Introduction to Mobile Computing

The rapidly expanding technology of cellular communication, wireless LANs, and satellite services will make information accessible anywhere and at any time. Regardless of size, most mobile computers will be equipped with a wireless connection to the fixed part of the network, and, perhaps, to other mobile computers. The resulting computing environment, which is often referred to as ***mobile or nomadic computing***, no longer requires users to maintain a fixed and universally known position in the network and enables almost unrestricted mobility. Mobility and portability will create an entire new class of applications and, possibly, new massive markets combining personal computing and consumer electronics.

Mobile Computing is an umbrella term used to describe technologies that enable people to access network services anyplace, anytime, and anywhere.

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

1. 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.
2. 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.
3. 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.
4. 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 OF MOBILE COMPUTING

In many fields of work, the ability to keep on the move is vital in order to utilise time efficiently. The importance of Mobile Computers has been highlighted in many fields of which a few are described below:

1. **Vehicles:** 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. 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. 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**: An ambulance with a high-quality wireless connection to a hospital can carry vital information about injured persons 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.
3. **Business**: Managers can use mobile computers say, critical presentations to major customers. They can access the latest market share information. At a small recess, they can revise the presentation to take advantage of this information. They can communicate with the office about possible new offers and call meetings for discussing responds to the new proposals. Therefore, mobile computers can leverage competitive advantages. 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.
4. **Credit Card Verification**: At Point of Sale (POS) terminals in shops and

   Supermarkets, when customers use credit cards for transactions, the intercommunication required between the bank central computer and the POS terminal, in order to effect verification of the card usage, can take place quickly and securely over cellular channels using a mobile computer unit. This can speed up the transaction process and relieve congestion at the POS terminals.
5. **Replacement of Wired Networks**: 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. Other examples for wireless networks are computers, sensors, or information displays in historical buildings, where excess cabling may destroy valuable walls or floors.
6. **Infotainment**: wireless networks can provide up-to-date information at any appropriate location. The travel guide might tell you something about the history of a building (knowing via GPS, contact to a local base station, or triangulation where you are) downloading information about a concert in the building at the same evening via a local wireless network. Another growing field of wireless network applications lies in entertainment and games to enable, e.g., ad-hoc gaming networks as soon as people

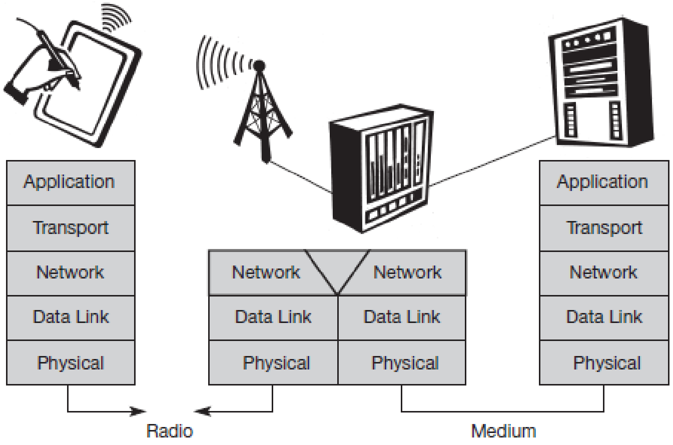
meet to play together.

Limitations of Mobile Computing

Resource constraints: Battery
Interference: Radio transmission cannot be protected against interference using shielding and result in higher loss rates for transmitted data or higher bit error rates respectively
Bandwidth: Although they are continuously increasing, transmission rates are still very low for wireless devices compared to desktop systems. Researchers look for more efficient communication protocols with low overhead.
Dynamic changes in communication environment: variations in signal power within a region, thus link delays and connection losses
Network Issues: discovery of the connection-service to destination and connection stability
Interoperability issues: the varying protocol standards
Security constraints: Not only can portable devices be stolen more easily, but the radio interface is also prone to the dangers of eavesdropping. Wireless access must always include encryption, authentication, and other security mechanisms that must be efficient and simple to use.

A SIMPLIFIED REFERENCE MODEL

The figure shows the **protocol stack** implemented in the system according to the reference model. **End-systems**, such as the PDA and computer in the example, need a full protocol stack comprising the application layer, transport layer, network layer, data link layer, and physical layer. Applications on the end-systems communicate with each other using the lower layer services. **Intermediate systems**, such as the interworking unit, do not necessarily need all of the layers.



A Simplified Reference Model

a) **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.

b) **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.

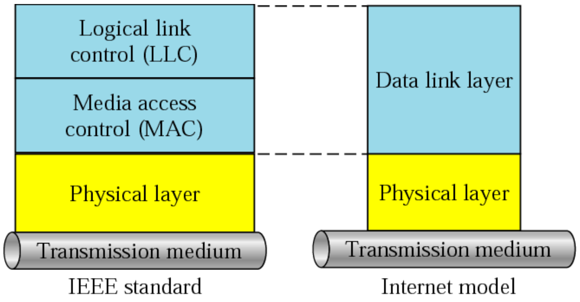
c) **Network layer**: This third layer is responsible for routing packets through a network or establishing a connection between two entities over many other intermediate systems. Important functions are addressing, routing, device location, and handover between different networks.

d) **Transport layer**: This layer is used in the reference model to establish an end-to-end connection

e) **Application layer**: Finally, the applications (complemented by additional layers that can support applications) are situated on top of all transmission oriented layers. Functions 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.

Media Access Control (**MAC**)

The **Media Access Control** (**MAC**) data communication protocol sub-layer, also known as the Medium Access Control, is a sublayer of the Data Link Layer specified in the seven-layer OSI model (layer 2). The hardware that implements the MAC is referred to as a **Medium Access Controller**. The MAC sub-layer acts as an interface between the Logical Link Control (LLC) sublayer and the network's physical layer. The MAC layer emulates a full-duplex logical communication channel in a multi-point network. This channel may provide unicast, multicast or broadcast communication service.



LLC and MAC sublayers

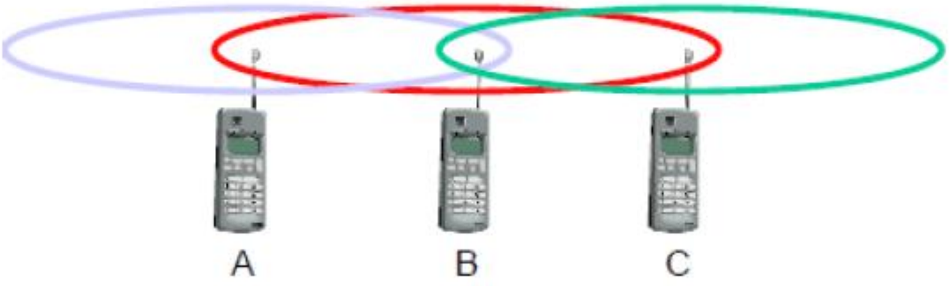
Motivation for a specialized MAC

One of the most commonly used MAC schemes for wired networks is carrier sense multiple access with collision detection (CSMA/CD). In this scheme, 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. But this scheme doest work well with wireless networks. The problems are:

a) Signal strength decreases proportional to the square of the distance
b) The sender would apply CS and CD, but the collisions happen at the receiver
c) It might be a case that a sender cannot "hear" the collision, i.e., CD does not work
d) Furthermore, CS might not work, if for e.g., a terminal is "hidden"

Hidden and Exposed Terminals

Consider the scenario with three mobile phones as shown below. 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.



Hidden terminals

a) A sends to B, C cannot hear A
b) C wants to send to B, C senses a "free" medium (CS fails) and starts transmitting
c) Collision at B occurs, A cannot detect this collision (CD fails) and continues with its transmission to B
d) A is "hidden" from C and vice versa

Exposed terminals

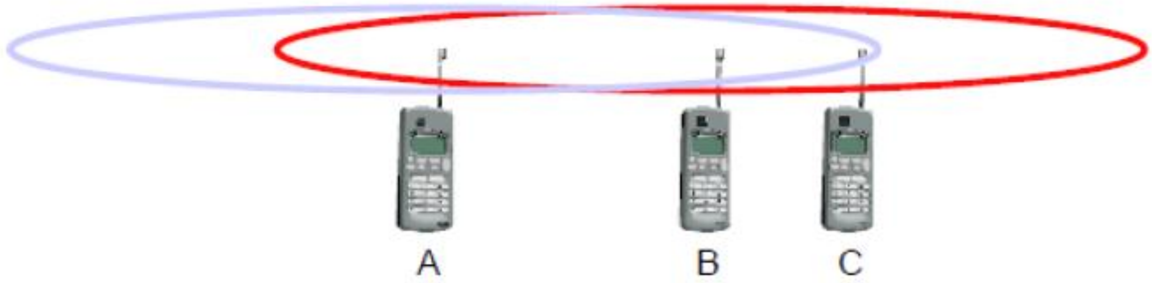
a) B sends to A, C wants to send to another terminal (not A or B) outside the range
b) C senses the carrier and detects that the carrier is busy.
c) C postpones its transmission until it detects the medium as being idle again but A is outside radio range of C, waiting is **not** necessary
d) C is "exposed" to B

Hidden terminals cause collisions, where as Exposed terminals causes unnecessary delay.

Near and far terminals

Consider the situation shown below. A and B are both sending with the same transmission power.

a) Signal strength decreases proportional to the square of the distance
b) So, B’s signal drowns out A’s signal making C unable to receive A’s transmission
c) If C is an arbiter for sending rights, B drown out A’s signal on the physical layer making C unable to hear out A.



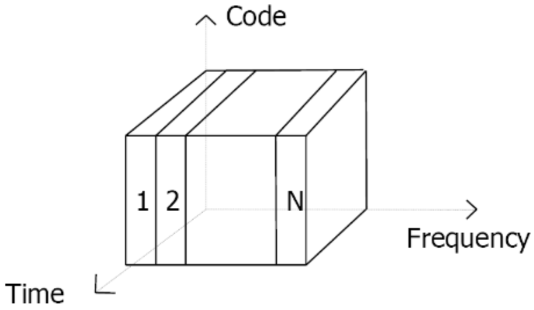
The **near/far effect** is a severe problem of wireless networks using CDM. All signals should arrive at the receiver with more or less the same strength for which Precise power control is to be implemented.

SDMA

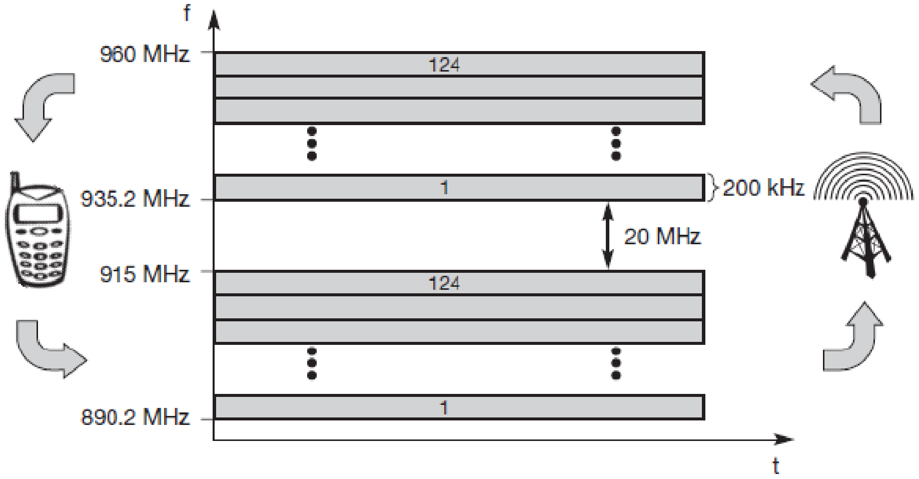
Space Division Multiple Access (SDMA) is used for allocating a separated space to users in wireless networks. A typical application involves assigning an optimal base station to a mobile phone user. The mobile phone may receive several base stations with different quality. A MAC algorithm could now decide which base station is best, taking into account which frequencies (FDM), time slots (TDM) or code (CDM) are still available. The basis for the SDMA algorithm is formed by cells and sectorized antennas which constitute the infrastructure implementing **space division multiplexing (SDM).** SDM has the unique advantage of not requiring any multiplexing equipment. It is usually combined with other multiplexing techniques to better utilize the individual physical channels.

FDMA

Frequency division multiplexing (FDM) describes schemes to subdivide the frequency dimension into several non-overlapping frequency bands.



Frequency Division Multiple Access is a method employed to permit several users to transmit simultaneously on one satellite transponder by assigning a specific frequency within the channel to each user. Each conversation gets its own, unique, radio channel. The channels are relatively narrow, usually 30 KHz or less and are defined as either transmit or receive channels. A full duplex conversation requires a transmit & receive channel pair. FDM is often used for simultaneous access to the medium by base station and mobile station in cellular networks establishing a duplex channel. A scheme called **frequency division duplexing (FDD)** in which the two directions, mobile station to base station and vice versa are now separated using different frequencies.



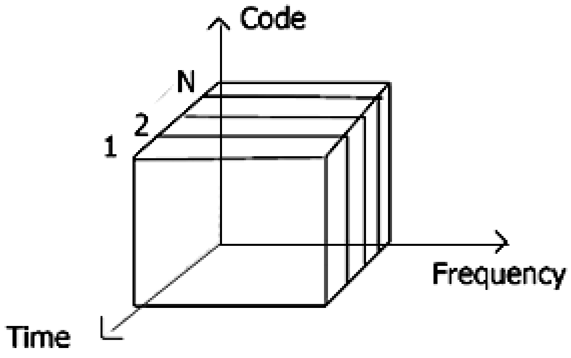
FDM for multiple access and duplex

The two frequencies are also known as **uplink**, i.e., from mobile station to base station or from ground control to satellite, and as **downlink**, i.e., from base station to mobile station or from satellite to ground control. The basic frequency allocation scheme for GSM is fixed and regulated by national authorities. All uplinks use the band between 890.2 and 915 MHz, all downlinks use 935.2 to 960 MHz. According to FDMA, the base station, shown on the right side, allocates a certain frequency for up- and downlink to establish a duplex channel with a mobile phone. Up- and downlink have a fixed relation. If the uplink frequency is fu = 890 MHz + n·0.2 MHz, the downlink frequency is fd = fu + 45 MHz, i.e**., fd = 935 MHz + n·0.2 MHz** for a certain channel n. The base station selects the channel. Each channel (uplink and downlink) has a bandwidth of 200 kHz.

This scheme also has disadvantages. While radio stations broadcast 24 hours a day, mobile 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

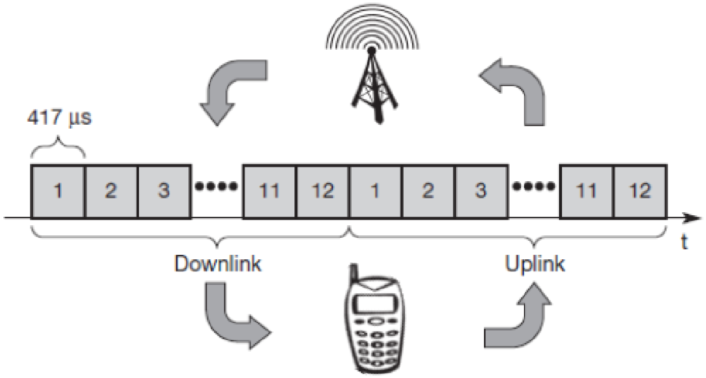
A more flexible multiplexing scheme for typical mobile communications is time division multiplexing (TDM). Compared to FDMA, time division multiple access (TDMA) offers a much more flexible scheme, which comprises all technologies that allocate certain time slots for communication. Now synchronization 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.



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. Fixed schemes do not need identification, but are not as flexible considering varying bandwidth requirements.

Fixed TDM

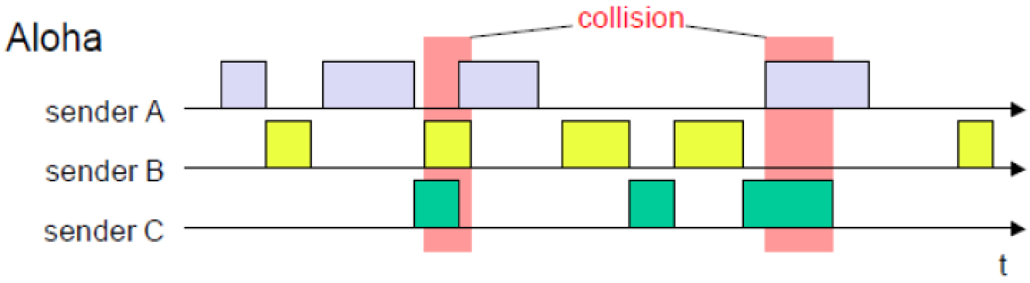
The simplest algorithm for using TDM is allocating time slots for channels in a fixed pattern. This results in a fixed bandwidth and is the typical solution for wireless phone systems. MAC is quite simple, as the only crucial factor is accessing the reserved time slot at the right moment. If this synchronization is assured, each mobile station knows its turn and no interference will happen. The fixed pattern can be assigned by the base station, where competition between different mobile stations that want to access the medium is solved.



The above figure shows how these fixed TDM patterns are used to implement multiple access and a duplex channel between a base station and mobile station. Assigning different slots for uplink and downlink using the same frequency is called **time division duplex (TDD)**. As shown in the figure, the base station uses one out of 12 slots for the downlink, whereas the mobile station uses one out of 12 different slots for the uplink. Uplink and downlink are separated in time. Up to 12 different mobile stations can use the same frequency without interference using this scheme. Each connection is allotted its own up- and downlink pair. This general scheme still wastes a lot of bandwidth. It is too static, too inflexible for data communication. In this case, connectionless, demand-oriented TDMA schemes can be used.

Classical Aloha

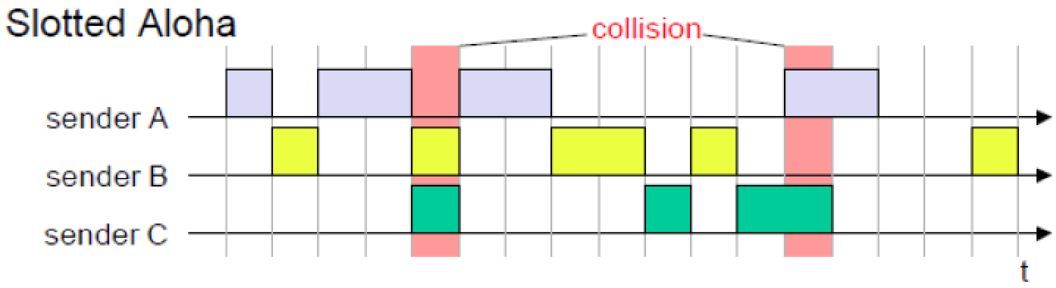
In this scheme, TDM is applied without controlling medium access. Here each station can access the medium at any time as shown below:



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.

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 below.



The introduction of slots raises the throughput from 18 per cent to 36 per cent, i.e., slotting doubles the throughput. Both basic Aloha principles occur in many systems that implement distributed access to a medium. Aloha systems work perfectly well under a light load, but they cannot give any hard transmission guarantees, such as maximum delay before accessing the medium or minimum throughput.

Carrier sense multiple access

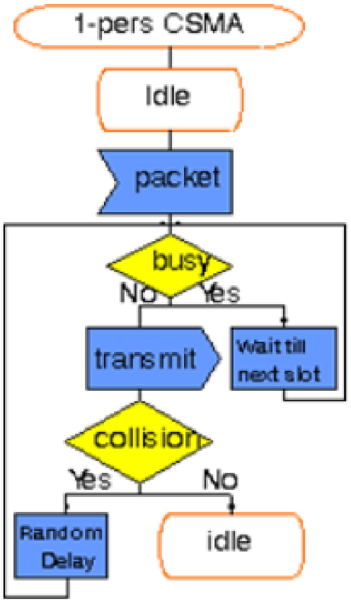
One improvement to the basic Aloha is sensing the carrier before accessing the medium. Sensing the carrier and accessing the medium only if the carrier is idle decreases the probability of a collision. But, as already mentioned in the introduction, hidden terminals cannot be detected, so, if a hidden terminal transmits at the same time as another sender, a collision might occur at the receiver. This basic scheme is still used in most wireless LANs. The different versions of CSMA are:

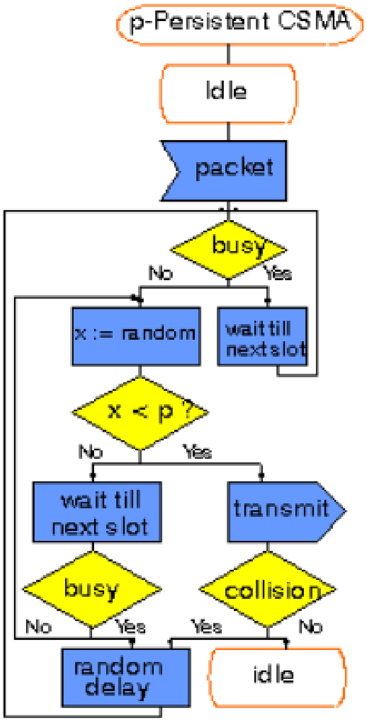
a) **1-persistent CSMA**: Stations sense the channel and listens if its busy and transmit immediately, when the channel becomes idle. It’s called 1-persistent CSMA because the host transmits with a probability of 1 whenever it finds the channel idle.

b) **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.

c) **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

CSMA with collision avoidance (**CSMA/CA**) is one of the access schemes used in wireless LANs following the standard IEEE 802.11. Here sensing the carrier is combined with a back-off scheme in case of a busy medium to achieve some fairness among competing stations.



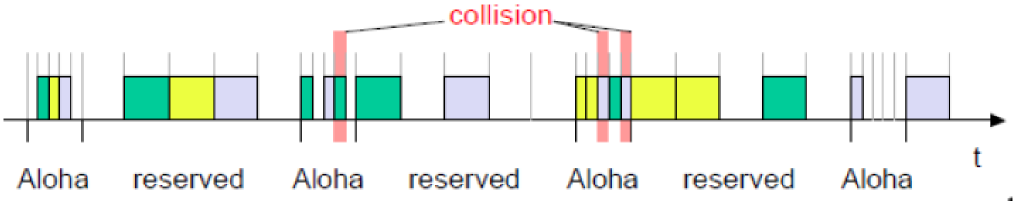


Demand assigned multiple access

Channel efficiency for Aloha is 18% and for slotted Aloha is 36%. It can be increased to 80% by implementing 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.

One basic scheme is **demand assigned multiple access (DAMA)** also called **reservation Aloha**, a scheme typical for satellite systems. It increases the amount of users in a pool of satellite channels that are available for use by any station in a network. It is assumed that not all users will need simultaneous access to the same communication channels. So that a call can be established, DAMA assigns a pair of available channels based on requests issued from a user. Once the call is completed, the channels are returned to the pool for an assignment to another call. Since the resources of the satellite are being used only in proportion to the occupied channels for the time in which they are being held, it is a perfect environment for voice traffic and data traffic in batch mode.

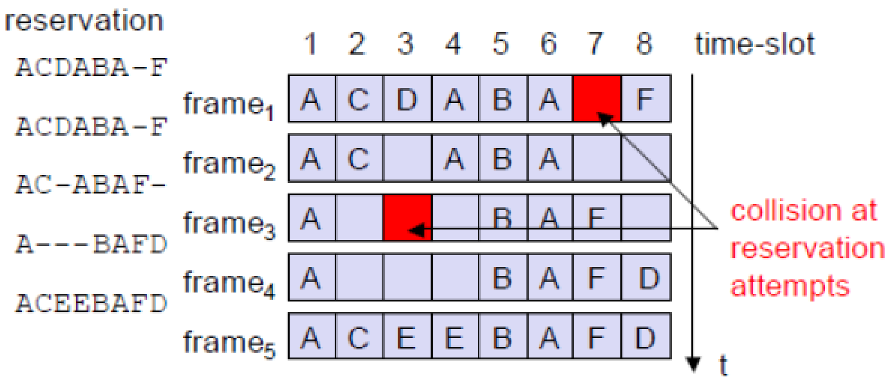
It has two modes as shown below.



During a contention phase following the slotted Aloha scheme; all stations can try to reserve future slots. 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.

PRMA packet reservation multiple access

It is a kind of implicit reservation scheme where, slots can be reserved implicitly. A certain number of slots form a frame. The frame is repeated in time i.e., a fixed TDM pattern is applied. A base station, which could be a satellite, now broadcasts the status of each slot to all mobile stations. All stations receiving this vector will then know which slot is occupied and which slot is currently free.

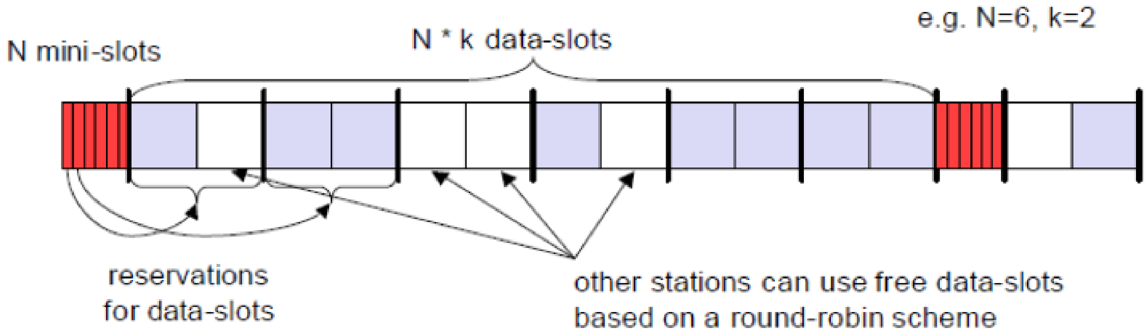


The base station broadcasts the reservation status ‘ACDABA-F’ to all stations, here A to F. This means that slots one to six and eight are occupied, but slot seven is free in the following transmission. All stations wishing to transmit can now compete for this free slot in Aloha fashion. The already occupied slots are not touched. In the example shown, more than one station wants to access this slot, so a collision occurs. The base station returns the reservation status ‘ACDABA-F’, indicating that the reservation of slot seven failed (still indicated as free) and that nothing has changed for the other slots. Again, stations can compete for this slot. Additionally, station D has stopped sending in slot three and station F in slot eight. This is noticed by the base station after the second frame. Before the third frame starts, the base station indicates that slots three and eight are now idle. Station F has succeeded in reserving slot seven as also indicated by the base station.

As soon as a station has succeeded with a reservation, all future slots are implicitly reserved for this station. This ensures transmission with a guaranteed data rate. The slotted aloha scheme is used for idle slots only; data transmission is not destroyed by collision.

Reservation TDMA

In a fixed TDM scheme N mini-slots followed by N·k data-slots form a frame that is repeated. Each station is allotted its own mini-slot and can use it to reserve up to k data-slots.


e.g. N=6, k=2
N mini-slots
N * k data-slots
reservations for data-slots
other stations can use free data-slots based on a round-robin scheme

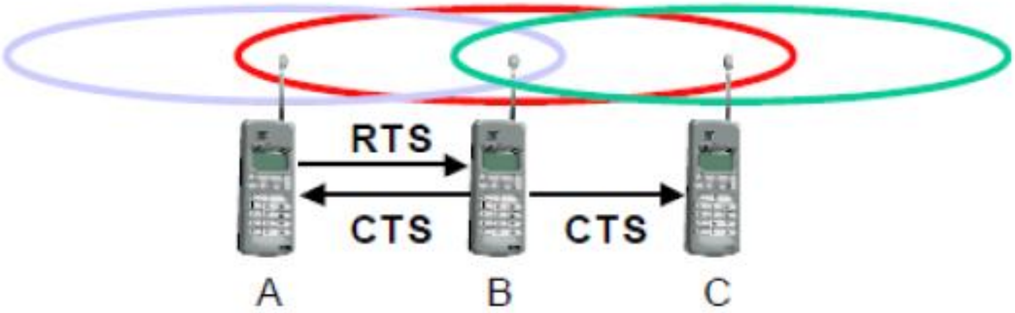
This guarantees each station a certain bandwidth and a fixed delay. Other stations can now send data in unused data-slots as shown. Using these free slots can be based on a simple round-robin scheme or can be uncoordinated using an Aloha scheme. This scheme allows for the combination of, e.g., isochronous traffic with fixed bitrates and best-effort traffic without any guarantees.

Multiple access with collision avoidance

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. Consider the hidden terminal problem scenario.

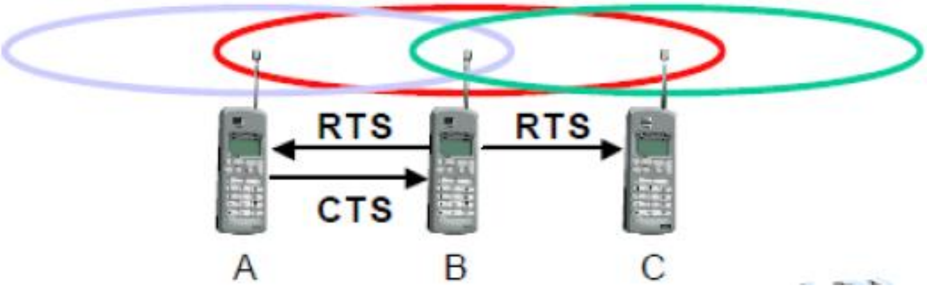
A starts sending to B, C does not receive this transmission. C also wants to send something to B and senses the medium. The medium appears to be free, the carrier sense fails. C also starts sending causing a collision at B. But A cannot detect this collision at B and continues with its transmission.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

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. Still collisions might occur when A and C transmits a RTS at the same time. B resolves this contention and acknowledges only one station in the CTS. No transmission is allowed without an appropriate CTS.

Now MACA tries to avoid the **exposed terminals** in the following way:



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 this 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.

Polling

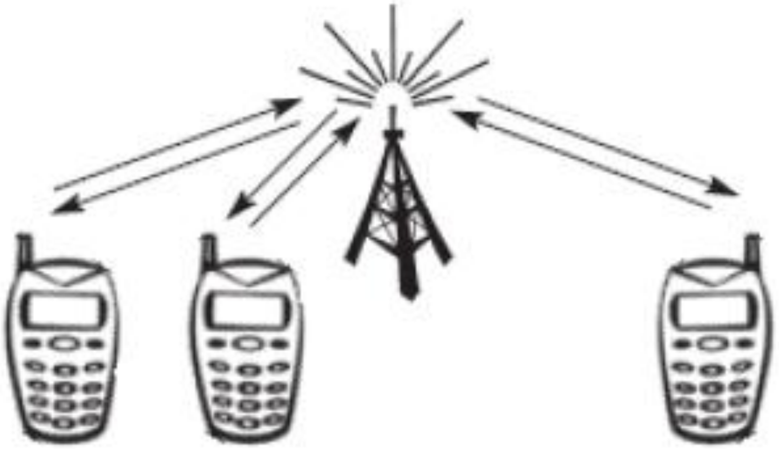
Polling schemes are used when one station wants to be heard by others. Polling is a strictly centralized scheme with one master station and several slave stations. The master can poll the slaves according to many schemes: round robin (only efficient if traffic patterns are similar over all stations), randomly, according to reservations (the classroom example with polite students) etc. The master could also establish a list of stations wishing to transmit during a contention phase. After this phase, the station polls each station on the list.

Example: Randomly Addressed Polling

- base station signals readiness to all mobile terminals
- terminals ready to send transmit random number without collision using CDMA or FDMA
- the base station chooses one address for polling from list of all random numbers (collision if two terminals choose the same address)
- the base station acknowledges correct packets and continues polling the next terminal
- this cycle starts again after polling all terminals of the list

Inhibit sense multiple access

This scheme, which is used for the packet data transmission service Cellular Digital Packet Data (CDPD) in the AMPS mobile phone system, is also known as **digital sense multiple access (DSMA)**. Here, the base station only signals a busy medium via a busy tone (called BUSY/IDLE indicator) on the downlink.



After the busy tone stops, accessing the uplink is not coordinated any further. The base station acknowledges successful transmissions; a mobile station detects a collision only via the missing positive acknowledgement. In case of collisions, additional back-off and retransmission mechanisms are implemented.

CDMA

Code division multiple access systems apply codes with certain characteristics to the transmission to separate different users in code space and to enable access to a shared medium without interference.

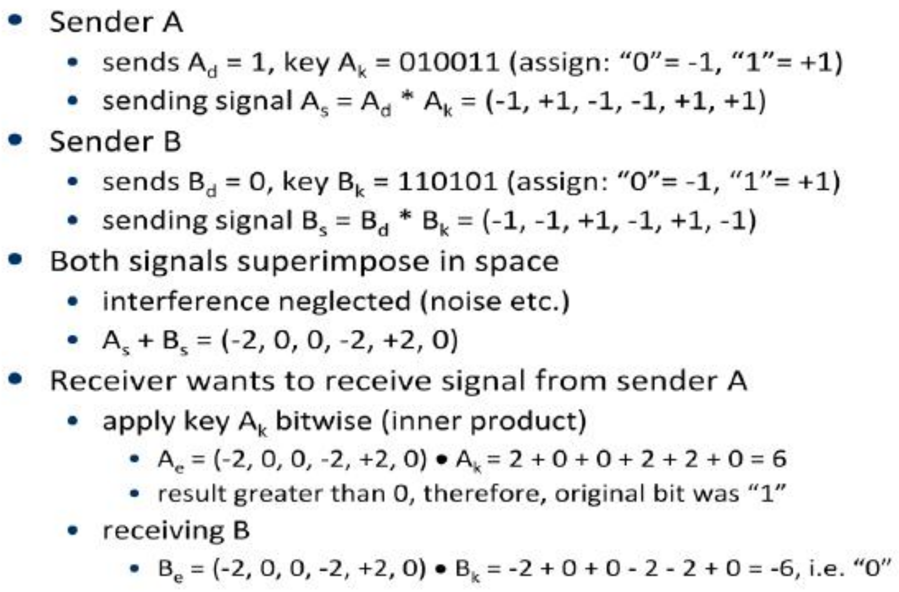
All terminals send on the same frequency probably at the same time and can use the whole bandwidth of the transmission channel. Each sender has a unique random number, the sender XORs the signal with this random number. The receiver can "tune" into this signal if it knows the pseudo random number, tuning is done via a correlation function

Disadvantages:

1. higher complexity of a receiver (receiver cannot just listen into the medium and start receiving if there is a signal)
2. all signals should have the same strength at a receiver

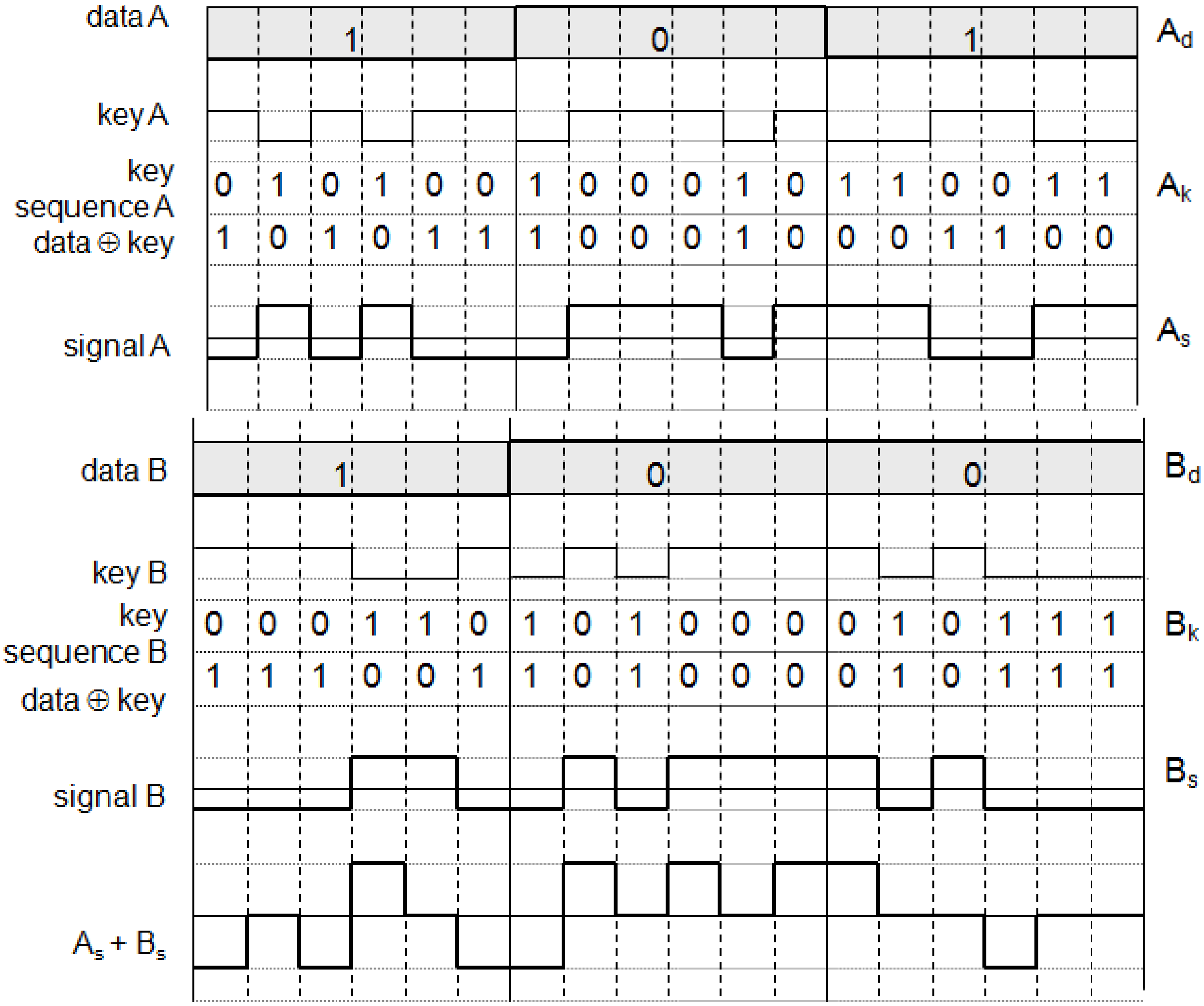
Advantages:

1. all terminals can use the same frequency, no planning needed
2. huge code space (e.g. 232) compared to frequency space
3. interferences (e.g. white noise) is not coded
4. forward error correction and encryption can be easily integrated



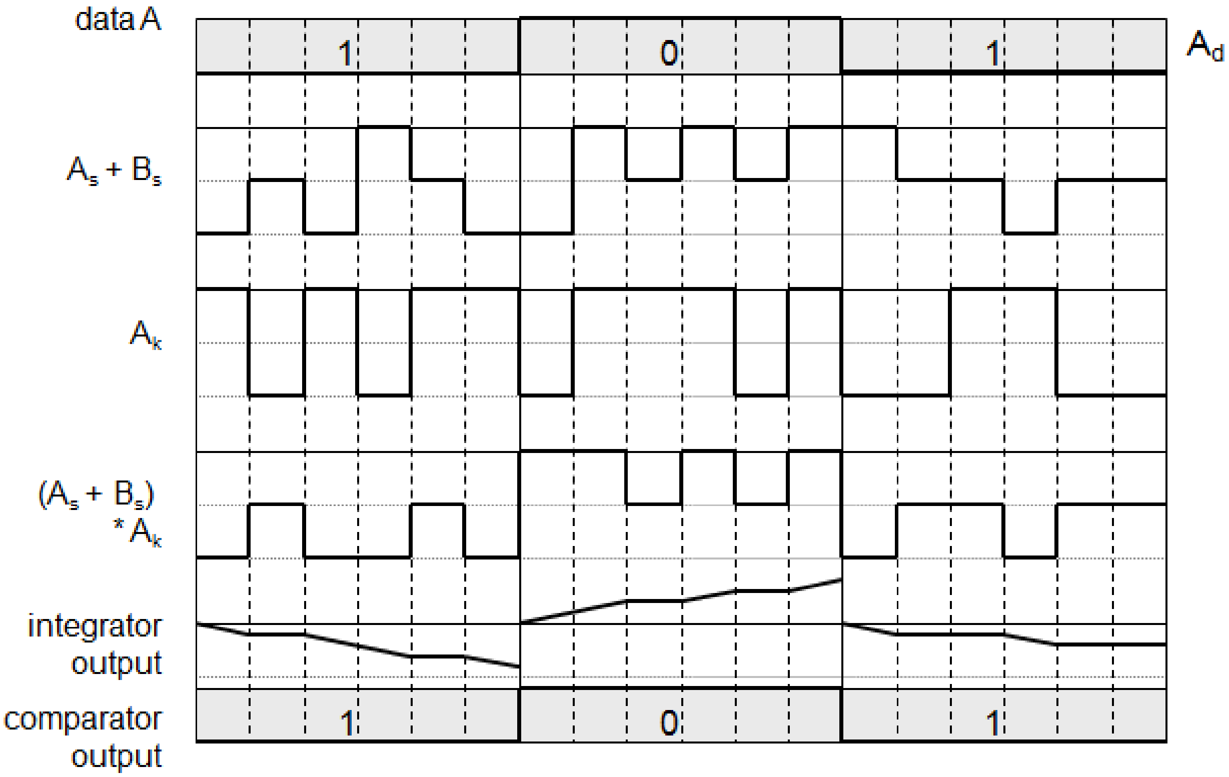
- Sender A
  - sends $A_d = 1$, key $A_k = 010011$ (assign: "0"= -1, "1"= +1)
  - sending signal $A_s = A_d * A_k = (-1, +1, -1, -1, +1, +1)$
- Sender B
  - sends $B_d = 0$, key $B_k = 110101$ (assign: "0"= -1, "1"= +1)
  - sending signal $B_s = B_d * B_k = (-1, -1, +1, -1, +1, -1)$
- Both signals superimpose in space
  - interference neglected (noise etc.)
  - $A_s + B_s = (-2, 0, 0, -2, +2, 0)$
- Receiver wants to receive signal from sender A
  - apply key A_k bitwise (inner product)
    - $A_e = (-2, 0, 0, -2, +2, 0) \bullet A_k = 2 + 0 + 0 + 2 + 2 + 0 = 6$
    - result greater than 0, therefore, original bit was "1"
  - receiving B
    - $B_e = (-2, 0, 0, -2, +2, 0) \bullet B_k = -2 + 0 + 0 - 2 - 2 + 0 = -6$, i.e. "0"

The following figure shows a sender A that wants to transmit the bits 101. The key of A is shown as signal and binary sequence Ak. The binary "0" is assigned a positive signal value, the binary "1" a negative signal value. After spreading, i.e., XORing Ad and Ak, the resulting signal is As.



Coding and spreading of data from sender A and sender B

The same happens with data from sender B with bits 100. The result is Bs. As and Bs now superimpose during transmission. The resulting signal is simply the sum As + Bs as shown above. A now tries to reconstruct the original data from Ad. The receiver applies A’s key, Ak, to the received signal and feeds the result into an integrator. The integrator adds the products, a comparator then has to decide if the result is a 0 or a 1 as shown below. As clearly seen, although the original signal form is distorted by B’s signal, the result is quite clear. The same happens if a receiver wants to receive B’s data.



Reconstruction of A’s data

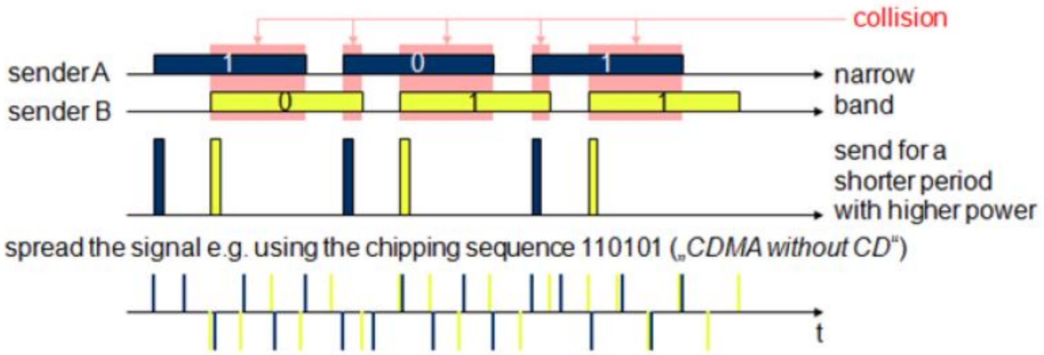
Soft handover or **soft handoff** refers to a feature used by the CDMA and WCDMA standards, where a cell phone is simultaneously connected to two or more cells (or cell sectors) during a call. If the sectors are from the same physical cell site (a sectorised site), it is referred to as **softer handoff**. This technique is a form of mobile-assisted handover, for IS-95/CDMA2000 CDMA cell phones continuously make power measurements of a list of neighboring cell sites, and determine whether or not to request or end soft handover with the cell sectors on the list.

Soft handoff is different from the traditional hard-handoff process. With hard handoff, a definite decision is made on whether to hand off or not. The handoff is initiated and executed without the user attempting to have simultaneous traffic channel communications with the two base stations. With soft handoff, a c*onditional* decision is made on whether to hand off. Depending on the changes in pilot signal strength from the two or more base stations involved, a hard decision will eventually be made to communicate with only one. This normally happens after it is evident that the signal from one base station is considerably stronger than those from the others. In the interim period, the user has simultaneous traffic channel communication with all candidate base stations. It is desirable to implement soft handoff in power-controlled CDMA systems because implementing hard handoff is potentially difficult in such systems.

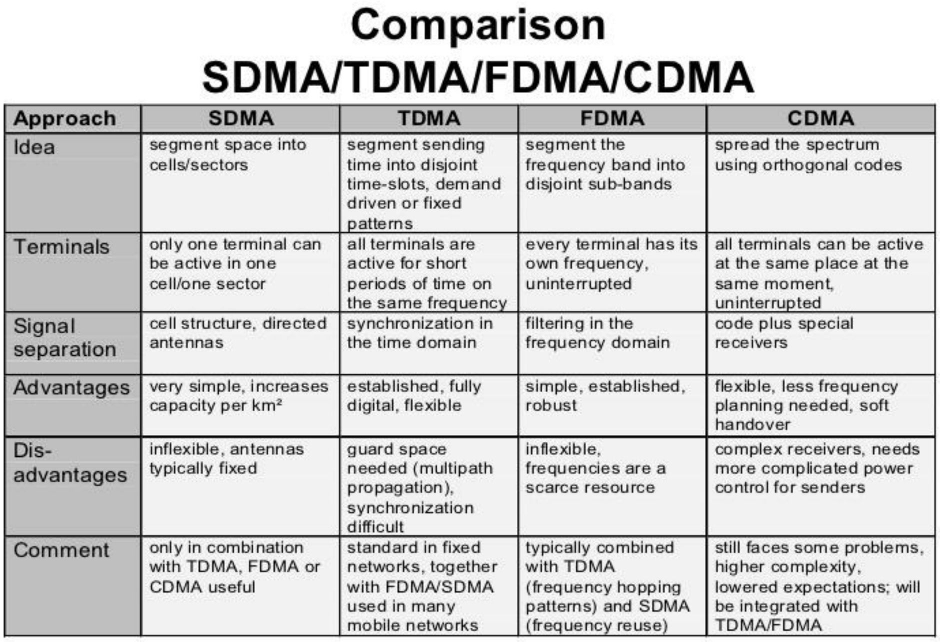
Spread Aloha multiple access (SAMA)

CDMA senders and receivers are not really simple devices. Communicating with *n* devices requires programming of the receiver to be able to decode *n* different codes. Aloha was a very simple scheme, but could only provide a relatively low bandwidth due to collisions. SAMA uses spread spectrum with only one single code (chipping sequence) for spreading for all senders accessing according to aloha.

In SAMA, each sender uses the same spreading code, for ex 110101 as shown below. Sender A and B access the medium at the same time in their narrowband spectrum, so that the three bits shown causes collisions. The same data could also be sent with higher power for shorter periods as show.



The main problem in using this approach is finding good chipping sequences. The maximum throughput is about 18 per cent, which is very similar to Aloha, but the approach benefits from the advantages of spread spectrum techniques: robustness against narrowband interference and simple coexistence with other systems in the same frequency bands.



Comparison SDMA/TDMA/FDMA/CDMA

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 km²	established, fully digital, flexible	simple, established, robust	flexible, less frequency planning needed, soft handover
Dis-advantages	inflexible, antennas typically fixed	guard space needed (multipath 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)	still faces some problems, higher complexity, lowered expectations; will be integrated with TDMA/FDMA

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):

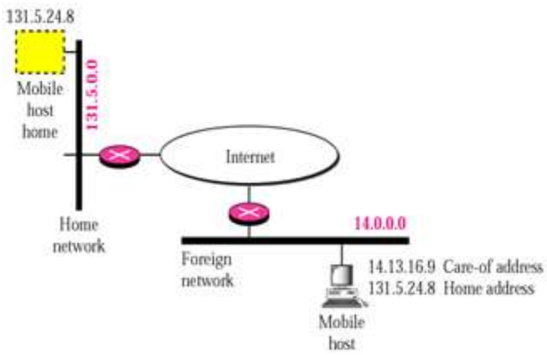
At least one partner is needed for communication. In the following the CN represents this partner for the MN. The CN can be a fixed or mobile node.

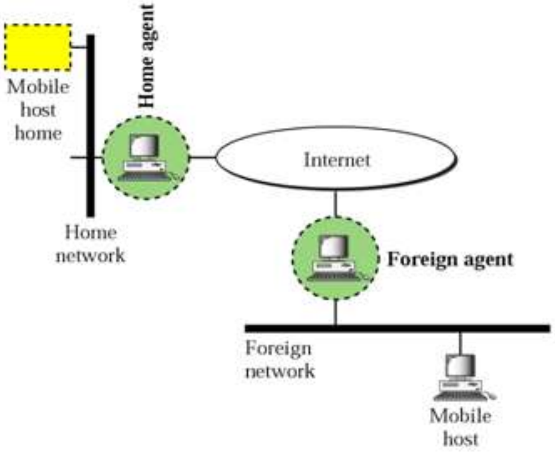
Home network:

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:

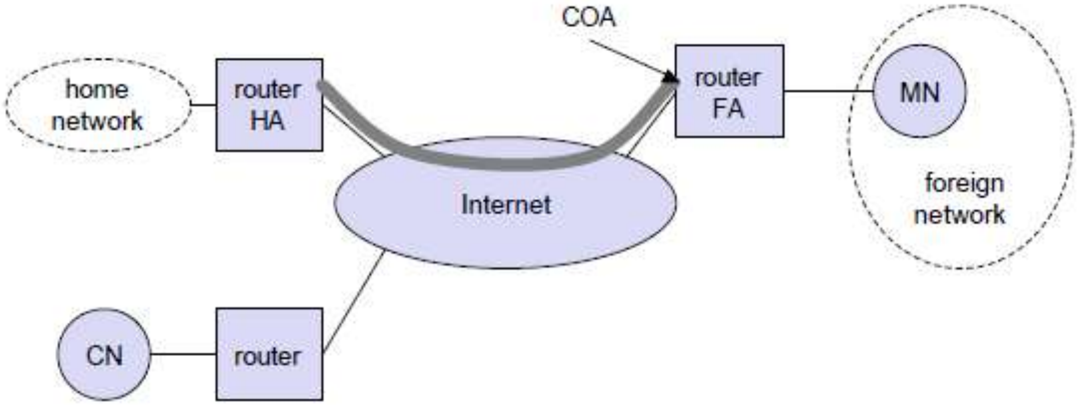
The COA is co-located if the MN temporarily acquired an additional IP address which acts as COA. This address is now topologically correct, and the tunnel endpoint is at the MN. Co-located addresses can be acquired using services such as DHCP.

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. 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.

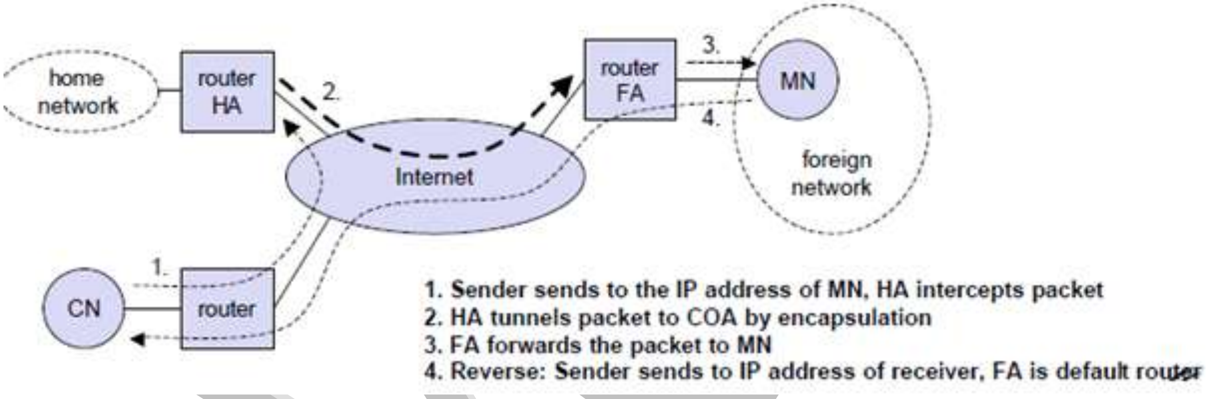


IP packet delivery

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 tunnelled 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.


home network
router HA
2.
router FA
3.
MN
4.
Internet
foreign network
1.
CN
router
1. Sender sends to the IP address of MN, HA intercepts packet
2. HA tunnels packet to COA by encapsulation
3. FA forwards the packet to MN
4. Reverse: Sender sends to IP address of receiver, FA is default router

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:-

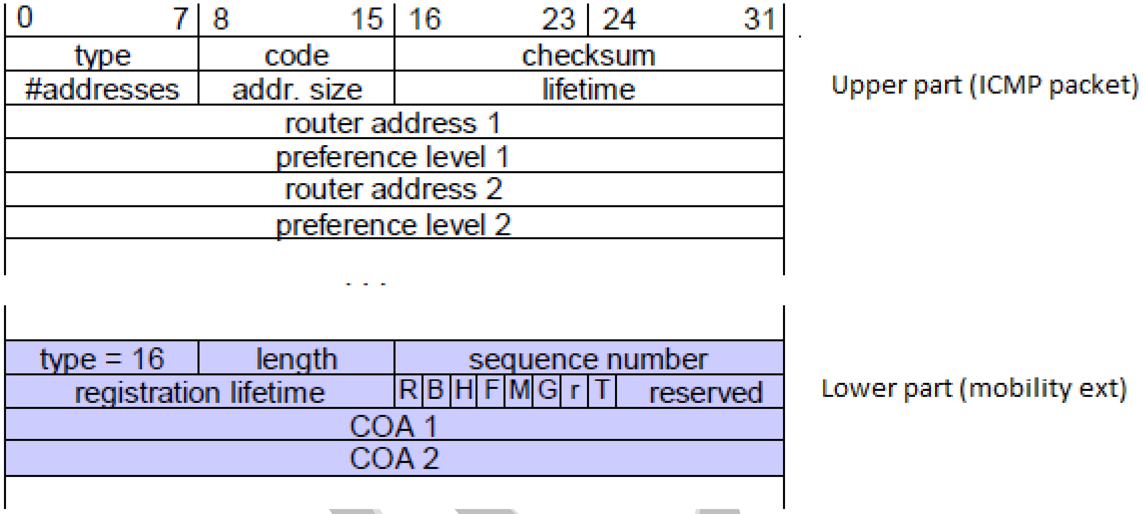
Mobile IP has two addresses for a mobile host: one home address and one care-of address. The home address is permanent; the care-of address changes as the mobile host moves from one network to another. To make the change of address transparent to the rest of the Internet requires a home agent and a foreign 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

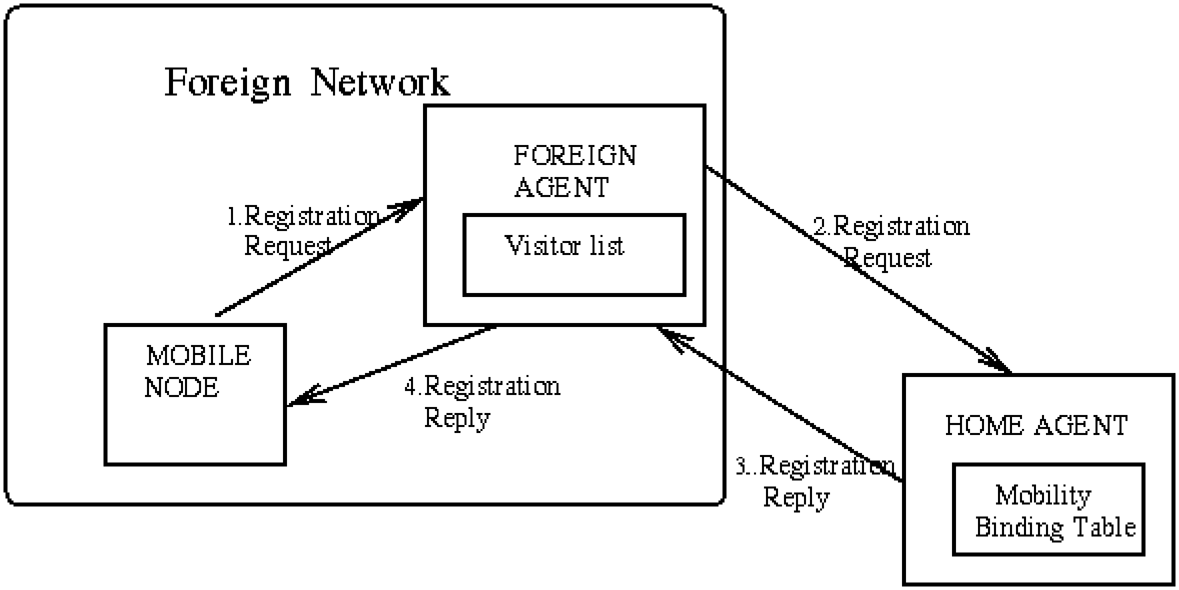
COAs advertised. A foreign agent setting the F bit must advertise at least one COA. A mobile node in a subnet can now receive agent advertisements from either its home agent or a foreign agent. This is one way for the MN to discover its location.

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.

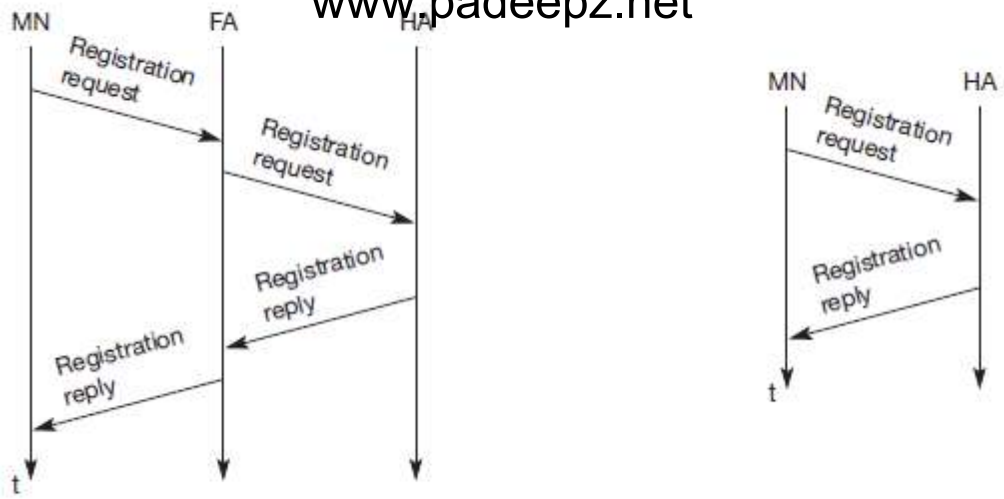
Agent Registration

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.



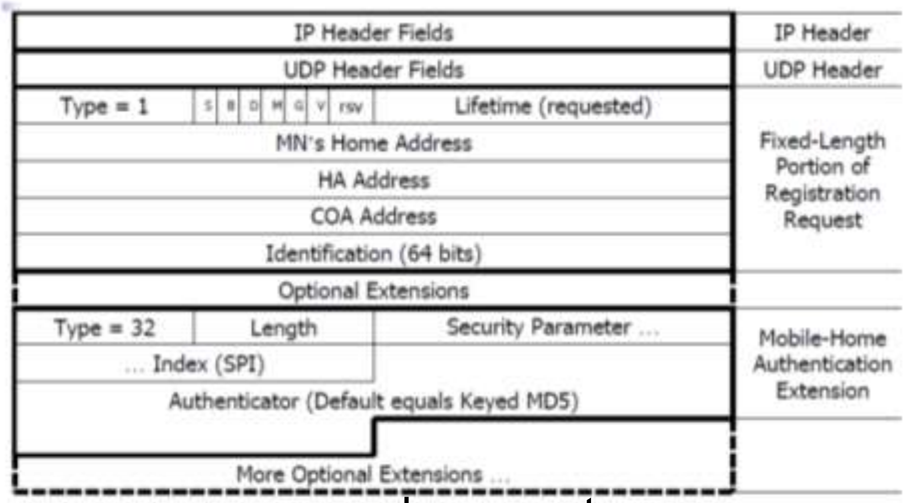
Registration of a mobile node via the FA or directly with the HA

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.

Registration Request

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.



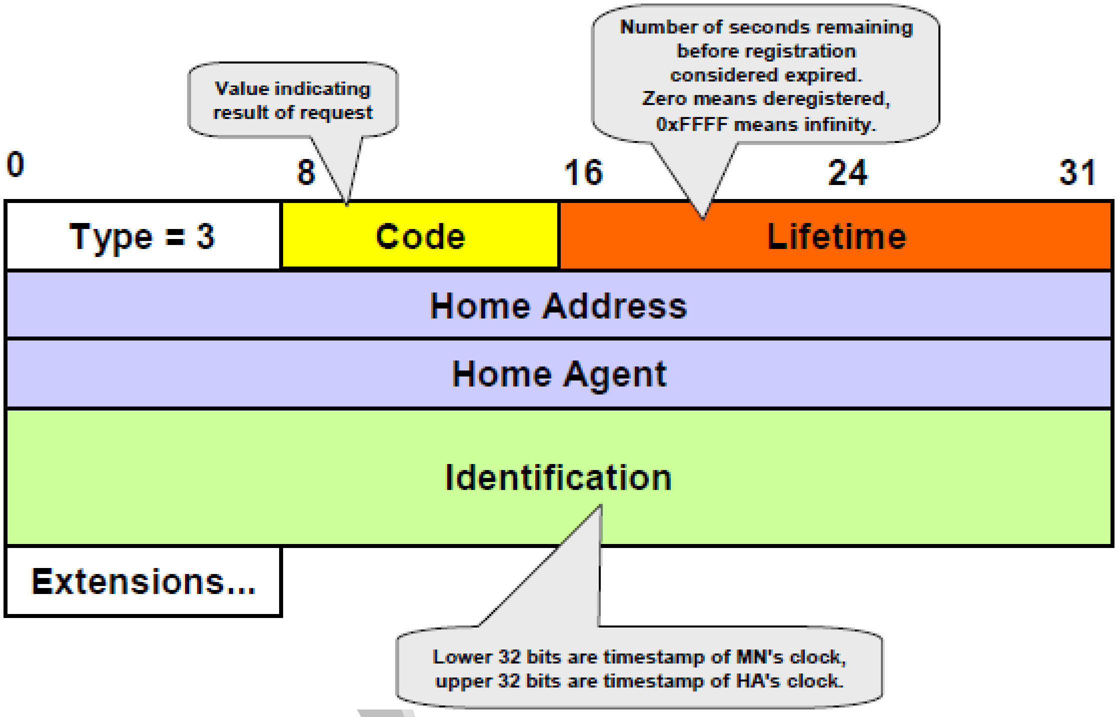
Registration Request

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.

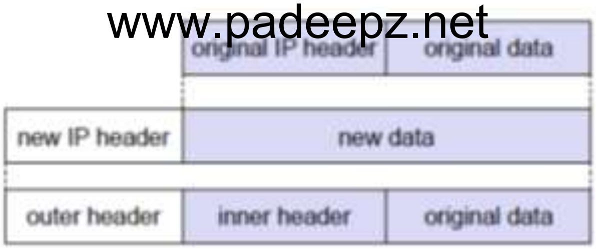
Registration Reply

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.


Number of seconds remaining before registration considered expired. Zero means deregistered, 0xFFFF means infinity.
Value indicating result of request
0
8
16
24
31
Type = 3
Code
Lifetime
Home Address
Home Agent
Identification
Extensions...
Lower 32 bits are timestamp of MN's clock, upper 32 bits are timestamp of HA's clock.

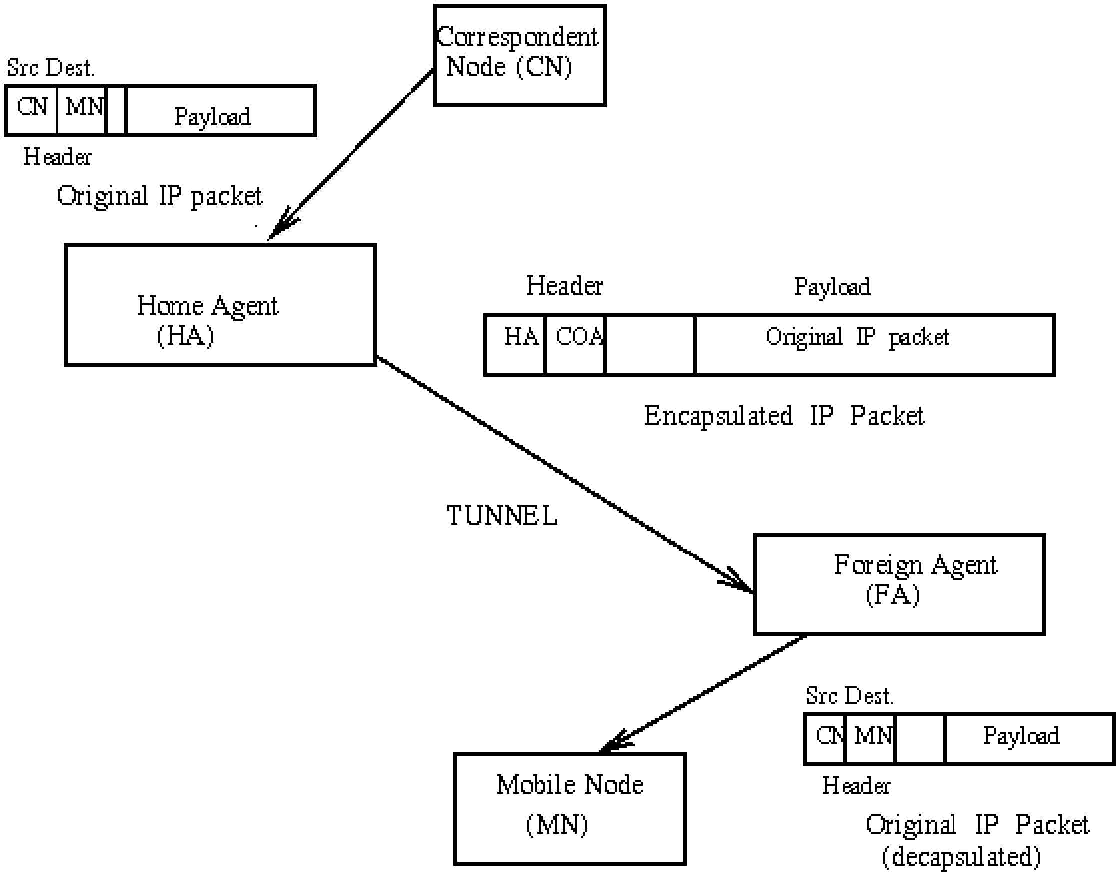
Tunnelling and encapsulation

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.

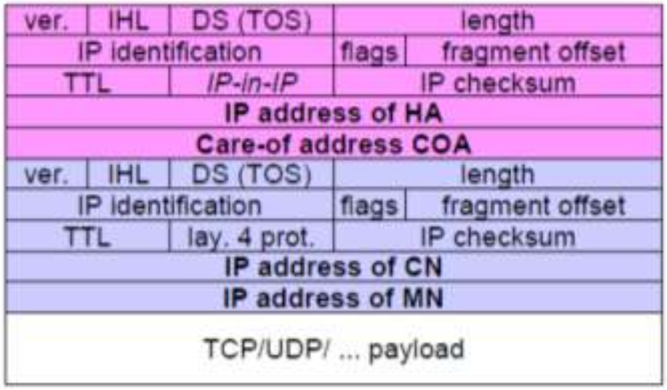
The HA takes the original packet with the MN as destination, puts it into the data part of a new packet and sets the new IP header so that the packet is routed to the COA. The new header is called outer header.



Mobile IP tunnelling

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.



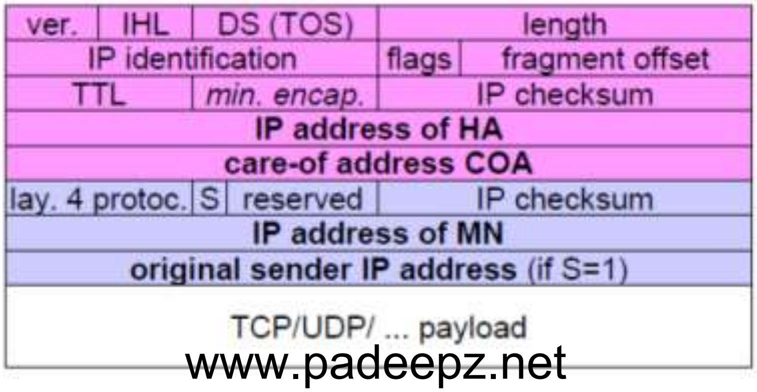
ver.	IHL	DS (TOS)	length	
IP identification			flags	fragment offset
TTL		*IP-in-IP*	IP checksum	
IP address of HA				
Care-of address COA				
ver.	IHL	DS (TOS)	length	
IP identification			flags	fragment offset
TTL		lay. 4 prot.	IP checksum	
IP address of CN				
IP address of MN				
TCP/UDP/ ... payload				

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.

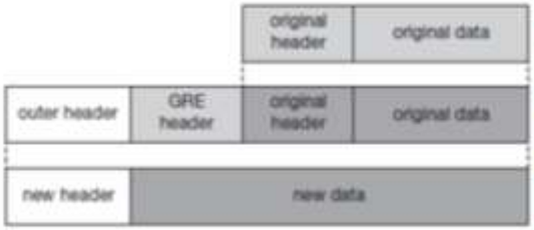


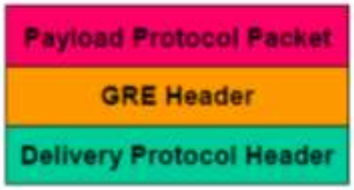
ver.	IHL	DS (TOS)	length	
IP identification			flags	fragment offset
TTL		*min. encap.*	IP checksum	
IP address of HA				
care-of address COA				
lay. 4 protoc.	S	reserved	IP checksum	
IP address of MN				
original sender IP address (if S=1)				
TCP/UDP/ ... payload				

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.

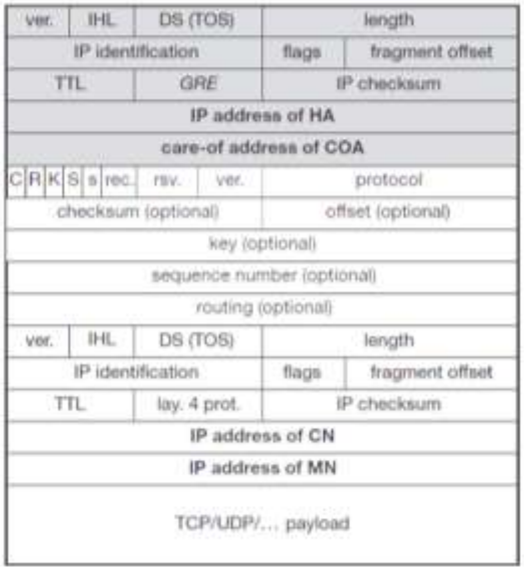
Generic Routing Encapsulation

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.





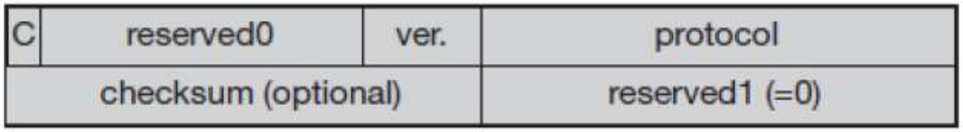
The packet of one protocol suite with the original packet header and data is taken and a new GRE header is prepended. Together this forms the new data part of the new packet. Finally, the header of the second protocol suite is put in front.The following figure shows the fields of a packet inside the 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.



<table>
<tr><td>C</td><td>reserved0</td><td>ver.</td><td>protocol</td></tr>
<tr><td colspan="3">checksum (optional)</td><td>reserved1 (=0)</td></tr>
</table>

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:

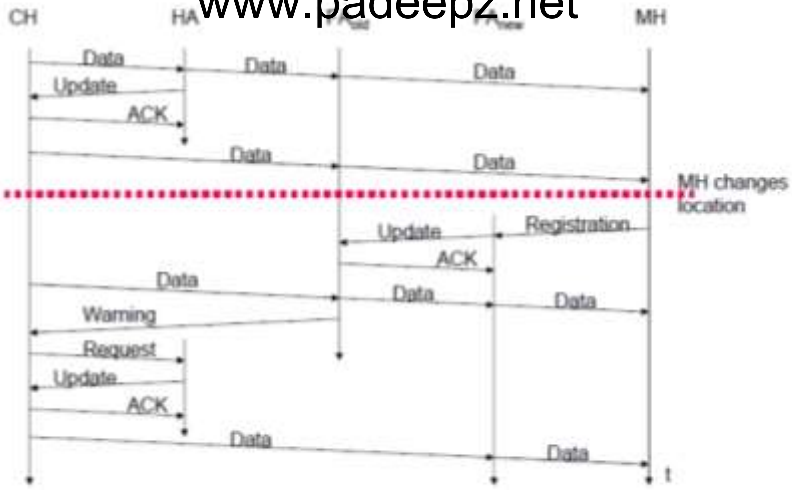
Binding Request: It is sent by the node that wants to know the current location of an MN to the HA. HA checks if it is allowed to reveal the location and then sends back a binding update

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, FAnew. This registration is also forwarded to the HA to update its location database. Furthermore, FAnew informs FAold about the new registration of MN. MN’s registration message contains the address of FAold for this purpose. Passing this information is achieved via an update message, which is acknowledged by FAold.

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, FAold. This FA now notices packets with destination MN, but also knows that it is not the current FA of MN. FAold might now forward these packets to the new COA of MN which is FAnew 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, FAold 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 FAnew, again avoiding triangular routing. Unfortunately, this optimization of mobile IP to avoid triangular routing causes several security problems

Reverse Tunnelling

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

Reverse Tunnelling

Packet Forwarding Reverse Tunnel

Reverse tunneling does not solve problems with *firewalls*, the reverse tunnel can be abused to circumvent security mechanisms (tunnel hijacking) and 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.

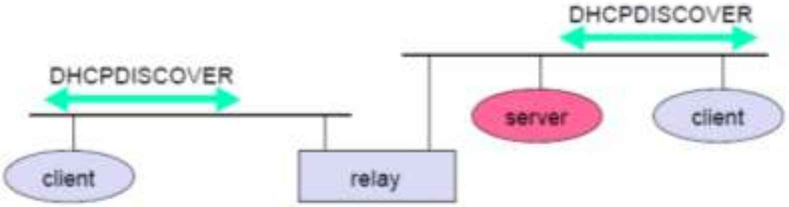
Mobile IPv6 is decoupled from any particular link layer, as it uses IPv6 Neighbor Discovery instead of ARP. This also improves the robustness of the protocol.

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.

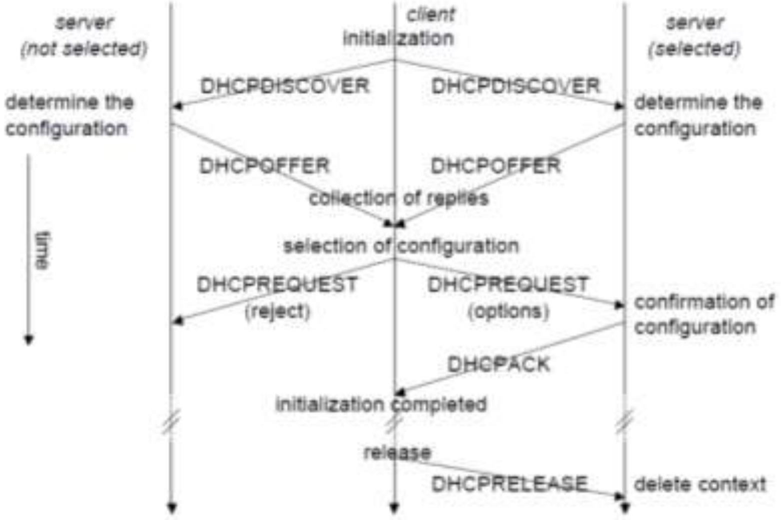
Dynamic Host Configuration Protocol (DHCP)

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 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…

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.

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.

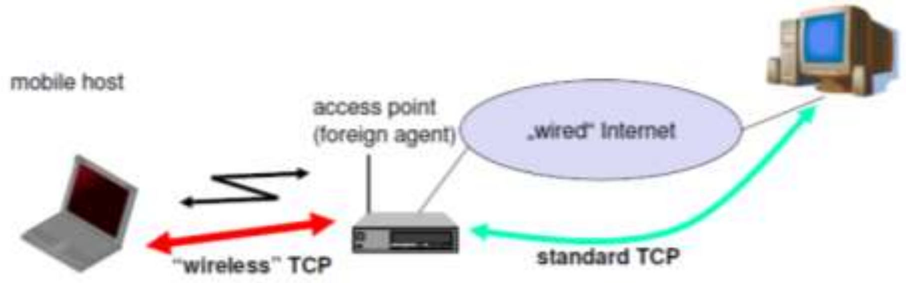
Standard TCP reacts with slow start if acknowledgements are missing, which does not help in the case of transmission errors over wireless links and which does not really help during handover. This behavior results in a severe performance degradation of an unchanged TCP if used together with wireless links or mobile nodes

Classical TCP Improvements

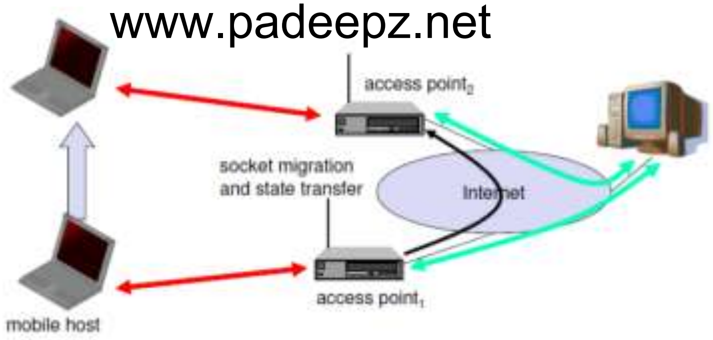
Indirect TCP (I-TCP)

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.

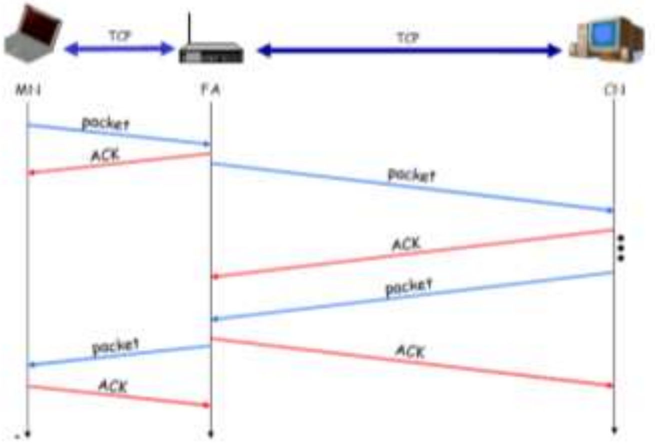


The foreign agent acts as a proxy and relays all data in both directions. If CH (correspondent host) sends a packet to the MH, the FA acknowledges it and forwards it to the MH. MH acknowledges on successful reception, but this is only used by the FA. If a packet is lost on the wireless link, CH doesn't 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

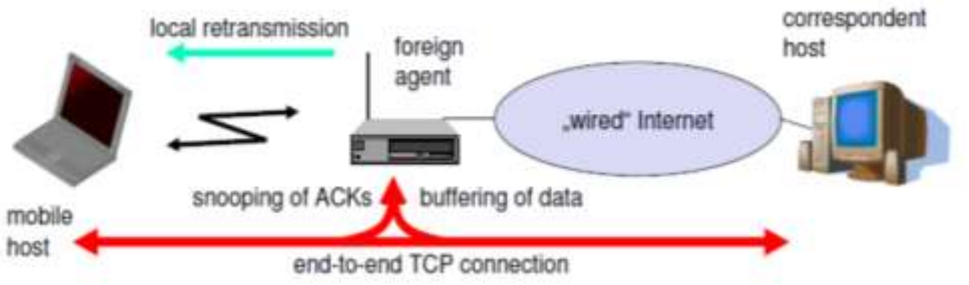
Loss of end-to-end semantics:- an acknowledgement to a sender no longer means that a receiver really has received a packet, foreign agents might crash.

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

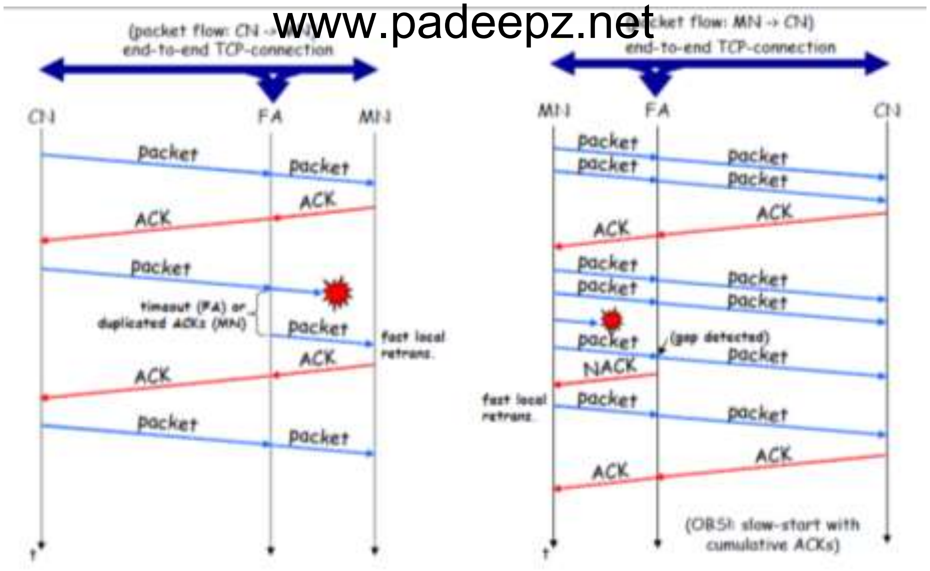
The main drawback of I-TCP is the segmentation of the single TCP connection into two TCP connections, which loses the original end-to-end TCP semantic. A new enhancement, which leaves the TCP connection intact and is completely transparent, is Snooping TCP. The main function is to buffer data close to 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 foreign agent could receive a duplicate ACK which also shows the loss of a packet. Now, the FA retransmits the packet directly from the buffer thus performing a faster retransmission compared to the CH. For transparency, the FA does not acknowledge data to the CH, which would violate end-to-end semantic in case of a FA failure. The foreign agent can filter the duplicate acknowledgements to avoid unnecessary retransmissions of data from the correspondent host. If the foreign agent now crashes, the time-out of the correspondent host still works and triggers a retransmission. The foreign agent may discard duplicates of packets already retransmitted locally and acknowledged by the mobile host. This avoids unnecessary traffic on the wireless link.

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

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. 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 and **Supplementary Services**.

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. Transmission quality can be improved with 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. Transparent bearer services do not try to recover lost data in case of, for example, shadowing or interruptions due to handover. **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)**, and special selective-reject mechanisms to trigger retransmission of erroneous data. Using transparent and non-transparent services, GSM specifies several bearer services for interworking with PSTN, ISDN, and packet switched public data networks (PSPDN) like X.25, which is available worldwide. Data transmission can be full-duplex, synchronous with data rates of 1.2, 2.4, 4.8, and 9.6 kbit/s or full-duplex, asynchronous from 300 to 9,600 bit/s.

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). The primary goal of GSM was the provision of high-quality digital voice transmission. Special codecs (coder/decoder) are used for voice transmission, while other codecs are used for the transmission of analog data for communication with traditional computer modems used in, e.g., fax machines. Another service offered by GSM is the **emergency number** (eg 911, 999). 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. Sending and receiving of SMS is possible during data or voice transmission. It can be used for "serious" applications such as displaying road conditions, e-mail headers or stock quotes, but it can also transfer logos, ring tones, horoscopes and love letters.

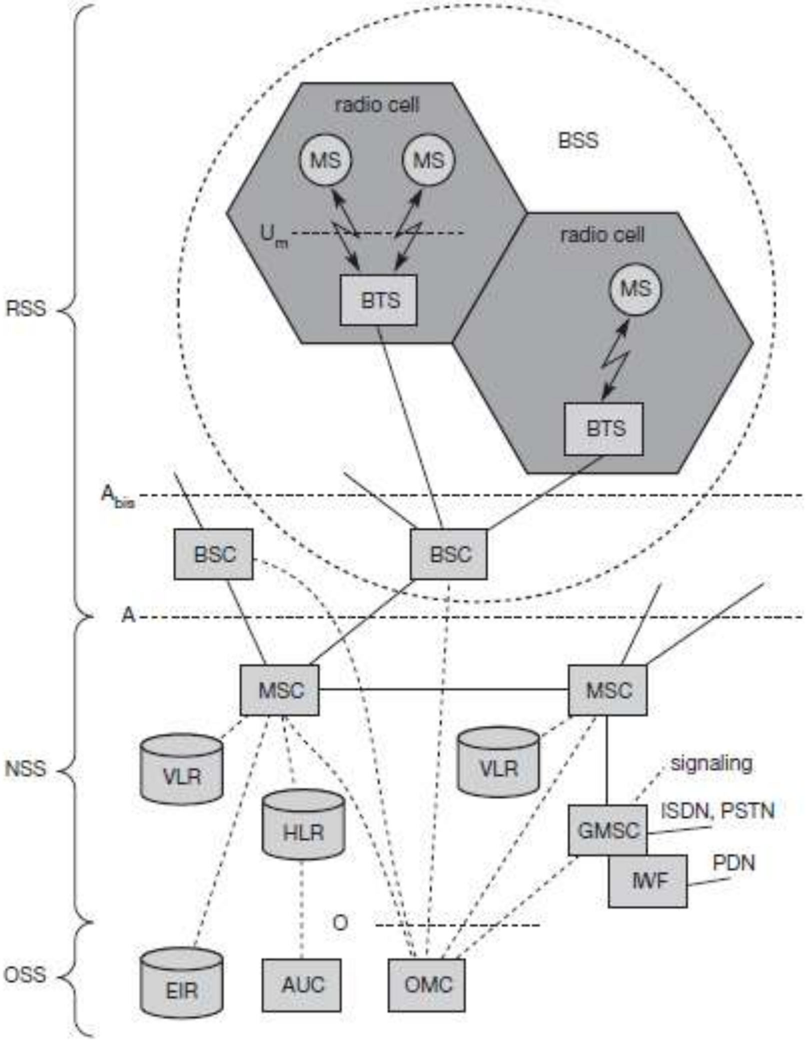
The successor of SMS, the **enhanced message service (EMS)**, offers a larger message size, formatted text, and the transmission of animated pictures, small images and ring tones in a standardized way. But with MMS, EMS was hardly used. MMS offers the transmission of larger pictures (GIF, JPG, WBMP), short video clips etc. and comes with mobile phones that integrate small cameras. Another non-voice tele service is **group 3 fax**, which is available worldwide. In this service, fax data is transmitted as digital data over the analog telephone network according to the ITU-T standards T.4 and T.30 using modems.

Supplementary services:

In addition to tele and bearer services, GSM providers can offer **supplementary services**. 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, barring of incoming/outgoing calls, Advice of Charge (AoC) etc. Standard ISDN features such as **closed user groups** and **multiparty** communication may be available.

GSM Architecture

A GSM system consists of three subsystems, the radio sub system (RSS), the network and switching subsystem (NSS), and the operation subsystem (OSS).



Functional Architecture of a GSM System

Network Switching Subsystem: The NSS is responsible for performing call processing and subscriber related functions. The switching system includes the following functional units:

Home location register (HLR): It is a database used for storage and management of subscriptions. HLR stores permanent data about subscribers, including a subscribers service profile, location information and activity status. When an individual buys a subscription from the PCS provider, he or she is registered in the HLR of that operator.

Visitor location register (VLR): It is a database that contains temporary information about subscribers that is needed by the MSC in order to service visiting subscribers. VLR is always integrated with the MSC. When a MS roams into a new MSC area, the VLR connected to that MSC will request data about the mobile station from the HLR. Later if the mobile station needs to make a call, VLR will be having all the information needed for call setup.

Authentication center (AUC): A unit called the AUC provides authentication and encryption parameters that verify the users identity and ensure the confidentiality of each call.

Equipment identity register (EIR): It is a database that contains information about the identity of mobile equipment that prevents calls from stolen, unauthorized or defective mobile stations.

Mobile switching center (MSC): The MSC performs the telephony switching functions of the system. It controls calls to and from other telephone and data systems.

Radio Subsystem (RSS):

The radio **subsystem (RSS)** comprises all radio specific entities, i.e., the mobile **stations (MS)** and the **base station subsystem (BSS)**. The figure 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 station controllers (BSC): The BSC provides all the control functions and physical links between the MSC and BTS. It is a high capacity switch that provides functions such as handover, cell configuration data, and control of radio frequency (RF) power levels in BTS. A number of BSC's are served by and MSC.

Base transceiver station (BTS): The BTS handles the radio interface to the mobile station. A BTS can form a radio cell or, using sectorized antennas, several and is connected to MS via the **Um interface**, and to the BSC via the **Abis interface**. The Um interface contains all the mechanisms necessary for wireless transmission (TDMA, FDMA etc.)The BTS is the radio equipment (transceivers and antennas) needed to service each cell in the network. A group of BTS's are controlled by an BSC.

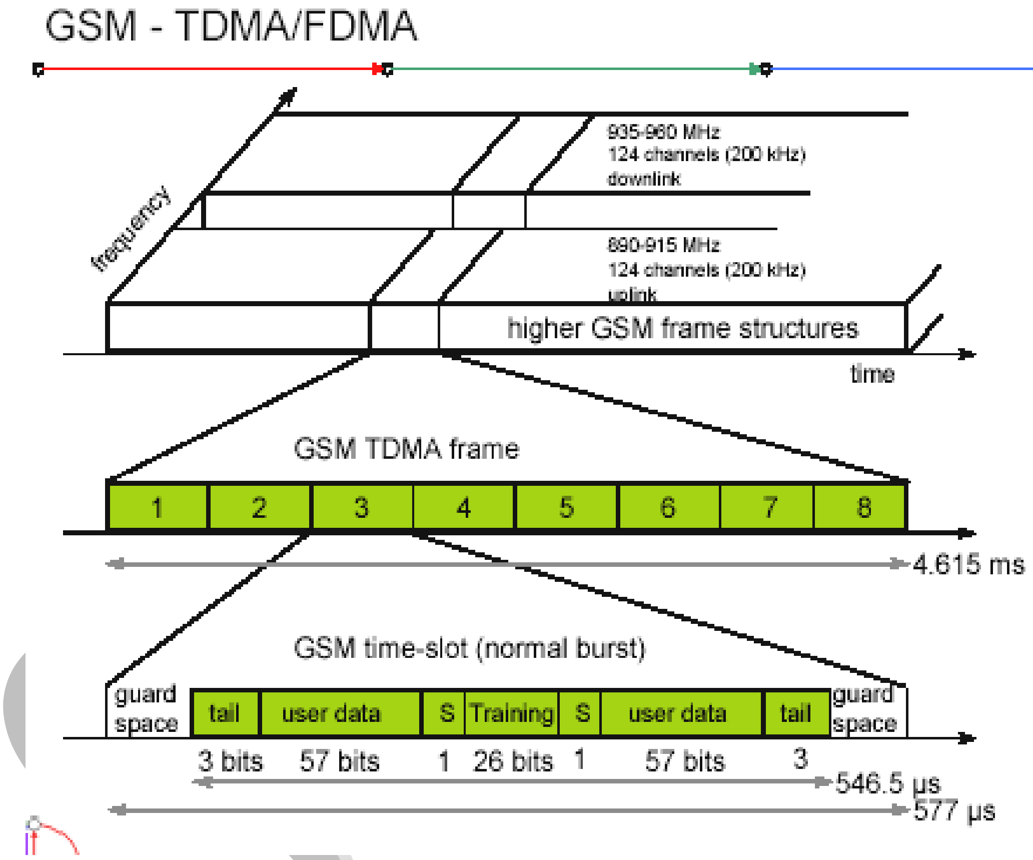
Operation and Support system:

The operations and maintenance center (OMC) is connected to all equipment in the switching system and to the BSC. Implementation of OMC is called operation and support system (OSS). The OSS is the functional entity from which the network operator monitors and controls the system. The purpose of OSS is to offer the customer cost-effective support for centralized, regional and local operational and maintenance activities that are required for a GSM network. OSS provides a network overview and allows engineers to monitor, diagnose and troubleshoot every aspect of the GSM network.

The mobile station (MS) consists of the mobile equipment (the terminal) and a smart card called the Subscriber Identity Module (SIM). The SIM provides personal mobility, so that the user can have access to subscribed services irrespective of a specific terminal. By inserting the SIM card into another GSM terminal, the user is able to receive calls at that terminal, make calls from that terminal, and receive other subscribed services. The mobile equipment is uniquely identified by the International Mobile Equipment Identity (IMEI). The SIM card contains the International Mobile Subscriber Identity (IMSI) used to identify the subscriber to the system, a secret key for authentication, and other information. The IMEI and the IMSI are independent, thereby allowing personal mobility. The SIM card may be protected against unauthorized use by a password or personal identity number.

Radio Interface

The most interesting interface in a GSM system is Um, the radio interface, as it comprises various multiplexing and media access mechanisms. GSM implements SDMA using cells with BTS and assigns an MS to a BTS.



GSM TDMA Frame, Slots and Bursts

Each of the 248 channels is additionally separated in time via a **GSM TDMA frame**, i.e., each 200 kHz carrier is subdivided into frames that are repeated continuously. The duration of a frame is 4.615 ms. A frame is again subdivided into 8 **GSM time slots**, where each slot represents a physical TDM channel and lasts for 577 μs. Each TDM channel occupies the 200 kHz carrier for 577 μs every 4.615 ms. Data is transmitted in small portions, called **bursts**. The following figure shows a so called **normal burst** as used for data transmission inside a time slot. As shown, the burst is only 546.5 μs long and contains 148 bits. The remaining 30.5 μs are used as **guard space** to avoid overlapping with other bursts due to different path delays and to give the transmitter time to turn on and off. The first and last three bits of a normal burst (**tail**) are all set to 0 and can be used to enhance the receiver performance. The **training** sequence in the middle of a slot is used to adapt the parameters of the receiver to the current path propagation characteristics and to select the strongest signal in case of multi-path propagation. A flag **S** indicates whether the **data** field contains user or network control data.

Apart from the normal burst, ETSI (1993a) defines four more bursts for data transmission: a **frequency correction** burst allows the MS to correct the local oscillator to avoid interference with neighbouring channels, a **synchronization burst** with an extended training sequence synchronizes the MS with the BTS in time, an **access burst** is used for the initial connection setup between MS and BTS, and finally a **dummy burst** is used if no data is available for a slot.

Logical channels and frame hierarchy

Two types of channels, namely physical channels and logical channels are present.

Physical channel: channel defined by specifying both, a carrier frequency and a TDMA timeslot number.

Logic channel: logical channels are multiplexed into the physical channels. Each logic channel performs a specific task. Consequently the data of a logical channel is transmitted in the corresponding timeslots of the physical channel. During this process, logical channels can occupy a part of the physical channel or even the entire channel.

Each of the frequency carriers is divided into frames of 8 timeslots of approximately 577s (15/26 s) duration with 156.25 bits per timeslot. The duration of a TDMA frame is 4.615ms (577s x 8 = 4.615 ms). The bits per timeslot and frame duration yield a gross bit rate of about 271kbps per TDMA frame.

TDMA frames are grouped into two types of multiframes:

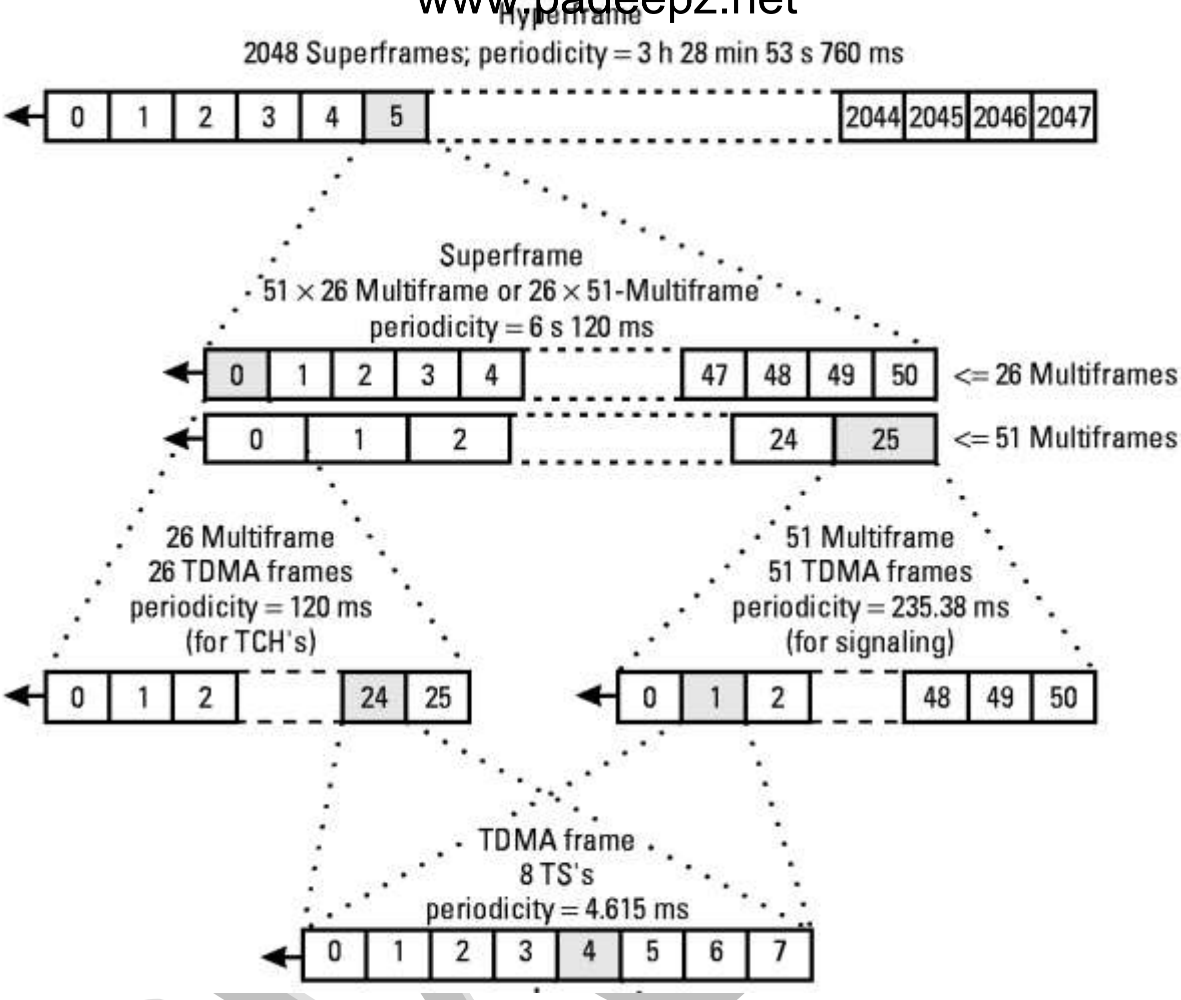
26-frame multiframe (4.615ms x 26 = 120 ms) comprising of 26 TDMA frames. This multiframe is used to carry traffic channels and their associated control channels.

51-frame multiframe (4.615ms x 51 □ □ 235.4 ms) comprising 51 TDMA frames. This multiframe is exclusively used for control channels. The multiframe structure is further multiplexed into a single superframe of duration of 6.12sec. This means a superframe consists of

51 multiframes of 26 frames.

26 multiframes of 51 frames.

The last multiplexing level of the frame hierarchy, consisting of 2048 superframes (2715648 TDMA frames), is a hyperframe. This long time period is needed to support the GSM data encryption mechanisms. The frame hierarchy is shown below:



GSM Frame Hierarchy

There are two different types of logical channel within the GSM system: Traffic channels (TCHs), Control channels (CCHs).

Traffic Channels:

Traffic channels carry user information such as encoded speech or user data. Traffic channels are defined by using a 26-frame multiframe. Two general forms are defined:

i. Full rate traffic channels (TCH/F), at a gross bit rate of 22.8 kbps (456bits / 20ms)

ii. Half rate traffic channels (TCH/H), at a gross bit rate of 11.4 kbps.

Uplink and downlink are separated by three slots (bursts) in the 26-multiframe structure.

This simplifies the duplexing function in mobile terminals design, as mobiles will not need to transmit and receive at the same time. The 26-frame multiframe structure, shown below multiplexes two types of logical channels, a TCH and a Slow Associated Control Channel (SACCH).

However, if required, a Fast Associated Control CHannel (FACCH) can steal TCH in order to transmit control information at a higher bit rate. This is usually the case during the handover process. In total 24 TCH/F are transmitted and one SACCH.

Control Channels:

Control channels carry system signalling and synchronisation data for control procedures such as location registration, mobile station synchronisation, paging, random access etc. between base station and mobile station. Three categories of control channel are defined: Broadcast, Common and Dedicated. Control channels are multiplexed into the 51-frame multiframe.

Broadcast control channel (BCCH):

A BTS uses this channel to signal information to all MSs within a cell. Information transmitted in this channel is, e.g., the cell identifier, options available within this cell (frequency hopping), and frequencies available inside the cell and in neighboring cells. The BTS sends information for frequency correction via the **frequency correction channel (FCCH)** and information about time synchronization via the **synchronization channel (SCH)**, where both channels are subchannels of the BCCH.

Common control channel (CCCH):

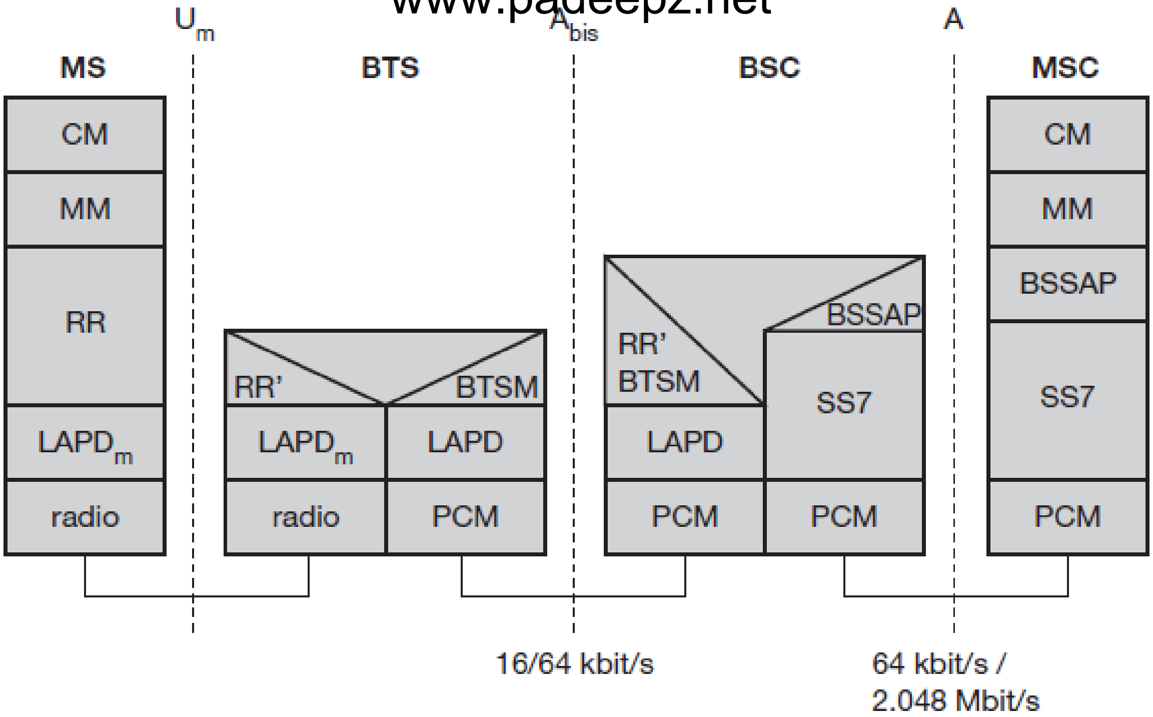
All information regarding connection setup between MS and BS is exchanged via the CCCH. For calls toward an MS, the BTS uses the **paging channel (PCH)** for paging the appropriate MS. If an MS wants to set up a call, it uses the **random access channel (RACH)** to send data to the BTS. The RACH implements multiple access (all MSs within a cell may access this channel) using slotted Aloha. This is where a collision may occur with other MSs in a GSM system. The BTS uses the **access grant channel (AGCH)** to signal an MS that it can use a TCH or SDCCH for further connection setup.

Dedicated control channel (DCCH):

While the previous channels have all been unidirectional, the following channels are bidirectional. As long as an MS has not established a TCH with the BTS, it uses the **stand-alone dedicated control channel (SDCCH)** with a low data rate (782 bit/s) for signaling. This can comprise authentication, registration or other data needed for setting up a TCH. Each TCH and SDCCH has a **slow associated dedicated control channel (SACCH)** associated with it, which is used to exchange system information, such as the channel quality and signal power level. Finally, if more signaling information needs to be transmitted and a TCH already exists, GSM uses a **fast associated dedicated control channel (FACCH)**. The FACCH uses the time slots which are otherwise used by the TCH. This is necessary in the case of handovers where BTS and MS have to exchange larger amounts of data in less time.

GSM Protocols

The signalling protocol in GSM is structured into three general layers depending on the interface, as shown below. Layer 1 is the physical layer that 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 qualit**y 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.



Protocol Architecture for Signaling

The main tasks of the physical layer comprise **channel coding** and **error detection/correction**, which is directly combined with the coding mechanisms. Channel coding makes extensive use of different **forward error correction (FEC)** schemes.

Signaling between entities in a GSM network requires higher layers. For this purpose, the **LAPDm** protocol has been defined at the Um interface for **layer two**. LAPDm has been derived from link access procedure for the D-channel (**LAPD**) in ISDN systems, which is a version of HDLC. LAPDm is a lightweight LAPD because it does not need synchronization flags or checksumming for error detection. LAPDm offers reliable data transfer over connections, resequencing of data frames, and flow control.

The network layer in GSM, layer three, comprises several sublayers. The lowest sublayer 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. Mobility management (MM) contains functions for registration, authentication, identification, location updating, and the provision of a temporary mobile subscriber identity (TMSI). Finally, the call management (CM) layer contains three entities: call control (CC), short message service (SMS), and supplementary service (SS). SMS allows for message transfer using the control channels SDCCH and SACCH, while SS offers the services like user identification, call redirection, or forwarding of ongoing calls. 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. This layer also provides functions to send in-band tones, called dual tone multiple frequency (DTMF), over the GSM network. These tones are used, e.g., for the remote control of answering machines or the entry of PINs in electronic banking and are, also used for dialing in traditional analog telephone systems. Additional protocols are used at the Abis and A interfaces.

Data transmission at the physical layer typically uses **pulse code modulation (PCM)** systems. LAPD is used for layer two at Abis, BTSM for BTS management. **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)**.

Localization and Calling

The fundamental feature of the GSM system is the automatic, worldwide localization of users for which, the system performs periodic location updates. The HLR always contains information about the current location and the VLR currently responsible for the MS informs the HLR about the location changes. Changing VLRs with uninterrupted availability is called roaming. Roaming can take place within a network of one provider, between two providers in a country and also between different providers in different countries.

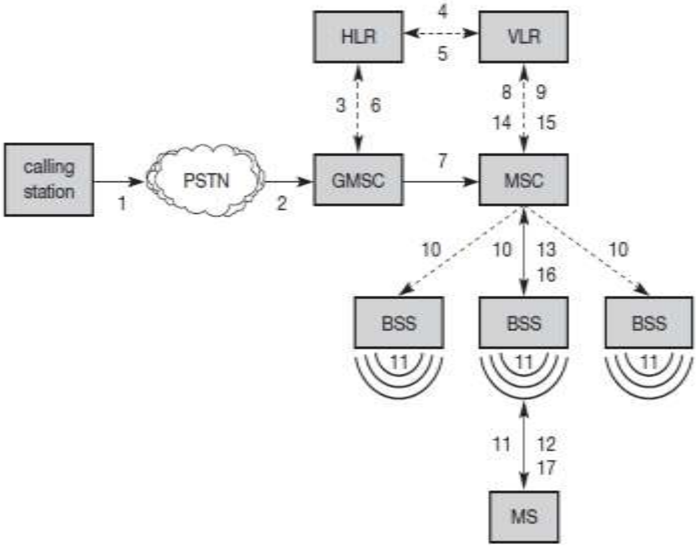
To locate and address an MS, several numbers are needed:

Mobile station international ISDN number (MSISDN):- The only important number for a user of GSM is the phone number. This number consists of the country code (CC), the national destination code (NDC) 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), the mobile network code (MNC), and finally the mobile subscriber identification number (MSIN).

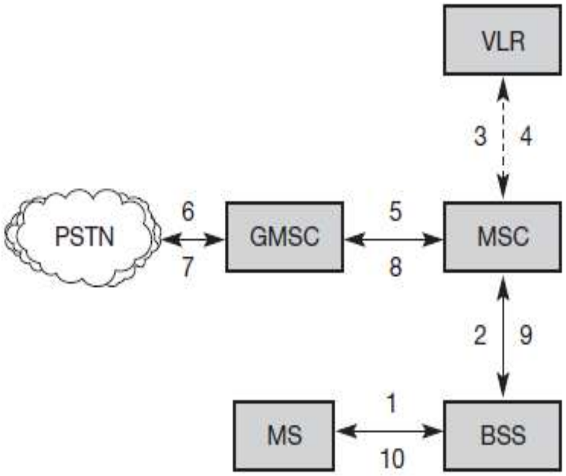
Temporary mobile subscriber identity (TMSI): To hide the IMSI, which would give away the exact identity of the user signalling over the air interface, GSM uses the 4 byte TMSI for local subscriber identification.

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. For *a mobile terminated call (MTC)*, the following figure shows the different steps that take place:

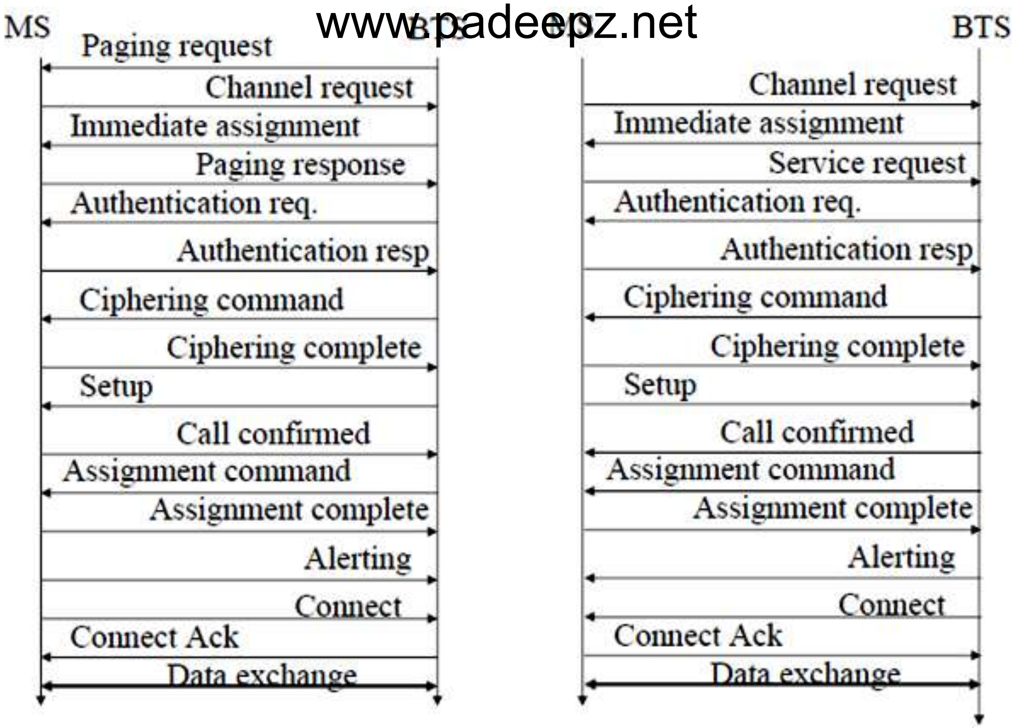


Mobile Terminated Call (MTC)

step 1: User dials the phone number of a GSM subscriber.
step 2: The fixed network (PSTN) identifies the number belongs to a user in GSM network and forwards the call setup to the Gateway MSC (GMSC).
step 3: The GMSC identifies the HLR for the subscriber and signals the call setup to HLR
step 4: The HLR checks for number existence and its subscribed services and requests an MSRN from the current VLR.
step 5: VLR sends the MSRN to HLR
step 6: Upon receiving MSRN, the HLR determines the MSC responsible for MS and forwards the information to the GMSC
step 7: The GMSC can now forward the call setup request to the MSC indicated
step 8: The MSC requests the VLR for the current status of the MS
step 9: VLR sends the requested information
step 10: If MS is available, the MSC initiates paging in all cells it is responsible for.
step 11: The BTSs of all BSSs transmit the paging signal to the MS
step 12: Step 13: If MS answers, VLR performs security checks
step 15: Till step 17: Then the VLR signals to the MSC to setup a connection to the MS. For a mobile originated call (MOC), the following steps take place:



step 1: The MS transmits a request for a new connection
step 2: The BSS forwards this request to the MSC
step 3: Step 4: The MSC then checks if this user is allowed to set up a call with the requested 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 connection setup (in either direction).



Message flow for MTC and MOC

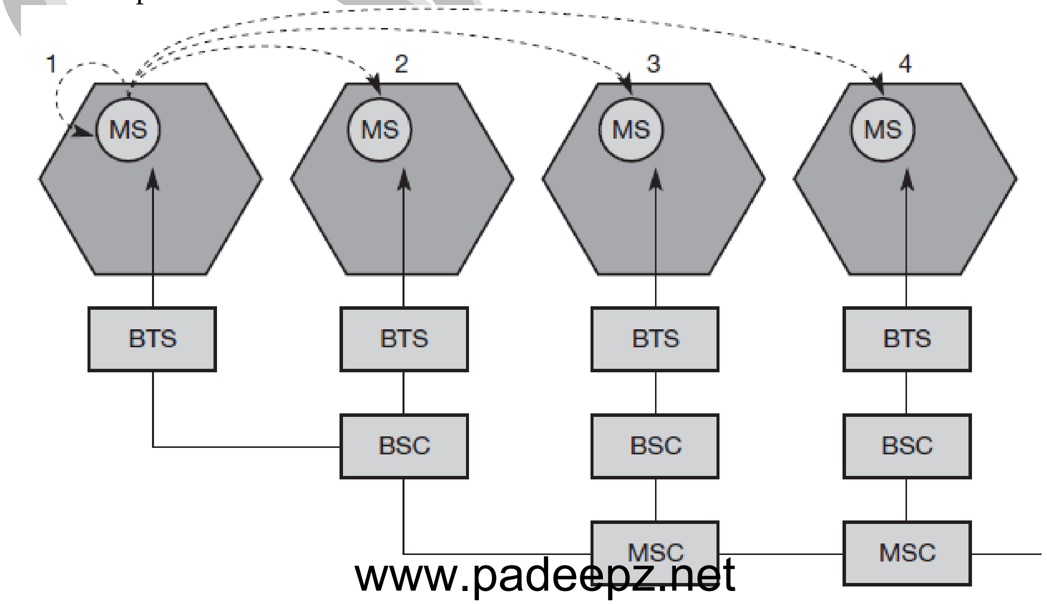
Handover

Cellular systems require **handover** procedures, as single cells do not cover the whole service area. However, a handover should not cause a cut-off, also called **call drop**. GSM aims at maximum handover duration of 60 ms. There are two basic reasons for a handover:

1. The mobile station **moves out of the range** of a BTS, decreasing the received **signal level** increasing the **error rate** thereby diminishing the **quality of the radio link.**

2. Handover may be due to **load balancing,** when an MSC/BSC decides the traffic is too high in one cell and shifts some MS to other cells with a lower load.

The four possible handover scenarios of GSM are shown below:



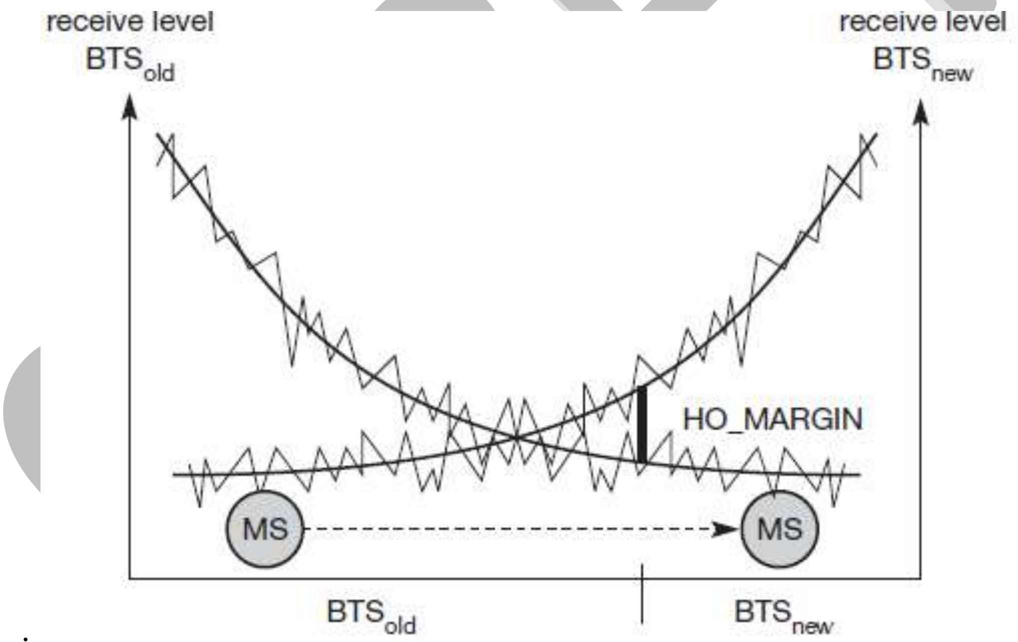
Intra-cell handover: Within a cell, narrow-band interference could make transmission at a certain frequency impossible. The BSC could then decide to change the carrier frequency (scenario 1).

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).

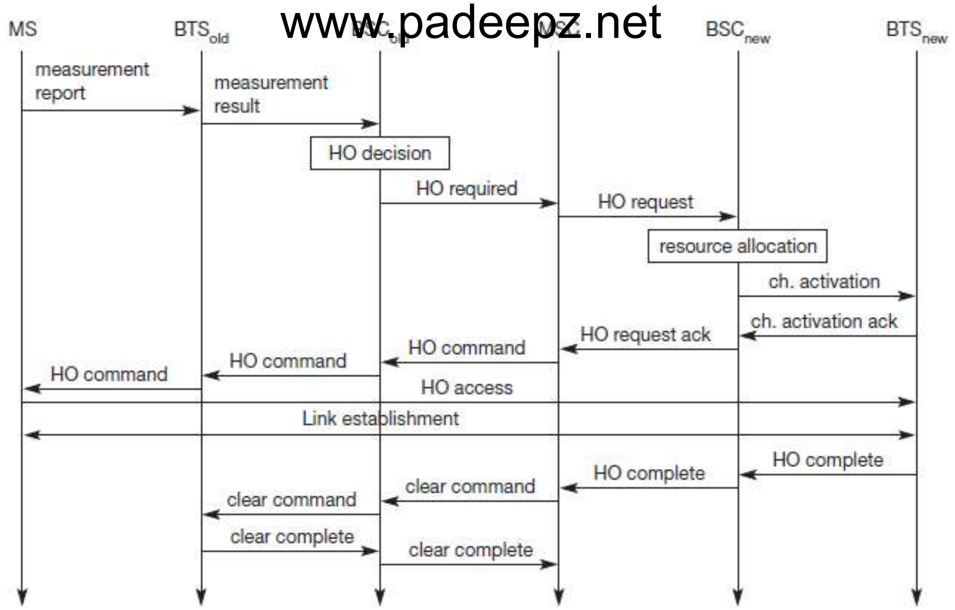
Inter MSC handover: A handover could be required between two cells belonging to different MSCs. Now both MSCs perform the handover together (scenario 4).

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 BCCHs).



.

Handover decision depending on receive level



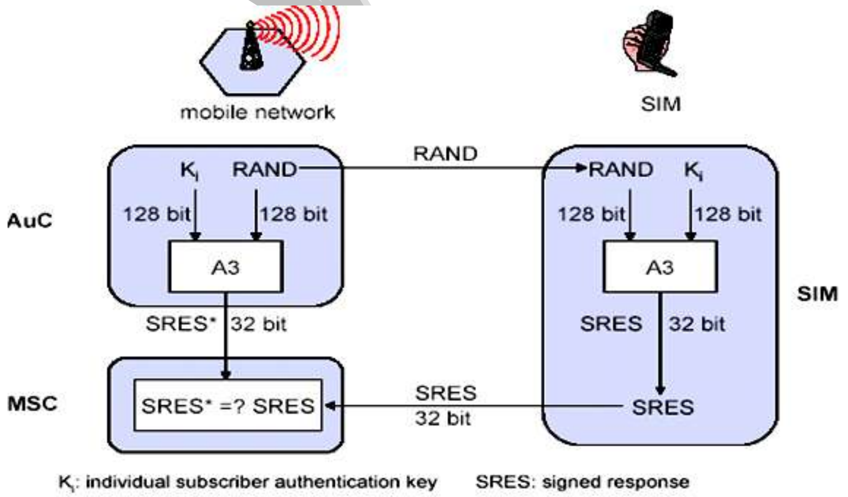
Intra-MSC handover

More sophisticated handover mechanisms are needed for seamless handovers between different systems.

Security

GSM offers several security services using confidential information stored in the AuC and in the individual SIM. The SIM stores personal, secret data and is protected with a PIN against unauthorized use. 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**. The various security services offered by GSM are:

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. This step is based on a challenge-response scheme as shown below:

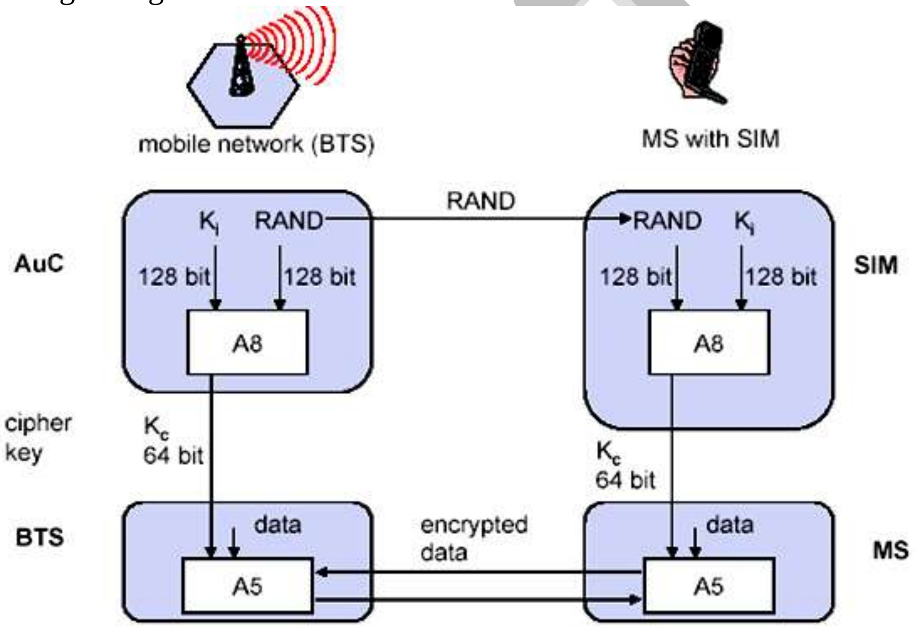


Subscriber Authentication

Authentication is based on the SIM, which stores the **individual authentication key Ki**, the **user identification IMSI**, and the algorithm used for authentication **A3**. The AuC performs the basic generation of random values RAND, signed responses SRES, and cipher keys Kc 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.

Confidentiality:

All user-related data is encrypted. After authentication, BTS and MS apply encryption to voice, data, and signalling as shown below.



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**, which 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.

Anonymity:

To provide user anonymity, all data is encrypted before transmission, and user identifiers 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.

New Data Services

To enhance the data transmission capabilities of GSM, two basic approaches are possible. As the basic GSM is based on connection-oriented traffic channels, e.g., with 9.6 kbit/s each, several channels could be combined to increase bandwidth. This system is called **HSCSD {high speed circuit switched data}.** A more progressive step is the introduction of packet oriented traffic in GSM, i.e., shifting the paradigm from connections/telephone thinking to packets/internet thinking. The system is called **GPRS {general packet radio service}**.

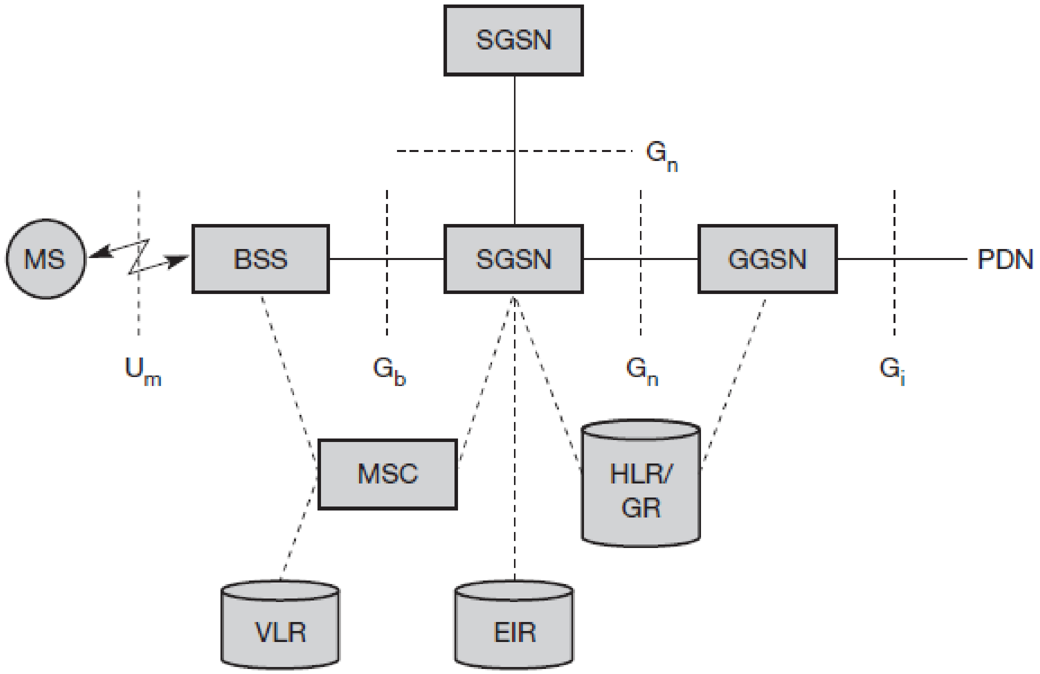
HSCD:

A straightforward improvement of GSM's data transmission capabilities is high speed circuit switched data (HSCSD) in which 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 behaviour of downloading more data compared to uploading. A major disadvantage of HSCD is that it still uses the connection-oriented mechanisms of GSM, which is not efficient for computer data traffic.

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. 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. The GPRS concept is independent of channel characteristics and of the type of channel (traditional GSM traffic or control channel), and does not limit the maximum data rate (only the GSM transport system limits the rate). All GPRS services can be used in parallel to conventional services. GPRS includes several **security services** such as authentication, access control, user identity confidentiality, and user information confidentiality.

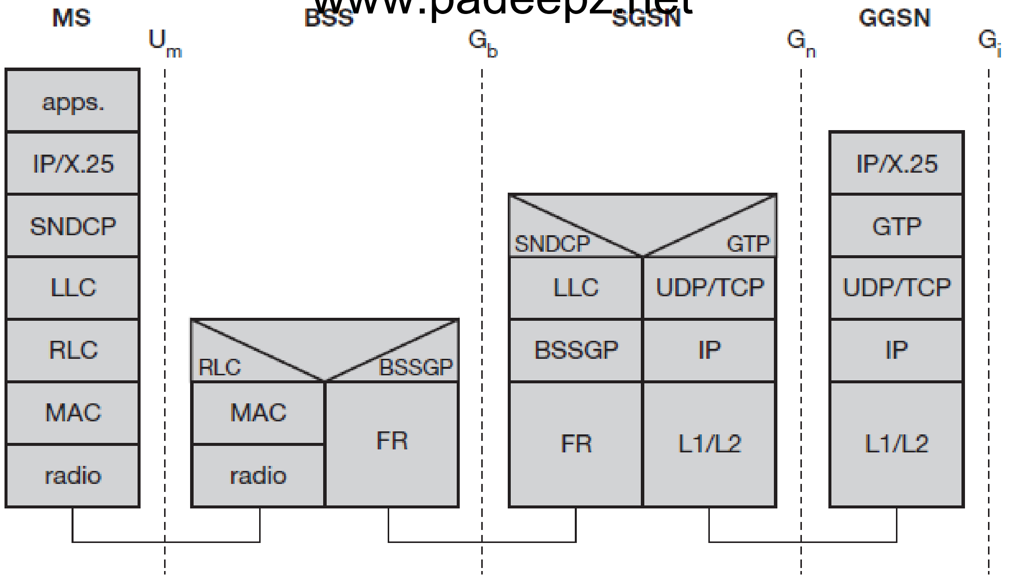
The GPRS architecture introduces two new network elements, which are called GPRS support nodes (GSN) and are in fact routers. All GSNs are integrated into the standard GSM architecture, and many new interfaces have been defined. The gateway GPRS support node (GGSN) is the interworking unit between the 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 IPbased 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.



GPRS Architecture Reference Model

As shown above, packet data is transmitted from a PDN, via the GGSN and SGSN directly to the BSS and finally to the MS. The MSC, which is responsible for data transport in the traditional circuit-switched GSM, is only used for signaling in the GPRS scenario. Before sending any data over the GPRS network, an MS must attach to it, following the procedures of the **mobility management**. The attachment procedure includes assigning a temporal identifier, called a **temporary logical link identity (TLLI)**, and a **ciphering key sequence number (CKSN)** for data encryption. For each MS, a **GPRS context** is set up and stored in the MS and in the corresponding SGSN. Besides attaching and detaching, mobility management also comprises functions for authentication, location management, and ciphering.

The following figure shows the protocol architecture of the transmission plane for GPRS. All data within the GPRS backbone, i.e., between the GSNs, is transferred using the **GPRS tunnelling protocol (GTP)**. GTP can use two different transport protocols, either the reliable **TCP** (needed for reliable transfer of X.25 packets) or the non-reliable **UDP** (used for IP packets). The network protocol for the GPRS backbone is **IP** (using any lower layers). To adapt to the different characteristics of the underlying networks, the **subnetwork dependent convergence protocol (SNDCP)** is used between an SGSN and the MS. On top of SNDCP and GTP, user packet data is tunneled from the MS to the GGSN and vice versa. To achieve a high reliability of packet transfer between SGSN and MS, a special LLC is used, which comprises ARQ and FEC mechanisms for PTP (and later PTM) services.



GPRS Transmission Plane Protocol Reference Model

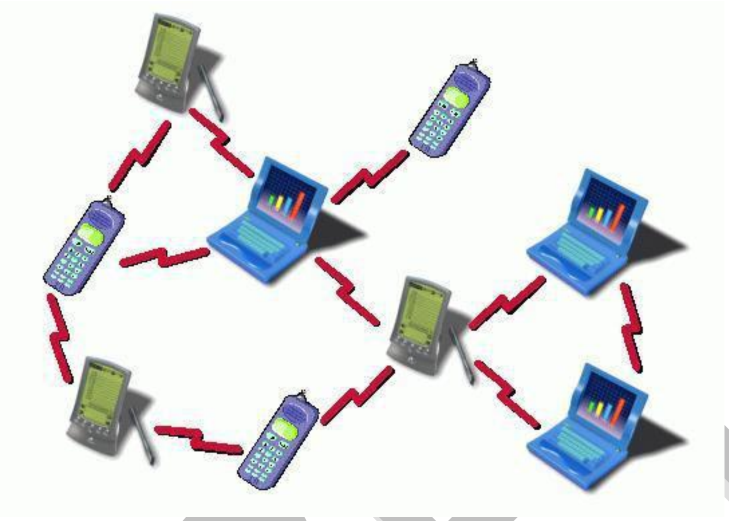
A base station subsystem GPRS protocol (BSSGP) is used to convey routing and QoS-related information between the BSS and SGSN. BSSGP does not perform error correction and works on top of a frame relay (FR) network. Finally, radio link dependent protocols are needed to transfer data over the Um interface. The radio link protocol (RLC) provides a reliable link, while the MAC controls access with signalling procedures for the radio channel and the mapping of LLC frames onto the GSM physical channels. The radio interface at Um needed for GPRS does not require fundamental changes compared to standard GSM.

HISTORICAL DEVELOPMENTS OF MANET

In early 1970s, the Mobile Ad hoc Network (MANET) was called packet radio network, which was sponsored by Defense Advanced Research Projects Agency (DARPA). They had a project named packet radio having several wireless terminals that could communication with each other on battlefields. “It is interesting to note that these early packet radio systems predict the Internet and indeed were part of the motivation of the original Internet Protocol suite” . The whole life cycle of Ad hoc networks could be categorized into the first, second, and the third generation Ad hoc networks systems. Present Ad hoc networks systems are considered the third generation . The first generation goes back to 1972. At the time, they were called PRNET (Packet Radio Networks). In conjunction with ALOHA (Arial Locations of Hazardous Atmospheres) and CSMA (Carrier Sense Medium Access), approaches for medium access control and a kind of distance-vector routing PRNET were used on a trial basis to provide different networking capabilities in a combat environment. The second generation of Ad hoc networks emerged in 1980s, when the Ad hoc network systems were further enhanced and implemented as a part of the SURAN (Survivable Adaptive Radio Networks) program. This provided a packet-switched network to the mobile battlefield in an environment without infrastructure. This On-Demand Routing In Multi-Hop Wireless Mobile Ad hoc Networks Overview of Mobile Ad hoc Networks 20 program proved to be beneficial in improving the radios' performance by making them smaller, cheaper, and resilient to electronic attacks. In the 1990s (Third generation), the concept of commercial Ad hoc networks arrived with notebook computers and other viable communication equipments. At the same time, the idea of a collection of mobile nodes was proposed at several researchers gatherings. The IEEE 802.11 subcommittee had adopted the term "Ad hoc networks" and the research community had started to look into the possibility of deploying Ad hoc networks in other areas of application

BASIC CONCEPTS OF MOBILE AD HOC NETWORKS

An Ad hoc network is a collection of mobile nodes, which forms a temporary network without the aid of centralized administration or standard support devices regularly available as conventional networks. These nodes generally have a limited transmission range and, so, each node seeks the assistance of its neighboring nodes in forwarding packets and hence the nodes in an Ad hoc network can act as both routers and hosts. Thus a node may forward packets between other nodes as well as run user applications. By nature these types of networks are suitable for situations where either no fixed infrastructure exists or deploying network is not possible. Ad hoc mobile networks have found many applications in various fields like military, emergency, conferencing and sensor networks. Each of these application areas has their specific requirements for routing protocols. Since the network nodes are mobile, an Ad hoc network will typically have a dynamic topology, which will have profound effects on network characteristics. Network nodes will often be battery powered, which limits the capacity of CPU, memory, and bandwidth. This will require network functions that are resource effective. Furthermore, the wireless (radio) media will also affect the behavior of the network due to fluctuating link bandwidths resulting from relatively high error rates. These unique desirable features pose several new challenges in the design of wireless Ad hoc networking protocols.



Network functions such as routing, address allocation, authentication and authorization must be designed to cope with a dynamic and volatile On-Demand Routing In Multi-Hop Wireless Mobile Ad hoc Networks Overview of Mobile Ad hoc Networks 21 network topology. In order to establish routes between nodes, which are farther than a single hop, specially configured routing protocols are engaged. The unique feature of these protocols is their ability to trace routes in spite of a dynamic topology. In the simplest scenarios, nodes may be able to communicate directly with each other, for example, when they are within wireless transmission range of each other. However, Ad hoc networks must also support communication between nodes that are only indirectly connected by a series of wireless hops through other nodes. For example, to establish communication between nodes A and C the network must enlist the aid of node B to relay packets between them. The circles indicate the nominal range of each node's radio transceiver. Nodes A and C are not in direct transmission range of each other, since A's circle does not cover C. A Mobil Ad hoc network of three nodes, where nodes A and C must discover the route through B in order to communicate. In general, an Ad hoc network is a network in which every node is potentially a router and every node is potentially mobile. The presence of wireless communication and mobility make an Ad hoc network unlike a traditional wired network and requires that the routing protocols used in an Ad hoc network be based on new and different principles. Routing protocols for traditional wired networks are designed to support tremendous numbers of nodes, but they assume that the relative position of the nodes will generally remain unchanged.

Characteristics of MANET

In MANET, each node act as both host and router. That is it is autonomous in behavior.

Multi-hop radio relaying- When a source node and destination node for a message is out of the radio range, the MANETs are capable of multi-hop routing.

Distributed nature of operation for security, routing and host configuration. A centralized firewall is absent here.

The nodes can join or leave the network anytime, making the network topology dynamic in nature.

Mobile nodes are characterized with less memory, power and light weight features.

The reliability, efficiency, stability and capacity of wireless links are often inferior when compared with wired links. This shows the fluctuating link bandwidth of wireless links.

Mobile and spontaneous behavior which demands minimum human intervention to configure the network.

All nodes have identical features with similar responsibilities and capabilities and hence it forms a completely symmetric environment.

High user density and large level of user
mobility. Nodal connectivity is intermittent.

MANETs Applications

1)Military battlefield: Ad-Hoc networking would allow the military to take advantage of commonplace network technology to maintain an information network between the soldiers, vehicles, and military information head quarter.

Collaborative work: For some business environments, the need for collaborative computing might be more important outside office environments than inside and where people do need to have outside meetings to cooperate and exchange information on a given project.

Local level: Ad-Hoc networks can autonomously link an instant and temporary multimedia network using notebook computers to spread and share information among participants at a e.g. conference or classroom. Another appropriate local level application might be in home networks where devices can communicate directly to exchange information.

Personal area network and bluetooth : A personal area network is a short range, localized network where nodes are usually associated with a given person. Short-range MANET such as Bluetooth can simplify the inter communication between various mobile devices such as a laptop, and a mobile phone.

***Commercial Sector*:** Ad hoc can be used in emergency/rescue operations for disaster relief efforts, e.g. in fire, flood, or earthquake. Emergency rescue operations must take place where non-existing or damaged communications infrastructure and rapid deployment of a communication network is needed.

Tactical Networks

- Military communication and operations
- Automated battlefields

Emergency services

- Search and rescue operations
- Disaster recovery
- Replacement of fixed infrastructure in case of environmental disasters
- Policing and fire fighting
- Supporting doctors and nurses in hospitals

Commercial and civilian

- E-commerce: electronic payments anytime and anywhere environments
- Business: dynamic database access, mobile offices
- Vehicular services: road or accident guidance, transmission of road and weather conditions, taxi cab network, inter-vehicle networks
- Sports stadiums, trade fairs, shopping malls
- Networks of visitors at airports

Home and enterprise

- Home/office wireless networking networking
- Conferences, meeting rooms
- Personal area networks (PAN), Personal networks (PN)
- Networks at construction sites Education
- Universities and campus settings
- Virtual classrooms
- Ad hoc communications during meetings or lectures

Entertainment

- Multi-user games
- Wireless P2P networking
- Outdoor Internet access
- Robotic pets
- Theme parks

Sensor networks

- Home applications: smart sensors and actuators embedded in consumer electronics
- Body area networks (BAN)
- Data tracking of environmental conditions, animal movements, chemical/biological detection

Context aware services

- Follow-on services: call-forwarding, mobile workspace
- Information services: location specific services, time dependent services
- Infotainment: touristic information

Coverage extension

- Extending cellular network access
- Linking up with the Internet, intranets, etc.

Design Issues in MANET

1) Limited bandwidth: Wireless link continue to have significantly lower capacity than infrastructure networks. In addition, the realized throughput of wireless communication after accounting for the effect of multiple access, fading, noise, and interference conditions, etc., is often much less than a radio's maximum transmission rate.

2) Dynamic topology: Dynamic topology membership may disturb the trust relationship among nodes. The trust may also be disturbed if some nodes are detected as compromised.

3) Routing Overhead: In wireless adhoc networks, nodes often change their location within network. So, some stale routes are generated in the routing table which leads to unnecessary routing overhead.
*4) Hidden terminal problem***:** The hidden terminal problem refers to the collision of packets at a receiving node due to the simultaneous transmission of those nodes that are not within the direct transmission range of the sender, but are within the transmission range of the receiver.

5) Packet losses due to transmission errors: Ad hoc wireless networks experiences a much higher packet loss due to factors such as increased collisions due to the presence of hidden terminals, presence of interference, uni-directional links, frequent path breaks due to mobility of nodes.

6) Mobility-induced route changes: The network topology in an ad hoc wireless network is highly dynamic due to the movement of nodes; hence an on-going session suffers frequent path breaks. This situation often leads to frequent route changes.

7) Battery constraints: Devices used in these networks have restrictions on the power source in order to maintain portability, size and weight of the device. *8) Security threats*: The wireless mobile ad hoc nature of MANETs brings new security challenges to the network design. As the wireless medium is vulnerable to eavesdropping and ad hoc network functionality is established through node cooperation, mobile ad hoc networks are intrinsically exposed to numerous security attacks.

Routing in MANET

Routing is the process of information exchange from one host to the other host in a network."[4].
Routing is the mechanism of forwarding packet towards its destination using most efficient path.
Efficiency of the path is measured in various metrics like, Number of hops, traffic, security, etc.
In Ad-hoc network each host node acts as specialized router itself

Different Strategies

Routing protocol for ad-hoc network can be categorized in three strategies.

Flat Vs Hierarchical architecture.

Pro- active Vs Re- active routing protocol.

Hybrid protocols.

Flat Vs. Hierarchical architecture

Hierarchical network architecture topology consists of multiple layers where top layers are more seen as master of their lower layer nodes. There are cluster of nodes and one gateway node among all clusters has a duty to communicate with the gateway node in other cluster. In this schema there is a clear distribution of task. Burden of storage of network topology is on gateway nodes, where communicating different control message is dependent on cluster nodes.

But this architecture breaks down when there is single node failure (Gateway node). Gateway nodes become very critical for successful operation of network. Examples include Zone-based Hierarchical Link State (ZHLS) routing protocol [6]. Where in flat architecture there is no layering of responsibility. Each and every node does follow the same routing algorithm as any other node in the network.

Proactive Vs Reactive routing protocol in MANET

Proactive routing protocol

In proactive routing scheme every node continuously maintains complete routing information of the network. This is achieved by flooding network periodically with network status information to find out any possible change in network topology.

Current routing protocol like Link State Routing (LSR) protocol (open shortest path first) and the Distance Vector Routing Protocol (Bellman-Ford algorithm) are not suitable to be used in mobile environment.
Destination Sequenced Distance Vector Routing protocol (DSDV) and Wireless routing protocols were proposed to eliminate counting to infinity and looping problems of the distributed Bellman-Ford Algorithm.

Examples of Proactive Routing Protocols are

- Global State Routing (GSR).
- Hierarchical State Routing (HSR).

- Destination Sequenced Distance Vector Routing (DSDV).

Reactive routing protocol

Every node in this routing protocol maintains information of only active paths to the destination nodes. A route search is needed for every new destination therefore the communication overhead is reduced at the expense of delay to search the route. Rapidly changing wireless network topology may break active route and cause subsequent route search.

Examples of reactive protocols are:

- Ad hoc On-demand Distance Vector Routing (AODV).
- Dynamic Source Routing (DSR).
- Location Aided Routing (LAR).
- Temporally Ordered Routing Algorithm (TORA).

Hybrid routing protocols in MANET

There exist a number of routing protocols of globally reactive and locally proactive states.
Hybrid routing algorithm is ideal for Zone Based Routing Protocol (ZRP)

Essential of Routing Protocols

The following is a list of quantitative metrics that can be used to assess the performance of any routing protocol.

End-to-end data throughput and delay: Statistical measures of data routing performance (e.g., means, variances, distributions) are important. These are the measures of a routing policy'seffectiveness--how well it does its job--as measured from the external perspective of other policies that make use of routing.

Route Acquisition Time: A particular form of *external* end- to-end delay measurement--of particular concern with "on demand" routing algorithms--is the time required to establish route(s) when requested.

Percentage Out-of-Order Delivery: An external measure of connectionless routing performance of particular interest to transport layer protocols such as TCP which prefer in-order delivery.

Efficiency: If data routing effectiveness is the external measure of a policy's performance, efficiency is the *internal* measure of its effectiveness. To achieve a given level of data routing performance, two different policies can expend differing amounts of overhead, depending on their internal efficiency.

Protocol efficiency may or may not directly affect data routing performance. If control and data traffic must share the same channel, and the channel's capacity is limited, then excessivecontrol traffic often impacts data routing performance.

It is useful to track several ratios that illuminate the *internal* efficiency of a protocol in doing its job (there may be others that the authors have not considered):

- Average number of data bits transmitted/data bit delivered--this can be thought of as a measure of the bit efficiency of delivering data within the network. Indirectly, it also give the average hop count taken by data packets.
- Average number of control bits transmitted/data bit delivered--this measures the bit efficiency of the protocol in expending control overhead to delivery data. Note that this should include not only the bits in the routing control packets, but also the bits in the header of the data packets. In other words, anything that is not data is control overhead, and should be counted in the control portion of the algorithm.

An improved mobile routing capability at the IP layer can provide a benefit similar to the intention of the original Internet, viz. "an interoperable internetworking capability over a heterogeneous networking infrastructure". In this case, the infrastructure is wireless, rather than hardwired, consisting of multiple wireless technologies, channel access protocols, etc. Improved IP routing and related networking services provide the glue to preserve the integrity of the mobile internetwork segment in this more dynamic environment.

In other words, a real benefit to using IP-level routing in a MANET is to provide network-level consistency for multihop networks composed of nodes using a *mixture* of physical-layer media; i.e. a mixture of what are commonly thought of as subnet technologies. A MANET node principally consists of a router, which may be physically attached to multiple IP hosts (or IP-addressable devices), which has potentially *multiple* wireless interfaces--each interface using a *different* wireless technology. Thus, a MANET node with interfaces using technologies A and B can communicate with any other MANET node possessing an interface with technology A or B. The multihop connectivity of technology A forms a physical-layer multihop topology, the multihop connectivity of technology B forms *another* physical-layer topology (which may differ from that of A's topology), and the *union* of these topologies forms another topology (in graph theoretic terms--a multigraph), termed the "IP routing fabric", of the MANET. MANET nodes making routing decisions using the IP fabric can intercommunicate using either or both physical-layer topologies simultaneously. As new physical-layer technologies are developed, new device drivers can be written and another physical-layer multihop topology can be seamlessly added to the IP fabric. Likewise, older technologies can easily be dropped. Such is the functionality and architectural flexibility that IP-layer routing can support, which brings with it hardware economies of scale.

Distance Vector Protocols

By using the distance vector protocols, each router over the internetwork send the neighboring routers, the information about destination that it knows how to reach. Moreover to say the routers sends two pieces of information first, the router tells, how far it thinks the destination is and secondly, it tells in what direction (vector) to use to get to the destination. When the router receives the information from the others, it could then develop a table of destination addresses, distances and associated neighboring routers, and from this table then select the shortest route to the destination. Using a distance vector protocol, the router simply forwards the packet to the neighboring host (or destination) with the available shortest path in the routing table and assumesthat the receiving router will know how to forward the packet beyond that point [9]. The best example for this is the routing information protocol (RIP).

Link-State Protocols

In link state protocols, a router doesn’t provide the information about the destination instead it provides the information about the topology of the network. This usually consist of the network segments and links that are attached to that particular router along with the state of the link i.e., whether the link is in active state or the inactive state. This information is flooded throughout the network and then every router in the network then builds its own picture of the current state of all the links in the network.

Dynamic Source Routing

Dynamic Source Routing (DSR) is a reactive protocol based on the source route approach [9]. In *Dynamic Source Routing (DSR)*, shown in Figure.2, the protocol is based on the link state algorithm in which source initiates route discovery on demand basis. The sender determines the route from source to destination and it includes the address of intermediate nodes to the route record in the packet. DSR was designed for multi hop networks for small Diameters. It is a beaconless protocol in which no HELLO messages are exchanged between nodes to notify them of their neighbours in the network.

In DSR the route paths are discovered after source sends a packet to a destination node in the ad-hoc network. The source node initially does not have a path to the destination when the first packet is sent. The DSR has two functions first is route discovery and the second is route maintenance

Different DSR Algorithms

Route discovery.

Route maintenance.

Assumptions:

- X , Y, Z , V and W form ad-hoc network.
- X is the source node.
- Z is the destination node.

Route Discovery is the mechanism by which a node **S** wishing to send a packet to a destination node **D**obtains a source route to **D**. Route Discovery is used only when **S** attempts to send a packet to **D** and does not already know a route to **D**.

Route Maintenance is the mechanism by which node **S** is able to detect, while using a source routeto **D**, if the network topology has changed such that it can no longer use its route to **D** because a link along the route no longer works. When Route Maintenance indicates a source route is broken, **S** can attempt to use any other route it happens to know to **D**, or can invoke Route Discovery again to find a new route. Route Maintenance is used only when **S** is actually sending packets to **D**.

Destination Sequenced Distance Vector (DSDV) Protocol

The destination sequenced distance vector routing protocol is a proactive routing protocol which is a modification of conventional Bellman-Ford routing algorithm. This protocol adds a new attribute, sequence number, to each route table entry at each node. Routing table is maintained at each node and with this table, node transmits the packets to other nodes in the network. This protocol was motivated for the use of data exchange along changing and arbitrary paths of interconnection which may not be close to any base station. 6.1 Protocol Overview and activities Each node in the network maintains routing table for the transmission of the packets and also for the connectivity to different stations in the network. These stations list for all the available destinations, and the number of hops required to reach each destination in the routing table.

The routing entry is tagged with a sequence number which is originated by the destination station. In order to maintain the consistency, each station transmits and updates its routing table periodically. The packets being broadcasted between stations indicate which stations are accessible and how many hops are required to reach that particular station. The packets may be transmitted containing the layer 2 or layer 3 address. Routing information is advertised by broadcasting or multicasting the packets which are transmitted periodically as when the nodes move within the network.

The DSDV protocol requires that each mobile station in the network must constantly, advertise to each of its neighbors, its own routing table. Since, the entries in the table my change very quickly, the advertisement should be made frequently to ensure that every node can locate its neighbors in the network. This agreement is placed, to ensure the shortest number of hops for a route to a destination; in this way the node can exchange its data even if there is no direct communication link. The data broadcast by each node will contain its new sequence number and the following information for each new route:

Destination Sequenced Distance Vector (DSDV) Protocol – The destination address – The number of hops required to reach the destination and – The new sequence number, originally stamped by the destination The transmitted routing tables will also contain the hardware address, network address of the mobile host transmitting them. The routing tables will contain the sequence number created by the transmitter and hence the most new destination sequence number is preferred as the basis for making forwarding decisions.

This new sequence number is also updated to all the hosts in the network which may decide on how to maintain the routing entry for that originating mobile host. After receiving the route information, receiving node increments the metric and transmits information by broadcasting. Incrementing metric is done before transmission because, incoming packet will have to travel one more hop to reach its destination. Time between broadcasting the routing information packets is the other important factor to be considered. When the new information is received by the mobile host it will be retransmitted soon effecting the most rapid possible dissemination ofrouting information among all the cooperating mobile hosts. The mobile host cause broken links as they move form place to place within the network. The broken link may be detected by the layer2 protocol, which may be described as infinity. When the route is broken in a network, then immediately that metric is assigned an infinity metric there by determining that there is no hop and the sequence number is updated.

Sequence numbers originating from the mobile hosts are defined to be even number and the sequence numbers generated to indicate infinity metrics are odd numbers. The broadcasting of the information in the DSDV protocol is of two types namely: full dump and incremental dump. Full dump broadcasting will carry all the routing information while the incremental dump will carry only information that has changed since last full dump. Irrespective of the two types, broadcasting is done in network protocol data units (NPDU). Full dump requires multiple

NPDU's while incremental requires only one NPDU to fit in all the information. When an information packet is received from another node, it compares the sequence number with the available sequence number for that entry.

If the sequence number is larger, then it will update the routing information with the new sequence number else if the information arrives with the same sequence number it looks for the metric entry and if the number of hops is less than the previous entry the new information is updated (if information is same or metric is more then it will discard the information). While the nodes information is being updated the metric is increased by 1 and the sequence number is also increased by 2. Similarly, if a new node enters the network, it will announce itself in the network and the nodes in the network update their routing information with a new entry for the new node. During broadcasting, the mobile hosts will transmit their routing tables periodically but due to the frequent movements by the hosts in the networks, this will lead to continuous burst of new routes transmessions upon every new sequence number from that destination. The solution for this is to delay the advertisement of such routes until it shows up a better metric.

Advantages of DSDV

– DSDV protocol guarantees loop free paths

– Count to infinity problem is reduced in DSDV

– We can avoid extra traffic with incremental updates instead of full dump updates.

– Path Selection: DSDV maintains only the best path instead of maintaining multiple paths to every destination. With this, the amount of space in routing table is reduced.

Limitations of DSDV

– Wastage of bandwidth due to unnecessary advertising of routing information even if there is no change in the network topology
– DSDV doesn't support Multi path Routing.

– It is difficult to determine a time delay for the advertisement of routes

– It is difficult to maintain the routing table's advertisement for larger network. Each and every host in the network should maintain a routing table for advertising. But for larger network this would lead to overhead, which consumes more bandwidth.

Ad Hoc On-Demand Distance Vector Routing (AODV)

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

The concepts of AODV that make it desirable for MANETs with limited bandwidth include the following: – Minimal space complexity: The algorithm makes sure that the nodes that are not in the active path do not maintain information about this route. After a node receives the RREQ and sets a reverse path in its routing table and propagates the RREQ to its neighbors, if it does not receive any RREP from its neighbors for this request, it deletes the routing info that it has recorded.

– Maximum utilization of the bandwidth: This can be considered the major achievement of the algorithm. As the protocol does not require periodic global advertisements, the demand on the available bandwidth is less. And a monotonically increased sequence number counter is maintained by each node in order to supersede any stale cached routes. All the intermediate nodes in an active path updating their routing tables also make sure of maximum utilization of the bandwidth. Since, these routing tables will be used repeatedly if that intermediate node receives any RREQ from another source for same destination. Also, any RREPs that are received by the nodes are compared with the RREP that was propagated last using the destination sequence numbers and are discarded if they are not better than the already propagated RREPs.
– Simple:

It is simple with each node behaving as a router, maintaining a simple routing table, and the source node initiating path discovery request, making the network self-starting.

– Most effective routing info: After propagating an RREP, if a node finds receives an RREP with smaller hop-count, it updates its routing info with this better path and propagates it.

– Most current routing info: The route info is obtained on demand. Also, after propagating an RREP, if a node finds receives an RREP with greater destination sequence number, it updates its routing info with this latest path and propagates it.

– Loop-free routes: The algorithm maintains loop free routes by using the simple logic of nodes discarding non better packets for same broadcast-id.

– Coping up with dynamic topology and broken links: When the nodes in the network move from their places and the topology is changed or the links in the active path are broken, the intermediate node that discovers this link breakage propagates an RERR packet. And the source node re-initializes the path discovery if it still desires the route. This ensures quick response to broken links.

– Highly Scalable: The algorithm is highly scalable because of the minimum space complexity and broadcasts avoided when it compared with DSDV.

The concepts of AODV that make it desirable for MANETs with limited bandwidth include the following: – Minimal space complexity: The algorithm makes sure that the nodes that are not in the active path do not maintain information about this route. After a node receives the RREQ and sets a reverse path in its routing table and propagates the RREQ to its neighbors, if it does not receive any RREP from its neighbors for this request, it deletes the routing info that it has recorded.

Advanced uses of AODV

– Because of its reactive nature, AODV can handle highly dynamic behavior of Vehicle Ad-hoc networks
– Used for both unicasts and multicasts using the ’J’ (Join multicast group) flag in the packets

Limitations/Disadvantages of AODV

– Requirement on broadcast medium: The algorithm expects/requires that the nodes in the broadcast medium can detect each others’ broadcasts.

– Because of its reactive nature, AODV can handle highly dynamic behavior of Vehicle Ad-hoc networks
– Used for both unicasts and multicasts using the ’J’ (Join multicast group) flag in the packe

The algorithm expects/requires that the nodes in the broadcast medium can detect each others’ broadcasts.

Multicast Routing Protocol

TREE-BASED MULTICASTING

A tree-based multicast routing protocol establishes and maintains a shared multicast routing tree to deliver data from a source to receivers of a multicast group. A well-known example of treebased multicast routing protocols are the Multicast Ad hoc Ondemand Distance Vector routing protocol (MAODV).

Multicast Ad-hoc On-Demand Distance Vector Routing Protocol (MAODV).

MAODV is a multicast extension for AODV protocol. MAODV based on shared trees on-demand to connect multicast group members. MAODV has capability of unicast, broadcast, and multicast. MAODV protocol can be route information obtained when searching for multicast; it can also increase unicast routing knowledge and vice-versa. When a node wishes to join a multicast group or it has data to send to the group but does not has a route to that group, it originates a route request (RREQ) message. Only the members of the multicast group respond to the join RREQ. If an intermediate node receives a join RREQ for a multicast group of which it is not a member or it receives a route RREQ and it does not have a route to that group, it rebroadcast the RREQ to its neighbors. But if the RREQ is not a join request any node of the multicast group may respond.

MESH-BASED MUTICASTING

A mesh-based multicast routing protocol sustains a mesh consisting of a connected component of the network containing all the receivers of a group. Example of mesh-based multicast routing approaches is On-Demand Multicast Routing Protocol (ODMRP).

On-Demand Multicast Routing Protocol (ODMRP)

ODMRP is an on-demand mesh based, besides it is a multicast routing protocol, ODMRP protocol can make use of unicast technique to send multicast data packet form the sender nodes toward the receivers in the multicasting group. To carry multicast data via scoped flooding it uses forwarding group concept. The source, in ODMRP, establishes and maintains group membership. If source wishes to send packet to a multicast group but has no route to that group, it simply broadcasts JOIN_DATA control packet to the entire network. When an intermediate node receives the JOIN_DATA packet it stores source address and sequence number in its cache to detect duplicate. It performs necessary routing table updates for reverse path back to the source.

Zone Routing Protocol (ZRP)

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

Vehicular Ad Hoc Networks

Vehicular Ad Hoc Networks (VANETs) are created by applying the principles of mobile ad hoc networks (MANETs) - the spontaneous creation of a wireless network for data exchange - to the domain of vehicles. They are a key component of intelligent transportation systems (ITS).

While, in the early 2000s, VANETs were seen as a mere one-to-one application of MANET principles, they have since then developed into a field of research in their own right. By 2015, the term VANET became mostly synonymous with the more generic term inter-vehicle communication (IVC), although the focus remains on the aspect of spontaneous networking, much less on the use of infrastructure like Road Side Units (RSUs) or cellular networks.

Vehicular Ad Hoc Networks (VANETs) have grown out of the need to support the growing number of wireless products that can now be used in vehicles. These products include remote keyless entry devices, personal digital assistants (PDAs), laptops and mobile telephones. As mobile wireless devices and networks become increasingly important, the demand for Vehicle-to-Vehicle (V2V) and Vehicle- S. Zeadally et al. to-Roadside (VRC) or Vehicle-to-Infrastructure (V2I) Communication will continue to grow. VANETs can be utilized for a broad range of safety and non-safety applications, allow for value added services such as vehicle safety, automated toll payment, traffic management, enhanced navigation, location-based services such as finding the closest fuel station, restaurant or travel lodge and infotainment applications such as providing access to the Internet.

Intelligent transportation systems (ITSs) In intelligent transportation systems, each vehicle takes on the role of sender, receiver, and router to broadcast information to the vehicular network or transportation agency, which then uses the information to ensure safe, free-flow of traffic. For communication to occur between vehicles and RoadSide Units (RSUs), vehicles must be equipped with some sort of radio interface or OnBoard Unit (OBU) that enables short-range wireless ad hoc networks to be formed. Vehicles must also be fitted with hardware that permits detailed position information such as Global Positioning System (GPS) or a Differential Global Positioning System (DGPS) receiver. Fixed RSUs, which are connected to the backbone network, must be in place to facilitate communication.

Inter-vehicle communication The inter-vehicle communication configuration uses multi-hop multicast/broadcast to transmit traffic related information over multiple hops to a group of receivers. In intelligent transportation systems, vehicles need only be concerned with activity on the road ahead and not behind (an example of this would be for emergency message dissemination about an imminent collision or dynamic route scheduling). There are two types of message forwarding in inter-vehicle communications: naïve broadcasting and intelligent broadcasting. In naïve broadcasting, vehicles send broadcast messages periodically and at regular intervals. Upon receipt of the message, the vehicle ignores the message if it has come from a vehicle behind it. If the message comes from a vehicle in front, the receiving vehicle sends its own broadcast message to vehicles behind it.

Vehicle-to-roadside communication The vehicle-to-roadside communication configuration (Fig. 2) represents a single hop broadcast where the roadside unit sends a broadcast message to all equipped vehicles in the vicinity. Vehicle-to-roadside communication configuration provides a high bandwidth link between vehicles and roadside units. The roadside units may be placed every kilometer or less, enabling high data rates to be maintained in heavy traf- fic. For instance, when broadcasting dynamic speed limits, the roadside unit will determine the appropriate speed limit according to its internal timetable and traffic conditions. The roadside unit will periodically broadcast a message containing the speed limit and will compare any geographic or directional limits with vehicle data to determine if a speed limit warning applies to any of the vehicles in the vicinity.

Routing-based communication The routing-based communication configuration is a multi-hop unicast where a message is propagated in a multi Routing-based communication hop fashion until the vehicle carrying the desired data is reached. When the query is received by a vehicle owning the desired piece of information, the application at that vehicle immediately sends a unicast message containing the information to the vehicle it received the request from, which is then charged with the task of forwarding it towards the query source.

MANET vs VANET

Mobile Ad Hoc Network and Vehicular AdHoc Networks are emerging area for research and evelopment. VANETs are subclass of MANETs. But unlike MANETs it does not have battery constraints and have high mobility. Unicast and Multicast protocols in MANETs and VANETs use broadcasting to provide important control and route establishment. Possible applications of VANETs relying on broadcast for sharing safety, weather, and road information among vehicles.

Mobile nodes temporary form a network for information sharing and require no need of routers and base stations is called Mobile Ad Hoc Network (MANET). They communicate with each other over multihop wireless links. MANETs have different features like dynamic topologies, limited security, bandwidth and energy constrains.

Vehicular Ad-Hoc Networks (VANETs) are special case of MANETs. Self Organized and distributed network, where fast moving vehicles have fixed movement along some path. VANETs have salient features (high speed, no battery constraints, limited movement, reliability and security problems) that discriminate it from other adhoc network. In wireless network, broadcasting is frequently used operation as compared to wired network. They are a lot issues and problems in wireless adhoc network because of node mobility and scattered resources. VANETs are promising network for intelligent systems having short communication range between the vehicles. Mostly in Vehicular adhoc network (VANETs), vehicles are interested in the same kind of information for example information about any accident, road block, parking, and fuel station or weather situation of particular route. So the broadcast is frequently used in vehicular adhoc network for information sharing.

Uses of MANET	USES of VANET
• For military and rescue use. • Information distribution (meetings, seminars etc.) • Internet / intranet hot spots (public transportation) • Localized advertising and shopping • New mobile devices are invented constantly and used various ways.	Comfort Applications: It improves the traffic efficiency and passenger comfort. Traffic information system, gas station and weather information are example of comfort application. Safety Applications: Sharing emergency and safety data among vehicles improves the safety of passengers. Safety application are Emergency warning system, road condition and traffic sign violation waning.

BROADCAST APPROACHES IN MANETS

Different unicast and multicast protocols like Dynamic Source Routing [9], Zone Routing Protocol , Ad Hoc On Demand Distance Vector , and Location Aided Routing use broadcasting to establish and maintain the route in MANETs. Brad Williams in presents the comparative study of broadcast approaches in MANETS using NS2 simulator. Impact of simple flooding, probability method, Area method and Neighbor knowledge method are analyzed with different network parameter like increasing network load, node mobility and traffic rate.

BROADCAST APPROACHES IN VANETS

In VANETs, broadcast is a most commonly used technique. Sharing safety, weather, and road information among vehicles depends on broadcast. Broadcast plays important role in VANETS, as it is used to establish and maintain the route for unicast and multicast protocols. Different broadcast schemes are designed for VANETs scenarios. Urban MultiHop Broadcast is based upon IEEE 802.11 protocol. It solves the Broadcast storm, Hidden node problem without sharing information among the neighbor nodes. Mobility Centric Data Dissemination Algorithm for Vehicular Networks is a mobility centric scheme that merge three techniques, Geographical, Opportunistic and Trajectory based forwarding. Multi-Hop Vehicular Broadcast disseminates the information to other vehicles & store it in the local database for safety use. It has two main features. i:e Congestion Detection and Backfire algorithm.

Security in MANETS

Mobile ad hoc networks have far more vulnerabilities than the traditional wired networks, security is much more difficult to maintain in the mobile ad hoc network than in the wired network. In this section, we discuss the various vulnerabilities that exist in the mobile ad hoc networks.

Lack of Secure Boundaries:

The meaning of this vulnerability is self-evident: there is not such a clear secure boundary in the mobile ad hoc network, which can be compared with the clear line of defense in the traditional wired network. This vulnerability originates from the nature of the mobile ad hoc network: freedom to join, leave and move inside the network.

Lack of Centralized Management Facility

Ad hoc networks do not have a centralized piece of management machinery such as a name server, which lead to some vulnerable problems. Now let us discuss this problem in a more detailed manner.

Restricted Power Supply

As we all know, due to the mobility of nodes in the ad hoc network, it is common that the 5 nodes in the ad hoc network will reply on battery as their power supply method. While nodes in the wired network do not need to consider the power supply problem because they can get electric power supply from the outlets, which generally mean that their power supply should be approximately infinite; the nodes in the mobile ad hoc network need to consider the restricted battery power, which will cause several problems.

Scalability

Finally, we need to address the scalability problem when we discuss the vulnerabilities in the mobile ad hoc network. Unlike the traditional wired network in that its scale is generally predefined when it is designed and will not change much during the use, the scale of the ad hoc network keeps changing all the time: because of the mobility of the nodes in the mobile ad hoc network, you can hardly predict how many nodes there will be in the network in the future.

Security Criteria

Availability The term Availability means that a node should maintain its ability to provide all the designed services regardless of the security state of it. This security criterion is challenged mainly during the denial-of-service attacks, in which all the nodes in the network can be the attack target and thus some selfish nodes make some of the network services unavailable, such as the routing protocol or the key management service.

Integrity: Integrity guarantees the identity of the messages when they are transmitted. Integrity can be compromised mainly in two ways

Malicious altering

Accidental altering

A message can be removed, replayed or revised by an adversary with malicious goal, which is regarded as malicious altering; on the contrary, if the message is lost or its content is changed due to some benign failures, which may be transmission errors in communication or hardware errors such as hard disk failure, then it is categorized as accidental altering.

Confidentiality: Confidentiality means that certain information is only accessible to those who have been authorized to access it. In other words, in order to maintain the confidentiality of some confidential information, we need to keep them secret from all entities that do not have the privilege to access them.

Authenticity: Authenticity is essentially assurance that participants in communication are genuine and not impersonators. It is necessary for the communication participants to prove their identities as what they have claimed using some techniques so as to ensure the authenticity. If there is not such an authentication mechanism, the adversary could impersonate a benign node and thus get access to confidential resources, or even propagate some fake messages to disturb the normal network operations.

Nonrepudiation: Nonrepudiation ensures that the sender and the receiver of a message cannot disavow that they have ever sent or received such a message. This is useful especially when we need to discriminate if a node with some abnormal behavior is compromised or not: if a node recognizes that the message it has received is erroneous, it can then use the incorrect message as an evidence to notify other nodes that the node sending out the improper message should have been compromised.

Authorization: Authorization is a process in which an entity is issued a credential, which specifies the privileges and permissions it has and cannot be falsified, by the certificate authority. Authorization is generally used to assign different access rights to different level of users. For instance, we need to ensure that network management function is only accessible by the network administrator. Therefore there should be an authorization process before the network administrator accesses the network management functions.

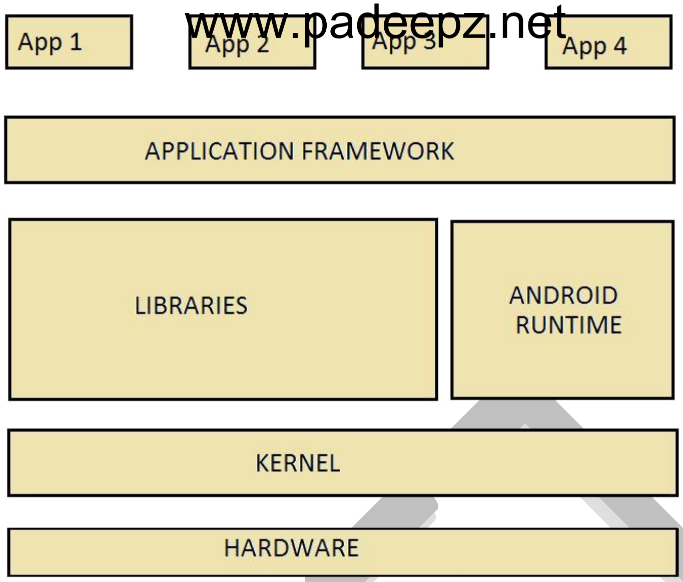
Mobile OS

A mobile operating system (or mobile OS) is an operating system for smartphones, tablets, PDAs, or other mobile devices. While computers such as the typical laptop are mobile, the operating systems usually used on them are not considered mobile ones as they were originally designed for bigger stationary desktop computers that historically did not have or need specific "mobile" features. This distinction is getting blurred in some newer operating systems that are hybrids made for both uses.

Mobile operating systems combine features of a personal computer operating system with other features useful for mobile or handheld use; usually including, and most of the following considered essential in modern mobile systems; a touchscreen, cellular, Bluetooth, Wi-Fi, GPS mobile navigation, camera, video camera, speech recognition, voice recorder, music player, near field communication and infrared blaster.

Mobile devices with mobile communications capabilities (e.g. smartphones) contain two mobile operating systems – the main user-facing software platform is supplemented by a second low-level proprietary real-time operating system which operates the radio and other hardware. Research has shown that these low-level systems may contain a range of security vulnerabilities permitting malicious base stations to gain high levels of control over the mobile device.

A mobile OS is a software platform on top of which other programs called application programs, can run on mobile devices such as PDA, cellular phones, smartphone and etc. A Mobile operating system is a System Software that is specifically designed to run on handheld devices such as Mobile Phones, PDA’s. It is a Platform on top of which the application programs run on mobile devices. Each Operating System follows its own Architecture. Mobile devices evolved the way users across the globe leverage services on the go from voice calls to smart devices which enables users to access value added services anytime and anywhere. At present, the mobile devices are able to provide various services to users but still suffers from issues include Performance, security and Privacy, Reliability and Band width costs. In this paper, we pointed out the issues, challenges, Advantages and Disadvantages of various Mobile Operating systems in terms of their Architectures.



Applications

The diagram shows four basic apps (App 1, App 2, App 3 and App 4), just to give the idea that there can be multiple apps sitting on top of Android. These apps are like any user interface you use on Android; for example, when you use a music player, the GUI on which there are buttons to play, pause, seek, etc is an application. Similarly, is an app for making calls, a camera app, and so on. All these apps are not necessarily from Google. Anyone can develop an app and make it available to everyone through Google Play Store. These apps are developed in Java, and are installed directly, without the need to integrate with Android OS.

Application Framework

Scratching further below the applications, we reach the application framework, which application developers can leverage in developing Android applications. The framework offers a huge set of APIs used by developers for various standard purposes, so that they don't have to code every basic task.The framework consists of certain entities; major ones are:

Activity Manager

This manages the activities that govern the application life cycle and has several states. An application may have multiple activities, which have their own life cycles. However, there is one main activity that starts when the application is launched. Generally, each activity in an application is given a window that has its own layout and user interface. An activity is stopped when another starts, and gets back to the window that initiated it through an activity callback.

- Notification Manager

This manager enables the applications to create customized alerts

- Views

Views are used to create layouts, including components such as grids, lists, buttons, etc.

- Resource Managers

Applications do require external resources, such as graphics, external strings, etc. All these resources are managed by the resource manager, which makes them available in a standardized way.

- Content Provider

Applications also share data. From time to time, one application may need some data from another application. For example, an international calling application will need to access the user's address book. This access to another application's data is enabled by the content providers.

Libraries

This layer holds the Android native libraries. These libraries are written in C/C++ and offer capabilities similar to the above layer, while sitting on top of the kernel. A few of the major native libraries include

- Surface Manager: Manages the display and compositing window-ing manager. - Media framework: Supports various audio and video formats and codecs including their playback and recording.
- System C Libraries: Standard C library like libc targeted for ARM or embedded devices.
- OpenGL ES Libraries : These are the graphics libraries for rendering 2D and 3D graphics.
- SQLite : A database engine for Android.

Kernel

The Android OS is derived from Linux Kernel 2.6 and is actually created from Linux source, compiled for mobile devices. The memory management, process management etc. are mostly similar. The kernel acts as a Hardware Abstraction Layer between hardware and the Android software stack.

Mobile OS Special Constraints:

Smaller screen size	Stay focused on the user's immediate task. Display only the information that users need at any given moment. For example, a customer relationship management system can provide a massive amount of information, but users only require a small amount of that information at one time. Design the UI so that users can perform tasks easily and access information quickly.
One screen appears at a time	Use a single screen if possible. If your application requires multiple screens to be open at the same time, use a split screen or rethink the flow of your application.
Shorter battery life	Try to handle data transmission efficiently. The less often the device needs to transmit data, the longer the battery lasts.
Wireless network connections	Try to simplify how your application creates network connections. Compared with standard LANs, longer latency periods that are inherent in some wireless network connections can influence how quickly users receive information that is sent over the network.
Slower processor speeds	Avoid processor-intensive tasks where possible. Slower processor speeds can affect how users perceive the responsiveness of an application.
Less available memory	Free up as much memory as possible. For example, while an application is not being used, try to keep it from using memory.

Special Service Requirements

Support for specific communication protocol

Support for a variety of input mechanisms

Compliance with open standards

Extensive library support

Support for Integrated Development Environment

Commercial Mobile OS

Smartphones are now participating nearly in each and every sphere of life like business, education, workplace and healthcare. The Worldwide Mobile Communications Device Open Operating System Sales (WMCDOOS) provides total market of 104,898 to End Users by OS. There are over 1.3 million active applications in Google Play App Store. Android is the first open source, Linux-based and modern mobile handset platform. Google developed it for handset manufacturers like T-Mobile, Sprint Nextel, Google, Intel, Samsung, etc.. It offers to consumers a richer, less expensive, better mobile experience and various features like 3D, SQLite, Connectivity, WebKit, Dalvik and FreeType etc. Since android provides open source operating system; users by Microsoft for smartphones and Pocket PCs.

Its origins dated back to Windows CE in 1996, though Windows Mobile itself first appeared in 2000 as *PocketPC 2000*. It was renamed "Windows Mobile" in 2003, at which point it came in several versions (similar to the desktop versions of Windows) and was aimed at business and enterprise consumers. By 2007, it was the most popular smartphone software in the U.S., but this popularity faded in the following years. In February 2010, facing competition from rival OSs including iOSand Android, Microsoft announced Windows Phone to supersede Windows Mobile. As a result, Windows Mobile has been deprecated. Windows Phone is incompatible with Windows Mobile devices and software. The last version of Windows Mobile, released after the announcement of Windows Phone, was 6.5.5. After this, Microsoft ceased development on Windows Mobile, in order to concentrate on Windows Phone.

Most versions of Windows Mobile have a standard set of features, such as multitasking and the ability to navigate a file system similar to that of Windows 9x andWindows NT, including support for many of the same file types. Similarly to its desktop counterpart, it comes bundled with a set of applications that perform basic tasks. Internet Explorer Mobile is the default web browser, and Windows Media Player is the default media player used for playing digital media. The mobile version of Microsoft Office, is the default office suite.

Internet Connection Sharing, supported on compatible devices, allows the phone to share its Internet connection with computers via USB and Bluetooth. Windows Mobile supports virtual private networking over PPTP protocol. Most devices with mobile connectivity also have a Radio Interface Layer. The Radio Interface Layer provides the system interface between the Cell Core layer within the Windows Mobile OS and the radio protocol stack used by the wireless modem hardware. This allows OEMs to integrate a variety of modems into their equipment.

The user interface changed dramatically between versions, only retaining similar functionality. The *Today Screen*, later called the *Home Screen*, shows the current date, owner information, upcoming appointments, e-mails, and tasks. The taskbar display the current time as well as the volume level. Devices with a cellular radio also show the signal strength on said taskbar.

Palm OS (also known as **Garnet OS**) is a mobile operating system initially developed by Palm, Inc., for personal digital assistants (PDAs) in 1996. Palm OS was designed for ease of use with a touchscreen-based graphical user interface. It is provided with a suite of basic applications for personal information management. Later versions of the OS have been extended to support smartphones. Several other licensees have manufactured devices powered by Palm OS.

Following Palm's purchase of the Palm trademark, the currently licensed version from ACCESS was renamed *Garnet OS*. In 2007, ACCESS introduced the successor to Garnet OS, called Access Linux Platform and in 2009, the main licensee of Palm OS, Palm, Inc., switched from Palm OS to webOS for their forthcoming devices.

Palm OS was originally developed under the direction of Jeff Hawkins at Palm Computing, Inc. Palm was later acquired by U.S. Robotics Corp., which in turn was later bought by 3Com, which made the Palm subsidiary an independent publicly traded company on March 2, 2000.

In January 2002, Palm set up a wholly owned subsidiary to develop and license Palm OS, which was named PalmSource. PalmSource was then spun off from Palm as an independent company on October 28, 2003. Palm (then called palmOne) became a regular licensee of Palm OS, no longer in control of the operating system.

In September 2005, PalmSource announced that it was being acquired by ACCESS.

In December 2006, Palm gained perpetual rights to the Palm OS source code from ACCESS.[9] With this Palm can modify the licensed operating system as needed without paying further royalties to ACCESS. Together with the May 2005 acquisition of full rights to the *Palm* brand name, only Palm can publish releases of the operating system under the name 'Palm OS'.

As a consequence, on January 25, 2007, ACCESS announced a name change to their current Palm OS operating system, now titled *Garnet OS.*

Palm OS is a proprietary mobile operating system. Designed in 1996 for Palm Computing, Inc.'s new Pilot PDA, it has been implemented on a wide array of mobile devices, including smartphones, wrist watches,handheld gaming consoles, barcode readers and GPS devices.

Palm OS versions earlier than 5.0 run on Motorola/Freescale DragonBall processors. From version 5.0 onwards, Palm OS runs on ARM architecture-based processors.

The key features of the current Palm OS Garnet are:

Simple, single-tasking environment to allow launching of full screen applications with a basic, common GUI set

Monochrome or color screens with resolutions up to 480x320 pixel

Handwriting recognition input system called Graffiti 2

HotSync technology for data synchronization with desktop computers

Sound playback and record capabilities

Simple security model: Device can be locked by password, arbitrary application records can be made private

TCP/IP network access

Serial port/USB, infrared, Bluetooth and Wi-Fi connections

Expansion memory card support

Defined standard data format for personal information management applications to store calendar, address, task and note entries, accessible by third-party applications.

Included with the OS is also a set of standard applications, with the most relevant ones for the four mentioned PIM operations.

Symbian was a closed-source mobile operating system (OS) and computing platform designed for smartphones.[6] Symbian was originally developed by Symbian Ltd., as a descendant of Psion's EPOC and runs exclusively on ARM processors, although an unreleased x86 port existed.

Symbian was previously an open-source platform developed by the now defunct Symbian Foundation in 2009, as the successor of the original **Symbian OS** before being transitioned to a non-open license in 2011. Symbian was used by many major mobile phone brands, like Samsung, Motorola, Sony Ericsson, and above all by Nokia. It was briefly the most popular smartphone OS on a worldwide average until the end of 2010 – at a time when smartphones were
in limited use, when it was overtaken byAndroid, as Google and its partners achieved wide adoption.

Symbian rose to fame from its use with the S60 platform built by Nokia, first released in 2002 and powering most Nokia smartphones. UIQ, another Symbian platform, ran in parallel, but these two platforms were not compatible with each other. Symbian^3 was officially released in Q4 2010 as the successor of S60 and UIQ, first used in theNokia N8, to use a single platform for the OS. In May 2011 an update, Symbian Anna, was officially announced, followed by Nokia Belle (previously Symbian Belle) in August 2011.

iOS (originally **iPhone OS**) is a mobile operating system created and developed by Apple Inc. and distributed exclusively for Apple hardware. It is the operating system that presently powers many of the company's mobile devices, including the iPhone, iPad, and iPod touch. It is the second most popular mobile operating system in the world by sales, after Android. iPad tablets are also the second most popular, by sales, against Android since 2013, when Android tablet sales increased by 127%.[7]

Originally unveiled in 2007, for the iPhone, it has been extended to support other Apple devices such as the iPod Touch (September 2007), iPad (January 2010), iPad Mini (November 2012) and second-generation Apple TV onward (September 2010). As of January 2015, Apple's App Store contained more than 1.4 million iOS applications, 725,000 of which are native for iPads.[8] These mobile apps have collectively been downloaded more than 100 billion times.[9]

The iOS user interface is based on the concept of direct manipulation, using multi-touch gestures. Interface control elements consist of sliders, switches, and buttons. Interaction with the OS includes gestures such as *swipe*, *tap*, *pinch*, and *reverse pinch*, all of which have specific definitions within the context of the iOS operating system and its multi-touch interface. Internal accelerometers are used by some applications to respond to shaking the device (one common result is the undo command) or rotating it in three dimensions (one common result is switching from portrait to landscape mode).

iOS shares with OS X some frameworks such as Core Foundation and Foundation Kit; however, its UI toolkit is Cocoa Touch rather than OS X's Cocoa, so that it provides the UIKit framework rather than the AppKit framework. It is therefore not compatible with OS X for applications. Also while iOS also shares the Darwinfoundation with OS X, Unix-like shell access is not available for users and restricted for apps, making iOS not fully Unix-compatible either.

Major versions of iOS are released annually. The current release, iOS 9.3, was released on March 21, 2016. In iOS, there are four abstraction layers: the Core OS layer, the Core Services layer, the Media layer, and the Cocoa Touch layer. The current version of the operating system (iOS 9), dedicates around 1.3 GB of the device's flash memory for iOS itself.[10] It runs on the iPhone 4S and later, iPad 2 and later, iPad Pro, all models of the iPad Mini, and the 5th-generation iPod Touch and later.

Android is a mobile operating system (OS) currently developed by Google, based on the Linux kernel and designed primarily for touchscreen mobile devices such as smartphones and tablets. Android's user interface is mainly based on direct manipulation, using touch gestures that loosely correspond to real-world actions, such as swiping, tapping and pinching, to manipulate on-screen objects, along with a virtual keyboard for text input. In addition to touchscreen devices, Google has further developed Android TV for televisions, Android Auto for cars, and Android Wear for wrist watches, each with a specialized user interface. Variants of Android are also used on notebooks, game consoles, digital cameras, and other electronics.

Android has the largest installed base of all operating systems of any kind. Android has been the best selling OS on tablets since 2013, and on smartphones it is dominant by any metric.

Initially developed by Android, Inc., which Google bought in 2005, Android was unveiled in 2007, along with the founding of the Open Handset Alliance – a consortium of hardware, software, and telecommunication companies devoted to advancing open standards for mobile devices. As of July 2013, the Google Play store has had over one million Android applications ("apps") published, and over 50 billion applications downloaded.[18] An April–May 2013 survey of mobile application developers found that 71% of developers create applications for Android,[19] and a 2015 survey found that 40% of full-time professional developers see Android as their priority target platform, which is comparable to Apple's iOS on 37% with both platforms far above others.[20] At Google I/O 2014, the company revealed that there were over one billion active monthly Android users, up from 538 million in June 2013.

Android's source code is released by Google under open source licenses, although most Android devices ultimately ship with a combination of open source and proprietary software, including proprietary software required for accessing Google services. Android is popular with technology companies that require a ready-made, low-cost and customizable operating system for high-tech devices. Its open nature has encouraged a large community of developers and enthusiasts to use the open-source code as a foundation for community-driven projects, which add new features for advanced users[23] or bring Android to devices originally shipped with other operating systems. At the same time, as Android has no centralised update system most Android devices fail to receive security updates: research in 2015 concluded that almost 90% of Android phones in use had known but unpatched security vulnerabilities due to lack of updates and support. The success of Android has made it a target for patent litigation as part of the so-called "smartphone wars" between technology companies

Software Development Kit

The iOS SDK (Software Development Kit) (formerly iPhone SDK) is a software development kit developed by Apple Inc. and released in February 2008 to developnative applications for iOS.

On October 17, 2007, in an open letter posted to Apple's "Hot News" weblog, Steve Jobs announced that a software development kit (SDK) would be made available to third-party developers in February 2008.[1] The SDK was released on March 6, 2008, and allows developers to make applications for the iPhone and iPod Touch, as well as test them in an "iPhone simulator". However, loading an application onto the devices is only possible after paying an iOS Developer Program fee, which is $99.00 USD per year.[2] Since the release of Xcode 3.1, Xcode is the development environment for the iOS SDK. iPhone applications, like OS Xapplications, are written in Swift and Objective-C,[3] with some elements of an application able to be written in C or C++.

Developers are able to set any price above a set minimum for their applications to be distributed through the App Store, of which they will receive a 70% share. Alternately, they may opt to release the application for free and need not pay any costs to release or distribute the application except for the membership fee.[25]

Since its release, there has been some controversy regarding the refund policy in the fine print of the Developer Agreement with Apple. According to the agreement that developers must agree to, if someone purchases an app from the app store, 30% of the price goes to Apple, and 70% to the developer. If a refund is granted to the customer (at Apple's discretion), the 30% is returned to the customer from Apple, and 70% from the developer; however, Apple can then take another 30% of the cost from the developer to make up for Apple's loss

Android software development is the process by which new applications are created for the Android operating system. Applications are usually developed in Javaprogramming language using the Android software development kit (SDK), but other development environments are also available.

The Android software development kit (SDK) includes a comprehensive set of development tools. These include a debugger, libraries, a handset emulator based onQEMU, documentation, sample code, and tutorials. Currently supported development platforms include computers running Linux (any modern desktop Linux distribution),Mac OS X 10.5.8 or later, and Windows XP or later. As of March 2015, the SDK is not available on Android itself, but the software development is possible by using specialized Android applications.

Until around the end of 2014, the officially supported integrated development environment (IDE) was Eclipse using the Android Development Tools (ADT) Plugin, thoughIntelliJ IDEA IDE (all editions) fully supports Android development out of the box,[7] and NetBeans IDE also supports Android development via a plugin. As of 2015,Android Studio, made by Google and powered by IntelliJ, is the official IDE; however, developers are free to use others. Additionally, developers may use any text editor to edit Java and XML files, then use command line tools (Java Development Kit and Apache Ant are required) to create, build and debug Android applications as well as control attached Android devices (e.g., triggering a reboot, installing software package(s) remotely).

Enhancements to Android's SDK go hand in hand with the overall Android platform development. The SDK also supports older versions of the Android platform in case developers wish to target their applications at older devices. Development tools are downloadable components, so after one has downloaded the latest version and platform, older platforms and tools can also be downloaded for compatibility testing.

Android applications are packaged in .apk format and stored under /data/app folder on the Android OS (the folder is accessible only to the root user for security reasons). APK package contains .dex files (compiled byte code files called Dalvik executables), resource files, etc.

BlackBerry OS

BlackBerry OS is a proprietary mobile operating system developed by BlackBerry Ltd for its BlackBerry line of smartphone handheld devices. The operating system provides multitasking and supports specialized input devices that have been adopted by BlackBerry Ltd. for use in its handhelds, particularly the trackwheel, trackball, and most recently, the trackpad and touchscreen.

The BlackBerry platform is perhaps best known for its native support for corporate email, through MIDP 1.0 and, more recently, a subset of MIDP 2.0, which allows complete wireless activation and synchronization with Microsoft Exchange, Lotus Domino, or Novell GroupWise email, calendar, tasks, notes, and contacts, when used with BlackBerry Enterprise Server. The operating system also supports WAP 1.2.Updates to the operating system may be automatically available from wireless carriers that support the BlackBerry over the air software loading (OTASL) service.

Third-party developers can write software using the available BlackBerry API classes, although applications that make use of certain functionality must be digitally signed.Research from June 2011 indicated that approximately 45% of mobile developers were using the platform at the time of publication. BlackBerry OS was discontinued after the release of BlackBerry 10 but BlackBerry will continue support for the BlackBerry OS.

The Windows Software Development Kit (SDK) for Windows 8 contains headers, libraries, and a selection of tools that you can use when you create apps that run on Windows operating systems. You can use the Windows SDK, along with your chosen development environment, to write Windows Store apps (only on Windows 8) using web technologies (such as HTML5, CSS3, and JavaScript), native (C++), and managed (C#, Visual Basic) code; desktop applications that use the native (Win32/COM) programming model; or desktop applications that use the managed (.NET Framework) programming model.

The Windows SDK also includes the Windows App Certification Kit (ACK) 2.2 to test your app for the Windows 8 Certification Program and the Windows 7 Logo Program. If you also want to test your app on Windows RT, use the Windows App Certification Kit for Windows RT .

The Windows SDK no longer ships with a complete command-line build environment. You must install a compiler and build environment separately. If you require a complete development environment that includes compilers and a build environment, you can download Visual Studio Express , which includes the appropriate components of the Windows SDK. To download the SDK and install it on another computer, click the download link and run the setup. Then in the Specify Location dialog box, click

Mobile commerce

The phrase mobile commerce was originally coined in 1997 by Kevin Duffey at the launch of the Global Mobile Commerce Forum, to mean "the delivery of electronic commerce capabilities directly into the consumer's hand, anywhere, via wireless technology."[1] Many choose to think of Mobile Commerce as meaning "a retail outlet in your customer's pocket

The Global Mobile Commerce Forum, which came to include over 100 organisations, had its fully minuted launch in London on 10 November 1997. Kevin Duffey was elected as the Executive Chairman at the first meeting in November 1997. The meeting was opened by Dr Mike Short, former chairman of the GSM Association, with the very first forecasts for mobile commerce from Kevin Duffey (Group Telecoms Director ofLogica) and Tom Alexander (later CEO of Virgin Mobile and then of Orange). Over 100 companies joined the Forum within a year, many forming mobile commerce teams of their own, e.g. MasterCard and Motorola. Of these one hundred companies, the first two were Logica and Cellnet (which later became O2). Member organisations such as Nokia, Apple, Alcatel, and Vodafone began a series of trials and collaborations.

Mobile commerce services were first delivered in 1997, when the first two mobile-phone enabled Coca Cola vending machines were installed in the Helsinki area in Finland. The machines accepted payment via SMStext messages. This work evolved to several new mobile applications such as the first mobile phone-based banking service was launched in 1997 by Merita Bank of Finland, also using SMS. Finnair mobile check-in was also a major milestone, first introduced in 2001

M-COMMERCE APPLICATIONS The general m-commerce applications are:

1. Mobile ticketing

Tickets can be sent to mobile phones using a variety of technologies. Users are then able to use their tickets immediately by presenting their phones at the venue.Tickets can be booked and cancelled on the mobile with the help of simple applicationdownloads or by accessing WAP portals of various Travel agents or direct service providers. Mobile ticketing for airports,

ballparks, and train stations, for example, will not only streamline unexpected metropolitan traffic surges, but also help users remotely secure parking spots (even while in their vehicles) and greatly facilitate mass surveillance at transport hubs.

2. Mobile vouchers, coupons and loyalty cards

Mobile ticketing technology can also be used for the distribution of vouchers, coupons and loyalty cards. The voucher, coupon, or loyalty card is represented by a virtual token that is sent to the mobile phone. Presenting a mobile phone with one of these tokens at the point of sale allows the customer to receive the same benefits as another customer who has a loyalty card or other paper coupon/voucher. Mobile delivery enables:

- economy of scale
- quicker and easier delivery
- effective target marketing
- privacy-friendly data mining on consumer behaviour
- environment-friendly and resources-saving efficacy

Content purchase and delivery

Currently, mobile content purchase and delivery mainly consists of the sale of ring-tones, wallpapers, and games for mobile phones. The convergence of mobile phones, mp3 players and video players into a single device will result in an increase in the purchase and delivery of full-length music tracks and video. Download speeds, if increased to 4G levels, will make it possible to buy a movie on a mobile device in a couple of seconds, while on the go.

4. Location-based services

Unlike a home PC, the location of the mobile phone user is an important piece of information used during mobile commerce transactions. Knowing the location of the user allows for location based services such as:

- local maps
- local offers
- local weather

people tracking and monitoring