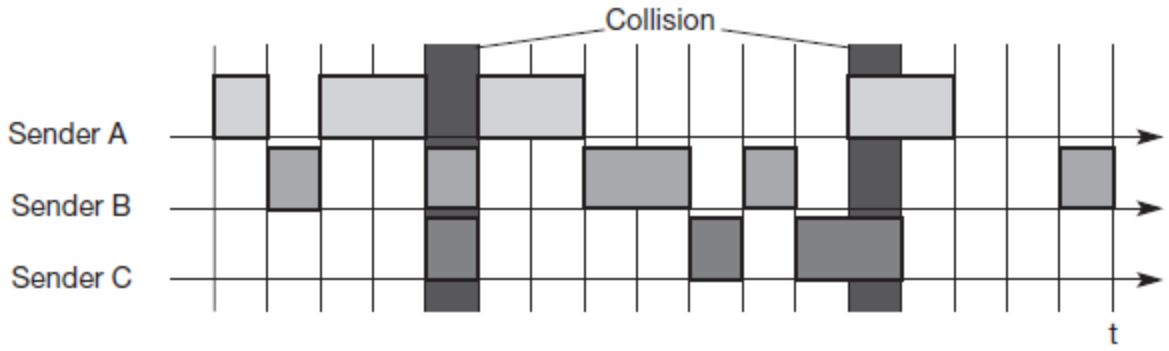


Fig: Slotted Aloha multiple access

Both basic Aloha principles occur in many systems that implement distributed access to a medium. Aloha systems work perfectly well under a light load (as most schemes do), but they cannot give any hard transmission guarantees, such as maximum delay before accessing the medium, or minimum throughput. Here one needs additional mechanisms, e.g., combining fixed schemes and Aloha schemes.

Carrier sense multiple access

One improvement to the basic Aloha is sensing the carrier before accessing the medium. This is what carrier sense multiple access (CSMA) schemes generally do. Sensing the carrier and accessing the medium only if the carrier is idle decreases the probability of a collision. Hidden terminals cannot be detected, so, if a hidden terminal transmits at the same time as another sender, a collision might occur at the receiver.

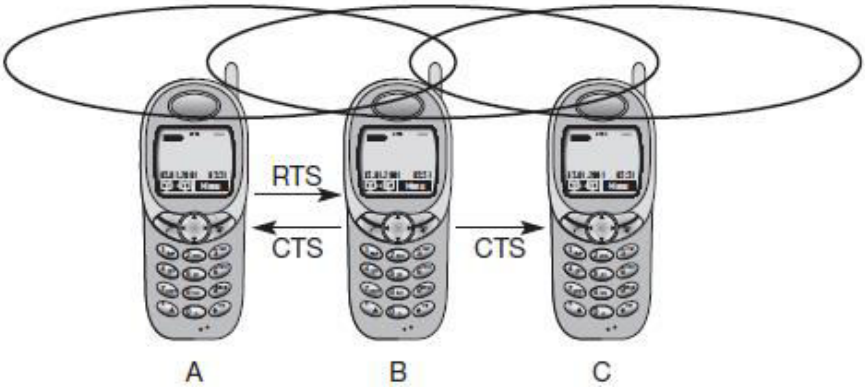
Several versions of CSMA exist. In **non-persistent CSMA**, stations sense the carrier and start sending immediately if the medium is idle. If the medium is busy, the station pauses a random amount of time before sensing the medium again and repeating this pattern. In **p-persistent CSMA** systems nodes also sense the medium, but only transmit with a probability of p, with the station deferring to the next slot with the probability 1-p, i.e., access is slotted in addition. In **1-persistent CSMA systems**, all stations wishing to transmit access the medium at the same time, as soon as it becomes idle. This will cause many collisions if many stations wish to send and block each other. To create some fairness for stations waiting for a longer time, back-off algorithms can be introduced, which are sensitive to waiting time as this is done for standard Ethernet.

Demand assigned multiple access

A general improvement of Aloha access systems can also be achieved by **reservation** mechanisms and combinations with some (fixed) TDM patterns. These schemes typically have a reservation period followed by a transmission period. During the reservation period, stations can reserve future slots in the transmission period. While, depending on the scheme, collisions may occur during the reservation period, the transmission period can then be accessed without collision. Alternatively, the transmission period can be split into periods with and without collision. In general, these schemes cause a higher delay under a light load (first the reservation has to take place), but allow higher throughput due to less collisions.

One basic scheme is **demand assigned multiple access (DAMA)** also called **reservation Aloha**, a scheme typical for satellite systems. DAMA has two modes. During a contention phase following the slotted Aloha scheme, all stations can try to reserve future slots. For example, different stations on earth try to reserve access time for satellite transmission. Collisions during the reservation phase do not destroy data transmission, but only the short requests for data transmission. If successful, a time slot in the future is reserved, and no other station is allowed to transmit during this slot. Therefore, the satellite collects all successful requests (the others are destroyed) and sends back a reservation list indicating access rights for future slots. All ground stations have to obey this list. To maintain the fixed TDM pattern of reservation and transmission, the stations have to be synchronized from time to time. DAMA is an **explicit reservation** scheme. Each transmission slot has to be reserved explicitly.

Multiple access with collision avoidance

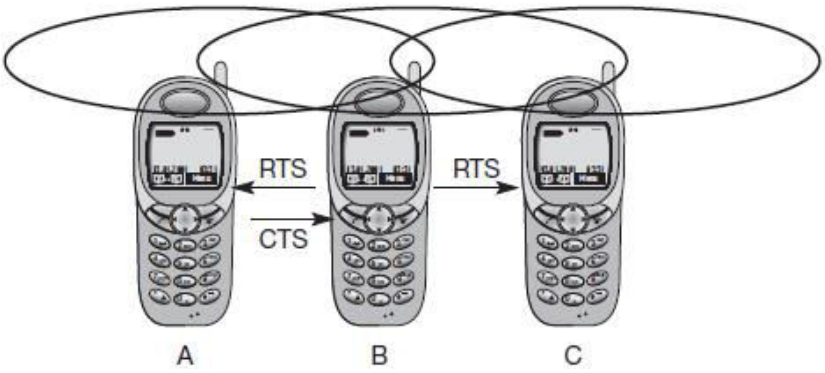


MACA can avoid hidden terminals

Multiple access with collision avoidance (MACA) presents a simple scheme that solves the hidden terminal problem, does not need a base station, and is still a random access Aloha scheme – but with dynamic reservation and C both want to send to B. A has already started the transmission, but is hidden for C, C also starts with its transmission, thereby causing a collision at B.

With MACA, A does not start its transmission at once, but sends a **request to send (RTS)** first. B receives the RTS that contains the name of sender and receiver, as well as the length of the future transmission. This RTS is not heard by C, but triggers an acknowledgement from B, called **clear to send (CTS)**. The CTS again contains the names of sender (A) and receiver (B) of the user data, and the length of the future transmission. This CTS is now heard by C and the medium for future use by A is now reserved for the duration of the transmission. After receiving a CTS, C is not allowed to send anything for the duration indicated in the CTS toward B. A collision cannot occur at B during data transmission, and the hidden terminal problem is solved – provided that the transmission conditions remain the same. (Another station could move into the transmission range of B after the transmission of CTS.)

Still, collisions can occur during the sending of an RTS. Both A and C could send an RTS that collides at B. RTS is very small compared to the data transmission, so the probability of a collision is much lower. B resolves this contention and acknowledges only one station in the CTS (if it was able to recover the RTS at all). No transmission is allowed without appropriate CTS.



MACA can avoid exposed terminals

MACA also help to solve the 'exposed terminal' problem. B wants to send data to A, C to someone else. But C is polite enough to sense the medium before transmitting, sensing a busy medium caused by the transmission from B. C defers, although C could never cause a collision at A.

With MACA, B has to transmit an RTS first containing the name of the receiver (A) and the sender (B). C does not react to this message as it is not the receiver, but A acknowledges using a CTS which identifies B as the sender and A as the receiver of the following data transmission. C does not receive these CTS and concludes that A is outside the detection range. C can start its transmission assuming it will not cause a collision at A. The problem with exposed terminals is solved without fixed access patterns or a base station.

Comparison of S/T/F/CDMA

The table shows the MAC schemes without combination with other schemes. However, in real systems, the MAC schemes always occur in combinations. A very typical combination is constituted by SDMA/TDMA/FDMA as used in IS-54, GSM, DECT, PHS, and PACS phone systems, or the Iridium and ICO satellite systems. CDMA together with SDMA is used in the IS-95 mobile phone system and the Globalstar satellite system.

Approach	SDMA	TDMA	FDMA	CDMA
Idea	Segment space into cells/sectors	Segment sending time into disjoint time-slots, demand driven or fixed patterns	Segment the frequency band into disjoint sub-bands	Spread the spectrum using orthogonal codes
Terminals	Only one terminal can be active in one cell/one sector	All terminals are active for short periods of time on the same frequency	Every terminal has its own frequency, uninterrupted	All terminals can be active at the same place at the same moment, uninterrupted
Signal separation	Cell structure directed antennas	Synchronization in the time domain	Filtering in the frequency domain	Code plus special receivers
Advantages	Very simple, increases capacity per km^2	Established, fully digital, very flexible	Simple, established, robust	Flexible, less planning needed, soft handover
Disadvantages	Inflexible, antennas typically fixed	Guard space needed (multi-path propagation), synchronization difficult	Inflexible, frequencies are a scarce resource	Complex receivers, needs more complicated power control for senders
Comment	Only in combination with TDMA, FDMA or CDMA useful	Standard in fixed networks, together with FDMA/SDMA used in many mobile networks	Typically combined with TDMA (frequency hopping patterns) and SDMA (frequency reuse)	Used in many 3G systems, higher complexity, lowered expectations; integrated with TDMA/FDMA

UNIT II: Telecommunication system

GSM

GSM is the most successful digital mobile telecommunication system in the world today. It is used by over 800 million people in more than 190 countries. In the early 1980s, Europe had numerous coexisting analog mobile phone systems, which were often based on similar standards (e.g., NMT 450), but ran on slightly different carrier frequencies. To avoid this situation for a second generation fully digital system, the group special mobile (GSM) was founded in 1982. This system was soon named the global system for mobile communications (GSM), with the specification process lying in the hands of ETSI (ETSI, 2002), (GSM Association, 2002).

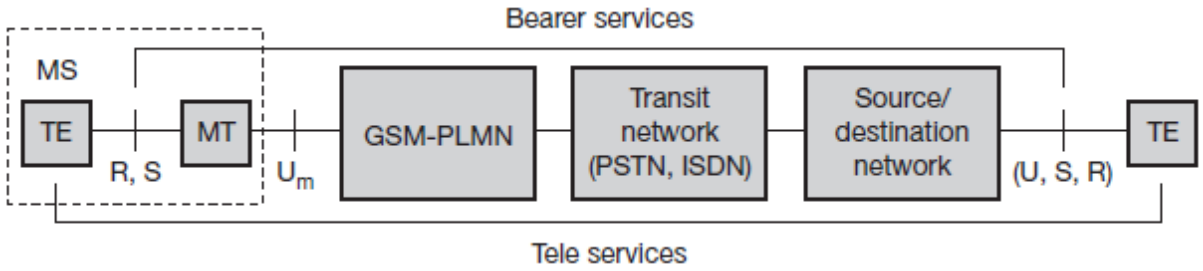
GSM Mobile services

GSM permits the integration of different voice and data services and the interworking with existing networks. Services make a network interesting for customers.

GSM has defined three different categories of services:

- Bearer services
- Tele services
- Supplementary services.

Figure shows a reference model for GSM services.



Bearer and Tele services reference model

A mobile station MS is connected to the GSM public land mobile network (PLMN) via the Um interface. (GSM-PLMN is the infrastructure needed for the GSM network.) This network is connected to transit networks, e.g., integrated services digital network (ISDN) or traditional public switched telephone network (PSTN). There might be an additional network, the source/destination network, before another terminal TE is connected. Bearer services now comprise all services that enable the transparent transmission of data between the interfaces to the network, i.e., S in case of the mobile station, and a similar interface for the other terminal (e.g., S0 for ISDN terminals). Interfaces like U, S, and R in case of ISDN have not been defined for all networks, so it depends on the specific network which interface is used as a reference for the transparent transmission of data. In the classical GSM model, bearer services are connection-oriented and circuit- or packet-switched. These services only need the lower three layers of the ISO/OSI reference model.

Within the mobile station MS, the mobile termination (MT) performs all network specific tasks (TDMA, FDMA, coding etc.) and offers an interface for data transmission (S) to the terminal TE which can then be network independent. Depending on the capabilities of TE, further interfaces

may be needed, such as R, according to the ISDN reference model (Halsall, 1996). Tele services are application specific and may thus need all seven layers of the ISO/OSI reference model. These services are specified end-to-end, i.e., from one terminal TE to another.

Bearer services

GSM specifies different mechanisms for data transmission, the original GSM allowing for data rates of up to 9600 bit/s for non-voice services. Bearer services permit transparent and non-transparent, synchronous or asynchronous data transmission. **Transparent bearer services** only use the functions of the physical layer (layer 1) to transmit data. Data transmission has a constant delay and throughput if no transmission errors occur. The only mechanism to increase transmission quality is the use of **forward error correction (FEC)**, which codes redundancy into the data stream and helps to reconstruct the original data in case of transmission errors.

Non-transparent bearer services use protocols of layers two and three to implement error correction and flow control. These services use the transparent bearer services, adding a **radio link protocol (RLP)**. This protocol comprises mechanisms of **high-level data link control (HDLC)**, (Halsall, 1996) and special selective-reject mechanisms to trigger retransmission of erroneous data.

Tele services

GSM mainly focuses on voice-oriented tele services. These comprise encrypted voice transmission, message services, and basic data communication with terminals as known from the PSTN or ISDN (e.g., fax).

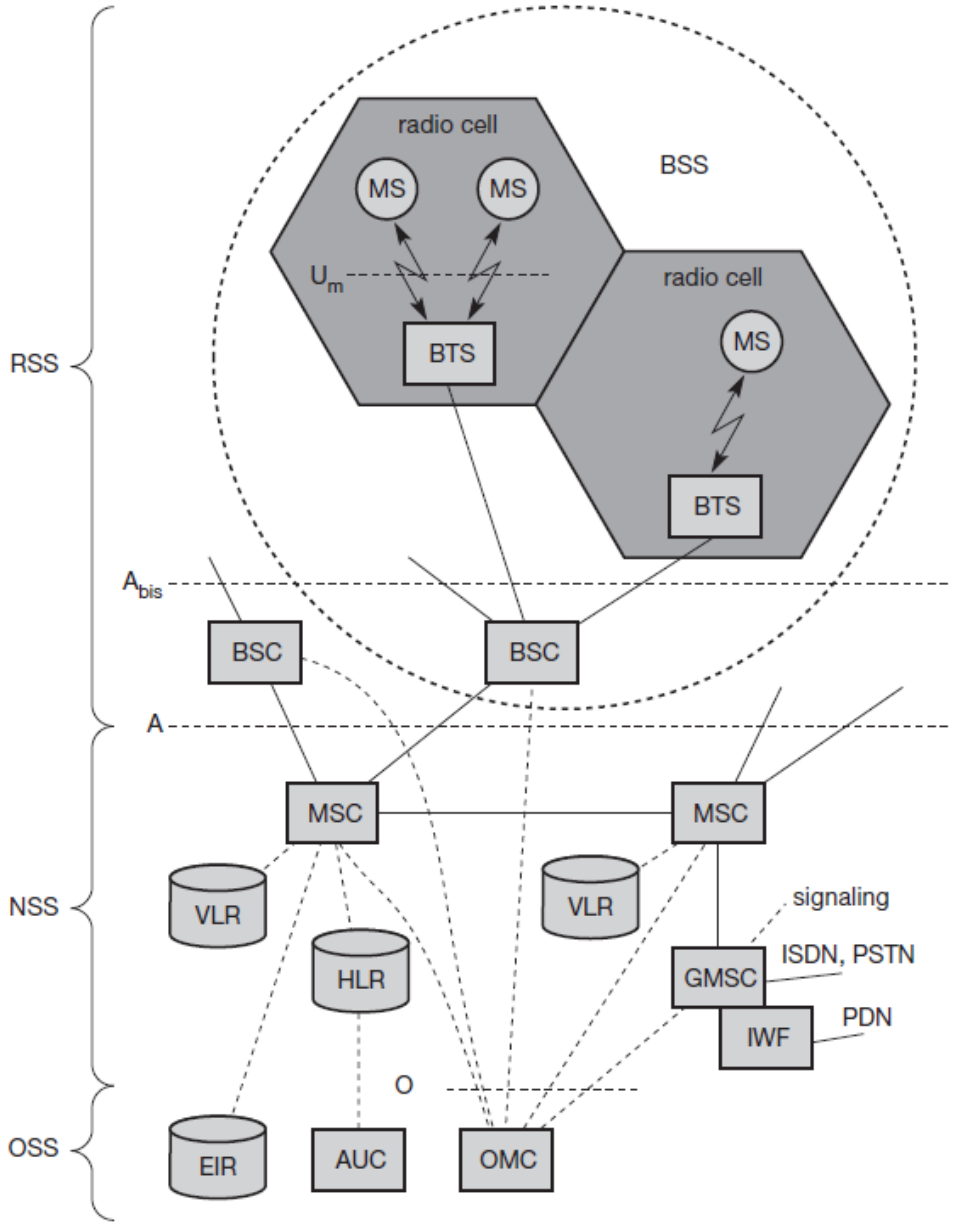
Another service offered by GSM is the **emergency number**. The same number can be used throughout country. This service is mandatory for all providers and free of charge. This connection also has the highest priority, possibly pre-empting other connections, and will automatically be set up with the closest emergency center.

A useful service for very simple message transfer is the **short message service (SMS)**, which offers transmission of messages of up to 160 characters.

Supplementary services

In addition to tele and bearer services, GSM providers can offer **supplementary services**. Similar to ISDN networks, these services offer various enhancements for the standard telephony service, and may vary from provider to provider. Typical services are user **identification**, call **redirection**, or **forwarding** of ongoing calls. Standard ISDN features such as **closed user groups** and **multiparty** communication may be available. Closed user groups are of special interest to companies because they allow, for example, a company-specific GSM sub-network, to which only members of the group have access

GSM ARCHITECTURE



A GSM system consists of three subsystems

- Radio sub system (RSS),
- Network and switching subsystem (NSS)
- Operation subsystem (OSS).

Radio subsystem

As the name implies, the radio subsystem (RSS) comprises all radio specific entities, i.e., the mobile stations (MS) and the base station subsystem (BSS). Figure 4.4 shows the connection

between the RSS and the NSS via the A interface (solid lines) and the connection to the OSS via the O interface (dashed lines).

- **Base station subsystem (BSS)**: A GSM network comprises many BSSs, each controlled by a base station controller (BSC). The BSS performs all functions necessary to maintain radio connections to an MS, coding/decoding of voice, and rate adaptation to/from the wireless network part. Besides a BSC, the BSS contains several BTSs.
- **Base transceiver station (BTS)**: A BTS comprises all radio equipment, i.e., antennas, signal processing, amplifiers necessary for radio transmission. A BTS can form a radio cell or, using sectorized antennas, several cells, and is connected to MS via the **Um interface** (ISDN U interface for mobile use), and to the BSC via the **Abis interface**. The Um interface contains all the mechanisms necessary for wireless transmission (TDMA, FDMA etc.) and will be discussed in more detail below. The Abis interface consists of 16 or 64 kbit/s connections. A GSM cell can measure between some 100 m and 35 km depending on the environment (buildings, open space, mountains etc.) but also expected traffic.
- **Base station controller (BSC)**: The BSC basically manages the BTSs. It reserves radio frequencies, handles the handover from one BTS to another within the BSS, and performs paging of the MS. The BSC also multiplexes the radio channels onto the fixed network connections at the A interface.
- **Mobile station (MS)**: The MS comprises all user equipment and software needed for communication with a GSM network. An MS consists of user independent hard- and software and of the **subscriber identity module (SIM)**, which stores all user-specific data that is relevant to GSM.3 While an MS can be identified via the **international mobile equipment identity (IMEI)**, a user can personalize any MS using his or her SIM, i.e., user-specific mechanisms like charging and authentication are based on the SIM, not on the device itself. Device-specific mechanisms, e.g., theft protection, use the device specific IMEI. Without the SIM, only emergency calls are possible. The SIM card contains many identifiers and tables, such as card-type, serial number, a list of subscribed services, a **personal identity number (PIN)**, a **PIN unblocking key (PUK)**, an **authentication key Ki**, and the **international mobile subscriber identity (IMSI)** (ETSI, 1991c). The PIN is used to unlock the MS. Using the wrong PIN three times will lock the SIM. In such cases, the PUK is needed to unlock the SIM.

Network and switching subsystem

The "heart" of the GSM system is formed by the **network and switching subsystem (NSS)**. The NSS connects the wireless network with standard public networks, performs handovers between different BSSs, comprises functions for worldwide localization of users and supports charging, accounting, and roaming of users between different providers in different countries. The NSS consists of the following switches and databases:

- **Mobile services switching center (MSC)**: MSCs are high-performance digital ISDN switches. They set up connections to other MSCs and to the BSCs via the A interface, and

form the fixed backbone network of a GSM system. Typically, an MSC manages several BSCs in a geographical region. A **gateway MSC (GMSC)** has additional connections to other fixed networks, such as **PSTN** and **ISDN**. Using additional **interworking functions (IWF)**, an MSC can also connect to **public data networks (PDN)** such as X.25. An MSC handles all signaling needed for connection setup, connection release and handover of connections to other. An MSC also performs all functions needed for supplementary services such as call forwarding, multi-party calls, reverse charging etc.

- **Home location register (HLR)**: The HLR is the most important database in a GSM system as it stores all user-relevant information. This comprises static information, such as the **mobile subscriber ISDN number (MSISDN)**, subscribed services (e.g., call forwarding, roaming restrictions, GPRS), and the **international mobile subscriber identity (IMSI)**. Dynamic information is also needed, e.g., the current **location area (LA)** of the MS, the **mobile subscriber roaming number (MSRN)**, the current VLR and MSC. As soon as an MS leaves its current LA, the information in the HLR is updated. This information is necessary to localize a user in the worldwide GSM network. All these user-specific information elements only exist once for each user in a single HLR, which also supports charging and accounting.
- **Visitor location register (VLR)**: The VLR associated to each MSC is a dynamic database which stores all important information needed for the MS users currently in the LA that is associated to the MSC (e.g., IMSI, MSISDN, HLR address). If a new MS comes into an LA the VLR is responsible for, it copies all relevant information for this user from the HLR. This hierarchy of VLR and HLR avoids frequent HLR updates and long-distance signaling of user information.

Operation subsystem

The third part of a GSM system, the operation subsystem (OSS), contains the necessary functions for network operation and maintenance.

- **Operation and maintenance center (OMC)**: The OMC monitors and controls all other network entities via the O interface. Typical OMC management functions are traffic monitoring, status reports of network entities, subscriber and security management, or accounting and billing.
- **Authentication centre (AuC)**: As the radio interface and mobile stations are particularly vulnerable, a separate AuC has been defined to protect user identity and data transmission. The AuC contains the algorithms for authentication as well as the keys for encryption and generates the values needed for user authentication in the HLR. The AuC may, in fact, be situated in a special protected part of the HLR.
- **Equipment identity register (EIR)**: The EIR is a database for all IMEIs, i.e., it stores all device identifications registered for this network. As MSs are mobile, they can be easily stolen. With a valid SIM, anyone could use the stolen MS. The EIR has a blacklist of stolen (or locked) devices. In theory an MS is useless as soon as the owner has reported a theft.

Unfortunately, the blacklists of different providers are not usually synchronized and the illegal use of a device in another operator's network is possible

Protocols

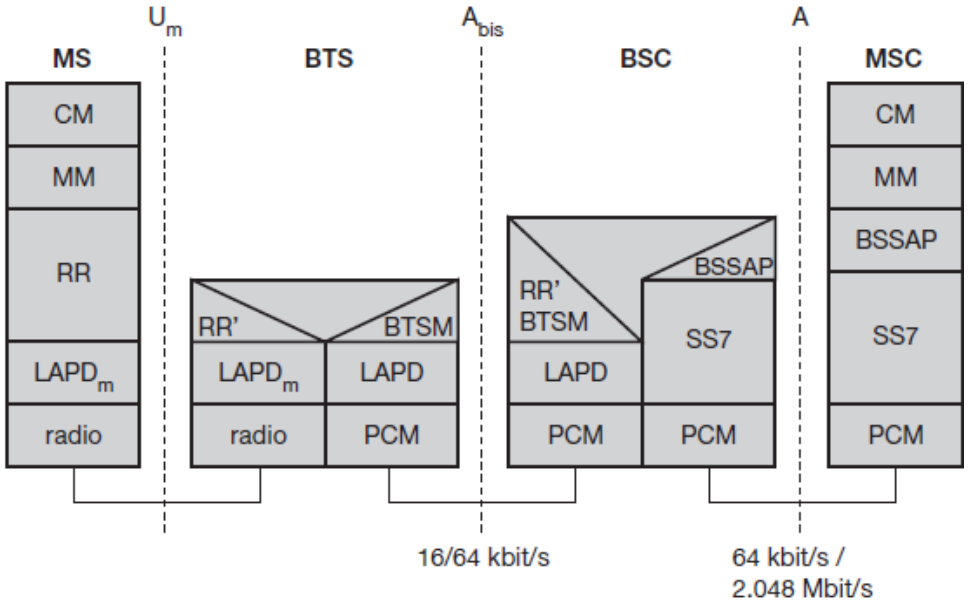


Figure shows the protocol architecture of GSM with signaling protocols, interfaces, as well as the entities. The main interest lies in the U_m interface, as the other interfaces occur between entities in a fixed network.

Layer 1, the physical layer, handles all radio-specific functions. This includes the creation of bursts according to the five different formats, multiplexing of bursts into a TDMA frame, synchronization with the BTS, detection of idle channels, and measurement of the channel quality on the downlink. The physical layer at Um uses GMSK for digital modulation and performs encryption/decryption of data, i.e., encryption is not performed end-to-end, but only between MS and BSS over the air interface. The main tasks of the physical layer comprise channel coding and error detection/correction.

Signaling between entities in a GSM network requires higher layers. For this purpose, the $LAPD_m$ protocol has been defined at the U_m interface for layer two. $LAPD_m$, as the name already implies, has been derived from link access procedure for the D-channel (LAPD) in ISDN systems,. $LAPD_m$ is a lightweight LAPD because it does not need synchronization flags or check summing for error detection. $LAPD_m$ offers reliable data transfer over connections, re-sequencing of data frames, and flow control. As there is no buffering between layer one and two, $LAPD_m$ has to obey the frame structures, recurrence patterns etc., defined for the Um interface. Further services provided by $LAPD_m$ include segmentation and reassembly of data and acknowledged/unacknowledged data transfer.

The network layer in GSM, **layer three**, comprises several sub layers. The lowest sub layer is the **radio resource management (RR)**.Only a part of this layer, **RR**', is implemented in the BTS, the remainder is situated in the BSC. The functions of RR' are supported by the BSC via the **BTS management (BTSM)**. The main tasks of RR are setup, maintenance, and release of radio channels. RR also directly accesses the physical layer for radio information and offers a reliable connection to the next higher layer.

Mobility management (MM) contains functions for registration, authentication, identification, location updating, and the provision of a temporary mobile subscriber identity (TMSI) that replaces the international mobile subscriber identity (IMSI) and which hides the real identity of an MS user over the air interface. While the IMSI identifies a user, the TMSI is valid only in the current location area of a VLR. MM offers a reliable connection to the next higher layer.

Finally, the call management (CM) layer contains three entities: call control (CC), short message service (SMS), and supplementary service (SS). CC provides a point-to-point connection between two terminals and is used by higher layers for call establishment, call clearing and change of call parameters.

Additional protocols are used at the A_{bis} and A interfaces. Data transmission at the physical layer typically uses pulse code modulation (PCM) systems.

Signaling system No. 7 (SS7) is used for signaling between an MSC and a BSC. This protocol also transfers all management information between MSCs, HLR, VLRs, AuC, EIR, and OMC. An MSC can also control a BSS via a **BSS application part (BSSAP)**.

DECT

Another fully digital cellular network is the digital enhanced cordless telecommunications (DECT) system specified by ETSI (2002, 1998j, k), (DECT Forum, 2002). Formerly also called digital European cordless telephone and digital European cordless telecommunications, DECT replaces older analog cordless phone systemssuch as CT1 and CT1+. These analog systems only ensured security to a limited extent as they did not use encryption for data transmission and only offered a relatively low capacity. DECT is also a more powerful alternative to the digital system CT2, which is mainly used in the UK (the DECT standard works throughout Europe), and has even been selected as one of the 3G candidates in the IMT-2000 family. DECT is mainly used in offices, on campus, at trade shows, or in the home. Furthermore, access points to the PSTN can be established within, e.g., railway stations, large government buildings and hospitals, offering a much cheaper telephone service compared to a GSM system. DECT could also be used to bridge the last few hundred meters between a new network operator and customers. Using this 'small range' local loop, new companies can offer their service without having their own lines installed in the streets. DECT systems offer many different interworking units, e.g., with GSM, ISDN, or data networks. Currently, over 100 million DECT units are in use (DECT, 2002). A big difference between DECT and GSM exists in terms of cell diameter and cell capacity. While GSM is designed for outdoor use with a cell diameter of up to 70 km, the range of DECT is limited to about 300 m from the base

station (only around 50 m are feasible inside buildings depending on the walls). Due to this limited range and additional multiplexing techniques, DECT can offer its service to some 10,000 people within one km2. This is a typical scenario within a big city, where thousands of offices are located in skyscrapers close together. DECT also uses base stations, but these base stations together with a mobile station are in a price range of €100 compared to several €10,000 for a GSM base station. GSM base stations can typically not be used by individuals for private networks. One reason is licensing as all GSM frequencies have been licensed to network operators. DECT can also handle handover, but it was not designed to work at a higher speed (e.g., up to 250 km/h like GSM systems). Devices handling GSM and DECT exist but have never been a commercial success. DECT works at a frequency range of 1880–1990 MHz offering 120 full duplex channels. Time division duplex (TDD) is applied using 10 ms frames. The frequency range is subdivided into 10 carrier frequencies using FDMA, each frame being divided into 24 slots using TDMA. For the TDD mechanism 12 slots are used as uplink, 12 slots as downlink (see Figure 3.4). The digital modulation scheme is GMSK – each station has an average transmission power of only 10 mW with a maximum of 250 mW.

System architecture

A DECT system, may have various different physical implementation depending on its actual use. Different DECT entities can be integrated into one physical unit; entities can be distributed, replicated etc. However, all implementations are based on the same logical reference model of the system architecture A **global network** connects the local communication structure to the outside world and offers its services via the interface D1. Global networks could be integrated services digital networks (ISDN), public switched telephone networks (PSTN), public land mobile networks (PLMN), e.g., GSM, or packet switched public data network (PSPDN). The services offered by these networks include transportation of data and the translation of addresses and routing of data between the local networks.

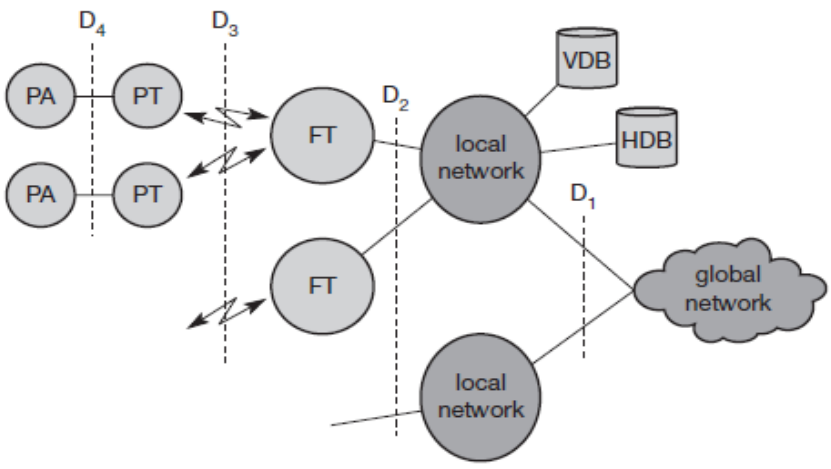


Figure 4.18
DECT system architecture reference model

Local networks in the DECT context offer local telecommunication services that can include everything from simple switching to intelligent call forwarding, address translation etc. Examples for such networks are analog or digital private branch exchanges (PBXs) or LANs, e.g., those following the IEEE 802.x family of LANs.

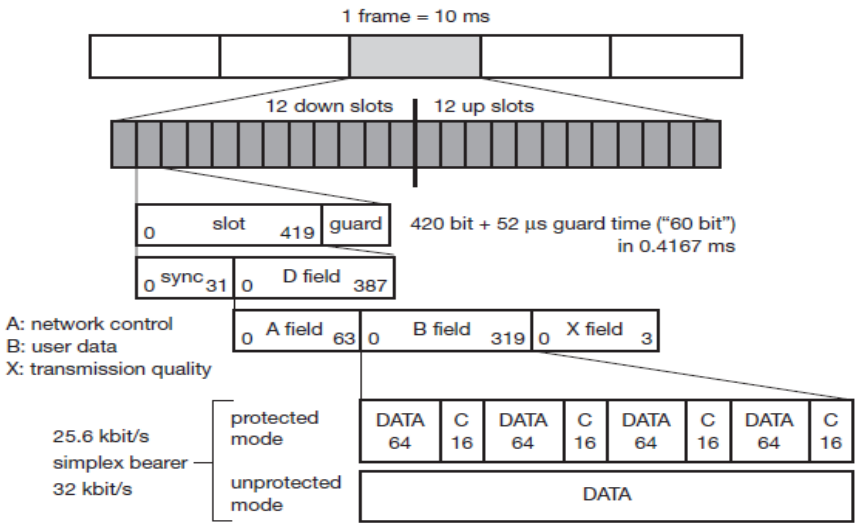


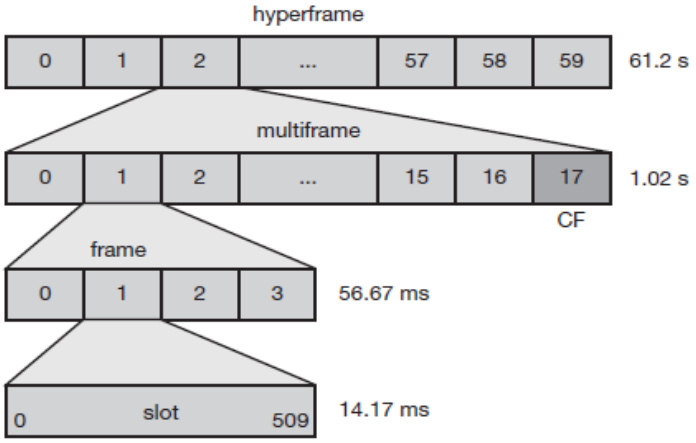
Figure 4.20
DECT multiplex and frame structure

As the core of the DECT system itself is quite simple, all typical network functions have to be integrated in the local or global network, where the databases **home data base (HDB)** and **visitor data base (VDB)** are also located. Both databases support mobility with functions that are similar to those in the HLR and VLR in GSM systems. Incoming calls are automatically forwarded to the current subsystem responsible for the DECT user, and the current VDB informs the HDB about changes in location. The DECT core network consists of the **fixed radio termination (FT)** and the **portable radio termination (PT)**, and basically only provides a multiplexing service. FT and PT cover layers one to three at the fixed network side and
mobile network side respectively. Additionally, several portable applications (PA) can be implemented on a device.

TETRA

Trunked radio systems constitute another method of wireless data transmission. These systems use many different radio carriers but only assign a specific carrier to a certain user for a short period of time according to demand. While, for example, taxi services, transport companies with fleet management systems and rescue teams all have their own unique carrier frequency in traditional systems, they can share a whole group of frequencies in trunked radio systems for better frequencyreuse via FDM and TDM techniques. These types of radio systems typically offer interfaces to the fixed telephone network, i.e., voice and data services, but are not publicly accessible. These systems are not only simpler than most other networks, they are also reliable and relatively cheap to set up and operate, as they only have to cover the region where the local users

Figure 4.21
TETRA frame structure



operate, e.g., a city taxi service. To allow a common system throughout Europe, ETSI standardized the **TETRA** system (**terrestrial trunked radio**)9 in 1991 (ETSI, 2002), (TETRA MoU, 2002). This system should replace national systems, such as MODACOM, MOBITEX and COGNITO in Europe that typically connect to an X.25 packet network. (An example system from the US is ARDIS.) TETRA offers two standards: the **Voice+Data (V+D)** service (ETSI, 1998l) and the **packet data optimized (PDO)** service (ETSI, 1998m). While V+D offers circuit-switched voice and data transmission, PDO only offers packet data transmission, either connection-oriented to connect to X.25 or connectionless for the ISO CLNS (connectionless network service). The latter service can be point-to-point or point-to-multipoint, the typical delay for a short message (128 byte) being less than 100 ms. V+D connection modes comprise unicast and broadcast connections, group communication within a certain protected group, and a direct ad hoc mode without a base station. However, delays for short messages can be up to 500 ms or higher depending on the priority. TETRA also offers bearer services of up to 28.8 kbit/s for unprotected data transmission and 9.6 kbit/s for protected transmission. Examples for end-to-end services are call forwarding, call barring, identification, call hold, call priorities, emergency calls and group joins. The system architecture of TETRA is very similar to GSM. Via the radio interface Um, the **mobile station (MS)** connects to the **switching and management infrastructure (SwMI)**, which contains the user data bases (HDB, VDB), the base station, and interfaces to PSTN, ISDN, or PDN. The system itself, however, is much simpler in real implementation compared to GSM, as typically no handover is needed. Taxis usually remain within a certain area which can be covered by one TETRA cell. Several frequencies have been specified for TETRA which uses FDD (e.g., 380–390 MHz uplink/390–400 MHz downlink, 410–420 MHz uplink/420–430 MHz downlink). Each channel has a bandwidth of 25 kHz and can carry

36 kbit/s. Modulation is DQPSK. While V+D uses up to four TDMA voice or data channels per carrier, PDO performs statistical multiplexing. For accessing a channel, slotted Aloha is used. typical **TDMA frame structure** of TETRA. Each **frame** consists of four slots (four channels in the V+D service per carrier), with a frame duration of 56.67 ms. Each **slot** carries 510 bits within 14.17 ms, i.e., 36 kbit/s. 16 frames together with one **control frame** (CF) form a **multiframe**, and finally,

a **hyperframe** contains 60 multiframes. To avoid sending and receiving at the same time, TETRA shifts the uplink for a period of two slots compared to the downlink. TETRA offers **traffic channels (TCH)** and **control channels (CCH)** similar to GSM. Typical TCHs are TCH/S for voice transmission, and TCH/7.2, TCH/4.8, TCH/2.4 for data transmission (depending on the FEC mechanisms required). However, in contrast to GSM, TETRA offers additional services like group call, acknowledged group call, broadcast call, and discreet listening. Emergency services need a sub-second group-call setup in harsh environments which possibly lack all infrastructure. These features are currently not available in GSM or other typical mobile telephone networks, so TETRA is complementary to other systems. TETRA has been chosen by many government organizations in Europe and China.

Localization and calling

One fundamental feature of the GSM system is the automatic, worldwide localization of users. The system always knows where a user currently is, and the same phone number is valid worldwide. To provide this service, GSM performs periodic location updates even if a user does not use the mobile station (provided that the MS is still logged into the GSM network and is not completely switched off).

The HLR always contains information about the current location (only the location area, not the precise geographical location), and the VLR currently responsible for the MS informs the HLR about location changes. As soon as an MS moves into the range of a new VLR (a new location area), the HLR sends all user data needed to the new VLR. **Changing VLRs with uninterrupted availability of all services is also called roaming**. Roaming can take place within the network of one provider, between two providers in one country (national roaming is, often not supported due to competition between operators), but also between different providers in different countries (international roaming).

To locate an MS and to address the MS, several numbers are needed:

- **Mobile station international ISDN number (MSISDN):** The only important number for a user of GSM is the phone number. The phone number is not associated with a certain device but with the SIM, which is personalized for a user. This number consists of the **country code (CC)** (e.g., +49 179 1234567 with 49 for Germany), the **national destination code (NDC)** (i.e., the address of the network provider, e.g., 179), and the **subscriber number (SN)**.
- **International mobile subscriber identity (IMSI):** GSM uses the IMSI for internal unique identification of a subscriber. IMSI consists of a mobile country code (MCC) (e.g., 240 for Sweden, 208 for France), the mobile network code (MNC) (i.e., the code of the network provider), and finally the mobile subscriber identification number (MSIN).
- **Temporary mobile subscriber identity (TMSI):** To hide the IMSI, which would give away the exact identity of the user signaling over the air interface, GSM uses the 4 byte TMSI for local subscriber identification. TMSI is selected by the current

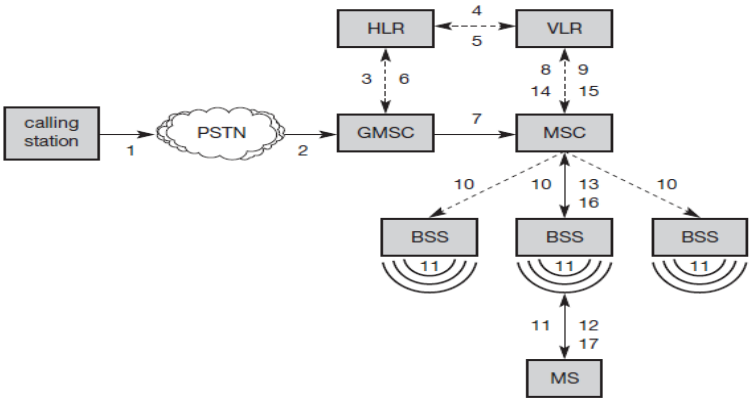
VLR and is only valid temporarily and within the location area of the VLR. Additionally, a VLR may change the TMSI periodically.

- **Mobile station roaming number (MSRN):** Another temporary address that hides the identity and location of a subscriber is MSRN. The VLR generates this address on request from the MSC, and the address is also stored in the HLR. MSRN contains the current **visitor country code (VCC)**, the **visitor national destination code (VNDC)**, the identification of the current MSC together with the subscriber number. The MSRN helps the HLR to find a subscriber for an incoming call.

Calling in GSM I of two types:

- Mobile terminated call (MTC)
- Mobile originated call (MOC)

Mobile terminated call (MTC)



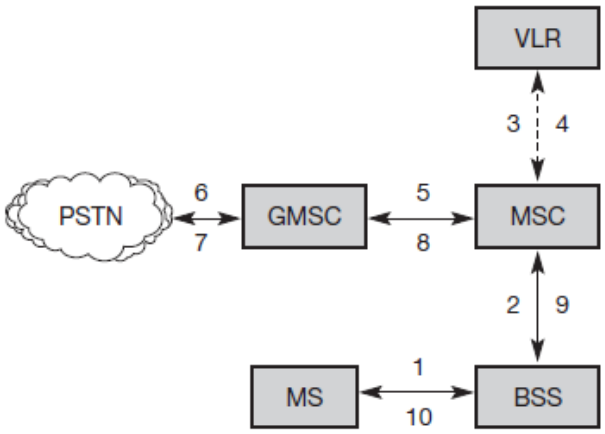
A situation in which a station calls a mobile station (the calling station could be outside the GSM network or another mobile station). Figure shows the basic steps needed to connect the calling station with the mobile user.

In step 1, a user dials the phone number of a GSM subscriber. The fixed network (PSTN) notices (looking at the destination code) that the number belongs to a user in the GSM network and forwards the call setup to the Gateway MSC (2). The GMSC identifies the HLR for the subscriber (which is coded in the phone number) and signals the call setup to the HLR (3). The HLR now checks whether the number exists and whether the user has subscribed to the requested services, and requests an MSRN from the current VLR (4). After receiving the MSRN (5), the HLR can determine the MSC responsible for the MS and forwards this information to the GMSC (6). The GMSC can now forward the call setup request to the MSC indicated (7).

From this point on, the MSC is responsible for all further steps. First, it requests the current status of the MS from the VLR (8). If the MS is available, the MSC initiates paging in all cells it is responsible for (10), as searching for the right cell would be too time consuming (but this approach puts some load on the signaling channels so optimizations

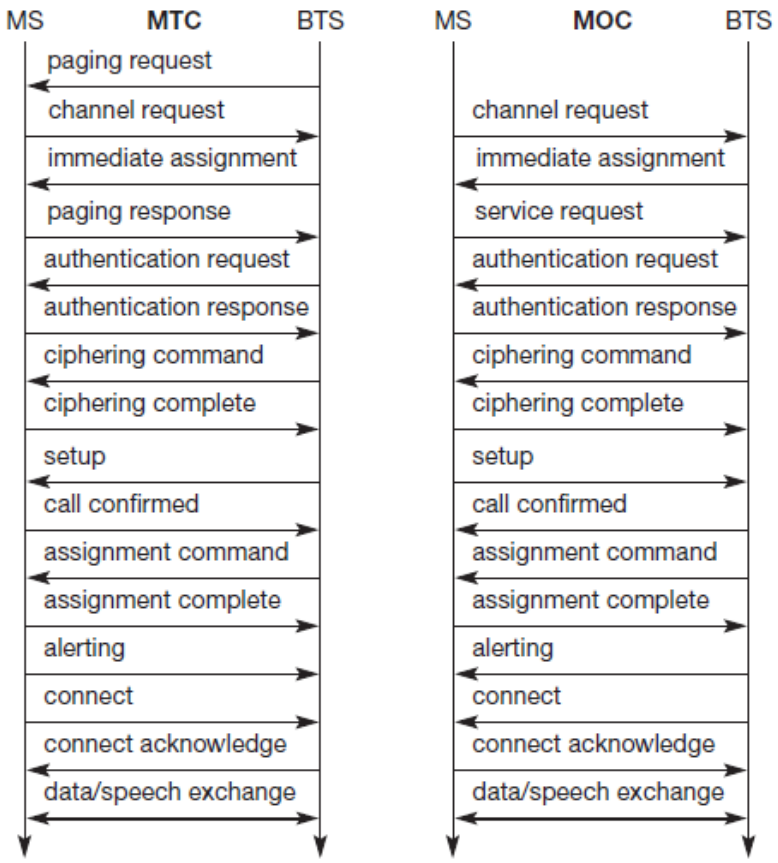
exist). The BTSs of all BSSs transmit this paging signal to the MS (11). If the MS answers (12 and 13), the VLR has to perform security checks (set up encryption etc.). The VLR then signals to the MSC to set up a connection to the MS (steps 15 to 17).

Mobile originated call (MOC)



It is much simpler to perform a **mobile originated call (MOC)** compared to a MTC. The MS transmits a request for a new connection (1), the BSS forwards this request to the MSC (2). The MSC then checks if this user is allowed to set up a call with the requested service (3 and 4) and checks the availability of resources through the GSM network and into the PSTN. If all resources are available, the MSC sets up a connection between the MS and the fixed network.

In addition to the steps mentioned above, other messages are exchanged between an MS and BTS during onnection setup (in either direction).



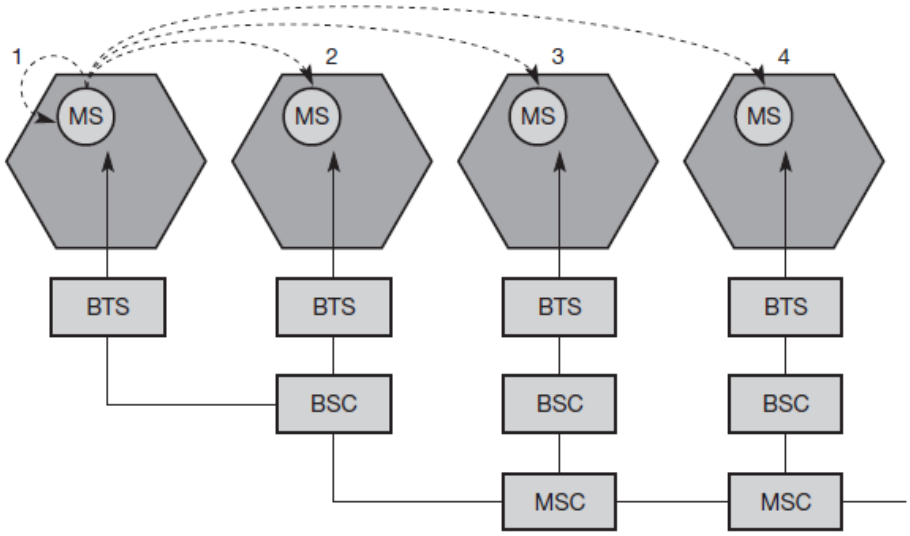
Message flow for MTC and MOC

Handover

Cellular systems require handover procedures, as single cells do not cover the whole service area, but, e.g., only up to 35 km around each antenna on the countryside and some hundred meters in cities. The smaller the cell size and the faster the movement of a mobile station through the cells (up to 250 km/h for GSM), the more handovers of ongoing calls are required. However, a handover should not cause a cut-off, also called call drop. GSM aims at maximum handover duration of 60ms.

There are two basic reasons for a handover:

- ✓ The mobile station **moves out of the range** of a BTS or a certain antenna of a BTS respectively. The received **signal level** decreases continuously until it falls below the minimal requirements for communication. The **error rate** may grow due to interference, the distance to the BTS may be too high (max. 35 km) etc. – all these effects may diminish the **quality of the radio link** and make radio transmission impossible in the near future.
- ✓ The wired infrastructure (MSC, BSC) may decide that the **traffic in one cell is too high** and shift some MS to other cells with a lower load (if possible). Handover may be due to **load balancing**.



Types of handover in GSM

There are 4 scenarios of handover in GSM:

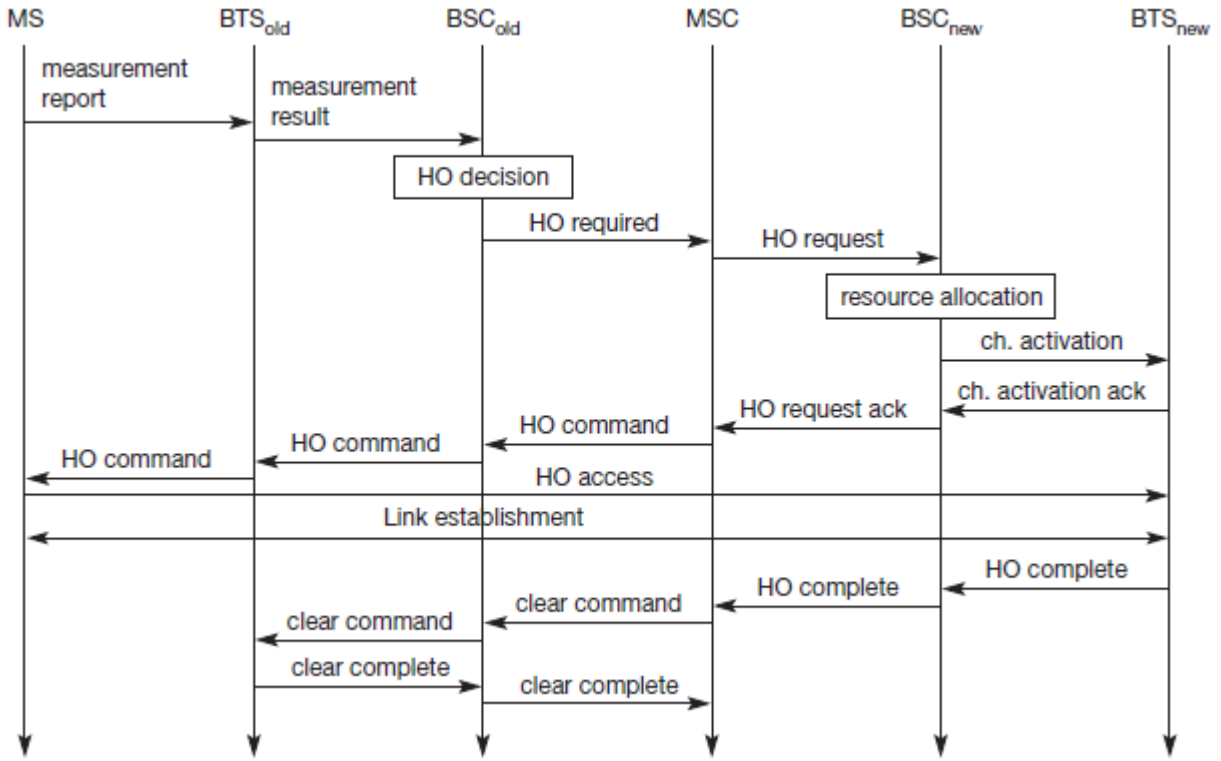
- **Intra-cell handover:** Within a cell, narrow-band interference could make transmission at a certain frequency impossible. The BSC could then decide to change the carrier frequency (scenario 1).
- **Inter-cell, intra-BSC handover:** This is a typical handover scenario. The mobile station moves from one cell to another, but stays within the control of the same BSC. The BSC then performs a handover, assigns a new radio channel in the new cell and releases the old one (scenario 2).
- **Inter-BSC, intra-MSC handover:** As a BSC only controls a limited number of cells; GSM also has to perform handovers between cells controlled by different BSCs. This handover then has to be controlled by the MSC (scenario 3).
- **Inter MSC handover:** A handover could be required between two cells belonging to different MSCs. Now both MSCs perform the handover together (scenario 4).

For example following is the mechanism of the intra MSC handover:

Intra-MSC handover

To provide all the necessary information for a handover due to a weak link, MS and BTS both perform periodic measurements of the downlink and uplink quality respectively. Measurement reports are sent by the MS about every half-second and contain the quality of the current link used for transmission as well as the quality of certain channels in neighboring cells. The MS sends its periodic measurements reports, the BTS_{old} forwards these reports to the BSC_{old} together with its own measurements. Based on these values and, e.g., on current traffic conditions, the BSC_{old} may decide to perform a handover and sends the message $HO_{required}$ to the MSC. The task of the MSC then comprises the request of the resources needed for the handover from the new BSC, BSC_{new}. This BSC checks if enough resources (typically frequencies or time

slots) are available and activates a physical channel at the BTS_{new} to prepare for the arrival of the MS.



The BTS_{new} acknowledges the successful channel activation, BSC_{new} acknowledges the handover request. The MSC then issues a handover command that is forwarded to the MS. The MS now breaks its old radio link and accesses the new BTS. The next steps include the establishment of the link. Basically, the MS has then finished the handover, but it is important to release the resources at the old BSC and BTS and to signal the successful handover using the handover and clear complete messages as shown.

Security

GSM offers several security services using confidential information stored in the AuC and in the individual SIM (which is plugged into an arbitrary MS). The SIM stores personal, secret data and is protected with a PIN against unauthorized use. (For example, the secret key K_i used for authentication and encryption procedures is stored in the SIM.)

- **Access control and authentication:** The first step includes the authentication of a valid user for the SIM. The user needs a secret PIN to access the SIM. The next step is the subscriber authentication.
- **Confidentiality:** All user-related data is encrypted. After authentication, BTS and MS apply encryption to voice, data, and signaling. This confidentiality exists only between MS and BTS, but it does not exist end-to-end or within the whole fixed GSM/telephone network.

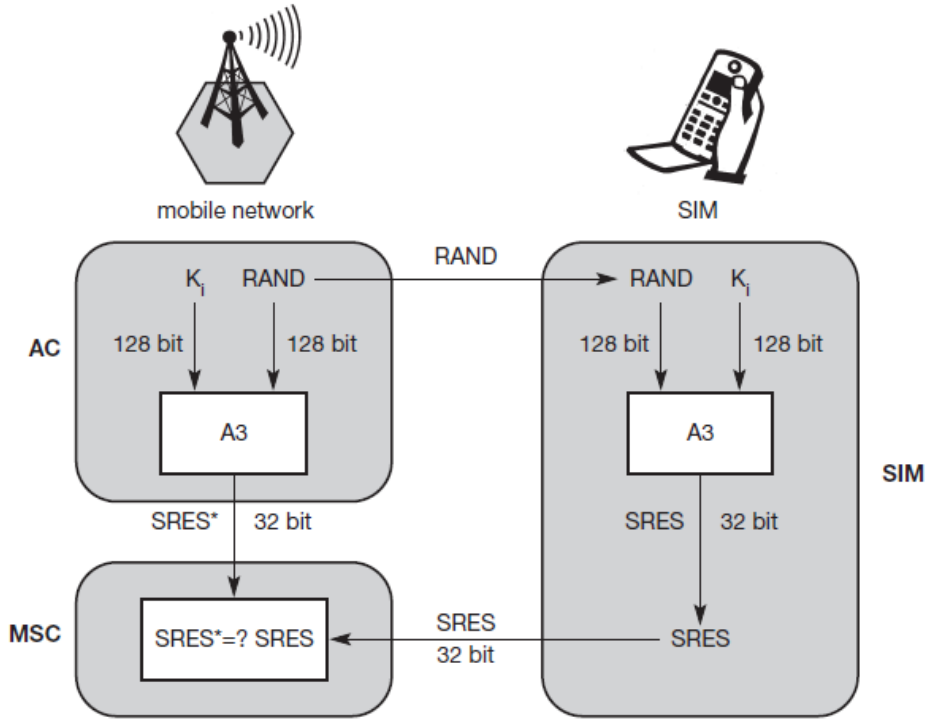
- **Anonymity:** To provide user anonymity, all data is encrypted before transmission, and user identifiers (which would reveal an identity) are not used over the air. Instead, GSM transmits a temporary identifier (TMSI), which is newly assigned by the VLR after each location update. Additionally, the VLR can change the TMSI at any time.

Three algorithms have been specified to provide security services in GSM. Algorithm A3 is used for authentication, A5 for encryption, and A8 for the generation of a cipher key

Authentication

Before a subscriber can use any service from the GSM network, he or she must be authenticated. Authentication is based on the SIM, which stores the individual authentication key K_i, the user identification IMSI, and the algorithm used for authentication A3. Authentication uses a challenge-response method: the access control AC generates a random number RAND as challenge, and the SIM within the MS answers with SRES (signed response) as response. The AuC performs the basic generation of random values RAND, signed responses SRES and cipher keys K_c for each IMSI, and then forwards this information to the HLR. The current VLR requests the appropriate values for RAND, SRES, and Kc from the HLR.

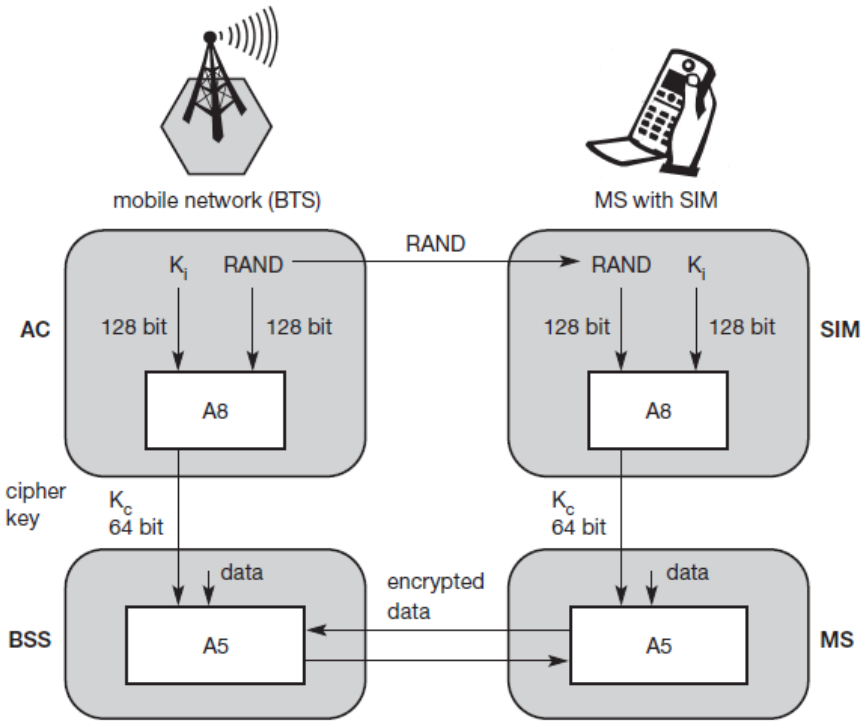
For authentication, the VLR sends the random value RAND to the SIM. Both sides, network and subscriber module, perform the same operation with RAND and the key Ki, called A3. The MS sends back the SRES generated by the SIM; the VLR can now compare both values. If they are the same, the VLR accepts the subscriber, otherwise the subscriber is rejected.



Encryption

To ensure privacy, all messages containing user-related information are encrypted in GSM over the air interface. After authentication, MS and BSS can start using encryption by applying the cipher key Kc. Kc is generated using the individual key Ki and a random value by applying the algorithm A8. Note that the SIM in the MS and the network both calculate the same Kc based on the random value RAND. The key Kc itself is not transmitted over the air interface.

MS and BTS can now encrypt and decrypt data using the algorithm A5 and the cipher key Kc. As Figure 4.15 shows, Kc should be a 64 bit key – which is not very strong, but is at least a good protection against simple eavesdropping. However, the publication of A3 and A8 on the internet showed that in certain implementations 10 of the 64 bits are always set to 0, so that the real length of the key is thus only 54 consequently, the encryption is much weaker.



New data services

The standard bandwidth of 9.6 kbit/s available for data transmission is not sufficient for the requirements of today's computers. When GSM was developed, not many people anticipated the tremendous growth of data communication compared to voice communication. At that time, 9.6 kbit/s was a lot, or at least enough for standard group 3 fax machines. But with the requirements of, e.g., web browsing, file download, or even intensive e-mail exchange with attachments, this is not enough.

To enhance the data transmission capabilities of GSM, two basic approaches are possible.

- HSCSD(High speed circuit switched data)
- GPRS(General packet radio service)

HSCSD

A straightforward improvement of GSM's data transmission capabilities is high speed circuit switched data (HSCSD), which is available with some providers. In this system, higher data rates are achieved by bundling several TCHs. An MS requests one or more TCHs from the GSM network, i.e., it allocates several TDMA slots within a TDMA frame. This allocation can be asymmetrical, i.e., more slots can be allocated on the downlink than on the uplink, which fits the typical user behavior of downloading more data compared to uploading. Basically, HSCSD only requires software upgrades in an MS and MSC.

In theory, an MS could use all eight slots within a TDMA frame to achieve an air interface user rate (AIUR) of, e.g., 8 TCH/F14.4 channels or 115.2 kbit/s. One problem of this configuration is that the MS is required to send and receive at the same time. Standard GSM does not require this capability – uplink and downlink slots are always shifted for three slots.

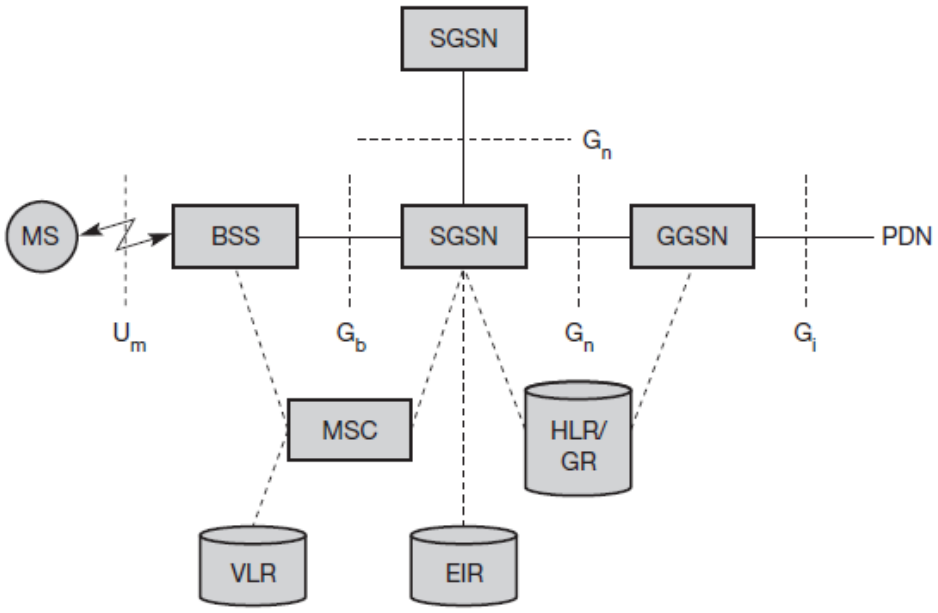
Although it appears attractive at first glance, HSCSD exhibits some major disadvantages. It still uses the connection-oriented mechanisms of GSM. These are not at all efficient for computer data traffic, which is typically bursty and asymmetrical. While downloading a larger file may require all channels reserved, typical web browsing would leave the channels idle most of the time. Allocating channels is reflected directly in the service costs, as once the channels have been reserved, other users cannot use them.

GPRS

The next step toward more flexible and powerful data transmission avoids the problems of HSCSD by being fully packet-oriented. The general packet radio service (GPRS) provides packet mode transfer for applications that exhibit traffic patterns such as frequent transmission of small volumes (e.g., typical web requests) or infrequent transmissions of small or medium volumes (e.g., typical web responses) according to the requirement specification.

The main concepts of GPRS are as follows. For the new GPRS radio channels, the GSM system can allocate between one and eight time slots within a TDMA frame. Time slots are not allocated in a fixed, pre-determined manner but on demand. All time slots can be shared by the active users; up- and downlink are allocated separately. Allocation of the slots is based on current load and operator preferences. Depending on the coding, a transfer rate of up to 170 kbit/s is possible. For GPRS, operators often reserve at least a time slot per cell to guarantee a minimum data rate.

GPRS architecture reference model



The **GPRS architecture** introduces two new network elements, which are called **GPRS support nodes (GSN)** and are in fact routers. All GSNs are integrated into the standard GSM architecture, and many new interfaces have been defined. The **gateway GPRS support node (GGSN)** is the interworking unit between the GPRS network and external **packet data networks (PDN)**. This node contains routing information for GPRS users, performs address conversion, and tunnels data to a user via encapsulation. The GGSN is connected to external networks (e.g., IP or X.25) via the Gi interface and transfers packets to the SGSN via an IP-based GPRS backbone network (Gn interface).

The other new element is the serving GPRS support node (SGSN) which supports the MS via the Gb interface. The SGSN, for example, requests user addresses from the GPRS register (GR), keeps track of the individual MSs' location, is responsible for collecting billing information (e.g., counting bytes), and performs several security functions such as access control. The SGSN is connected to a BSC via frame relay and is basically on the same hierarchy level as an MSC. The GR, which is typically a part of the HLR, stores all GPRS-relevant data. GGSNs and SGSNs can be compared with home and foreign agents, respectively, in a mobile IP network.

Broadcast systems

. Typical broadcast systems, such as radio and television, distribute information regardless of the needs of individual users. As an addition to two-way communication technologies, broadcasting information can be very cost effective. Just imagine the distribution of a movie trailer to millions

of potential customers and compare it with the abilities of 3G base stations to provide 10–20 simultaneous users with a 128 kbit/s video stream. The distribution of the trailer would block the whole mobile network for a long time even if tens of thousand base stations are assumed. In the future, television and radio transmissions will be fully digital. Already several radio stations produce and transmit their programmes digitally via the internet or digital radio (Digital television is on its way. Besides transmitting video and audio, digital transmission allows for the distribution of arbitrary digital data, i.e., multimedia information can accompany radio and TV programmes at very low cost compared to individual wireless connections.

Cyclical repetition of data

A broadcast sender of data does not know when a receiver starts to listen to the transmission. While for radio or television this is no problem (if you do not listen you will not get the message), transmission of other important information, such as traffic or weather conditions, has to be repeated to give receivers a chance to receive this information after having listened for a certain amount of time (like the news every full hour). The cyclical repetition of data blocks sent via broadcast is often called a broadcast disk according to the project in Acharya (1995) or data carousel, e.g., according to the DAB/DVB standards (ETSI, 2002).

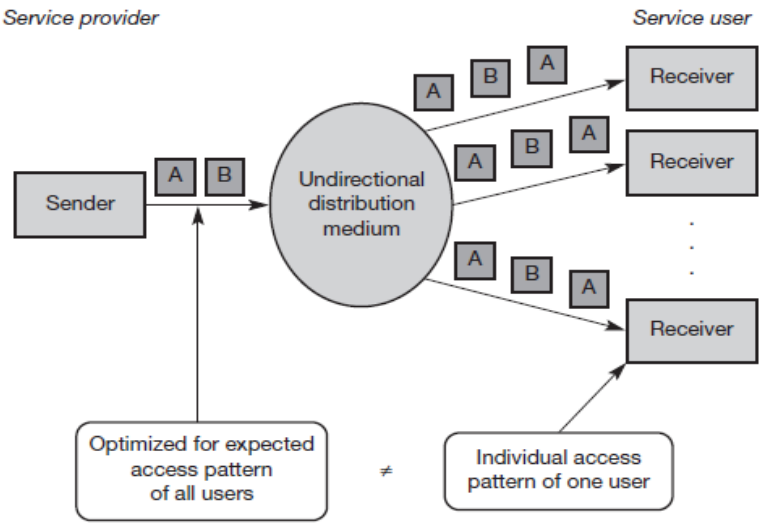


Figure 6.1
Broadcast transmission

Different patterns are possible (The sender repeats the three data blocks A, B, and C in a cycle. Using a flat disk, all blocks are repeated one after another. Every block is transmitted for an equal amount of time, the average waiting time for receiving a block is the same for A, B, and C.

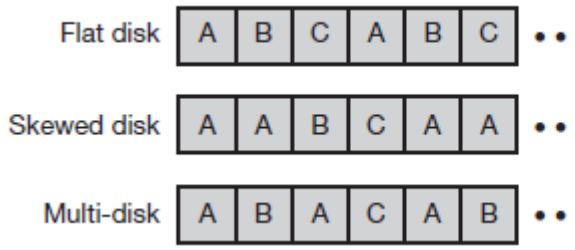


Figure 6.2
Different broadcast patterns

Skewed disks favor one or more data blocks by repeating them once or several times. This raises the probability of receiving a repeated block (here A) if the block was corrupted the first time. Finally, multi-disks distribute blocks that are repeated more often than others evenly over the cyclic pattern. This minimizes the delay if a user wants to access, e.g., block A.

Digital audio broadcasting

Today's analog radio system still follows the basic principle of frequency modulation invented back in 1933. In addition to audio transmission, very limited information such as the station identification can accompany the program. Transmission quality varies greatly depending on multi-path effects and interference. The fully digital DAB system does not only offer sound in a CD-like quality, it is also practically immune to interference and multi-path propagation effects (ETSI, 2001a), (DAB, 2002). DAB systems can use single frequency networks (SFN), i.e., all senders transmitting the same radio program operate at the same frequency. Today, different senders have to use different frequencies to avoid interference although they are transmitting the same radio program. Using an SFN is very frequency efficient, as a single radio station only needs one frequency throughout the whole country. Additionally, DAB transmission power per antenna is orders ofmagnitude lower compared to traditional FM stations. DAB uses VHF and UHF frequency bands (depending on national regulations), e.g., the terrestrial TV channels 5 to 12 (174–230 MHz) or the L-band (1452–1492 MHz). The modulation scheme used is DQPSK. DAB is one of the systems using COFDM (with 192 to 1536 carriers (the so-called ensemble) within a DAB channel of 1.5 MHz. Additionally, DAB uses FEC to reduce the error rate and introduces guard spaces between single symbols during transmission. COFDM and the use of guard spaces reduce ISI to a minimum. DAB can even benefit from multipath propagation by recombining the signals from different paths Within every frequency block of 1.5 MHz, DAB can transmit up to six stereo audio programmes with a data rate of 192 kbit/s each. Depending on the redundancy coding, a data service with rates up to 1.5 Mbit/s is available as an alternative. For the DAB transmission system, audio is just another type of data *(besides different coding schemes). DAB uses two basic transport mechanisms:*

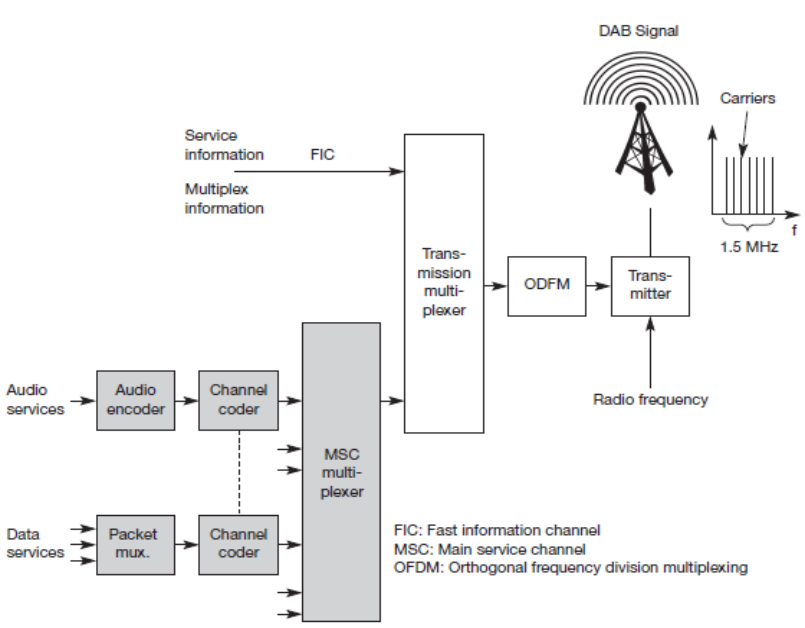


Figure 6.4
Components of a DAB sender (simplified)

● Main service channel (MSC): The MSC carries all user data, e.g. audio, multimedia data. The MSC consists of common interleaved frames (CIF), i.e., data fields of 55,296 bits that are sent every 24 ms (this interval depends on the transmission mode (ETSI, 2001a)). This results in a data rate of 2.304 Mbit/s. A CIF consists of capacity units (CU) with a size of 64 bits, which form the smallest addressable unit within a DAB system.

● Fast information channel (FIC): The FIC contains fast information blocks (FIB) with 256 bits each (16 bit checksum). An FIC carries all control information which is required for interpreting the configuration and content of the MSC. Two transport modes have been defined for the MSC. The stream mode offers a transparent data transmission from the source to the destination with a fixed bit rate in a sub channel. A sub channel is a part of the MSC and comprises several CUs within a CIF. The fixed data rate can be multiples of 8 kbit/s. The packet mode transfers data in addressable blocks (packets). These blocks are used to convey MSC data within a sub channel. DAB defines many service information structures accompanying an audio stream. This program associated data (PAD) can contain program information, control information, still pictures for display on a small LCD, title display etc. Audio coding uses PCM with a sampling rate of 48 kHz and MPEG audio compression. Each frame consists of three parts. The synchronization channel (SC) marks the start of a frame. It consists of a null symbol and a phase reference symbol to synchronize the receiver. The fast information channel (FIC) follows, containing control data in the FIBs. Finally, the main service channel (MSC) carries audio and data service components. . Audio services are encoded (MPEG compression) and coded for transmission (FEC). All data services are multiplexed and also coded with redundancy. The MSC multiplexer combines all user data streams and forwards them to the transmission multiplexer. This unit creates the frame structure by interleaving the FIC. Finally, OFDM coding is applied and the DAB signal is transmitted.

Digital video broadcasting

The logical consequence of applying digital technology to radio broadcasting is doing the same for the traditional television system. The analog system used today has basically remained unchanged for decades. The only invention worth mentioning was the introduction of color TV for the mass market back in the 1960s. Television still uses the low resolution of 625 lines for the European PAL system or only 525 lines for the US NTSC respectively2. The display is interlaced with 25 or 30 frames per second respectively. So, compared with today's computer displays with resolutions of 1,280 × 1,024 and more than 75 Hz frame rate, non-interlaced, TV performance is not very impressive. There have been many attempts to change this and to introduce digital TV with higher resolution, better sound and additional features, but no approach has yet been truly successful. One reason for this is the huge number of old systems that are installed and cannot be replaced as fast as computers (we can watch the latest movie on an old TV, but it is impossible to run new software on older computers!).

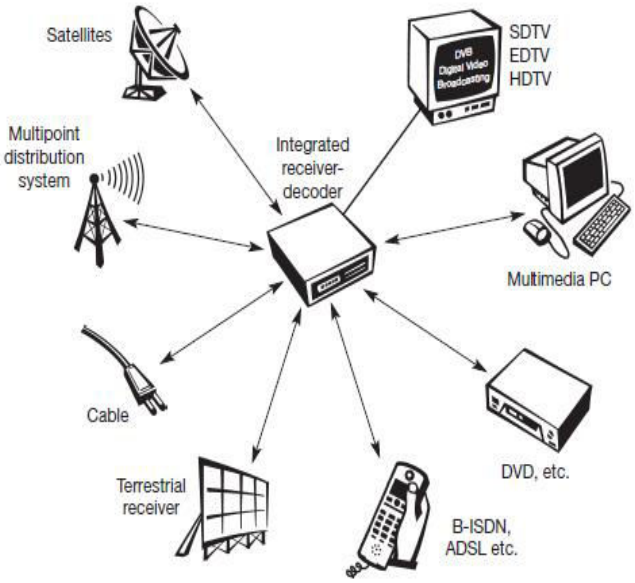


Figure 6.7 Digital video broadcasting scenario

Varying political and economic interests are counterproductive to a common standard for digital TV. One approach toward such a standard, which may prove useful for mobile communication, too, is presented in the following sections. After some national failures in introducing digital TV, the so-called European Launching Group was founded in 1991 with the aim of developing a common digital television system for Europe. In 1993 these common efforts
were named digital video broadcasting (DVB) (Reimers, 1998), (DVB, 2002). Although the name shows a certain affinity to DAB, there are some fundamental differences regarding the transmission technology, frequencies, modulation etc.

The goal of DVB is to introduce digital television broadcasting using satellite transmission (DVB-S, (ETSI, 1997)), cable technology (DVB-C, (ETSI, 1998)), and also terrestrial transmission (DVB-T, (ETSI, 2001b)). components that should be integrated into the DVB architecture. The center point is

an integrated receiver-decoder (set-top box) connected to a high-resolution monitor. This set-top box can receive DVB signals via satellites, terrestrial local/regional senders (multi-point distribution systems, terrestrial receiver), cable, B-ISDN, ADSL, or other possible future technologies. Cable, ADSL, and B-ISDN connections also offer a return channel, i.e., a user can send data such as channel selection, authentication information, or a shopping list. Audio/video streams can be recorded, processed, and replayed using digital versatile disk (DVD) or multimedia PCs. Different levels of quality are envisaged: standard definition TV (SDTV), enhanced definition TV (EDTV), and high definition TV (HDTV) with a resolution of up to 1,920 × 1,080 pixels. Similar to DAB, DVB also transmits data using flexible containers. These containers are basically MPEG-2 frames that do not restrict the type of information. DVB sends service information contained in its data stream, which specifies the content of a container.

Convergence of broadcasting and mobile communications

To enable the convergence of digital broadcasting systems and mobile communication systems ETSI (2000) and ETSI (1999d) define interaction channels through GSM for DAB and DVB, respectively. An interaction channel is not only common to DAB and DVB but covers also different fixed and mobile systems (UMTS, DECT, ISDN, PSTN etc.). 3G systems are typically characterized by very small cells, especially in densely populated areas. Although 3G systems offer higher data rates than 2G systems, their design has not fully taken into consideration the integration of broadcast quality audio and TV services onto 3G terminals. This is true from a technical point of view (capacity per cell in bit/s) as well as from an economic point of view (very high deployment cost for full coverage, typically low return on invest for video services). High bandwidth audio and video is sent together with IP data via the broadcast channel. IP data could use multi-casting, data carousels etc. as described above. For example, IP data in a DVB-T carousel could contain the top hundred web pages of the ISP's portal. Individual pages for single users are then additionally sent via GRPS or UMTS (DRiVE, 2002).

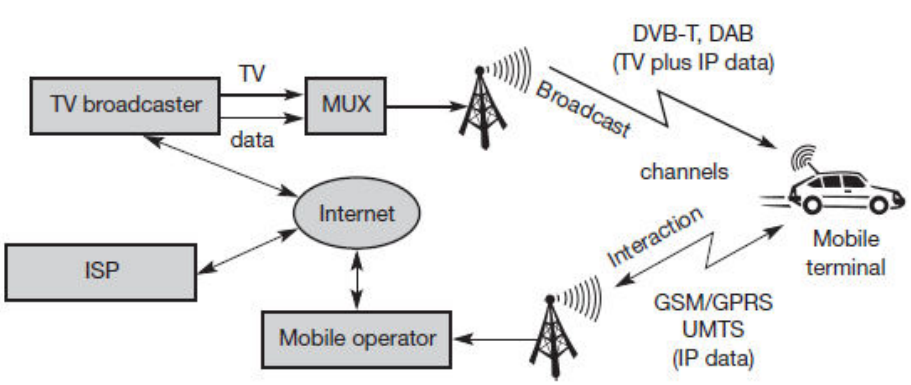


Figure 6.10
Mobile Internet services using IP over GSM/GPRS or UMTS as interaction channel for DAB or DVB

UNIT III

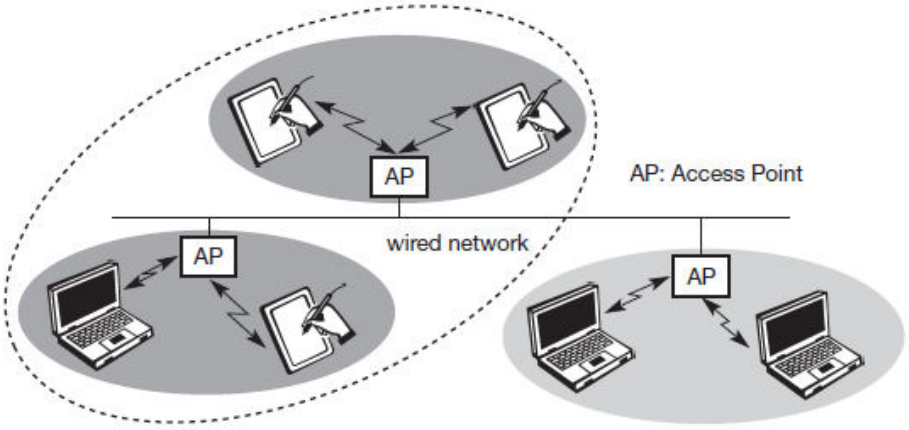
Some of the fundamental differences between wired networks & ad-hoc networks are:

 Asymmetric links: - Routing information collected for one direction is of no use for the other direction. Many routing algorithms for wired networks rely on a symmetric scenario.

 Redundant links: - In wired networks, some redundancy is present to survive link failures and this redundancy is controlled by a network administrator. In ad-hoc networks, nobody controls redundancy resulting in many redundant links up to the extreme of a complete meshed topology.

 Interference: - In wired networks, links exist only where a wire exists, and connections are planned by network administrators. But, in ad-hoc networks links come and go depending on transmission characteristics, one transmission might interfere with another and nodes might overhear the transmission of other nodes.

Figure 7.1 Example of three infrastructure-based wireless networks



 Dynamic topology: - The mobile nodes might move in an arbitrary manner or medium characteristics might change. This result in frequent changes in topology, so snapshots are valid only for a very short period of time. So, in ad-hoc networks, routing tables must somehow reflect these frequent changes in topology and routing algorithms have to be adopted.

BLUETOOTH

"Bluetooth" was the nickname of Harald Blåtland II, king of Denmark from 940 to 981, who united all of Denmark and part of Norway under his rule. **Bluetooth** is a proprietary open wireless technology standard for exchanging data over short distances (using short wavelength radio transmissions in the ISM band from 2400-2480 MHz) from fixed and mobile devices, creating personal area networks (PANs) with high levels of security. The Bluetooth technology aims at so-

called **ad-hoc piconets**, which are local area networks with a very limited coverage and without the need for an infrastructure.

Bluetooth Features

 Bluetooth is wireless and automatic. You don't have to keep track of cables, connectors, and connections, and you don't need to do anything special to initiate communications. Devices find each other automatically and start conversing without user input, expect where authentication is required; for example, users must log in to use their email accounts.

 Bluetooth is inexpensive. Market analysts peg the cost to incorporate Bluetooth technology into a PDA, cell phone, or other product at a minimum cost.

 The ISM band that Bluetooth uses is regulated, but unlicensed. Governments have converged on a single standard, so it's possible to use the same devices virtually wherever you travel, and you don't need to obtain legal permission in advance to begin using the technology.

 Bluetooth handles both data and voice. Its ability to handle both kinds of transmissions simultaneously makes possible such innovations as a mobile hands-free headset for voice with applications that print to fax, and that synchronize the address books on your PDA, your laptop, and your cell phone.

 Signals are omni-directional and can pass through walls and briefcases. Communicating devices don't need to be aligned and don't need an unobstructed line of sight like infrared.

 Bluetooth uses frequency hopping. Its spread spectrum approach greatly reduces the risk that communications will be intercepted.

Bluetooth Applications

 File transfer.

 Ad-hoc networking: Communicating devices can spontaneously form a community of networks that persists only as long as it's needed

 Device synchronization: Seamless connectivity among PDAs, computers, and mobile phones allows applications to update information on multiple devices automatically when data on any one device changes.

 Peripheral connectivity.

 Car kits: Hands-free packages enable users to access phones and other devices without taking their hands off the steering wheel

 Mobile payments: Your Bluetooth-enabled phone can communicate with a Bluetooth-enabled vending machine to buy a can of Diet Pepsi, and put the charge on your phone bill.

The 802.11b protocol is designed to connect relatively large devices with lots of power and speed, such as desktops and laptops, where devices communicate at up to 11 Mbit/sec, at greater distances (up to 300 feet, or 100 meters). By contrast, Bluetooth is designed to connect small devices like

PDAs, mobile phones, and peripherals at slower speeds (1 Mbit/sec), within a shorter range (30 feet, or 10 meters), which reduces power requirements. Another major difference is that 802.11b wasn't designed for voice communications, while any Bluetooth connection can support both data and voice communications.

User scenarios

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

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

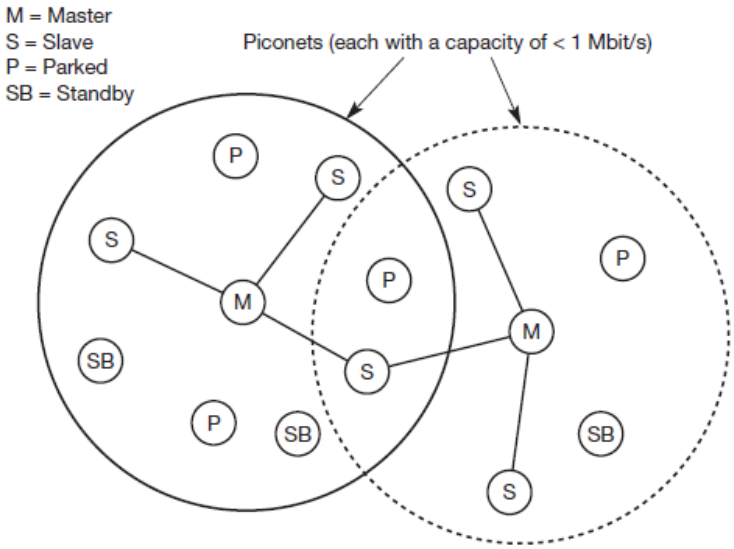
Support of ad-hoc networking: Imagine several people coming together, discussing issues, exchanging data (schedules, sales figures etc.). For instance, students might join a lecture, with the teacher distributing data to their personal digital assistants (PDAs). Wireless networks can support this type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

Bridging of networks: Using wireless piconets, a mobile phone can be connected to a PDA or laptop in a simple way. Mobile phones will not have full WLAN adapters built in, but could have a Bluetooth chip. The mobile phone can then act as a bridge between the local piconet and, e.g., the global GSM network.

Networking in Bluetooth

Bluetooth operates on 79 channels in the 2.4 GHz band with 1 MHz carrier spacing. Each device performs frequency hopping with 1,600 hops/s in a pseudo random fashion. A piconet is a collection of Bluetooth devices which are synchronized to the same hopping sequence. One device in the piconet can act as **master** (M), all other devices connected to the master must 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. A typical piconet is shown below:

Figure 7.43
Bluetooth scatternet



Parked devices (P) can not actively participate in the piconet (i.e., they do not have a connection), but are known and can be reactivated within some milliseconds. Devices in stand-by (SB) do not participate in the piconet. Each piconet has exactly one master and up to seven simultaneous slaves. More than 200 devices can be parked. The first step in forming a piconet involves a master sending its clock and device ID. All the Bluetooth devices have the same capability to become a master or a slave and two or three devices are sufficient to form a piconet. The unit establishing the piconet automatically becomes the master, all other devices will be slaves. The hopping pattern is determined by the device ID, a 48-bit worldwide unique identifier.
The phase in the hopping pattern is determined by the master's clock. After adjusting the internal clock according to the master a device may participate in the piconet. All active devices are assigned a 3-bit **active member address** (AMA). All parked devices use an 8-bit **parked member address** (PMA). Devices in stand-by do not need an address.

A device in one piconet can communicate to another device in another piconet, forming a **scatternet**. A master in one piconet may be a slave in another piconet. Both piconets use a different hopping sequence, always determined by the master of the piconet. Bluetooth applies **FH-CDMA** for separation of piconets. A collision occurs if two or more piconets use the same carrier frequency at the same time. This will probably happen as the hopping sequences are not coordinated. If a device wants to participate in more than one piconet, it has to synchronize to the hopping sequence of the piconet it wants to take part in. If a device acts as slave in one piconet, it simply starts to synchronize with the hopping sequence of the piconet it wants to join. After synchronization, it acts as a slave in this piconet and no longer participates in its former piconet. To enable synchronization, a slave has to know the identity of the master that determines the hopping sequence of a piconet.

Before leaving one piconet, a slave informs the current master that it will be unavailable for a certain amount of time. The remaining devices in the piconet continue to communicate as usual.

Bluetooth Protocol Stack

The Bluetooth protocol stack can be divided into a **core specification**, which describes the protocols from physical layer to the data link control together with management functions, and **profile specifications** describing many protocols and functions needed to adapt the wireless Bluetooth technology to legacy and new applications.

A high-level view of the architecture is shown. The responsibilities of the layers in this stack are as follows:

 The radio layer is the physical wireless connection. To avoid interference with other devices that communicate in the ISM band, the modulation is based on fast frequency hopping. Bluetooth divides the 2.4 GHz frequency band into 79 channels 1 MHz apart (from 2.402 to 2.480 GHz), and uses this spread spectrum to hop from one channel to another, up to 1600

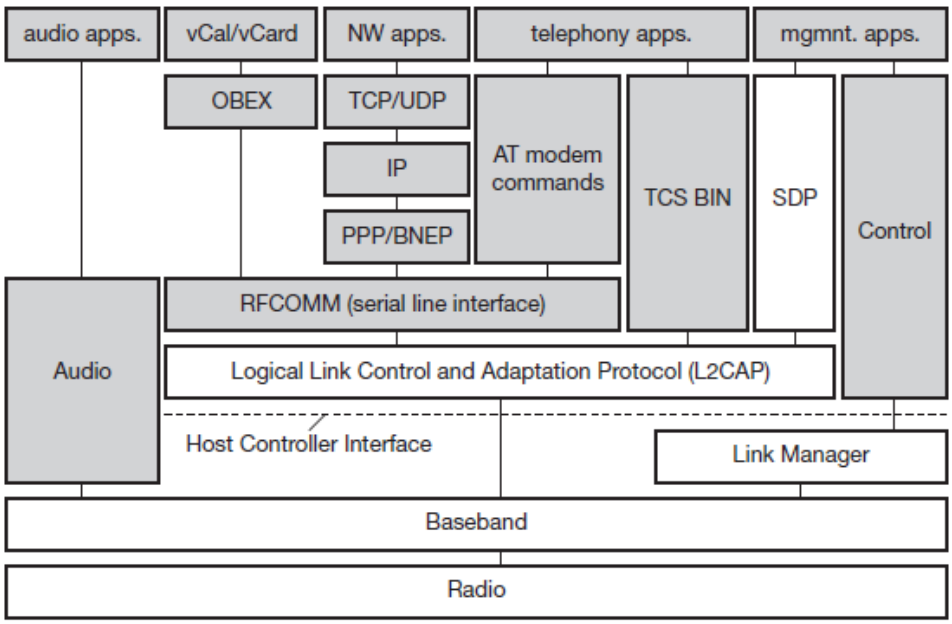


Figure 7.44 Bluetooth protocol stack

times a second. The standard wavelength range is 10 cm to 10 m, and can be extended to 100 m by increasing transmission power.

Bluetooth Protocol Stack

 The baseband layer is responsible for controlling and sending data packets over the radio link. It provides transmission channels for both data and voice. The baseband layer maintains Synchronous Connection-Oriented (SCO) links for voice and Asynchronous Connectionless (ACL) links for data. SCO packets are never retransmitted but ACL packets are, to ensure data integrity.

SCO links are point-to-point symmetric connections, where time slots are reserved to guarantee timely transmission. A slave device is allowed to respond during the time slot immediately following an SCO transmission from the master. A master can support up to three SCO links to a single slave or to multiple slaves, and a single slave can support up to two SCO links to different slaves. Data transmissions on ACL links, on the other hand, are established on a per-slot basis (using slots not reserved for SCO links). ACL links support point-to-multipoint transmissions. After an ACL transmission from the master, only a slave addressed specifically may respond during the next time slot; if no device is addressed, the message is treated as a broadcast.

 The Link Manager Protocol (LMP) uses the links set up by the baseband to establish connections and manage piconets. Responsibilities of the LMP also include authentication and security services, and monitoring of service quality.

 The Host Controller Interface (HCI) is the dividing line between software and hardware. The L2CAP and layers above it are currently implemented in software, and the LMP and lower layers are in hardware. The HCI is the driver interface for the physical bus that connects these two components. The HCI may not be required. The L2CAP may be accessed directly
by the application, or through certain support protocols provided to ease the burden on application programmers.

 The Logical Link Control and Adaptation Protocol (L2CAP) receives application data and adapts it to the Bluetooth format. Quality of Service (QoS) parameters are exchanged at this layer.

Link Manager Protocol

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

 Authentication, pairing, and encryption: Although basic authentication is handled in the baseband, LMP has to control the exchange of random numbers and signed responses. LMP is not directly involved in the encryption process, but sets the encryption mode (no encryption, point-to-point, or broadcast), key size, and random speed.

 Synchronization: Precise synchronization is of major importance within a Bluetooth network. The clock offset is updated each time a packet is received from the master.

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

 Quality of service negotiation:

Different parameters control the QoS of a Bluetooth device at these lower layers. The poll interval, i.e., the maximum time between transmissions from a master to a particular slave, controls the

latency and transfer capacity. A master can also limit the number of slots available for slaves' answers to increase its own bandwidth.

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

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

L2CAP

The logical link control and adaptation protocol (L2CAP) is a data link control protocol on top of the baseband layer offering logical channels between Bluetooth devices with QoS properties. L2CAP is available for ACLs only.

L2CAP provides three different types of logical channels that are transported via the ACL between master and slave:

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

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

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

Each channel can be identified by its **channel identifier (CID)**. Signaling channels always use a CID value of 1, a CID value of 2 is reserved for connectionless channels. For connection-oriented channels a unique CID (>= 64) is dynamically assigned at each end of the channel to identify the connection.

The following figure shows the three packet types belonging to the three logical channel types.

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

Besides protocol multiplexing, flow specification, and group management, the L2CAP layer also provides segmentation and reassembly functions. Depending on the baseband capabilities, large packets have to be chopped into smaller segments.

Security

The main security features offered by Bluetooth include a challenge response routine for authentication, a stream cipher for encryption, and a session key generation. Each connection may require a one-way, two-way, or no authentication using the challenge-response routine. The security algorithms use the public identity of a device, a secret private user key, and an internally generated random key as input parameters. For each transaction, a new random number is generated on the Bluetooth chip. Key management is left to higher layer software. The following figure shows several steps in the security architecture of Bluetooth.

The first step, called **pairing**, is necessary if two Bluetooth devices have never met before. To set up trust between the two devices a user can enter a secret PIN into both devices. This PIN can have a length of up to 16 byte. Based on the PIN, the device address, and random numbers, several keys can be computed which can be used as link key for **authentication**. The authentication is a challenge-response process based on the link key, a random number generated by a verifier (the device that requests authentication), and the device address of the claimat (the device that is authenticated

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

All Bluetooth-enabled devices must implement the Generic Access Profile, which contains all the Bluetooth protocols and possible devices. This profile defines a security model that includes three security modes:

- *Mode 1* is an insecure mode of operation. No security procedures are initiated.
- *Mode 2* is known as *service-level enforced security*. When devices operate in this mode, no security procedures are initiated before the channel is established. This mode enables applications to have different access policies and run them in parallel.
- *Mode 3* is known as *link-level enforced security*. In this mode, security procedures are initiated before link setup is complete.

Though Bluetooth offers a better security than WER in 802.11, it has several limitations. The PIN's are often fixed and some keys are permanently stored on the devices. The quality of the random number generators has not been specified.

SDP

To find new services available in the radio proximity, Bluetooth defined the **service discovery protocol (SDP)**. SDP defines only the discovery of services, not their usage. Discovered services

can be cached and gradual discovery is possible. All the information an SDP server has about a service is contained in a **service record**. This consists of a list of service attributes and is identified by a 32-bit service record handle.
A service attribute consists of an attribute ID and an attribute value. The 16-bit attribute ID distinguishes each service attribute from other service attributes within a service record. The attribute ID also identifies the semantics of the associated attribute value. The attribute value can be an integer, a UUID (universally unique identifier), a string, a Boolean, a URL (uniform resource locator) etc.

HiperLAN (High Performance Radio LAN) is a Wireless LAN standard It is a European alternative for the IEEE 802.11standards (the IEEE is an international organization). It is defined by the European Telecommunications Standards Institute(ETSI). In ETSI the standards are defined by the BRAN project (Broadband Radio Access Networks). The HiperLAN standard family has four different versions.

Planning for the first version of the standard, called HiperLAN/1, started 1991, when planning of 802.11 was already going on. The goal of the HiperLAN was the high data rate, higher than 802.11. The standard was approved in 1996. The functional specification is EN300652, the rest is in ETS300836.

The standard covers the Physical layer and the Media Access Control part of the Data link layer like 802.11. There is a new sublayer called Channel Access and Control sublayer (CAC). This sublayer deals with the access requests to the channels. The accomplishing of the request is dependent on the usage of the channel and the priority of the request.

CAC layer provides hierarchical independence with Elimination-Yield Non-Preemptive Multiple Access mechanism (EY-NPMA). EY-NPMA codes priority choices and other functions into one variable length radio pulse preceding the packet data. EY-NPMA enables the network to function with few collisions even though there would be a large number of users.Multimedia applications work in HiperLAN because of EY-NPMA priority mechanism. MAC layer defines protocols for routing, security and power saving and provides naturally data transfer to the upper layers.

On the physical layer FSK and GMSK modulations are used in HiperLAN/1.

HiperLAN features:

- range 50 m
- slow mobility (1.4 m/s)
- supports asynchronous and synchronous traffic
- Bit rate - 23.2 Mbit/s
- Description- Wireless Ethernet
- Frequency range- 5 GHz

HiperLAN does not conflict with microwave and other kitchen appliances, which are on 2.4 GHz. An innovative feature of HIPERLAN 1, which may other wireless networks do not offer, is its ability to forward data packets using several relays. Relays can extend the communication on the MAC layer beyond the radio range. For power conservation, a node may set up a specific wake up pattern. This pattern determines at what time the node is ready to receive, so that at other times, the node can turn off its receiver and save energy. These nodes are called p-savers and need so called p-supporters that contain information about wake up patterns of all the p-savers they are responsible for. A p-supporter only forwards data to a p-saver at the moment p-saver is awake. This action also requires buffering mechanisms for packets on p-supporting forwaders.

HiperLAN/2

HiperLAN/2 functional specification was accomplished February 2000. Version 2 is designed as a fast wireless connection for many kinds of networks. Those are UMTS back bone network, ATM and IP networks. Also it works as a network at home like HiperLAN/1. HiperLAN/2 uses the 5 GHz band and up to 54 Mbit/s data rate.

The physical layer of HiperLAN/2 is very similar to IEEE 802.11a wireless local area networks. However, the media access control (the multiple access protocol) is Dynamic TDMA in HiperLAN/2, while CSMA/CA is used in 802.11a/n.

Basic services in HiperLAN/2 are data, sound, and video transmission. The emphasis is in the quality of these services (QoS)

The standard covers Physical, Data Link Control and Convergence layers. Convergence layer takes care of service dependent functionality between DLC and Network layer (OSI 3). Convergence sublayers can be used also on the physical layer to connect IP, ATM or UMTS networks. This feature makes HiperLAN/2 suitable for the wireless connection of various networks.

On the physical layer BPSK, QPSK, 16QAM or 64QAM modulations are used.

HiperLAN/2 offers security measures. The data are secured with DES or Triple DES algorithms. The wireless access pointand the wireless terminal can authenticate each other.

The present document, the term "HIPERLAN" is used to refer to HIPERLAN, Type 1.

A HIPERLAN is a Radio Local Area Network (RLAN) in which all nodes communicate using a single shared communication channel. A HIPERLAN has the following properties:

- it provides a service that is compatible with the ISO MAC service definition in ISO/IEC 15 802-1

- its operations are compatible with the ISO MAC bridges specification in ISO/IEC 10 038 for interconnection with other LANs;

- it may be deployed in a pre-arranged or an ad-hoc fashion;

- it supports node mobility;

- it may have a coverage beyond the radio range limitation of a single node;

- it supports both asynchronous and time-bounded communication by means of a

Channel Access Mechanism (CAM) with priorities providing hierarchical independence of performance;

- its nodes may attempt to conserve power in communication by arranging when they need to be active for reception.

The HIPERLAN MAC service:

- is based on, and therefore is compatible with, the ISO MAC service definition;

- defines the communication service over a single HIPERLAN;

- allows the timing requirements of the MSDU transfer to be specified; and

- allows exploration of available HIPERLANs for dynamic HIPERLAN access.

The HIPERLAN CAC service:

- defines the communication service over a single shared communication channel;

- allows the channel access priority requirements of the HCSDU transfer to be specified; and

- frees the HCS-user from the concerns of the characteristics peculiar to any particular communication channel.

The HIPERLAN MAC protocol:

- provides the HIPERLAN MAC service;

- specifies the behaviour of a HM-entity in a given HIPERLAN;

- is compatible with the ISO MAC bridges specification in ISO/IEC 10 038 [8]; and

- uses the HIPERLAN CAC service.

The HIPERLAN CAC protocol:

- provides the HIPERLAN CAC service;

- specifies, for a particular set of one or more shared radio channels, the appropriate hierarchically independent

channel access mechanism used by a HC-entity in a given HIPERLAN; and

- uses the transmission and reception facilities specified by the HIPERLAN physical layer.

HIPERLAN addressing

The HIPERLAN addressing requirements are elaborated in the following subclauses.

MAC Service Access Point (MSAP) addressing

In order to be compatible with the ISO MAC service definition, the HIPERLAN MAC service uses the 48-bit LAN

MAC address for MSAP identification.

A HIPERLAN MAC entity (HM-entity) shall be attached to a single MSAP, through which the HM-entity provides the HIPERLAN MAC service to a single HMS-user; and it shall be attached to a single HIPERLAN CAC Service Access Point (HCSAP), through which the HM-entity uses the HIPERLAN CAC service provided by the HIPERLAN CAC service provider (HCS-provider).

An individual 48-bit LAN MAC address is used, as an individual-MSAP-address, to identify a single MSAP and its attached HMS-user and HM-entity. On the other hand, a group 48-bit LAN MAC address is used, as a group-MSAP-address, to identify a group of MSAPs and their attached HMS-users. Individual-MSAP-address and group-MSAP-address assignment is outside the scope of the present document and is governed by other relevant LAN standards. HCSAP addressing

A HIPERLAN CAC entity (HC-entity) shall be attached to a single HCSAP, through which the HC-entity provides the HIPERLAN CAC service to a single HCS-user. As a result, a HC-entity is attached to a single HM-entity. For practical reasons, a HM-entity's attached HCSAP shall also be identified by the same individual 48-bit LAN MAC address assigned to its attached MSAP. Therefore, an individual 48-bit LAN MAC address is inherited, as an individual-HCSAP-address, to identify a single HCSAP and its attached HCS-user and HC-entity. A group 48-bit LAN MAC address is then used, as a group-HCSAP-address, to identify a group of HCSAPs and their attached HCS-users. The group-HCSAP-address assignment from the entire group 48-bit LAN MAC address space is independent of the group-MSAP-address assignment.

HIPERLAN MAC sublayer features

Typical features of the HIPERLAN MAC sublayer are elaborated in the following subclauses.

HIPERLAN differentiation

Since a HIPERLAN's shared radio channel is not readily bounded, the HIPERLAN overlap situation may occur, in which multiple HIPERLANs' radio ranges overlap in the same radio channel. While wired LANs are implicitly distinct, the HIPERLAN overlap situation does not make a HIPERLAN implicitly distinct. On the other hand, due to limited radio range, mobile HIPERLAN nodes and adverse propagation conditions, the HIPERLAN fragmentation situation may occur, in which a HIPERLAN is effectively partitioned into multiple disjoint

communication subsets. Therefore, a HIPERLAN needs to be identifiable so that a fragmented HIPERLAN can re-merge automatically whenever the radio environment allows. Both the HIPERLAN overlap and fragmentation situations call for globally unique HIPERLAN identification or indistinguishably different HIPERLANs may mingle their communication. Unfortunately, globally unique HIPERLAN identification inevitably requires some kind of administrative co-ordination that makes ad-hoc or private HIPERLAN deployments impractical. In contrast, although mingled HIPERLAN communication may raise concerns of communication confidentiality, it does not introduce a particularly new problem because communication confidentiality is always an issue in the radio environment.

UNIT-IV

Need for Mobile IP:

The IP addresses are designed to work with stationary hosts because part of the address defines the network to which the host is attached**.** A host cannot change its IP address without terminating on-going sessions and restarting them after it acquires a new address. Other link layer mobility solutions exist but are not sufficient enough for the global Internet.

Mobility is the ability of a node to change its point-of-attachment while maintaining all existing communications and using the same IP address.

Nomadicity allows a node to move but it must terminate all existing communications and then can initiate new connections with a new address.

Mobile IP is a network layer solution for homogenous and heterogeneous mobility on the global Internet which is scalable, robust, secure and which allows nodes to maintain all ongoing communications while moving.

Design Goals: Mobile IP was developed as a means for transparently dealing with problems of mobile users. Mobile IP was designed to make the size and the frequency of required routing updates as small as possible. It was designed to make it simple to implement mobile node software. It was designed to avoid solutions that require mobile nodes to use multiple addresses.

Requirements: There are several requirements for Mobile IP to make it as a standard. Some of them are:

1. *Compatibility*: The whole architecture of internet is very huge and a new standard cannot introduce changes to the applications or network protocols already in use. Mobile IP is to be integrated into the existing operating systems. Also, for routers also it may be possible to enhance its capabilities to support mobility instead of changing the routers which is highly impossible. Mobile IP must not require special media or MAC/LLC protocols, so it must use the same interfaces and mechanisms to access the lower layers as IP does. Finally, end-systems enhanced with a mobile IP implementation should still be able to communicate with fixed systems without mobile IP.

2. *Transparency*: Mobility remains invisible for many higher layer protocols and applications. Higher layers continue to work even if the mobile computer has changed its point of attachment to the network and even notice a lower bandwidth and some interruption in the service. As many of today's applications have not been designed to use in mobile environments, the effects of mobility will be higher delay and lower bandwidth.

3. *Scalability and efficiency*: The efficiency of the network should not be affected even if a new mechanism is introduced into the internet. Enhancing IP for mobility must not generate many new messages flooding the whole network. Special care is necessary to be taken considering the lower bandwidth of wireless links. Many mobile systems have a wireless link to an attachment point. Therefore, only some additional packets must be necessary between a mobile system and a node in

the network. It is indispensable for a mobile IP to be scalable over a large number of participants in the whole internet, throughout the world.

4. *Security*: Mobility possesses many security problems. A minimum requirement is the authentication of all messages related to the management of mobile IP. It must be sure for the IP layer if it forwards a packet to a mobile host that this host really is the receiver of the packet. The IP layer can only guarantee that the IP address of the receiver is correct. There is no way to prevent faked IP addresses and other attacks.

The goal of a mobile IP can be summarized as: 'supporting end-system mobility while maintaining scalability, efficiency, and compatibility in all respects with existing applications and Internet protocols'.

Entities and terminology

The following defines several entities and terms needed to understand mobile IP as defined in RFC 3344.

Mobile Node (MN): A mobile node is an end-system or router that can change its point of attachment to the internet using mobile IP. The MN keeps its IP address and can continuously communicate with any other system in the internet as long as link-layer connectivity is given. Examples are laptop, mobile phone, router on an aircraft etc.

Correspondent node (CN): At least one partner is needed for communication. In the following the CN represents this partner for the MN. The CN can be a fixed or mobile node.

Home network: The home network is the subnet the MN belongs to with respect to its IP address. No mobile IP support is needed within the home network.

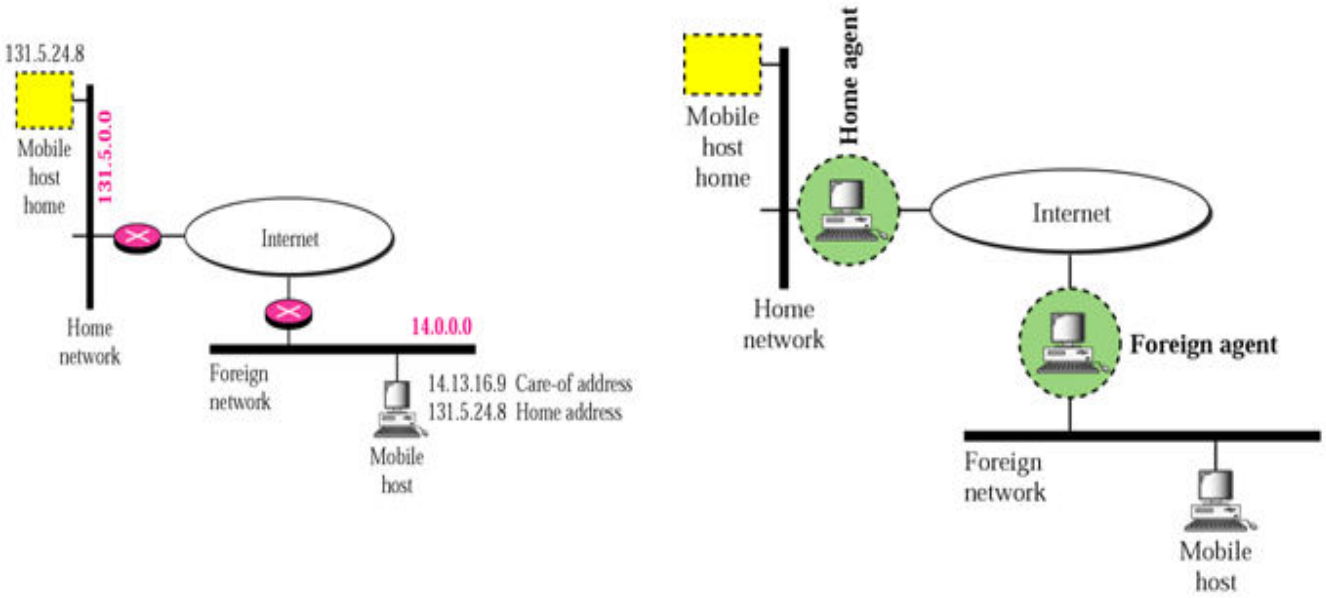
Foreign network: The foreign network is the current subnet the MN visits and which is not the home network.

Foreign agent (FA): The FA can provide several services to the MN during its visit to the foreign network. The FA can have the COA, acting as tunnel endpoint and forwarding packets to the MN. The FA can be the default router for the MN. FAs can also provide security services because they belong to the foreign network as opposed to the MN which is only visiting. FA is implemented on a router for the subnet the MN attaches to.

Care-of address (COA): The COA defines the current location of the MN from an IP point of view. All IP packets sent to the MN are delivered to the COA, not directly to the IP address of the MN. Packet delivery toward the MN is done using a tunnel, i.e., the COA marks the tunnel endpoint, i.e., the address where packets exit the tunnel. There are two different possibilities for the location of the COA:

Foreign agent COA: The COA could be located at the FA, i.e., the COA is an IP address of the FA. The FA is the tunnel end-point and forwards packets to the MN. Many MN using the FA can share this COA as common COA.

Co-located COA: The COA is co-located if the MN temporarily acquired an additional IP address which acts as COA. This address is now topologically correct, and the tunnel endpoint is at the MN. Co-located addresses can be acquired using services such as DHCP.

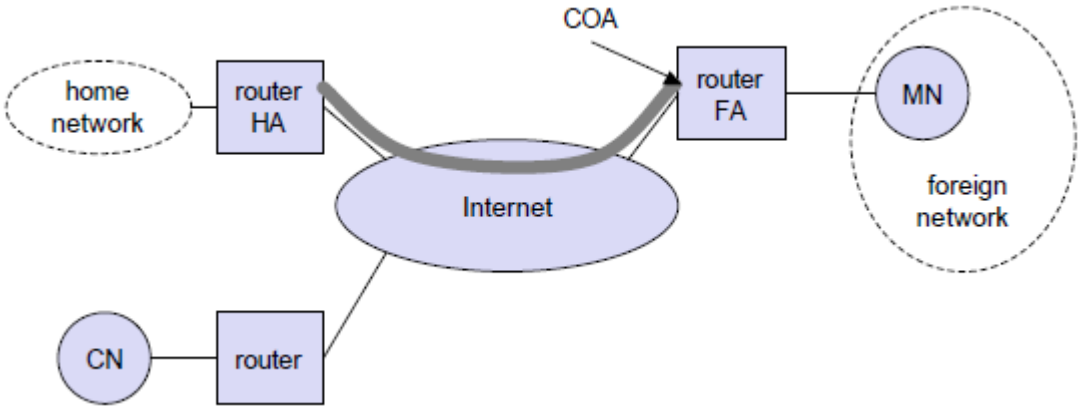


Home agent (HA): The HA provides several services for the MN and is located in the home network. The tunnel for packets toward the MN starts at the HA. The HA maintains a location registry, i.e., it is informed of the MN's location by the current COA. Three alternatives for the implementation of an HA exist.

1. The HA can be implemented on a router that is responsible for the home network. This is obviously the best position, because without optimizations to mobile IP, all packets for the MN have to go through the router anyway.

2. If changing the router's software is not possible, the HA could also be implemented on an arbitrary node in the subnet. One disadvantage of this solution is the double crossing of the router by the packet if the MN is in a foreign network. A packet for the MN comes in via the router; the HA sends it through the tunnel which again crosses the router.

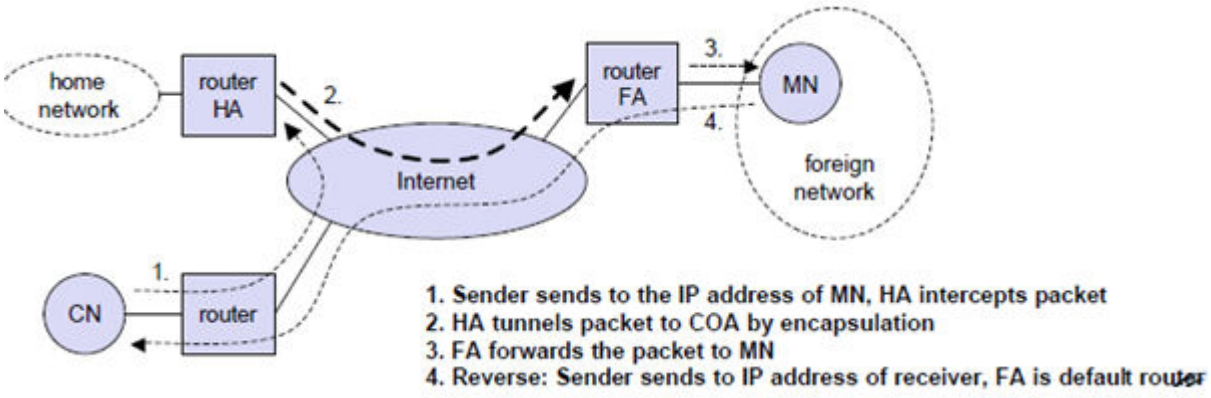
3. Finally, a home network is not necessary at all. The HA could be again on the 'router' but this time only acting as a manager for MNs belonging to a virtual home network. All MNs are always in a foreign network with this solution.



A CN is connected via a router to the internet, as are the home network and the foreign network. The HA is implemented on the router connecting the home network with the internet, an FA is implemented on the router to the foreign network. The MN is currently in the foreign network. The tunnel for packets toward the MN starts at the HA and ends at the FA, for the FA has the COA in the above example.

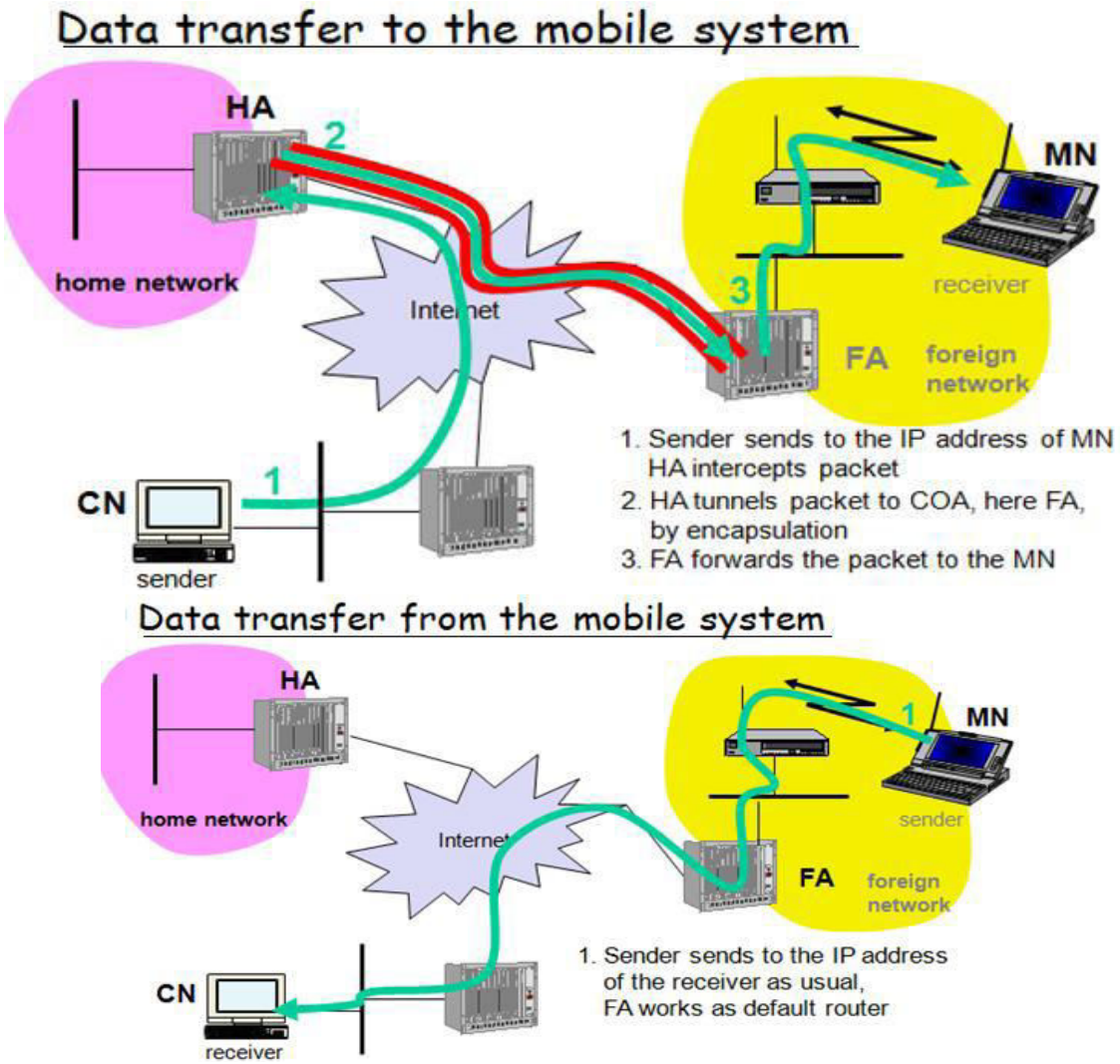
IP packet delivery

Consider the above example in which a correspondent node (CN) wants to send an IP packet to the MN. One of the requirements of mobile IP was to support hiding the mobility of the MN. CN does not need to know anything about the MN's current location and sends the packet as usual to the IP address of MN as shown below.



CN sends an IP packet with MN as a destination address and CN as a source address. The internet, not having information on the current location of MN, routes the packet to the router responsible for the home network of MN. This is done using the standard routing mechanisms of the internet. The HA now intercepts the packet, knowing that MN is currently not in its home network. The packet is not forwarded into the subnet as usual, but encapsulated and tunneled to the COA. A new header is put in front of the old IP header showing the COA as new destination and HA as source of the encapsulated packet (step 2).

The foreign agent now decapsulates the packet, i.e., removes the additional header, and forwards the original packet with CN as source and MN as destination to the MN (step 3). Again, for the MN mobility is not visible. It receives the packet with the same sender and receiver address as it would have done in the home network.



Sending packets from the mobile node (MN) to the CN is comparatively simple. The MN sends the packet as usual with its own fixed IP address as source and CN's address as destination (step 4). The router with the FA acts as default router and forwards the packet in the same way as it would do for any other node in the foreign network. As long as CN is a fixed node the remainder is in the fixed internet as usual. If CN were also a mobile node residing in a foreign network, the same mechanisms as described in steps 1 through 3 would apply now in the other direction.

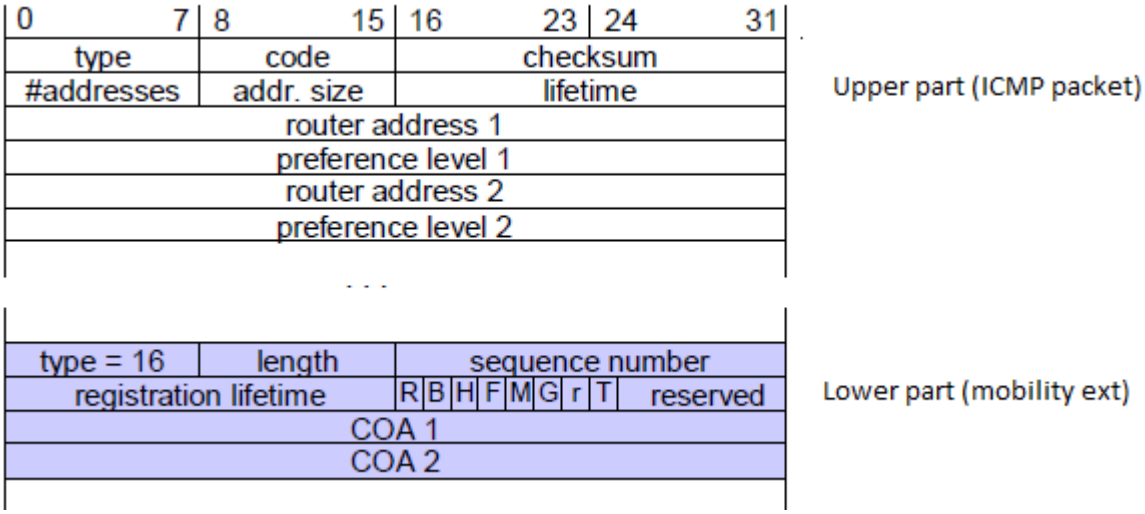
Working of Mobile IP: - Mobile IP has two addresses for a mobile host: one home address and one care-of address. The home address is permanent; the care-of addresses changes as the mobile host moves from one network to another. To make the change of address transparent to the rest of the Internet requires a home agent and a foreign agent. The specific function of an agent is performed in the application layer. When the mobile host and the foreign agent are the same, the care-of address is called a co-located care-of address. To communicate with a remote host, a mobile host goes through three phases: agent discovery, registration, and data transfer.

Agent Discovery

A mobile node has to find a foreign agent when it moves away from its home network. To solve this problem, mobile IP describes two methods: agent advertisement and agent solicitation.

Agent advertisement

For this method, foreign agents and home agents advertise their presence periodically using special **agent advertisement** messages, which are broadcast into the subnet. Mobile IP does not use a new packet type for agent advertisement; it uses the router advertisement packet of ICMP, and appends an agent advertisement message. The agent advertisement packet according to RFC 1256 with the extension for mobility is shown below:



The TTL field of the IP packet is set to 1 for all advertisements to avoid forwarding them. The **type** is set to 9, the **code** can be 0, if the agent also routes traffic from non-mobile nodes, or 16, if it does not route anything other than mobile traffic. The number of addresses advertised with this packet is in **#addresses** while the **addresses** themselves follow as shown. **Lifetime** denotes the length of time this advertisement is valid. **Preference** levels for each address help a node to choose the router that is the most eager one to get a new node.

The extension for mobility has the following fields defined: **type** is set to 16, **length** depends on the number of COAs provided with the message and equals 6 + 4*(number of addresses). The **sequence number** shows the total number of advertisements sent since initialization by the agent. By the **registration lifetime** the agent can specify the maximum lifetime in seconds a node can request during registration. The following bits specify the characteristics of an agent in detail.

The **R** bit (registration) shows, if a registration with this agent is required even when using a colocated COA at the MN. If the agent is currently too busy to accept new registrations it can set the **B** bit. The following two bits denote if the agent offers services as a home agent (**H**) or foreign agent (**F**) on the link where the advertisement has been sent. Bits M and G specify the method of encapsulation used for the tunnel. While IP-in-IP encapsulation is the mandatory standard, **M** can specify minimal encapsulation and **G** generic routing encapsulation. In the first

version of mobile IP (RFC 2002) the **V** bit specified the use of header compression according to RFC 1144. Now the field **r** at the same bit position is set to zero and must be ignored. The new field **T** indicates that reverse tunneling is supported by the FA. The following fields contain the **COAs** advertised. A foreign agent setting the F bit must advertise at least one COA. A mobile node in a subnet can now receive agent advertisements from either its home agent or a foreign agent. This is one way for the MN to discover its location.

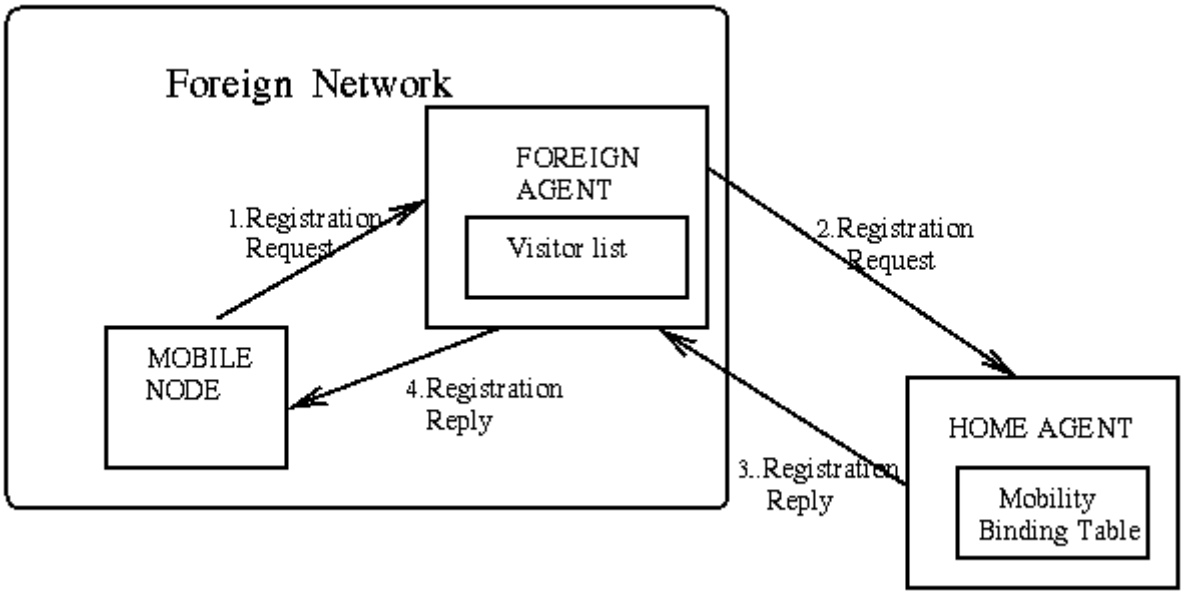
Agent Solicitation

If no agent advertisements are present or the inter-arrival time is too high, and an MN has not received a COA by other means, the mobile node must send **agent solicitations**. Care must be taken to ensure that these solicitation messages do not flood the network, but basically an MN can search for an FA endlessly sending out solicitation messages. If a node does not receive an answer to its solicitations it must decrease the rate of solicitations exponentially to avoid flooding the network until it reaches a maximum interval between solicitations (typically one minute). Discovering a new agent can be done anytime, not just if the MN is not connected to one.

After these steps of advertisements or solicitations the MN can now receive a COA, either one for an FA or a co-located COA.

Agent Registration

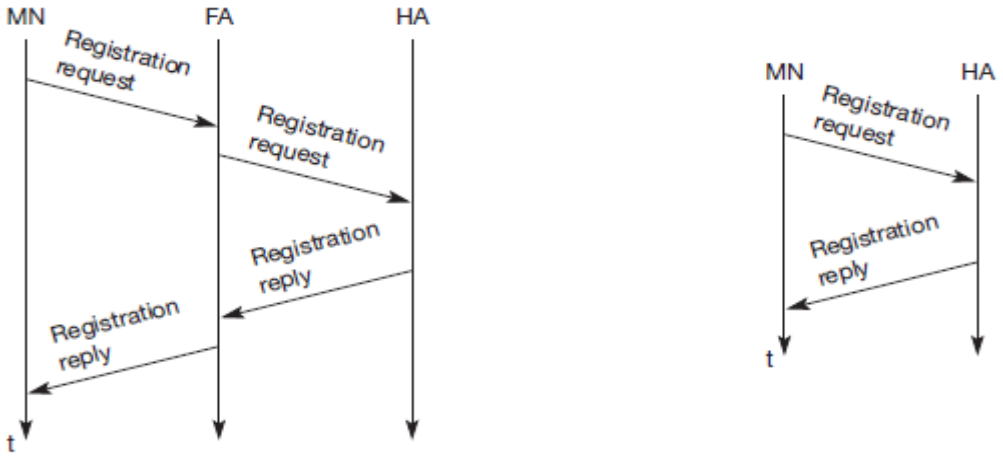
Having received a COA, the MN has to register with the HA. The main purpose of the registration is to inform the HA of the current location for correct forwarding of packets.



Registration can be done in two different ways depending on the location of the COA.

- If the COA is at the FA, the MN sends its registration request containing the COA to the FA which forwards the request to the HA. The HA now sets up a **mobility binding,** containing the mobile node's home IP address and the current COA. It also contains the lifetime of the registration which is negotiated during the registration process. Registration expires automatically after the lifetime and is deleted; so, an MN should reregister before expiration.

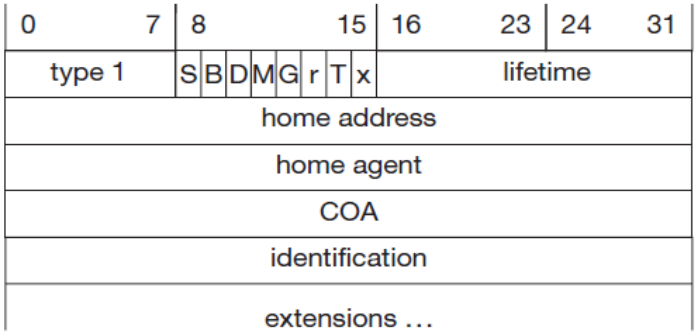
This mechanism is necessary to avoid mobility bindings which are no longer used. After setting up the mobility binding, the HA sends a reply message back to the FA which forwards it to the MN.

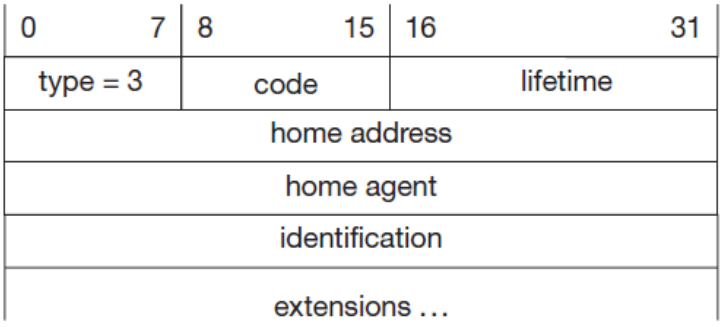


Registration of a mobile node via the FA or directly with the HA

- If the COA is co-located, registration can be simpler, the MN sends the request directly to the HA and vice versa. This is also the registration procedure for MNs returning to their home network to register directly with the HA.

UDP packets are used for the registration requests using the port no 434. The IP source address of the packet is set to the interface address of the MN, the IP destination address is that of the FA or HA.

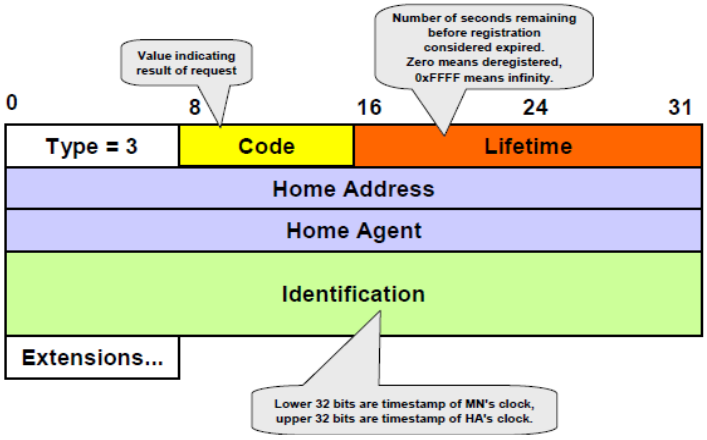




The first field **type** is set to 1 for a registration request. With the **S** bit an MN can specify if it wants the HA to retain prior mobility bindings. This allows for simultaneous bindings. Setting the **B** bit generally indicates that an MN also wants to receive the broadcast packets which have been received by the HA in the home network. If an MN uses a co-located COA, it also takes care of the decapsulation at the tunnel endpoint. The **D** bit indicates this behavior. As already defined for agent advertisements, the bits **M** and **G** denote the use of minimal encapsulation or generic routing encapsulation, respectively. **T** indicates reverse tunneling, **r** and **x** are set to zero.

Lifetime denotes the validity of the registration in seconds. A value of zero indicates deregistration; all bits set indicates infinity. The **home address** is the fixed IP address of the MN, **home agent** is the IP address of the HA, and **COA** represents the tunnel endpoint. The 64 bit **identification** is generated by the MN to identify a request and match it with registration replies. This field is used for protection against replay attacks of registrations. The **extensions** must at least contain parameters for authentication

A **registration reply**, which is conveyed in a UDP packet, contains a **type** field set to 3 and a **code** indicating the result of the registration request.

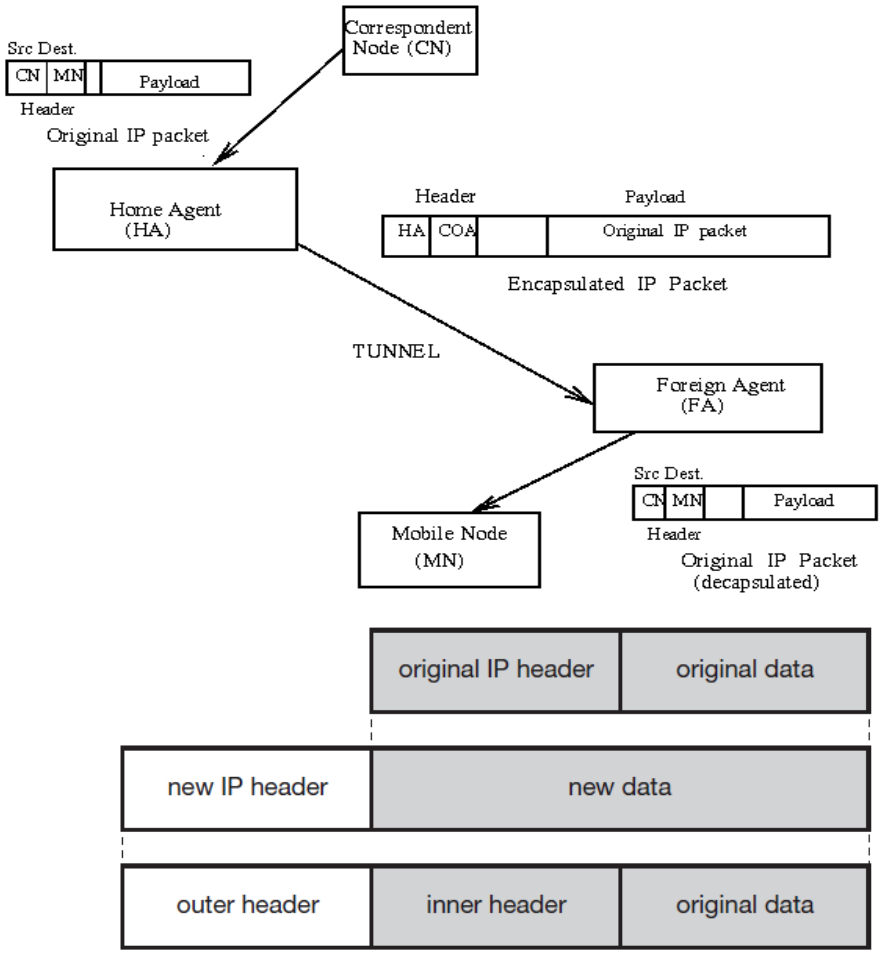


The **lifetime** field indicates how many seconds the registration is valid if it was successful. **Home address** and **home agent** are the addresses of the MN and the HA, respectively. The 64-bit **identification** is used to match registration requests with replies. The value is based on the identification field from the registration and the authentication method. Again, the **extensions** must at least contain parameters for authentication.

Registration	Code	Explanation
successful	0	registration accepted
	1	registration accepted, but simultaneous mobility bindings unsupported
denied by FA	65	administratively prohibited
	66	insufficient resources
	67	mobile node failed authentication
	68	home agent failed authentication
	69	requested lifetime too long
denied by HA	129	administratively prohibited
	130	insufficient resources
	131	mobile node failed authentication
	132	foreign agent failed authentication
	133	registration identification mismatch
	135	too many simultaneous mobility bindings

Tunneling and encapsulation

A **tunnel** establishes a virtual pipe for data packets between a tunnel entry and a tunnel endpoint. Packets entering a tunnel are forwarded inside the tunnel and leave the tunnel unchanged. Tunneling, i.e., sending a packet through a tunnel is achieved by using encapsulation.

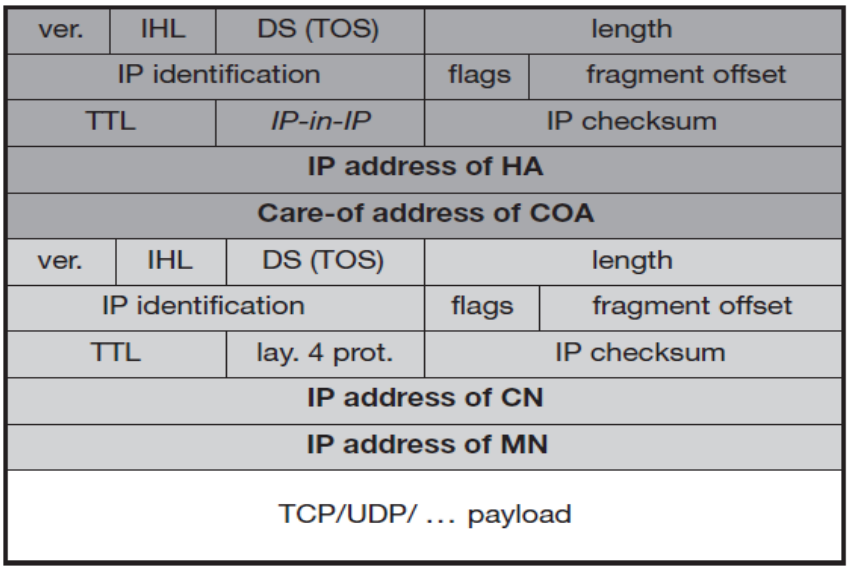


Encapsulation is the mechanism of taking a packet consisting of packet header and data and putting it into the data part of a new packet. The reverse operation, taking a packet out of the data part of another packet, is called **decapsulation**. Encapsulation and decapsulation are the operations typically performed when a packet is transferred from a higher protocol layer to a lower layer or from a lower to a higher layer respectively.

The HA takes the original packet with the MN as destination, puts it into the data part of a new packet and sets the new IP header so that the packet is routed to the COA. The new header is called outer header.

IP-in-IP encapsulation

There are different ways of performing the encapsulation needed for the tunnel between HA and COA. Mandatory for mobile IP is **IP-in-IP encapsulation** as specified in RFC 2003. The following fig shows a packet inside the tunnel.



The version field **ver** is 4 for IP version 4, the internet header length (**IHL**) denotes the length of the outer header in 32 bit words. **DS(TOS)** is just copied from the inner header, the **length** field covers the complete encapsulated packet. The fields up to TTL have no special meaning for mobile IP and are set according to RFC 791. **TTL** must be high enough so the packet can reach the tunnel endpoint. The next field, here denoted with **IP-in-IP**, is the type of the protocol used in the IP payload. This field is set to 4, the protocol type for IPv4 because again an IPv4 packet follows after this outer header. IP **checksum** is calculated as usual. The next fields are the tunnel entry as source address (the **IP address of the HA**) and the tunnel exit point as destination address (the **COA**).

If no options follow the outer header, the inner header starts with the same fields as above. This header remains almost unchanged during encapsulation, thus showing the original sender CN and the receiver MN of the packet. The only change is TTL which is decremented by 1. This means that the whole tunnel is considered a single hop from the original packet's point of view. This is a very important feature of tunneling as it allows the MN to behave as if it were attached to the home network. No matter how many real hops the packet has to take in the tunnel, it is just one (logical) hop away for the MN. Finally, the payload follows the two headers.

Minimal encapsulation

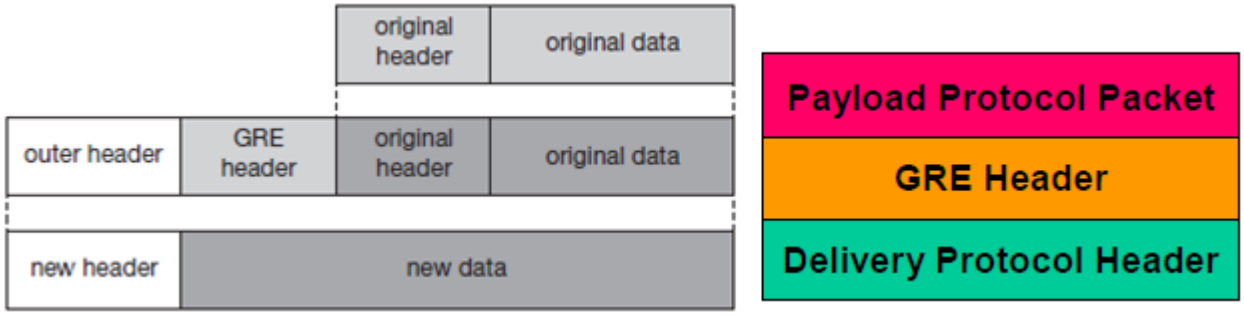
Minimal encapsulation (RFC 2004) as shown below is an optional encapsulation method for mobile IP which avoids repetitions of identical fields in IP-in-IP encapsulation. The tunnel entry point and endpoint are specified.

ver.	IHL	DS (TOS)	length
IP identification		flags	fragment offset
TTL		*min. encap*	IP checksum
IP address of HA			
care-of address of COA			
lay. 4 protoc.	S	reserved	IP checksum
IP address of MN			
original sender IP address (if S=1)			
TCP/UDP/ … payload			

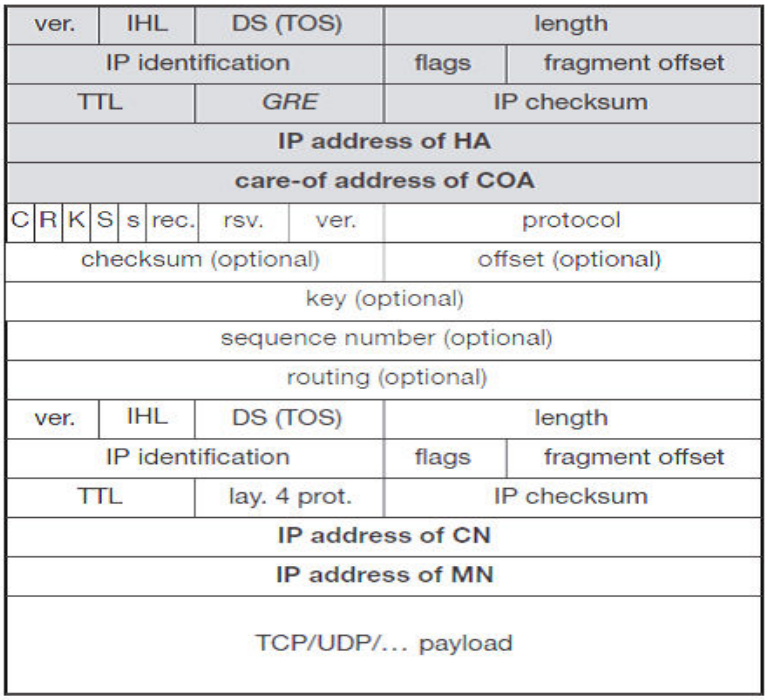
The field for the type of the following header contains the value 55 for the minimal encapsulation protocol. The inner header is different for minimal encapsulation. The type of the following protocol and the address of the MN are needed. If the **S** bit is set, the original sender address of the CN is included as omitting the source is quite often not an option. No field for fragmentation offset is left in the inner header and minimal encapsulation does not work with already fragmented packets.

Generic Routing Encapsulation

Unlike IP-in-IP and Minimal encapsulation which work only for IP packets, **Generic routing encapsulation** (GRE) allows the encapsulation of packets of one protocol suite into the payload portion of a packet of another protocol suite as shown below.



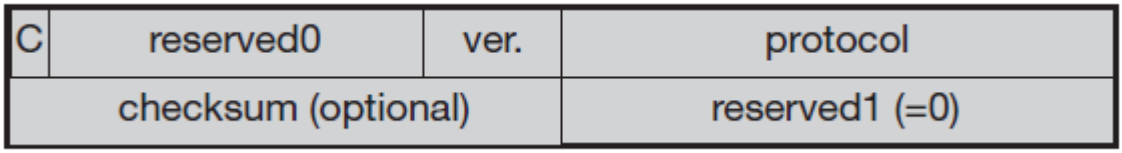
The packet of one protocol suite with the original packet header and data is taken and a new GRE header is prepended. Together this forms the new data part of the new packet. Finally, the header of the second protocol suite is put in front. The following figure shows the fields of a packet inside the tunnel between HA and COA using GRE as an encapsulation scheme according to RFC 1701. The outer header is the standard IP header with HA as source address and COA as destination address. The protocol type used in this outer IP header is 47 for GRE.



The GRE header starts with several flags indicating if certain fields are present or not. A minimal GRE header uses only 4 bytes. The **C** bit indicates if the checksum field is present and contains valid information. If **C** is set, the **checksum** field contains a valid IP checksum of the GRE header and the payload. The **R** bit indicates if the offset and routing fields are present and contain valid information. The **offset** represents the offset in bytes for the first source **routing** entry. The routing field, if present, has a variable length and contains fields for source routing. GRE also offers a **key** field which may be used for authentication. If this field is present, the **K** bit is set. The sequence number bit **S** indicates if the **sequence** number field is present, if the s bit is set, strict source routing is used.

The **recursion control** field (rec.) is an important field that additionally distinguishes GRE from IP-in-IP and minimal encapsulation. This field represents a counter that shows the number of allowed recursive encapsulations. The default value of this field should be 0, thus allowing only one level of encapsulation. The following **reserved** fields must be zero and are ignored on reception. The **version** field contains 0 for the GRE version. The following 2 byte **protocol** field represents the protocol of the packet following the GRE header. The standard header of the original packet follows with the source address of the correspondent node and the destination address of the mobile node.

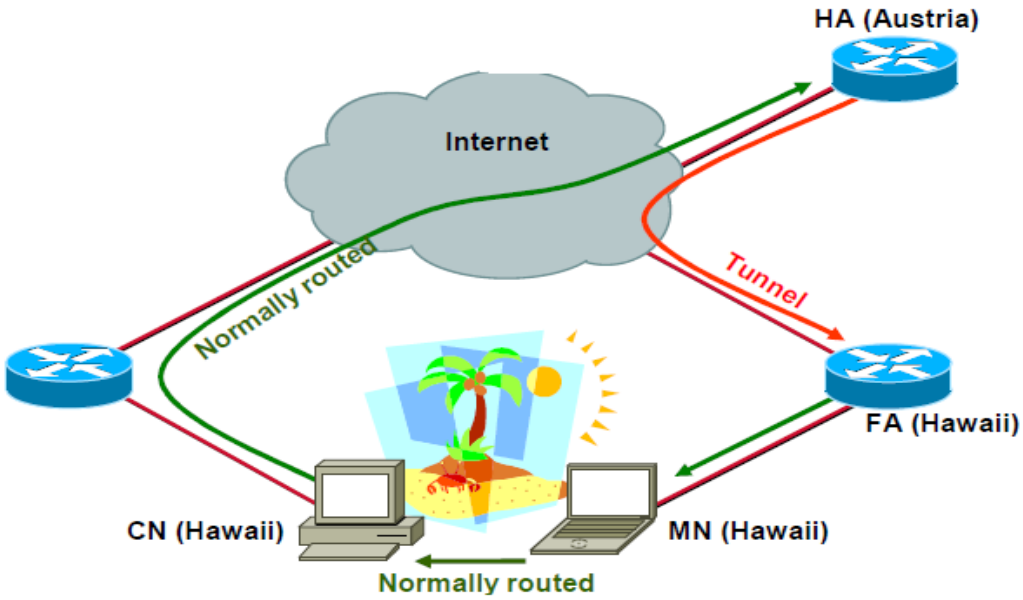
A simplified header of GRE following RFC 2784 is shown below.



The field **C** indicates again if a checksum is present. The next 5 bits are set to zero, then 7 reserved bits follow. The **version** field contains the value zero. The **protocol** type, again, defines the protocol of the payload following RFC 3232. If the flag C is set, then **checksum** field and a field called reserved1 follows. The latter field is constant zero set to zero follow.

Optimizations:

If a scenario occurs, where if the MN is in the same subnetwork as the node to which it is communicating and HA is on the other side of the world. It is called triangular routing problem as it causes unnecessary overheads for the network between CN and the HA.



A solution to this problem is to inform the CN of the current location of the MN. The CN can learn the location by caching it in a binding cache, which is a part of the routing table for the CN. HA informs the CN of the location. It needs four additional messages:

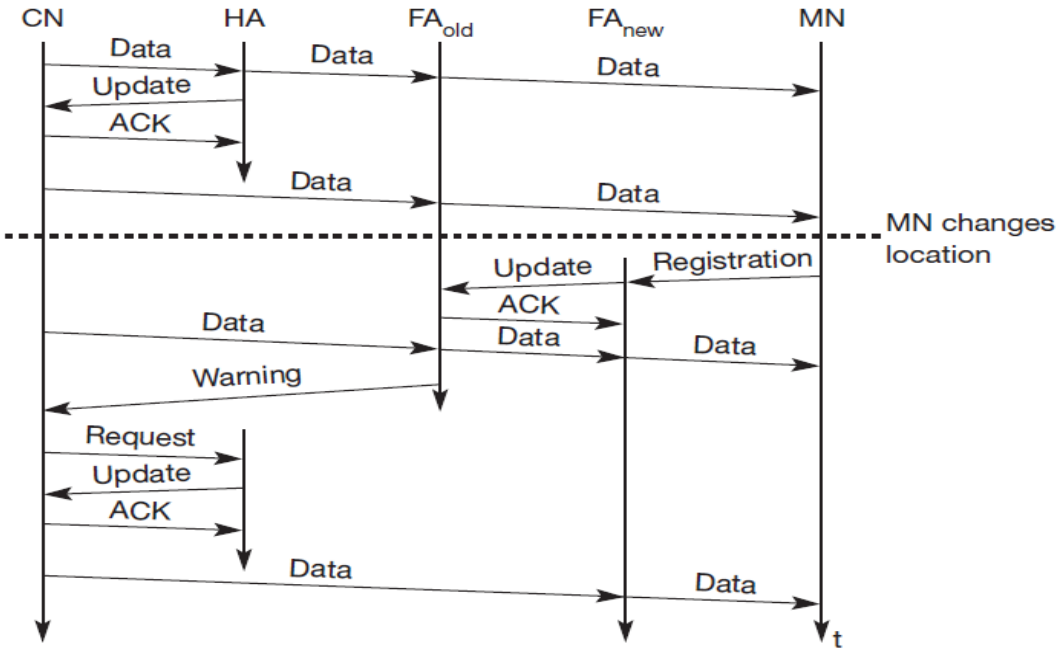
Binding Request: It is sent by the node that wants to know the current location of an MN to the HA. HA checks if it is allowed to reveal the location and then sends back a binding update

Binding update: It is sent by the HA to the CN revealing the current location of an MN. It contains the fixed IP address of the MN and the COA. This message can request an acknowledgement.

Binding acknowledgement: If requested, a node returns this acknowledgement after receiving a binding update message

Binding warning: A node sends a binding warning if it decapsulates a packet for an MN, but it is note the current FA of this MN. It contains MN's home address and a target nodes address. The recipient can be the HA, so the HA now sends a binding update to the node that obviously has a wrong COA for the MN.

The following figure shows how the four additional messages are used together if an MN changes its FA.



The CN can request the current location from the HA. If allowed by the MN, the HA returns the COA of the MN via an update message. The CN acknowledges this update message and stores the mobility binding. Now the CN can send its data directly to the current foreign agent FAold. FAold forwards the packets to the MN. This scenario shows a COA located at an FA. Encapsulation of data for tunneling to the COA is now done by the CN, not the HA.

The MN might now change its location and register with a new foreign agent, FA_{new}. This registration is also forwarded to the HA to update its location database. Furthermore, FA_{new} informs FA_{old} about the new registration of MN. MN's registration message contains the address of FA_{old} for this purpose. Passing this information is achieved via an update message, which is acknowledged by $FA_{old.}$

Without the information provided by the new FA, the old FA would not get to know anything about the new location of MN. In this case, CN does not know anything about the new location, so it still tunnels its packets for MN to the old FA, FA_{old}. This FA now notices packets with destination MN, but also knows that it is not the current FA of MN. FA_{old} might now forward these packets to the new COA of MN which is FA_{new} in this example. This forwarding of packets is another optimization of the basic Mobile IP providing **smooth handovers**. Without this optimization, all packets in transit would be lost while the MN moves from one FA to another.

To tell CN that it has a stale binding cache, FA_{old} sends, a binding warning message to CN. CN then requests a binding update. (The warning could also be directly sent to the HA triggering an update). The HA sends an update to inform the CN about the new location, which is acknowledged. Now CN can send its packets directly to FA_{new}, again avoiding triangular routing. Unfortunately, this optimization of mobile IP to avoid triangular routing causes several security problems.

Reverse Tunneling

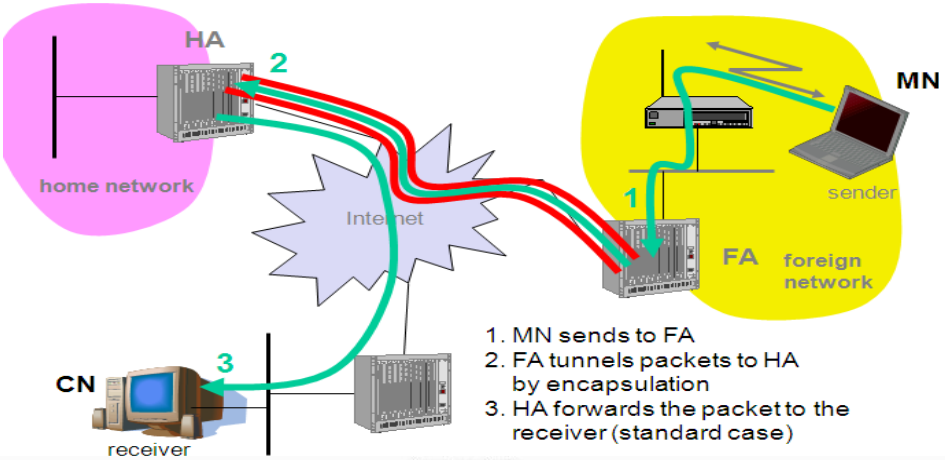
The reverse path from MS to the CN looks quite simple as the MN can directly send its packets to the CN as in any other standard IP situation. The destination address in the packets is that of CN. But it has some problems explained below:-

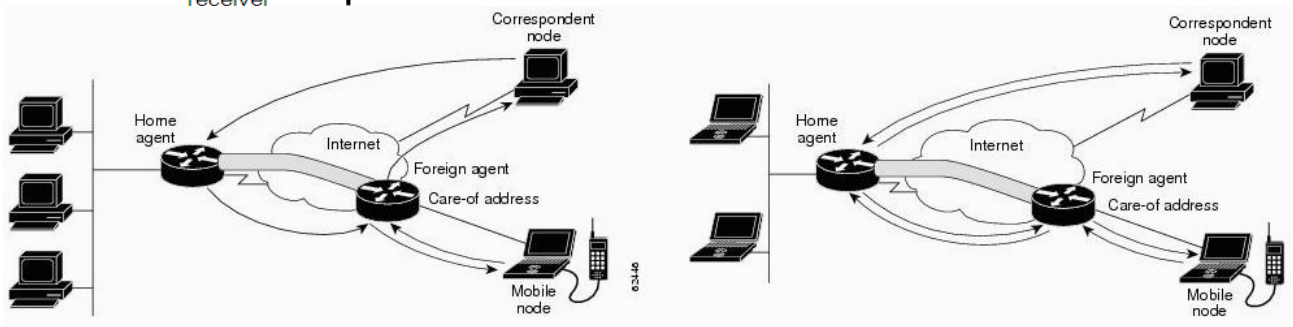
Quite often firewalls are designed to only allow packets with topologically correct addresses to pass to provide simple protection against misconfigured systems of unknown addresses. However, MN still sends packets with its fixed IP address as source which is not topologically correct in a foreign network. Firewalls often filter packets coming from outside containing a source address from computers of the internal network. This also implies that an MN cannot send a packet to a computer residing in its home network.

While the nodes in the home network might participate in a multi-cast group, an MN in a foreign network cannot transmit multi-cast packets in a way that they emanate from its home network without a reverse tunnel. The foreign network might not even provide the technical infrastructure for multi-cast communication (multi-cast backbone, Mbone).

If the MN moves to a new foreign network, the older TTL might be too low for the packets to reach the same destination nodes as before. Mobile IP is no longer transparent if a user has to adjust the TTL while moving. A reverse tunnel is needed that represents only one hop, no matter how many hops are really needed from the foreign to the home network

Based on the above considerations, reverse tunneling is defined as an extension to mobile IP (per RFC 2344). It was designed backward compatible to mobile IP and defines topologically correct reverse tunneling to handle the above stated problems.





Reverse tunneling does not solve:

Problems with *firewalls*, the reverse tunnel can be abused to circumvent security mechanisms (tunnel hijacking)

Optimization of data paths, i.e. packets will be forwarded through the tunnel via the HA to a sender (double triangular routing)

IPv6

The design of Mobile IP support in IPv6 (Mobile IPv6) benefits both from the experiences gained from the development of Mobile IP support in IPv4, and from the opportunities provided by IPv6. Mobile IPv6 thus shares many features with Mobile IPv4, but is integrated into IPv6 and offers many other improvements. This section summarizes the major differences between Mobile IPv4 and Mobile IPv6:

There is no need to deploy special routers as "foreign agents", as in Mobile IPv4. Mobile IPv6 operates in any location without any special support required from the local router.

Support for route optimization is a fundamental part of the protocol, rather than a nonstandard set of extensions.

Mobile IPv6 route optimization can operate securely even without pre-arranged security associations. It is expected that route optimization can be deployed on a global scale between all mobile nodes and correspondent nodes.

Support is also integrated into Mobile IPv6 for allowing route optimization to coexist efficiently with routers that perform "ingress filtering"

The IPv6 Neighbor Unreachability Detection assures symmetric reachability between the mobile node and its default router in the current location.

Most packets sent to a mobile node while away from home in Mobile IPv6 are sent using an IPv6 routing header rather than IP encapsulation, reducing the amount of resulting overhead compared to Mobile IPv4.

Mobile IPv6 is decoupled from any particular link layer, as it uses IPv6 Neighbor Discovery instead of ARP. This also improves the robustness of the protocol.

The use of IPv6 encapsulation (and the routing header) removes the need in Mobile IPv6 to manage "tunnel soft state".

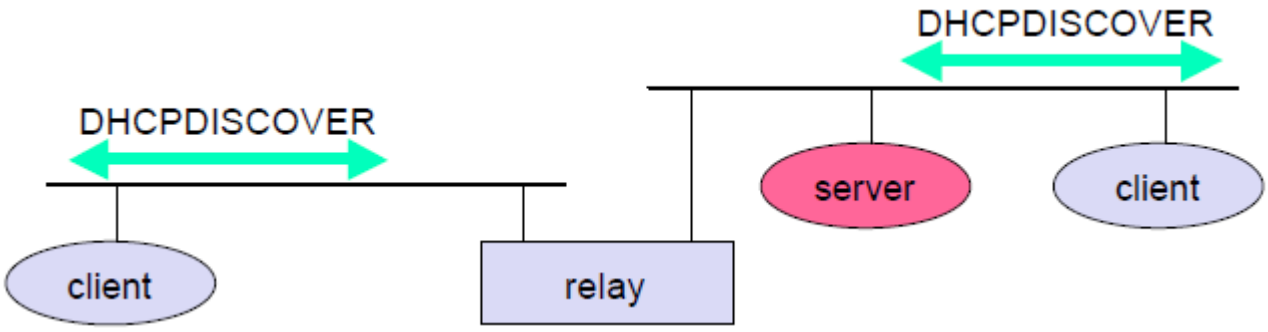
The dynamic home agent address discovery mechanism in Mobile IPv6 returns a single reply to the mobile node. The directed broadcast approach used in IPv4 returns separate replies from each home agent.

Dynamic Host Configuration Protocol (DHCP)

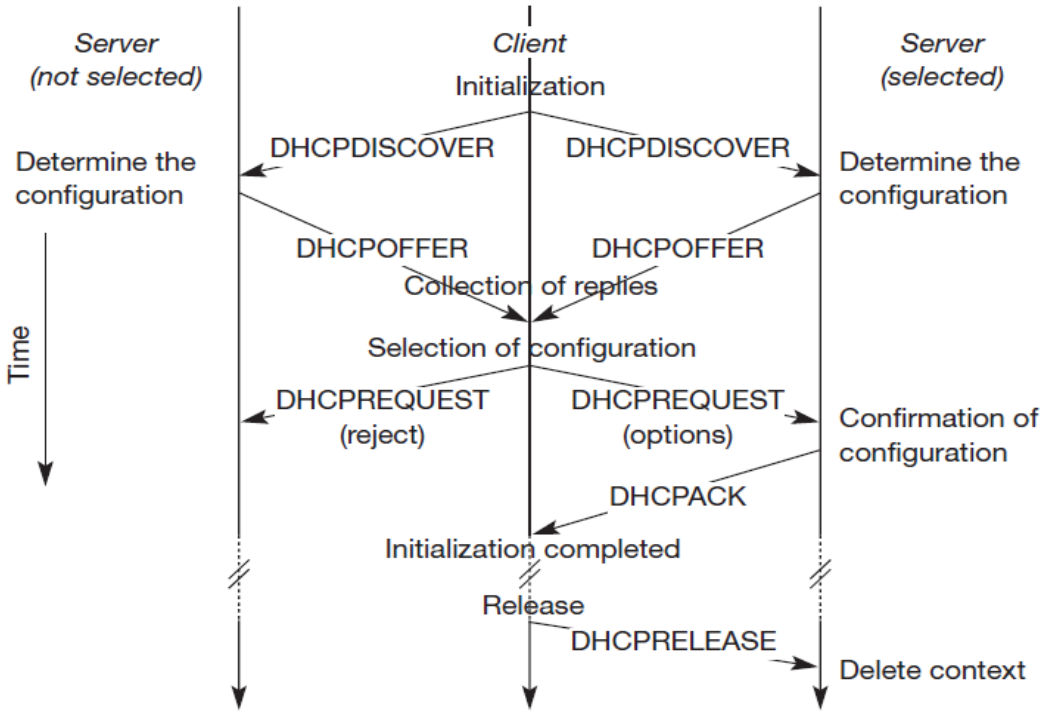
DHCP is an automatic configuration protocol used on IP networks. **DHCP** allows a computer to join an IP-based network without having a pre-configured IP address. DHCP is a protocol that assigns unique IP addresses to devices, then releases and renews these addresses as devices leave and re-join the network. If a new computer is connected to a network, DHCP can

provide it with all the necessary information for full system integration into the network, e.g., addresses of a DNS server and the default router, the subnet mask, the domain name, and an IP address. Providing an IP address makes DHCP very attractive for mobile IP as a source of care-of-addresses.

DHCP 20 **Mukesh Chinta Asst Prof, CSE, VNRVJIET** DHCP is based on a client/server model as shown below. DHCP clients send a request to a server (DHCPDISCOVER in the example) to which the server responds. A client sends requests using MAC broadcasts to reach all devices in the LAN. A DHCP relay might be needed to forward requests across inter-working units to a DHCP server.



Consider the scenario where there is one client and two servers are present. A typical initialization of a DHCP client is shown below:



The client broadcasts a DHCPDISCOVER into the subnet. There might be a relay to forward this broadcast. In the case shown, two servers receive this broadcast and determine the configuration they can offer to the client. Servers reply to the client's request with DHCPOFFER and offer a list of configuration parameters. The client can now choose one of the configurations offered. The client in turn replies to the servers, accepting one of the configurations and rejecting the others using DHCPREQUEST.

If a server receives a DHCPREQUEST with a rejection, it can free the reserved configuration for other possible clients. The server with the configuration accepted by the client now confirms the configuration with DHCPACK. This completes the initialization phase. If a client leaves a subnet, it should release the configuration received by the server using DHCPRELEASE. Now the server can free the context stored for the client and offer the configuration again. The configuration a client gets from a server is only leased for a certain amount of time; it has to be reconfirmed from time to time. Otherwise the server will free the configuration. This timeout of configuration helps in the case of crashed nodes or nodes moved away without releasing the context.

DHCP is a good candidate for supporting the acquisition of care-of addresses for mobile nodes. The same holds for all other parameters needed, such as addresses of the default router, DNS servers, the timeserver etc. A DHCP server should be located in the subnet of the access point of the mobile node, or at least a DHCP relay should provide forwarding of the messages. RFC 3118 specifies authentication for DHCP messages so as to provide protection from malicious DHCP servers. Without authentication, a DHCP server cannot trust the mobile node and vice versa…

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

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

MANET- Characteristics

- Dynamic network topology
- Bandwidth constraints and variable link capacity
- Energy constrained nodes
- Multi-hop communications

 Limited security

 Autonomous terminal

 Distributed operation

 Light-weight terminals

Need for Ad Hoc Networks

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

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

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

Properties of MANETs

 MANET enables fast establishment of networks. When anew network is to be established, the only requirement is to provide a new set of nodes with limited wireless communication range. A node has limited capability, that is, it can connect only to the nodes which are nearby. Hence it consumes limited power.

 A MANET node has the ability to discover a neighboring node and service. Using a service discovery protocol, a node discovers the service of a nearby node and communicates to a remote node in the MANET.

 MANET nodes have peer-to-peer connectivity among themselves.

 MANET nodes have independent computational, switching (or routing), and communication capabilities.

 The wireless connectivity range in MANETs includes only nearest node connectivity.

 The failure of an intermediate node results in greater latency in communicating with the remote server.

 Limited bandwidth available between two intermediate nodes becomes a constraint for the MANET. The node may have limited power and thus computations need to be energy-efficient.

 There is no access-point requirement in MANET. Only selected access points are provided for connection to other networks or other MANETs.

 MANET nodes can be the iPods, Palm handheld computers, Smartphones, PCs, smart labels, smart sensors, and automobile-embedded systems\

 MANET nodes can use different protocols, for example, IrDA, Bluetooth, ZigBee, 802.11, GSM, and TCP/IP.MANET node performs data caching, saving, and aggregation.

 MANET mobile device nodes interact seamlessly when they move with the nearby wireless nodes, sensor nodes, and embedded devices in automobiles so that the seamless connectivity is maintained between the devices.

Transmission Control Protocol (TCP

The **Transmission Control Protocol** (**TCP**) is one of the core protocols of the Internet protocol suite, often simply referred to as TCP/IP. TCP is reliable, guarantees in-order delivery of data and incorporates congestion control and flow control mechanisms.

TCP supports many of the Internet's most popular application protocols and resulting applications, including the World Wide Web, e-mail, File Transfer Protocol and Secure Shell. In the Internet protocol suite, TCP is the intermediate layer between the Internet layer and application layer.

The major responsibilities of TCP in an active session are to:

• **Provide reliable in-order transport of data**: to not allow losses of data.

• **Control congestions in the networks**: to not allow degradation of the network performance,

• **Control a packet flow between the transmitter and the receiver**: to not exceed the receiver's capacity.

TCP uses a number of mechanisms to achieve high performance and avoid 'congestion collapse', where network performance can fall by several orders of magnitude. These mechanisms control the rate of data entering the network, keeping the data flow below a rate that would trigger collapse. There are several mechanisms of TCP that influence the efficiency of TCP in a mobile environment. Acknowledgments for data sent, or lack of acknowledgments, are used by senders to implicitly interpret network conditions between the TCP sender and receiver.

The five-layer TCP/IP model
5. Application layer
DHCP • DNS • FTP • Gopher • HTTP • IMAP4 • IRC • NNTP • XMPP • MIME • POP3 • SIP • SMTP • SNMP • SSH • TELNET • RPC • RTP • RTCP • TLS/SSL • SDP • SOAP • VPN • PPTP • L2TP • GTP • ...
4. Transport layer
TCP • UDP • DCCP • SCTP • ...
3. Internet layer
IP (IPv4 • IPv6) • IGMP • ICMP • RSVP • BGP • RIP • OSPF • ISIS • IPsec • ARP • RARP • ...
2. Data link layer
802.11 • ATM • DTM • Ethernet • FDDI • Frame Relay • GPRS • EVDO • HSPA • HDLC • PPP • ...
1. Physical layer
Ethernet physical layer • ISDN • Modems • PLC • SONET/SDH • G.709 • WiMAX • ...

Congestion Control

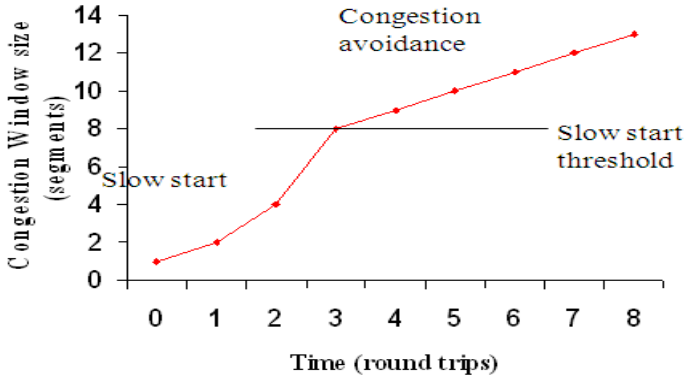
A transport layer protocol such as TCP has been designed for fixed networks with fixed end-systems. Congestion may appear from time to time even in carefully designed networks. The packet buffers of a router are filled and the router cannot forward the packets fast enough because the sum of the input rates of packets destined for one output link is higher than the capacity of the output link. The only thing a router can do in this situation is to drop packets.

A dropped packet is lost for the transmission, and the receiver notices a gap in the packet stream. Now the receiver does not directly tell the sender which packet is missing, but continues to acknowledge all in-sequence packets up to the missing one. The sender notices the missing acknowledgement for the lost packet and assumes a packet loss due to congestion. Retransmitting the missing packet and continuing at full sending rate would now be unwise, as this might only increase the congestion.

To mitigate congestion, TCP slows down the transmission rate dramatically. All other TCP connections experiencing the same congestion do exactly the same so the congestion is soon resolved. Slow start TCP's reaction to a missing acknowledgement is quite drastic, but it is necessary to get rid of congestion quickly. The behavior TCP shows after the detection of congestion is called **slow start.** The sender always calculates a **congestion window** for a receiver. The start size of the congestion window is one segment (TCP packet).

The sender sends one packet and waits for acknowledgement. If this acknowledgement arrives, the sender increases the congestion window by one, now sending two packets (congestion window = 2). This scheme doubles the congestion window every time the acknowledgements come back, which takes one round trip time (RTT). This is called the exponential growth of the congestion window in the slow start mechanism.

But doubling the congestion window is too dangerous. The exponential growth stops at the **congestion threshold**. As soon as the congestion window reaches the congestion threshold, further increase of the transmission rate is only linear by adding 1 to the congestion window each time the acknowledgements come back.



Linear increase continues until a time-out at the sender occurs due to a missing acknowledgement, or until the sender detects a gap in transmitted data because of continuous acknowledgements for the same packet. In either case the sender sets the congestion threshold to half of the current congestion window. The congestion window itself is set to one segment and the

sender starts sending a single segment. The exponential growth starts once more up to the new congestion threshold, then the window grows in linear fashion.

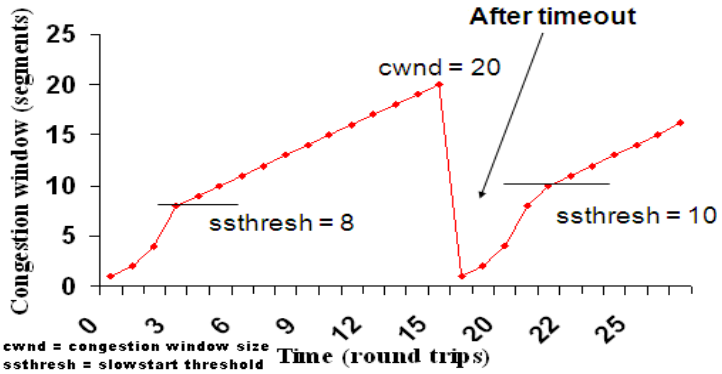
Fast retransmit/fast recovery

The congestion threshold can be reduced because of two reasons. First one is if the sender receives continuous acknowledgements for the same packet. It informs the sender that the receiver has got all the packets upto the acknowledged packet in the sequence and also the receiver is receiving something continuously from the sender. The gap in the packet stream is not due to congestion, but a simple packet loss due to a transmission error. The sender can now retransmit the missing packet(s) before the timer expires. This behavior is called **fast retransmit**. It is an early enhancement for preventing slow-start to trigger on losses not caused by congestion. The receipt of acknowledgements shows that there is no congestion to justify a slow start.

The sender can continue with the current congestion window. The sender performs a **fast recovery** from the packet loss. This mechanism can improve the efficiency of TCP dramatically. The other reason for activating slow start is a time-out due to a missing acknowledgement. TCP using fast retransmit/fast recovery interprets this congestion in the network and activates the slow start mechanism.

The advantage of this method is its simplicity. Minor changes in the MH's software results in performance increase. No changes are required in FA or CH.

The disadvantage of this scheme is insufficient isolation of packet losses. It mainly focuses on problems regarding Handover. Also it effects the efficiency when a CH transmits already delivered packets.



Problems with Traditional TCP in wireless environments

Slow Start mechanism in fixed networks decreases the efficiency of TCP if used with mobile receivers or senders.

Error rates on wireless links are orders of magnitude higher compared to fixed fiber or copper links. This makes compensation for packet loss by TCP quite difficult.

Mobility itself can cause packet loss. There are many situations where a soft handover from one access point to another is not possible for a mobile end-system.

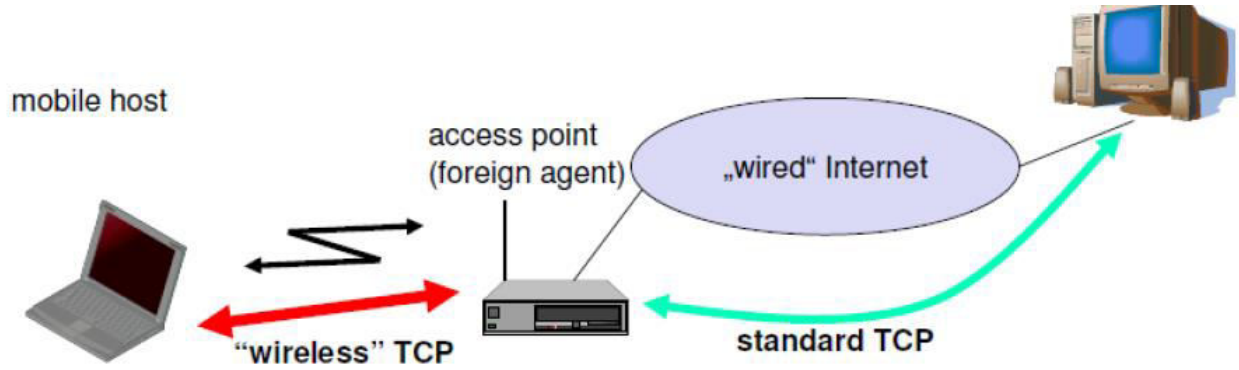
Standard TCP reacts with slow start if acknowledgements are missing, which does not help in the case of transmission errors over wireless links and which does not really help during handover. This

behavior results in a severe performance degradation of an unchanged TCP if used together with wireless links or mobile nodes

Classical TCP Improvements:

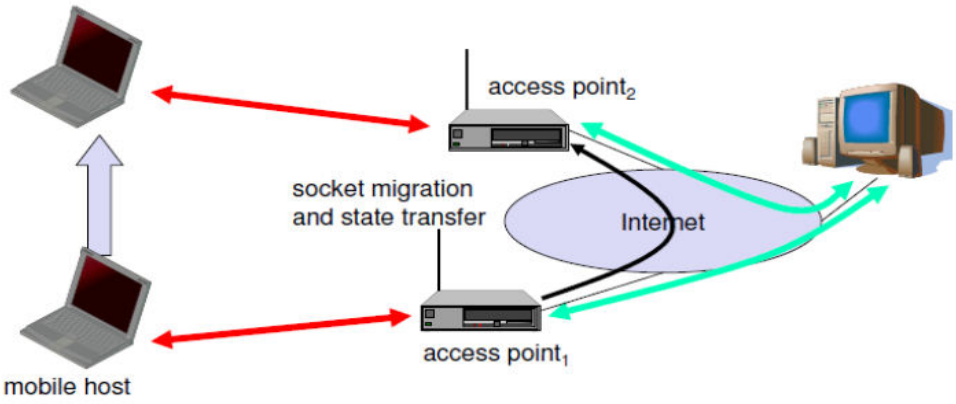
Indirect TCP (I-TCP)

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



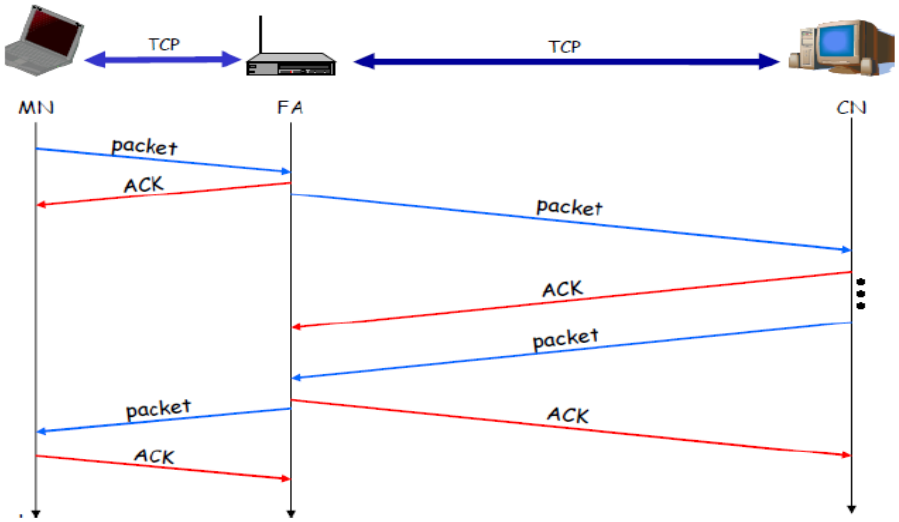
Standard TCP is used between the fixed computer and the access point. No computer in the internet recognizes any changes to TCP. Instead of the mobile host, the access point now terminates the standard TCP connection, acting as a proxy. This means that the access point is now seen as the mobile host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

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



Socket and state migration after handover of a mobile host

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



Advantages of I-TCP

- No changes in the fixed network necessary, no changes for the hosts (TCP protocol) necessary, all current optimizations to TCP still work
- Simple to control, mobile TCP is used only for one hop between, e.g., a foreign agent and mobile host

1. transmission errors on the wireless link do not propagate into the fixed network
2. therefore, a very fast retransmission of packets is possible, the short delay on the mobile hop s known

- It is always dangerous to introduce new mechanisms in a huge network without knowing exactly how they behave.

- New optimizations can be tested at the last hop, without jeopardizing the stability of the Internet.
- It is easy to use different protocols for wired and wireless networks.

Disadvantages of I-TCP

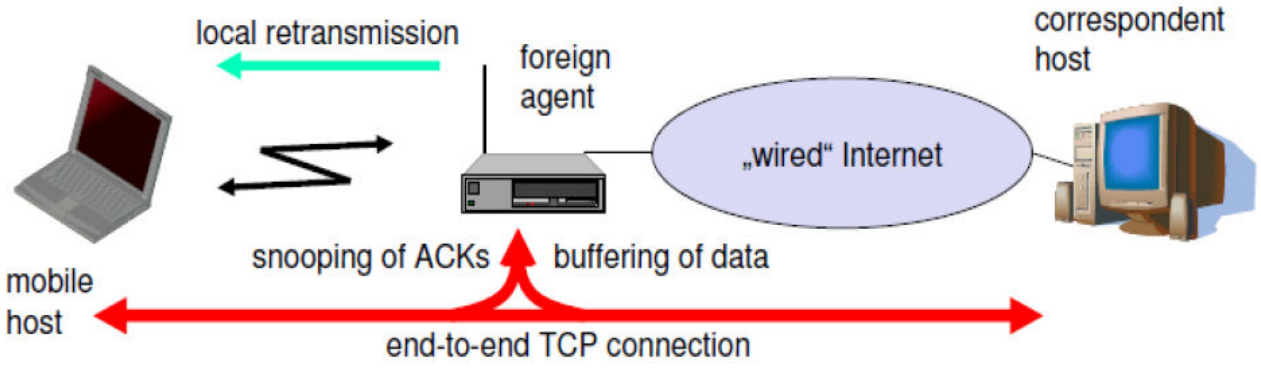
Loss of end-to-end semantics: - an acknowledgement to a sender no longer means that a receiver really has received a packet, foreign agents might crash.

Higher latency possible: - due to buffering of data within the foreign agent and forwarding to a new foreign agent

Security issue: - The foreign agent must be a trusted entity

Snooping TCP:

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

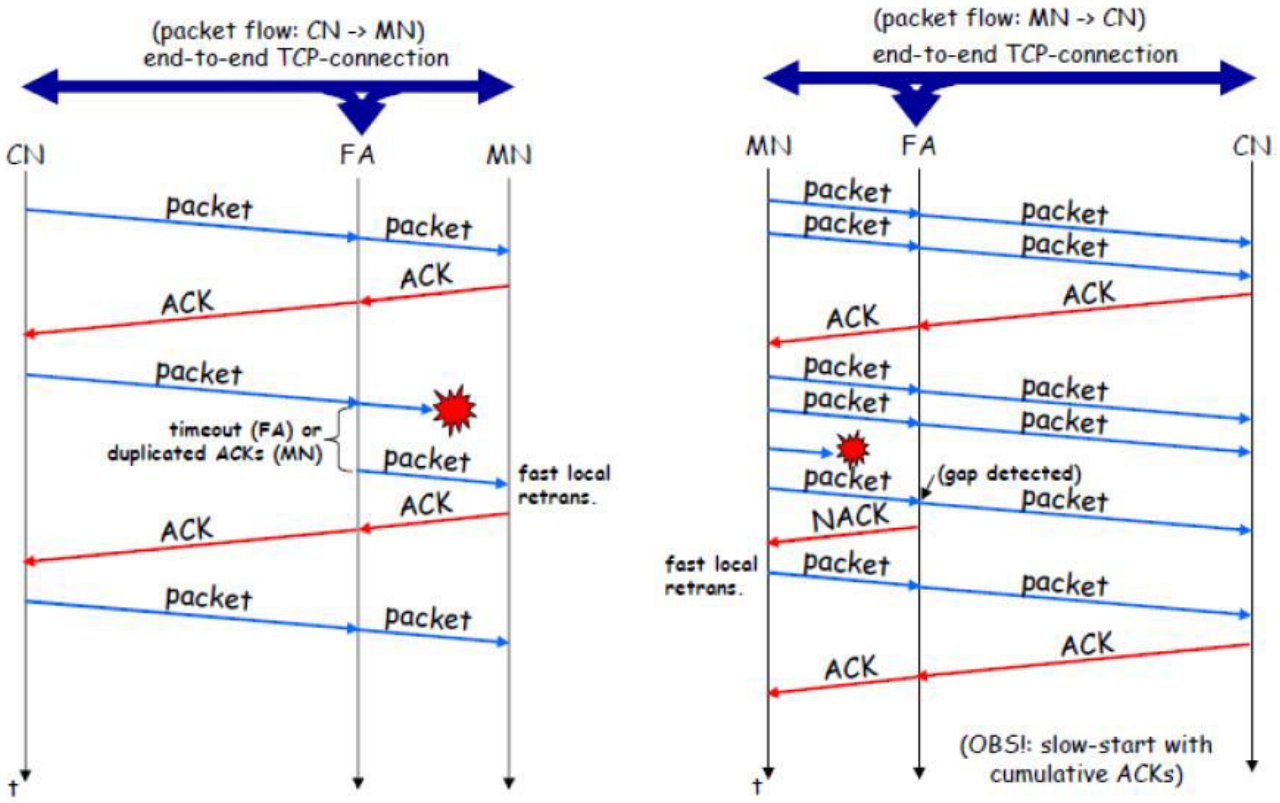


Snooping TCP as a transparent TCP extension

Here, the foreign agent buffers all packets with destination mobile host and additionally ‘snoops’ the packet flow in both directions to recognize acknowledgements. The foreign agent buffers every packet until it receives an acknowledgement from the mobile host. If the FA does not receive an acknowledgement from the mobile host within a certain amount of time, either the packet or the acknowledgement has been lost. Alternatively, the foreignagent could receive a duplicate ACK which also shows the loss of a packet. Now, the FAretransmits the packet directly from the buffer thus performing a faster retransmissioncompared to the CH. For transparency, the FA does not acknowledge data to the CH, whichwould violate end-to-end semantic in case of a FA failure. The foreign agent can filter theduplicate acknowledgements to avoid unnecessary retransmissions of data from thecorrespondent host. If the foreign agent now crashes, the time-out of the correspondent hoststill works and triggers a retransmission. The foreign agent may discard duplicates of packets

already retransmitted locally and acknowledged by the mobile host. This avoids unnecessary traffic on the wireless link.

For data transfer from the mobile host with destination correspondent host, the FA snoops into the packet stream to detect gaps in the sequence numbers of TCP. As soon as the foreign agent detects a missing packet, it returns a negative acknowledgement (NACK) to the mobile host. The mobile host can now retransmit the missing packet immediately. Reordering of packets is done automatically at the correspondent host by TCP.



Snooping TCP: Packet delivery

Advantages of snooping TCP:

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

Disadvantages of snooping TCP

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

Mobile TCP:

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

The SH monitors all packets sent to the MH and ACKs returned from the MH. If the SH does not receive an ACK for some time, it assumes that the MH is disconnected. It then chokes the sender by setting the sender's window size to 0. Setting the window size to 0 forces the sender to go into **persistent mode**, i.e., the state of the sender will not change no matter how long the receiver is disconnected. This means that the sender will not try to retransmit data. As soon as the SH (either the old SH or a new SH) detects connectivity again, it reopens the window of the sender to the old value. The sender can continue sending at full speed. This mechanism does not require changes to the sender's TCP. The wireless side uses an adapted TCP that can recover from packet loss much faster. This modified TCP does not use slow start, thus, M-TCP needs a **bandwidth manager** to implement fair sharing over the wireless link.

Advantages of M-TCP:

- It maintains the TCP end-to-end semantics. The SH does not send any ACK itself but forwards the ACKs from the MH.
- If the MH is disconnected, it avoids useless retransmissions, slow starts or breaking connections by simply shrinking the sender's window to 0.
- As no buffering is done as in I-TCP, there is no need to forward buffers to a new SH. Lost packets will be automatically retransmitted to the SH.

Disadvantages of M-TCP:

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

Transmission/time-out freezing

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

As soon as the MAC layer detects connectivity again, it signals TCP that it can resume operation at exactly the same point where it had been forced to stop. For TCP time simply does not advance, so no timers expire.

Advantages:

- It offers a way to resume TCP connections even after long interruptions of the connection.
- It can be used together with encrypted data as it is independent of other TCP mechanisms such as sequence no or acknowledgements

Disadvantages:

Lots of changes have to be made in software of MH, CH and FA.

Selective retransmission:

A very useful extension of TCP is the use of selective retransmission. TCP acknowledgements are cumulative, i.e., they acknowledge in-order receipt of packets up to a certain packet. A single acknowledgement confirms reception of all packets upto a certain packet. If a single packet is lost, the sender has to retransmit everything starting from the lost packet (go-back-n retransmission). This obviously wastes bandwidth, not just in the case of a mobile network, but for any network. Using selective retransmission, TCP can indirectly request a selective retransmission of packets. The receiver can acknowledge single packets, not only trains of in-sequence packets. The sender can now determine precisely which packet is needed and can retransmit it.

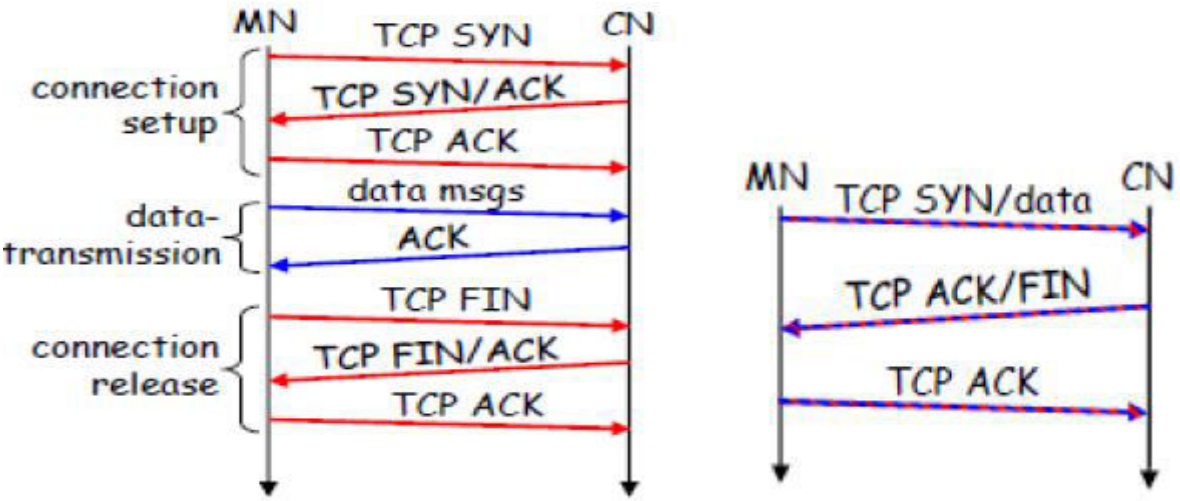
The **advantage** of this approach is obvious: a sender retransmits only the lost packets. This lowers bandwidth requirements and is extremely helpful in slow wireless links. The

disadvantage is that a more complex software on the receiver side is needed. Also more buffer space is needed to resequence data and to wait for gaps to be filled.

Transaction-oriented TCP

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

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



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

Classical Enhancements to TCP for mobility: A comparison

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

UNIT V

Features of windows CE

Applications and Services Development

Describes the operating system functionality that is available in Windows CE for developing applications and services.

Applications - End User

Describes the operating system functionality that is available for developing end user applications.

Communication Services and Networking

Describes the networking and communications capabilities in Windows CE that enable devices to connect and communicate with other devices and people over both wireless and wired networks.

Core OS Services

Describes the core operating system (OS) services that are available in Windows CE. Core OS services contain information on the Windows CE kernel and other features common to all Windows CE OS designs. The core OS services enable low-level tasks such as process, thread, and memory management.

File Systems and Data Store

Provides an overview of the file systems and data store architecture in Windows CE.

Fonts

Provides an overview of fonts and font technologies that are supported in Windows CE. Describes how you can replace fonts, specify a directory from which the OS should load fonts, and change the font size for the Help system. Also describes how you can enable ClearType, antialiased fonts, linked fonts, end-user-defined-characters (EUDC), and line breaking for Asian fonts.

Graphics and Multimedia Technologies

Describes the graphics and multimedia technologies that are supported in Windows CE. Includes detailed descriptions of the audio, graphics, and media support in Windows CE.

International

Describes the International support in Windows CE. The International technologies in Windows CE are comprised of a collection of functionality that provides general locale services and locale-specific support for certain key capabilities.

Internet Client Services

Describes the support for Internet client services in Windows CE. Windows CE provides support for browser applications, technologies that enable you to create custom browsers, and run-time engines for parsing and translating scripting languages.

Security

Provides an overview of the security technologies that enable you to enhance the security of your devices or applications.

Shell and User Interface

Provides a description of the shell and user interface technologies in Windows CE. These include the functionality that is necessary for a user to interact with a Windows CE-based device and the underlying OS.

Voice over IP Phone Services

Describes the technologies that are available in Windows CE to build IP phone devices.

Windows CE Error Reporting

Describes the Windows CE Error Reporting technology. Windows CE Error Reporting allows a device to save key information about the state of the machine at the time of a program crash.

PalmOS

Palm OS uses multitasking, but only one task is for applications. The user uses one application at a time, one application program must finish before the next can be selected. This constraint allows the operating system to devote full attention to the application that is open. The space needed by the system for any application that is running is kept in dynamic, reusable random access memory (RAM). The application and its related database are kept in what is called permanent storage, but here the permanent storage is RAM (rather than a hard disk) that cannot be reused as the dynamic RAM can. Palm OS divides an application into runnable code and different types of data elements, such as user interface elements and icons. The data elements can be easily changed without necessarily having to rewrite code.

Palm OS comes with these applications built-in: Dates, Address Book, To Do List, Memo Pad, Calculator, and Password Protection. New applications can be written and added using several facilities that accelerate development.

Palm supports Metrowerks' CodeWarrior as the official software development kit (SDK), using a Macintosh or Windows environment. UNIXplatform users can use a kit called GCC, which is available through the Free Software Foundation. Programmers can use C, C++, assembler, or scripting. The Palm user interface is emulated within a window in the desktop environment, encouraging rapid application development. Simpler applications can be developed using Palm's forms interface.

Palm OS comes with communication interfaces to infrared transmissiondevices, TCP/IP (for Web connection through wireless or wireline devices), and, optionally, barcode recognition scanners.

WWW

The World Wide Web abbreviated as WWW or W3, commonly known as the Web) is a system of interlinked hypertextdocuments that are accessed via the Internet. With a web browser, one can view web pages that may contain text, images, videos, and other multimedia and navigate between them via hyperlinks.

Wireless Application Protocol

WAP, Wireless Application Protocol aims to provide Internet content and advanced telephony services to digital mobile phones, pagers and other wireless terminals. The protocol family works across different wireless network environments and makes web pages visible on low-resolution and low-bandwidth devices. WAP phones are "smart phones" allowing their users to respond to e-mail, access computer databases and to empower the phone to interact with Internet-based content and e-mail.

WAP specifies a Wireless application Environment and Wireless Protocols. The Wireless application environment (WAE) is based on WSP (Wireless Session Protocol) and WTP (Wireless Transaction Protocol).

The basic construction of WAP architecture can be explained using the following model. The order of the independent levels – which are a hierarchy - has the advantage that the system is very flexible and can be scaled up or down. Because of the different levels – or stacks - this is called the "WAP Stack", which is divided into 5 different levels.

- Application Layer: Wireless Application Environment (WAE).
- Session Layer: Wireless Session Protocol (WSP).

- Transaction Layer: Wireless Transaction Protocol (WTP).
- Security Layer: Wireless Transport Layer Security (WTLS).

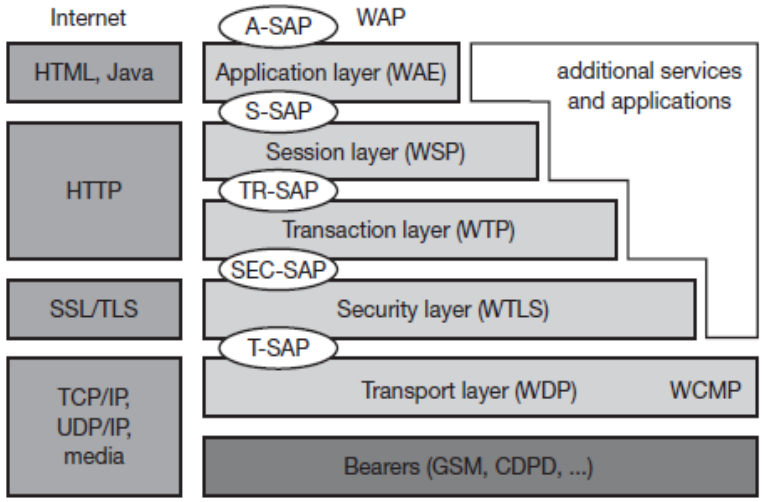


Figure 10.9 Components and interface of the WAP 1.x architecture

- Transport Layer: Wireless Datagram Protocol (WDP).

Each stack overlaps with the stack below. This stack architecture makes it possible for software manufacturers to develop applications and services for certain stacks. They may even develop services for stacks which are not specified yet.

The WAP stack is an entity of protocols which cover the wireless data transfer. The diagram above shows the order of the different stacks and their protocols. This includes the stacks responsible for the layout as well as the stacks resposible for the actual data transfer. The highest level or stack is the one which deals with the layout. A lower stack is responsible for the transfer and the security through WTLS (Wireless Transport Layer Security). All stacks lower than this one are being called network stack. Due to this hierarchy of stacks any changes made in the network stacks will have no influence over the stacks above

Application Layer (WAE and WTA)

The environment for wireless applications (Wireless Application Environment WAE) and the application for wireless phones (Wireless Telephony Application WTA) are the highest layer in the hierarchy of WAP architechture. These two are the main interface to the client device, which gives and controls the description language, the script language of any application and the specifics of the telephony. WAE and WTA have only a few easy functions on the client device, like the maintenance of a history list, for example.

Session Layer (Wireless Session Protocol WSP)

The Wireless Session Protocol (WSP) has all the specifications for a session. It is the interface between the application layer and the transfer layer and delivers all functions that are needed for wireless connections. A session mainly consists of 3 phases: start of the session, transfering information back and forth and the end of the session. Additionally, a session can be interrupted and started again (from the point where it was interrupted.)

Transaction Layer (Wireless Transaction Protocol WTP)

The specifications for the transfer layer are in the Wireless Transaction Protocol (WTP). Like the User Datagramm Protocol (UDP), the WTP runs at the head of the datagramm service. Both the UDP and the WTP are a part of the standard application from the TCP/IP to make the simplified protocol compatible to mobile terminals. WTP supports chaining together protocol data and the delayed response to reduce the number of transmissions. The protocol tries to optimize user interaction in order that information can be received when needed.

Wireless Transport Layer Security WTLS

The Wireless Transport Layer Security (WTLS) is a optional layer or stack which consists of description devices. A secure transmission is crucial for certain applications such as e-commerce or WAP-banking and is a standard in these days. Furthermore WTLS contains a check for data integrity, user authentification and gateway security.

Transport Layer (Wireless Datagram Protocol WDP)

The Wireless Datagram Protocol (WDP) represents the transfer or transmission layer and is also the interface of the network layer to all the above stacks/layers. With the help of WDP the transmission layer can be assimilated to the specifications of a network operator. This means that WAP is completely independent from any network operator. The transmission of SMS, USSD, CSD, CDPD, IS-136 packet data and GPRS is supported. The Wireless Control Message Protocol (WCMP) is an optional addition to WAP, which will inform users about occurred errors.

WTLS

Wapforum version 11/99

Wireless Transport Layer Security is a protocol based on the TLS protocol. It is used with the WAP transport protocols and has been optimised for use over narrow-band communication channels. The WTLs layer is above the transport protocol layer. The required security layer of the protocol determines whether it is used or not. It provides a secure transport service interface that preserves

the transport service interface below; additionally it provides an interface for managing secure connections. WTLS aims to provide privacy, data integrity and authentication between two communication applications. Among its features are datagram support, optimised handshaking and dynamic key refreshing. It is optimised for low-bandwidth bearer networks with relatively long latency.

The WTLS Record Protocol is a layered protocol. The Record Protocol takes messages to be transmitted, optionally compresses the data, applies a MAC, encrypts, and transmits the result. Received data is decrypted, verified, and decompressed, then delivered to higher-level clients. Four record protocol clients are described in the WTLS standard; the change cipher spec protocol, the handshake protocol, the alert protocol and the application data protocol. If a WTLS implementation receives a record type it does not understand, it ignores it. Several records can be concatenated into one transport SDU. For example, several handshake messages can be transmitted in one transport SDU. This is particularly useful with packet-oriented transports such as GSM short messages.

<table>
<tr><td>Handshake protocols</td><td>Alert Protocol</td><td>Application Protocol</td><td>Change Cipher Spec Protocol</td></tr>
<tr><td colspan="4">Record protocol</td></tr>
</table>

The handshake protocol is made up of 3 sub-protocols. All messages are encapsulated in a plaintext structure.

WTP

The Wireless Transaction Protocol provides the services necessary for interactive browsing applications. During a browsing session the client requests information from a server and the server responds with the information. This is referred to as a transaction. WTP runs on a datagram service and possible a security service.

Advantages of WTP include:

- Improved reliability over datagram services
- Imported efficiency over connection oriented services
- As a message oriented protocol, it is designed for services oriented towards transactions.

Main features:

- 3 kinds of transaction services.
  - Class 0 Unreliable invoke messages with no result messages
  - Class 1: Reliable invoke messages with no result messages
  - Class 2: Reliable invoke messages with exactly one reliable result message.

- Reliability achieved by using unique transaction identifiers, acknowledgements, duplicate removal; and retransmissions.
- No explicit set up or tear down phases.
- Optional user-to-user reliability.
- Optionally the last acknowledgement of the transaction may contain out-of-band information.
- Concatenation may be used to convey multiple PDUs in one service data unit of the datagram transport.
- The basic unit of interchange is an entire message, not a stream of bytes.
- Mechanisms are provided to minimize the number of transactions replayed as a result of duplicate packets.
- Abort of outstanding transactions.
- For reliable invoke messages, both success and failure reported.
- Asynchronous transactions allowed.

The protocol data unit (PDU) consists of the header and data (if present). The header contains a fixed part and a variable part; The variable parts are carried in the Transport Information Item (TPI). Each PDU has its own fixed header (the fixed headers vary slightly in structure). As an example, the structure of the invoke PDU fixed header appears below:

1	2-5			6	7	8
Con	PDU Type			GTR	TTR	RID
TID						
Version	TIDnew	U/P	RES	RES	TCL	

ONcontinueflag (1bit):
The continue flag indicates the presence of any TPIs in the variable part. If the flag is set, there are one or more TPIs in the variable portion of the header. If the flag is clear, the variable part of the header is empty. This flag is also used as the first bit of a TPI, and indicates whether the TPI is the last of the variable header. If the flag is set, another TPI follows this TPI. If the flag is clear, the octet after this TPI is the first octet of the user data.

PDUtype
The PDU type determines the length and structure of the header and dictates what type of WTP PDU the PDU is (Invoke, Ack, etc). This provides information to the receiving WTP provider as to how the PDU data should be interpreted and what action is required.

The following PDU types are defined:

PDU Code	PDU Type
0x01	Invoke
0x02	Result
0x03	Ack
0x04	Abort
0x05	Segmented Invoke
0x06	Segmented Result
0x07	Negative Ack

Group trailer (GTR) and Transmission trailer (TTR) flag (2 bit):
When segmentation and re-assembly is implemented, the TTR flag is used to indicate the last packet of the segmented message. The GTR flag is used to indicate the last packet of a packet group.

GTR/TTR flag combinations:

GTR	TTR	Description
00		Not last packet
01		Last packet of message
10		Last packet of packet group
11		Segmentation and Re-assembly NOT supported.

The default setting should be GTR=1 and TTR=1, that is, WTP segmentation and re-assembly not supported.

RIDRe-transmissionIndicator (1bit):
Enables the receiver to differentiate between packets duplicated by the network and packets re-transmitted by the sender. In the original message the RID is clear. When the message gets re-transmitted the RID is set.

TIDTransactionidentifier (16bit):
The TID is used to associate a packet with a particular transaction.

Version
The current version is 0X00

TIDnew flag
This bit is set when the Initiator has wrapped the TID value, i.e. set it to be lower than the previous TID value.

U/P
When this flag is set it indicates that the Initiator requires a User acknowledgement from the server WTP user. The WTP user confirms every received message.

RES
This is a reserved bit and its value should be set to 0.

TCL
The transaction class shows the desired transaction class in the invoke message.

WSP

WAP WSP 5/11/99

The Session layer protocol family in the WAP architecture is called the Wireless Session Protocol, WSP. WSP provides the upper-level application layer of WAP with a consistent interface for two session services. The first is a connection-mode service that operates above a transaction layer protocol WTP, and the second is a connectionless service that operates above a secure or non-secure datagram transport service.

The Wireless Session Protocols currently offer services most suited for browsing applications. WSP provides HTTP 1.1 functionality (it is a binary form of HTTP) and incorporates new features such as long-lived sessions, a common facility for data push, capability negotiation and session suspend/resume. The protocols in the WSP family are optimized for low-bandwidth bearer networks with relatively long latency. Requests and responses can include both headers and data. WSP provides push and pull data transfer WSP functions on the transaction and datagram services.

Messages can be in connection mode or connectionless. Connection mode messages are carried over WTP. In this case the protocol consists of WTP protocol messages with WSP PDUs as their data. Connectionless messages consist only of the WSP PDUs.

The general structure of the WSP PDU is as follows:

1 bite1 bite	

TID/PIDPDU Type	Type Specific Contents

TID/PID

Transaction ID or Push ID. The TID field is used to associate requests with replies in the connectionless session service. The presence of the TID is conditional. It is included in the connectionless WSP PDUs, and is not included in the connection-mode PDUs. In connectionless WSP, the TID is passed to and from the session user as the "Transaction Id" or "Push Id" parameters of the session primitive

PDU type

The Type field specifies the type and function of the PDU. The type numbers for the various PDUs are defined below. The rest of the PDU is type-specific information, referred to as the contents.