Iot

Gabriella Trasciatti

March 2021

1 Read me

This notes are made from the lessons of IoT by professor Petrioli, year 2020/2021, and the italian book GSM-GPRS; i included personal comments that are denoted with square brackets or text super script. Enjoy and feel free to share these notes.

2 3/03

slide iot 1

2.1 About transmission and waves

2.1.1 Friis transmission equation

The power received at a destination in a distance d in case of no obstacle and LOS (line of sight, the signal hasn't been modified [scattering, reflection etc.]) can be described as:

$$P_r = P_t g_t g_r \left(\frac{\lambda}{4\pi d}\right)^2 \frac{1}{L} \tag{1}$$

The power received is influenced by the power transmitted, multiplied for both the gain in transmission and reception; $\frac{1}{L}$ represents hardware loss, λ the wavelength (c/f, speed of light divided by frequency), and d the distance between the transmitter and the receiver.

2.1.2 Path loss with respect to distance

Path loss, which is the rate between the transmitted power and the received, can be described as:

$$PL = \frac{P_t}{P_r} = \left(\frac{\lambda}{4\pi d}\right)^{-2} \tag{2}$$

Intuitively, the path loss will increase with higher frequencies ($\lambda = c/f$, with high f, λ will be smaller and thus will be the whole fraction) and greater distance; this is assuming that the gains and hardware loss are equal to one.

2.1.3 Multi-path fading

It's the phenomenon that a signal experience when received as a copy of multiple delayed signals. Multiple copy of the signal will arrive at the destination because the signal can be scattered (the wave is divided in multiple waves; this happens when obstacles smaller than the wave are met), reflected, and diffracted (the wave is bent by obstacles with sharp edges); however, the different signals will experience different delays, and for this reason the received signal will experience Intersymbol Interference (the various copy of the signal will blur together). To fix this, equalization is used, and it works roughly as: a computer is given the original signal, and it will train itself with different disturbed copies of the signal; in this way it will learn to untangle the original signal from them.

2.2 Features of wireless networks

1. Broadcast medium: each terminal is overhead by all the others
2. Shared channel: the limited resources are shared among every user and MAC is required
3. High Bit Error Rate (BET): error detection, correction and retransmission are needed for reliable communication
4. Portable devices rely on batteries: energy efficient protocols must be deployed, communication vs computation trade off must be reasoned, and HW techniques to avoid wasting energy must be used.

Two models of wireless communication (are shown in the lesson): infrastructured networks, where the mobile user communicate to the access point and vice-versa, and ad-hoc wireless networks, where users communicate with each others.

2.2.1 Wireless vs wired

The unique features of the transmission medium have a big impact on the design; the things that must be accounted for are: low reliability, hidden terminal problem, broadcast feature, etc. this demands different solution in the data link and transport layer. The advantages of the wireless are: no cabling, anywhere anytime, cost vs performance.

3 5/03

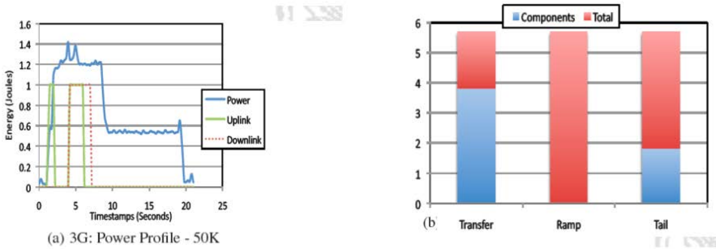
slide iot 2_5

4 10/03

slide 3_4

4.0.1 Energy consumption in 3G

In 3G, a large fraction (nearly 60%) of the energy, referred to as the tail energy, is wasted in high-power states after the completion of a typical transfer (the tail time is decided by the operator and not the devices; it's used to reduce latency and signal overhead, because in this idle state, if more packets are received, it's not needed to repeat the procedure of searching a channel etc.); the ramp energy on the other end, isn't this much consuming, and is described as the energy spent to switch to a high power state. These waste are amortized over larger file transfer, so a solution might be to transfer big batch of data (and it's applicable for delay tolerant application such as emails). The tail ender is a protocol that has the task to reduce the waste from tail energy, and it does that in 3 ways: combine the usage of 3G and wifi, predicting when a wifi network might be available, wait to transfer bigger batch of data for delay tolerant applications, and prefetching in application that can benefit from it (web surfing, we search).



(a) 3G: Power Profile - 50K

4.1 Energy consumption of 802.11 devices

4.1.1 Cross factor

The energy consumed by a frame that is delivered across the protocol stack, from the operative system to the network interface controller (NIC); it has no correlation on the size of a packet, but it represents the fact that it is simply processed.

5 12/03

5.1 Splitting resources to users

Radio spectrum is controlled by the government, and cellular providers pay to use part of the spectrum.

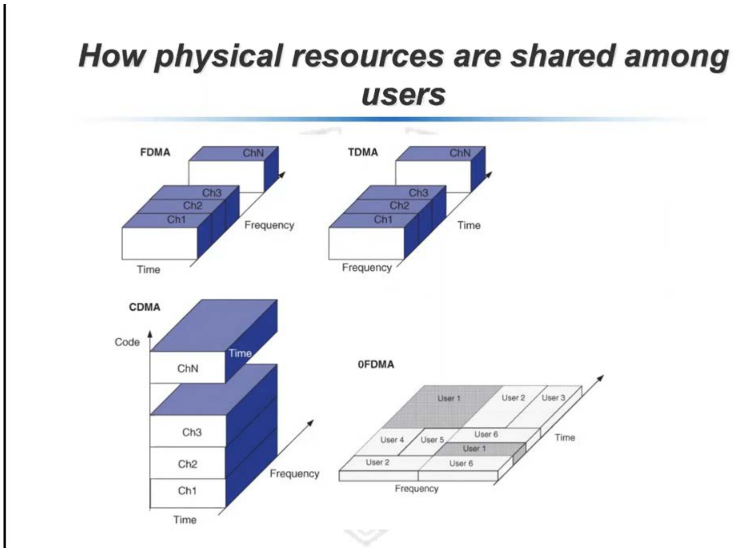

How physical resources are shared among users

Figure 1:

5.1.1 FDMA

In this case the frequency is split in different channels, and each channel is assigned to one user, so he can use $\frac{1}{N}$ of the bandwidth; first come first served policy is applied, and there are no collisions since the resources are dedicated to each user. To receive and transmit, two radio frequency are required (one from the calling and one from the called); the mobile station and the radio base station communicates on the same channel at the same time. For this reason the mobile station must have a duplexer (a circuit on the antenna) to filter and distinguish the received signal from the transmission (and do both simultaneously). The handover (change of base station) is perceived by the user since the transmission is continuous. This is adopted in analogical systems.

5.1.2 TDMA

The channels are created allocating all the bandwidth for a limited amount of time; there's no collision since the resources are dedicated to each user. The signal is digital (the voice is digitalized by the mobile station); the reception and transmission are typically done in different time slots and frequencies, and for this reason no duplexer is required. The handover is not perceived by the user; some synchronizing is required since the propagation of signal could take time and interfere with the time division.

5.1.3 CDMA

The bandwidth is used by all the users simultaneously; each user is given a (unique) code, and when receiving information, a mobile station is capable of extract them by applying its code; the information with a different code are considered noise. Notice that even if there's no limit to the amount of users, too many will reduce the quality, since interference will increase. It's used in 3G; it's costly in terms of power management.

5.1.4 0FDMA

The users are given resources blocks that are calculated dynamically.

5.2 Frequency reuse

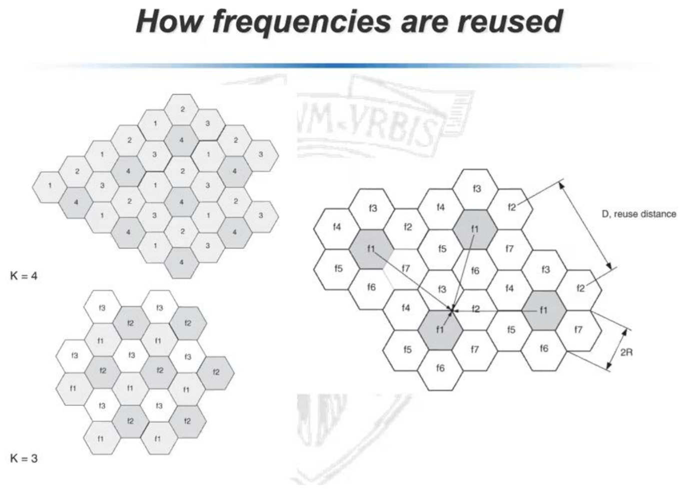

How frequencies are reused

Figure 2: Frequency reuse

.

Each hexagon is the range of a base station (located in the center). Cellular system operators will give each cell a channel different from the ones of adjacent cells, to avoid interference; the frequency reuse is the fact that far away cells will use the same channel without problems. This approach is vastly used and it's very scalable.

The cells can be of different size and a cluster of cells can use the same channel. To be very efficient, lots of small cells must be deployed in very populated areas, so the traffic capacity is stable, and larger cells must be used to coverage large but low-density areas; this is called splitting. Many cell concentrated in a small area are an advantage because in this way the same frequency can be used by many users without interfering with each others.

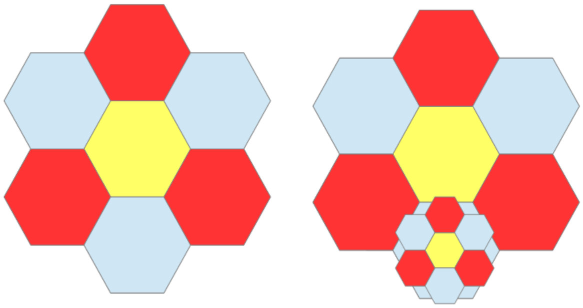


Figure 3: Illustration of using many cell in a small area

This figure roughly illustrate the concept; let's take the lower azure hexagon: it represent a cell who transmits in one of the three available frequency. If it's split in other little cells, different frequencies can be used without interference, so instead of a single frequency split among many user, there will be multiple frequencies split among a fraction of the original users (this fraction will be the one covered by the little cell). Keep in mind that the use of many cell instead of a bigger one is costly, and handovers will be more frequent; however, the power consumption will be lower.

5.3 GSM general features

GSM is a standard to describe the protocols of the 2° generation cellular system. In this system the signal is digital (an advantage is that forward error correction coding can be applied, and also the transmission can be encrypted). The carrier bandwidth is 200KHz (capable of transmitting at 270,7 kbit/s) [passo di canalizzazione, the division of frequency in radio channels], and a TDMA/FDMA mix is used. Frequency reuse is used. Services: telephony, circuit switching, packet switching (GPRS). The voice can be coded at 13 kbit/s (full rate) or 6,5kbit/s (half rate, not really used but is decent enough and potentially can double the number of channels). There are different bands of frequencies depending on the task (eg. railway connection), but the main 2 standards in europe are the one operating at 900 MHz (Primary-GSM) and the one at 1800MHz (DCS); they are divided again between uplink (from mobile station to base station) and downlink (from base station to mobile station); notice that since transmitting at higher frequencies absorbs more power, the uplink is on the lower-half of the frequency; the [passo di duplice, space between transmitting frequencies and receiving frequenciesy] is 45MHz (P-GSM) or 95MHz (DSM). Power control is used.

5.4 GSM Architecture

1. The base station subsystem is the access network, which interfaces the mobile users with the cellular system operator network; it will provide the wireless access to the mobile users with base stations, and it's made by BTS (base transmission stations, which physically transmit information) and BSC (base station controller)

2. The network switching subsystem is a part of the core component; MSC (mobile switching centers) relay the information on the network to GMSC, the connection to external network. There are also databases, the VLR (visitors location register) which provides information about the users currently resident inside cells controlled by a specific MSC; the HLR (home location register) mantains information about all the users that had contact with the cellular system operator. The EIR (equipment identity register) maintain information about the hardware of mobile stations, used to prevent unauthorized mobile stations from using the network. The AUT

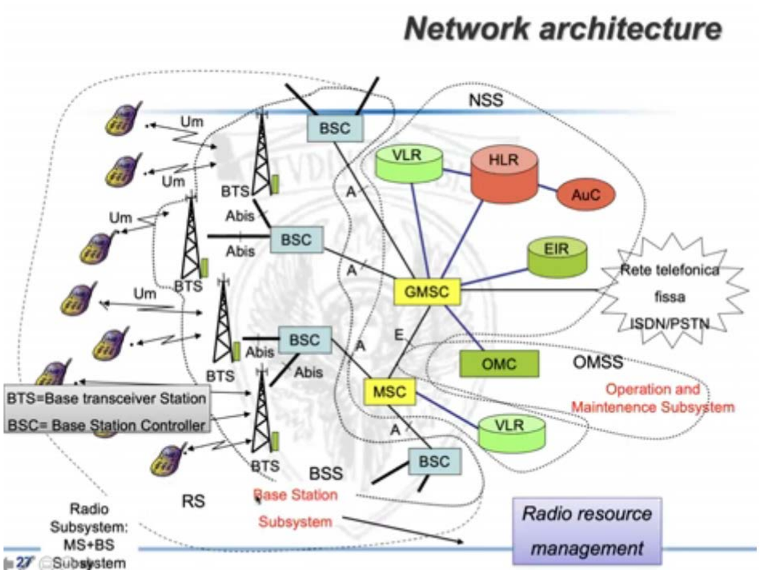

Network architecture
NSS
Um
Um
Um
Um
BTS
Abis
Abis
BTS
BTS
Abis
Abis
BTS
BSC
BSC
BSC
BSC
A
A
A
A
E
VLR
HLR
AuC
EIR
GMSC
Rete telefonica fissa ISDN/PSTN
OMC
OMSS
Operation and Maintenence Subsystem
MSC
VLR
BTS=Base transceiver Station
BSC= Base Station Controller
BSS
Base Station Subsystem
RS
Radio Subsystem: MS+BS Subsystem
Radio resource management
27

Figure 4: Network architecture

(authentication center) is used to generate parameters used to authenticate and encrypt the user

There are different areas of a GSM:

1. PLMN (Public Land Mobile Network): service area of a cellular network; it is split in sub-areas each managed by MSC

2. MSC/VLR: each sub-area of PLMN is managed by an MSC, and data regarding these users are temporary stored in VLR databases associated with each MSC.

3. Location area: a logical division of MSC/VLR area; when a user change LA, he must perform a location update. LA are identified by LAI , each transmitted by the BTS of the LA.

4. Cells: area covered by a BTS; it has a identifier BSIC (base station identity code), which is broadcasted by the BTS.

Mobile stations are divided in three categories depending on the emission power capacity: 20W (vehicular), 8W (Laptops), 2W (mobile phone); some mobile stations can operate multi-band (both GSM 800mhz, DCS 1800mhz and GSM-USA 1900mhz), multi-slot (can operate over different channels in different slots, used to exchange data, GPRS). They are composed of: a ME (mobile equipment) and a SIM(subscriber identity module). The ME is the terminal to access the cellular network, and it's composed by an hardware, the HW/SW interface and the user interface; it is identified by an IMEI. The SIM activates the terminal for a user and stores information, and has a serial number (which identify the SIM), an international mobile subscriber identity IMSI (uniquely identifies the user in the network, and it's stored in the VLR too to know which service offer), security authentication and cyphering information (keys and algorithms), and temporary network information (LAI, last visited area information, TMSI, temporary identifier assigned by the network, used instead of IMSI to avoid eavesdropping, since TMSI can be randomly generated); other information stored on the SIM are: the list of services, the telephone number, pin puc, access rights, etc.. A ME without a SIM is capable of making emergency calls.

5.4.1 Base Station Subsystem (BSS)

BSS is the unit that manage radio communication and radio resources; it also must manages the data of configuration of cells, and must handle handovers. It is formed by BTS (base transceiver stations), whose task is to provide radio coverage on the cell, and BSC (base station controller), who monitors and manages groups of BTS; BSC tells the BTS when initiate a call or perform an handover, and reserves/relases radio channels.

- BTS – Base Transceiver Station.
- BSC – Base Station Controller.

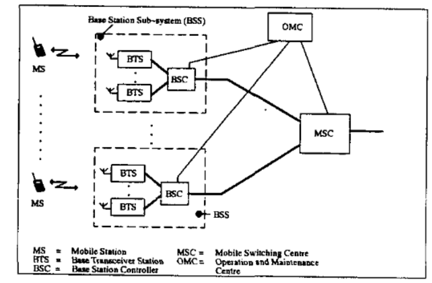

Base Station Sub-system (BSS)
OMC
MS
BTS
BTS
BSC
MSC
BTS
BTS
BSC
MS
BSS
MS = Mobile Station
BTS = Base Transceiver Station
BSC = Base Station Controller
MSC = Mobile Switching Centre
OMC = Operation and Maintenance Centre

Fig. 5.5 Elementi che compongono il BSS – Base Station System.

Figure 5: BSS architecture

BTS is formed by transceivers (TRX) who transmits, receives, carries out signal processing and exchange information with BSC (TRX controllers); and BCF (common base function), who contains function such as synchronization and algorithm for frequency hopping. Frequency hopping is the process of hopping from one channel to others, applying a pseudo-random algorithm. BTSs apply low-level protocols of the radio interface; they transmit and receive signals from MS and do frequency hopping (if enabled) and encryption. They broadcast System Information, which are information needed by MSs to access the network (such as cell identifier, local area identifier etc). They perform quality measurement (and receive measurement made by MSs) of physical channels and send them to BSCs, who will decide whether perform an handover or not. BTSs have also the task to locate users using paging messages.

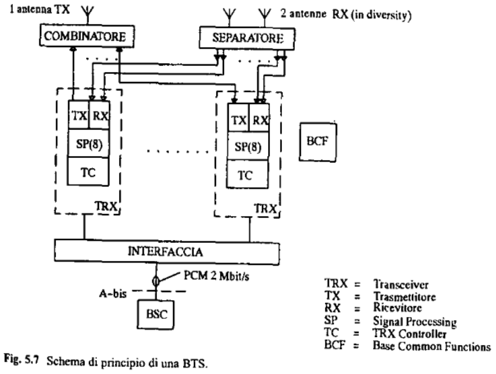


Fig. 5.7 Schema di principio di una BTS.

Figure 6: BTS architecture

TRAU Transcoder Rate Adaption Unit is applied to enable communication between BTS and BSC, since they are physically linked by a PCM channel, who communicates with channels at 64kbit/s, instead of the 13kbit/s of GSM. On each PCM channel, 4 channels of GSM are multiplexed, after being transformed in 16kbit/s adding redundancy. A total of 3 PCM channels are used, one for LAPD protocol and 2 to multiplex the 8 channels of GSM.

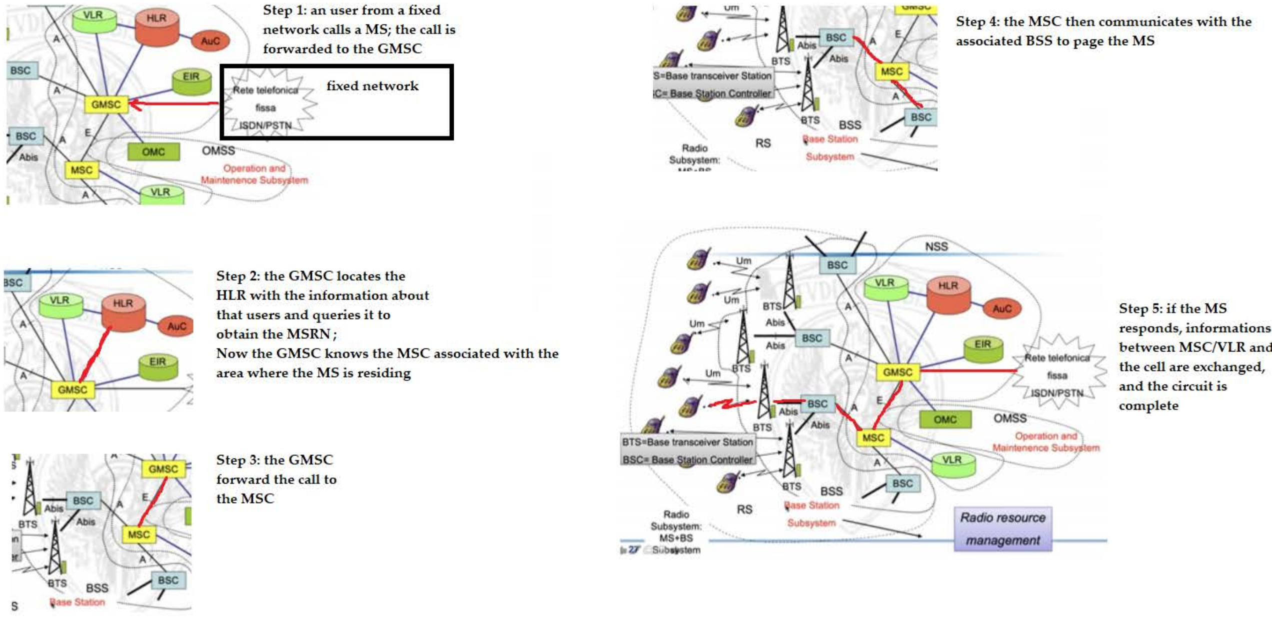
BSC controls various numbers of BTS. Their main tasks are: the configuration of each cell in terms of traffic control and control channels, manages handovers, the paging messages (the ones used to locate users) and performs analysis on quality measurements of BTSs and MSs.

5.4.2 Network Switching Subsystem

it's formed by:

1. MSC whose task is to manage mobility and to route calls.
2. VLR is a database that stores information of users in an MSC area
3. HLR stores information about all the users
4. AuC computes keys to identify and encrypt users; it's usually associated with a HLR
5. EIR contains the IMEI of all the authorized devices
6. CGR

MSC Mobile-switching services centre is a switching element (performs routing) who is connected to the BSC of its area and to other MSCs via PCM channels. One or more MSC (Gateway MSC - GSMC) for each cellular network (or PLMN Public land mobile network) is interfaced with the fixed telephone network ("rete telefonica fissa" - is an external network, such as internet). A MS user can be reached by fixed users (users connected to a fixed network - "rete fissa"$^{\text{sorry in english it sounds very bad}}$) using telephone number (whose acronym is MSISDN); the call is forwarded to GMSC, which identifies the HLR (home location register) that contains information about the user with that MSISDN, and queries it to determinate how to route to the mobile user current MSC. The HLR returns the MSRN (mobile station roaming number, the HLR knows it because it knows the VLR/MSC associated with that user); this MSRN is a temporary number given by the visited VLR, and it allows GMSC to route the call to the MSC area where the user is connected. Now the MSC/VLR will page the user to obtain information about the location area the user is residing, and then control information between the residing cell and the MSC/VLR will be exchanged to finish the circuit (the circuit is the concrete link from the calling user to the called user); once the circuit is complete, the conversation can begin.


VLR
HLR
AuC
BSC
EIR
Rete telefonica fissa ISDN/PSTN
GMSC
MSC
OMC
OMSS
Operation and Maintenence Subsystem
Abis
BTS
BSS
Base Station Subsystem
Um
BTS=Base transceiver Station
BSC= Base Station Controller
RS
Radio Subsystem: MS+BS Subsystem
NSS
Radio resource management
A
E

Step 1: an user from a fixed network calls a MS; the call is forwarded to the GMSC

fixed network

Step 2: the GMSC locates the HLR with the information about that users and queries it to obtain the MSRN;
Now the GMSC knows the MSC associated with the area where the MS is residing

Step 3: the GMSC forward the call to the MSC

Step 4: the MSC then communicates with the associated BSS to page the MS

Step 5: if the MS responds, informations between MSC/VLR and the cell are exchanged, and the circuit is complete

Figure 7: The call process [explained in detail in 7.3]

[please notice that the above image is made by me and i divided it in 5 steps, but they may not be 100% accurate]
The MSC provides the following functionalities:

1. Connection Management: originating calls, terminating call, gateway

2. Mobility Management: location updating, periodic registration, authentication, ecc..

The MSC/VLR implements various protocols to exchange information with other network elements according to the CCSS7 signaling system.

HLR is a permanent database uniquely associated with a GMSC; it stores information about all MS whose default location is the considered GMSC. The stored information are: the IMSI, the authentication number of the SIM and the associated authentication key, additional services subscribed by the user, etc.. Some temporary information are also stored, such as the address of the VLR where the user can be found, parameters for identification and encryption, list of phone numbers for call forwarding, etc..
Its main tasks are:

1. managing localization storing the VLR number of each registered user

2. sending routing information (MSRN) to the GMSC

3. registration, cancellation, activation/deactivation of additional services

4. storage and supply to the VLR with the parameters of autehentication and encryption

5. management of user data

Notice that HLR is a database and while it stores security parameters, it doesn't generates them.

VLR is a temporary database that contains all the data necessary to serve a MS that is under the jurisdiction of the associated MSC/VLR; it contains a copy of every permanent data (stored also in the HLR) minus the IMSI, who is mapped to a TMSI to avoid transmitting in clear the IMSI; the TMSI is changed frequently and is associated with the location of the MS.

5.4.3 Security

The security is guaranteed because every user is authenticated, and the conversation are encrypted via encryption algorithms and frequency hopping; encryption can be performed because the signal is digital. The anonymity is ensured trough the use of temporary identification numbers TMSI. Now let's see which parameters are used in authentication:

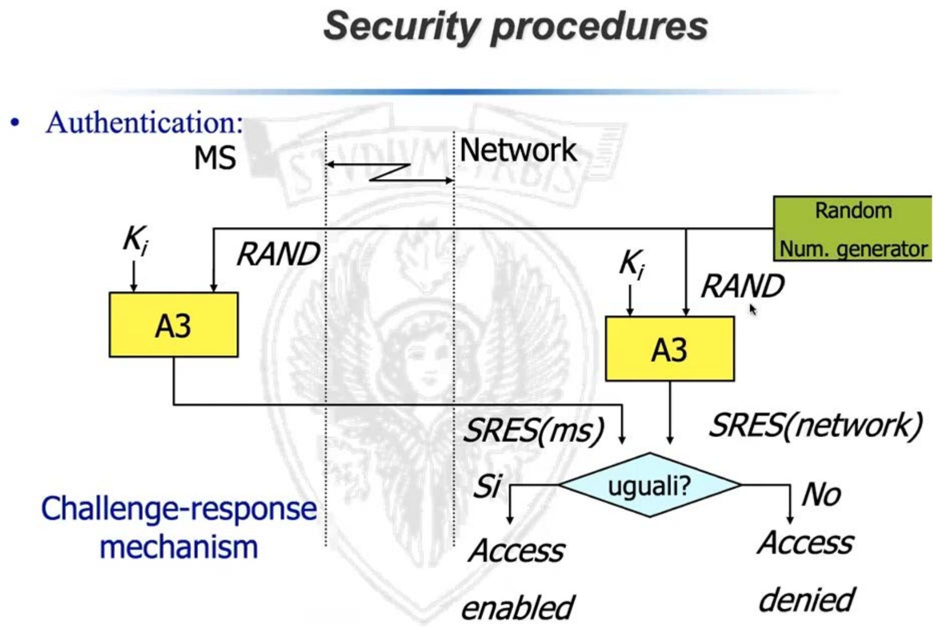
1. user's identification key K_i; it's 128 bits long and stored both in the SIM and AuC

2. a random number long 128 bits called $RAND$; it's generated by the AuC and sent to the MS after the access request,

3. an identification algorithm called A_3; it is stored both in the AuC and the SIM. The AuC can use it to determinate the respond (called SRES) that the MS must send to the network to obtain the access

4. algorithm A_8, who determinates the decipher key K_c using $RAND$ and K_i; it's stored both in the SIM and AuC.

In the network, the set ($RAND$, $SRES$, K_c) is known as triplet and is always associated with an IMSI; this triplets are generated frequently as different ones must be used in every access. These triplets are generated by the AuC, stored in the HLR, and must be sent to the VLRs that requests them. Notice that the only transmitted parameter over the network is $RAND$.



Security procedures

- Authentication:

Figure 8: The process of authentication

To start a ciphered conversation, the procedure is the following:

1. the network sends a ”ciphering mode” command to the MS
2. when the MS receive the ciphered mode message, it sends a ciphered message to the network
3. if the network can correctly decipher the message, then every other message between them is ciphered

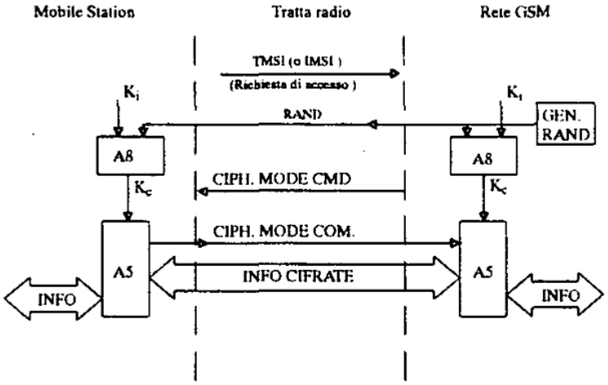


Figure 9: Ciphering process

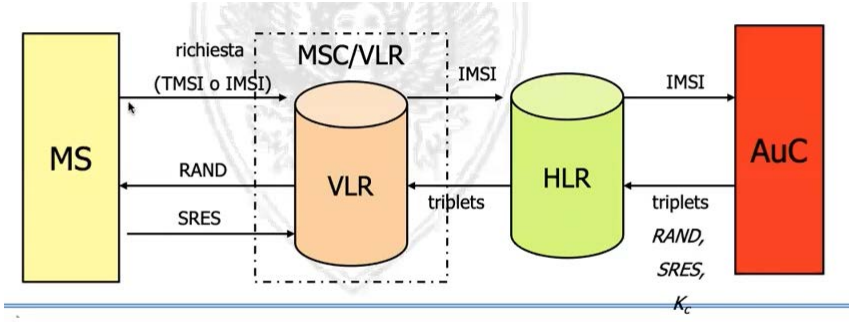


Figure 10: Protagonists involved in the process of authentication

TMSI allocation when a MS wants to start communication, it sends its IMSI to be identified; the VLR then allocate a TMSI, and does it every time there's a location update. TMSI is stored in the SIM, and the only two cases where it's not used is when a MS doesn't have one or when the networks specifically requests it. The TMSI is changed at every location update because it uniquely identifies a user only in a given Location Area.

IMSI constists in 3 fields: the MCC (Mobile Country Code), who identifies the country, the MNC (Mobile Network Code), who identifies the cellular operator, and the MSIC (Mobile Subscriber Identification Number), which identifies the SIM. Notice that the telephone number (MSISDN) is completely independent of this, and their prefix identifies the HLR and GMSC which the device is connected. Remember that the IMSI is the number used to identify users in databases.

EIR the Equipment Identity Register contains the identification and characteristics of GSM Terminal Equipment, together with the manufacturer, the country of manufacturer etc.. It can be used to protect the network from the use of stolen or not compliant to standard equipment; the use it at the discretion of the operator. It's used to check the IMEI of MS; wehn the MS send a connection request to the MSC/VLR, this latter send an IMEI request and, when the IMEI is received by the MS, a "check IMEI" is sent to the EIR. There are three type of IMEI lists: white, where valid IMEIs are stored and safely to use, black, where there are IMEIs banned from accessing the network (but not if they're making emergency calls) and gray, whose IMEIs can access the network at the discretion of the operator, but must be tracked. These lists are updated periodically, and the black one is exchanged among operators.

5.4.4 Operation and Maintenance Subsystem (OMSS)

Deals with all aspects of monitoring, maintenance and remote management of the network. Notice that the GSM network's management is based on the concept of Telecommunication Management Network (TMN), and thus is divided hierarchically as follows:

1. Network Element Level: it's the level that comprehends all the functional units that compose a GSM network, and that must be checked

2. Element Management Level: it comprehends a number of regional control centres named Operation and Maintenance Centre (OMC); their number varies as the size of the network varies

3. Network Management Level: it's the highest level; it provides a global vision of the network and thus all the activities of OMCs; It's called NMC and it coordinates the various OMC to avoid discrepancies.

The NMC provides a vision over all the activities of Operation Administration and Maintenance (OA&M), and it's composed by two main functional units: the GSM Management and Operation Centre (GMC), who manage all the aspects related to faults, performances, configuration, etc.; and GSM Support Centre (GSC), who manages the administrative aspects such as management of users, taxation etc.. Inside the Network Management Level there's also the PCS (Personalization Centre for SIM), who provides the mean to personalize the SIM; it can upload private user's data such as the IMSI and the personal identification key K_i.
The principal tasks of an OMC are:

1. Management of failures and maintenance of the network

2. Manage the configuration of network elements

3. Manage the performance of network elements (definition, acquisition, memorization, presentation etc.)

4. Manage the security aspects

5. Acquisition of billing data, whom are defined by the records of call documentation emitted by MSC; they are provided to centres that bill users of the GSM

6. Management of accounting data, whom permit to subdivide the taxation of a call between the GSM operator and other possible networks that intervened (ISDN/PTSN $^{\text{rete fissa}}$, PLMN)

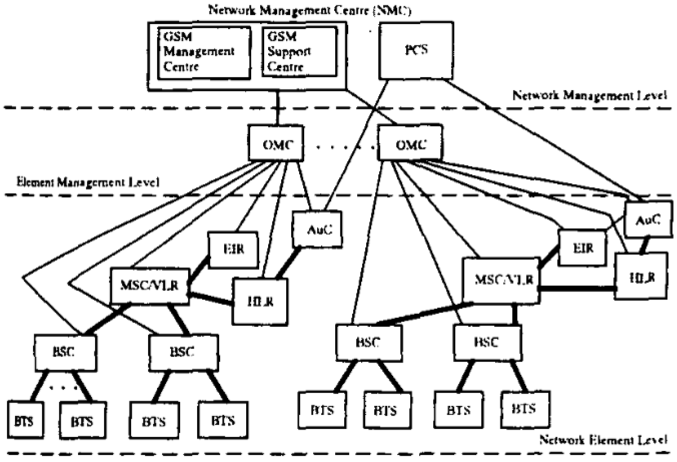


Figure 11: The hierarchical structure of OMSS

5.5 Identification numbers in GSM

There will be recapped all the acronyms and numbers used in the GSM system.

MSISDN Mobile Station ISDN Number is the telephone number of MSs; it's called like this because GSM networks can be interfaced with PTSN/ISDN networks ($^{\text{rete fissa}}$), so MS need numbers conforming with the ones of these networks (the reason is that ISDN networks are substituting [or they already substituted, the book is from 2003] PTSN ones). In this way, calls from MS can be taxed in the same way as the ones in PTSN/ISDN (hooray). The number is composed by: Country Code (identifies the country of that number), National Destination Code (identifies a specific PLMN [Public Land Mobile Network, the service area of a cellular network of a provider]; there can be different NDC for the same PLMN if said PLMN has more than one HLR), Subscriber Number (identifies the number inside a PLMN GSM).

MSRN Mobile Station Roaming Number it's needed because unlike in PTSN network, a MS can move wherever it likes inside the GSM, so a MSISDN can only provide information about the GMSC and not the current MSC/VLR. MSRN is a temporary telephone number assigned to the MS, with a structure similar to the MSISDN (Country code, National destination code, subscriber number), but it contains information that permits the routing to the MSC/VLR where the user is residing; this number is assigned by VLRs and it's valid only in the area of that VLR, and permits to complete the circuit from the GMSC to the MS.

Handover Number it's a number used when a MS in a conversation moves from a MSC/VLR service area to another one; in this case an inter MSC/VLR handover must be performed. The MSC/VLR target (the one the user has moved to) has to provide an handover number to the initial MSC/VLR, and must also complete the route to the MS.

IMSI is a number that uniquely identifies a user inside a GSM system; it's permanently memorized inside the SIM and the HLR, and temporarly memorized inside the VLR that is momentarily in charge of the user.

TMSI is a number that temporarily identifies the user, to avoid transmitting the IMSI in clear via radio; it's periodically assigned by VLRs, and, with exception of rare cases, it's the number by which the MS presents itself to the system.

IMEI is a number that uniquely identifies a ME (mobile equipment); it's stored in the EIR and checked frequently by the GSM system.

LAI Location Area Identity identifies location areas. The area serviced by a MSC/VLR is subdivided in various location areas, that are in turn subdivided in different cells. A location area is defined as "the area where the MS can freely move without updating its location"; this means that a MS is localized by looking at the residing location area. For this reasons, and the fact that a different location area may be residing in another MSC/VLR (and an handover could be required), they need an identification, which is the LAI.

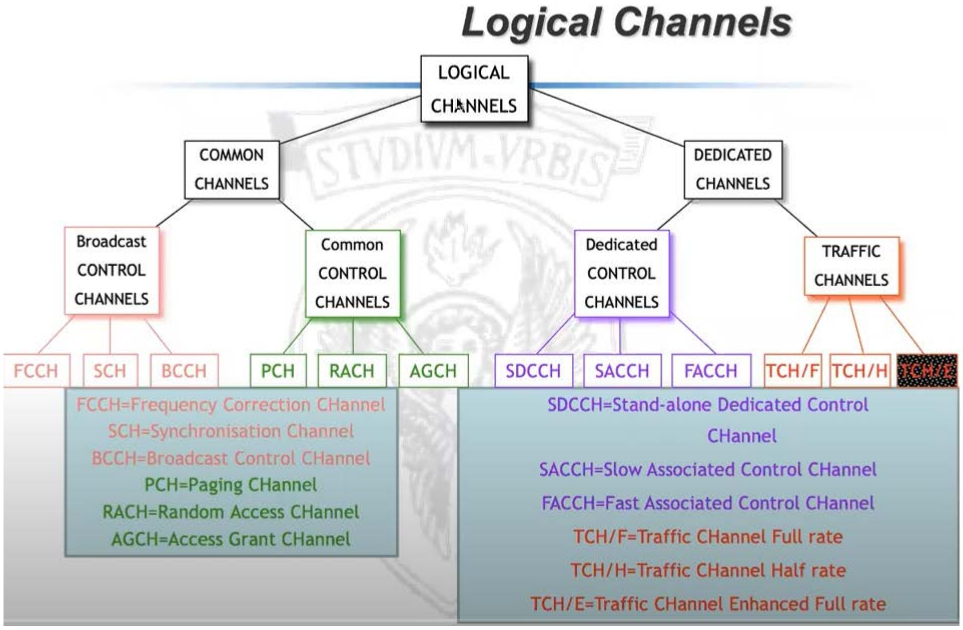
CGI it's a number that identifies cells, and its composed by the LAI and a CI (cell identifier).

BSIC Base Station Identity Code it's a (color) code that allows MS to distinguish adjacent BTS, and it's broadcasted by them. It's crucial for BSCs that decides handovers, because the MS performs measurements of signal intensity, and must report these measurements along with the BSIC to decide on which cell the handover must be done.

6 17/03

6.1 Logical channels

In addition to the typical data exchanged in GSM systems (voice), system data must be exchanged as well, such as: signaling, data traffic, etc. to support and guarantee all the services. This information flow is splitted into logical channels, and they're grouped depending on their needing. The first big split that can be done is differentiate between the channels that must carry traffic data and ones that carry control information; another split is between dedicated channels (that serve only one user, e.g the call circuit) and common channels (who serve different users at the same time, e.g paging channel, where paging messages are broadcasted to multiple users).



Logical Channels

LOGICAL CHANNELS

COMMON CHANNELS

DEDICATED CHANNELS

Broadcast CONTROL CHANNELS

Common CONTROL CHANNELS

Dedicated CONTROL CHANNELS

TRAFFIC CHANNELS

FCCH SCH BCCH PCH RACH AGCH SDCCH SACCH FACCH TCH/F TCH/H TCH/E

FCCH=Frequency Correction CHannel
SCH=Synchronisation Channel
BCCH=Broadcast Control CHannel
PCH=Paging CHannel
RACH=Random Access CHannel
AGCH=Access Grant CHannel

SDCCH=Stand-alone Dedicated Control CHannel
SACCH=Slow Associated Control CHannel
FACCH=Fast Associated Control CHannel
TCH/F=Traffic CHannel Full rate
TCH/H=Traffic CHannel Half rate
TCH/E=Traffic CHannel Enhanced Full rate

Figure 12: Logical channels' splitting

6.1.1 Broadcast control channels

Channels used to broadcast to all the users of a cell.

FCCH Frequency Correction CHannel allows the MS in a cell to get synchronized to the carrier frequency; a sequence of 148 bits ir's transmitted, and it represents the offset in the frequency (the sinusoidal pattern) of the carrier. This channel is unidirectional downlink.

SCH Synchronization CHannel carries 25bit of information to allow the synchronization of the MS and the identification of the BTS; 6 bits are used for the BSIC, and 19 are used to transmit the TDMA frame number [19 is a small number, so it's not exactly the frame number that is sent but some numbers to calculate it]. The channel is unidirectional downlink.

BCCH Broadcast Control CHannel that carries information for each cell to all the users served by that BTS; it carries a large number of information (184bits in total) and error correction is adequately applied. The information carried are: the number of common channels, a bit to flag if common channels are associated to dedicated channels on the same physical channel, the number of blocks reserved for the AGCH on each common control channel, the distance of two consecutive page messages to the same MS, and the parameters requested by the frequency hopping algorithm. This channel is unidirectional downlink.

6.1.2 Common control channels

Channels used to share information to some actors.

PCH Paging CHannel used by the BTS to notify a MS about an incoming call [as strange as it sounds, it's true] and paging messages; it's downlink to every cell in a location area.

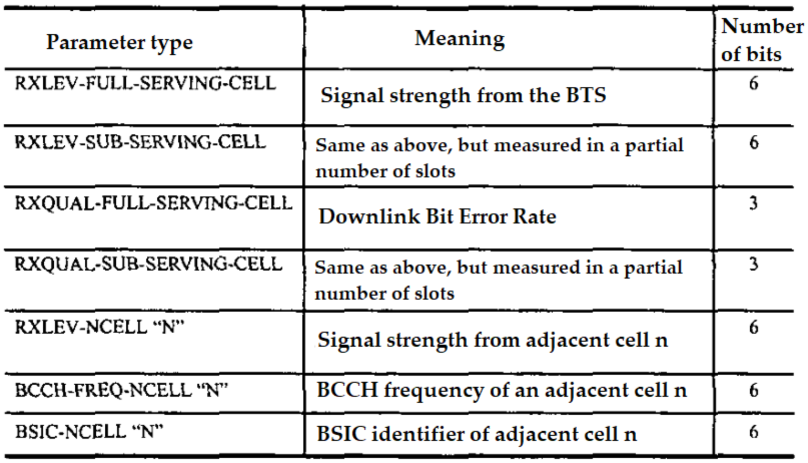
RACH Random Access CHannel used by MS to send signals to request the access to the network and to request a call start (in this case the request is to have a dedicated channel); it's also used for the procedure of updating the location of MSs. The channel is unidirectional uplink, and subject to collisions; for this reason, the transmission on RACH uses the slotted-ALOHA protocol.

AGCH Access Grant CHannel used to respond to a RACH request of a dedicated channel. Unidirectional downlink.

6.1.3 Dedicated control channels

Channels used to communicate to single actors.

SACCH Slow Associated Control CHannel carries information between mobile means and a fixed network during communication. In downlink, this channel carries information about measurements of BTS, control of transmission power and all the information of BCCH that would be lost by the MS that has settled down on its own traffic channel. On the uplink, it'll carry the information collected by the MS about the surrounding radio area; this channel is multiplexed with the user's traffic.



Parameter type	Meaning	Number of bits
RXLEV-FULL-SERVING-CELL	Signal strength from the BTS	6
RXLEV-SUB-SERVING-CELL	Same as above, but measured in a partial number of slots	6
RXQUAL-FULL-SERVING-CELL	Downlink Bit Error Rate	3
RXQUAL-SUB-SERVING-CELL	Same as above, but measured in a partial number of slots	3
RXLEV-NCELL "N"	Signal strength from adjacent cell n	6
BCCH-FREQ-NCELL "N"	BCCH frequency of an adjacent cell n	6
BSIC-NCELL "N"	BSIC identifier of adjacent cell n	6

Figure 13: Parameters transmitted in uplink

FACCH Fast Associated Control CHannel carries parameters that must be transmitted fast and so the SACCH can't be used, e.g an immediate handover caused by a fault in the channel. This channel "steals" time slots that should be used to transmit traffic in order to deliver the messages as soon as possible.

SDCCH Standalone Dedicated Control CHannel it's the dedicated channel given to a MS after a call request on RACH channel; it's received by message in the AGCH channel. It's used to exchange the parameters for authentification, identification and call set-up procedures.

CBCH Cell Broadcast CHAnnel it's a channel that permits the diffusion of small messages with low periodicity in a cell; it's capable of sending messages of 80 bytes every two seconds.

6.1.4 Traffic channels

These channels carry both codified speech and user's data; they're called TCH (Traffic CHannels) and are divided into two principal types.

TCH/F TCH Full-rate, carries information at full rate, 22,8 kbit/s.

TCH/H TCH Half-rate, carries information at half rate, 6,4 kbit/s.
Each channel can carry voice (speech) or data at different rates:

1. Voice TCH/F: TCH/FS
2. Voice TCH/H: TCH/HS
3. Data TCH/F at 9,6 kbit/s: TCH/F9.6
4. Data TCH/F at 4,8 kbit/s: TCH/F4.8
5. Data TCH/H at 4,8 kbit/s: TCH/H4.8
6. Data TCH/H at 2,4 kbit/s: TCH/H2.4
7. Data TCH/F at 2,4 kbit/s: TCH/F2.4

Half rate channels are multiplexed two-by-two in the same time interval but in alternate frames; every channel carries data that is ciphered by FEC (forward error control).

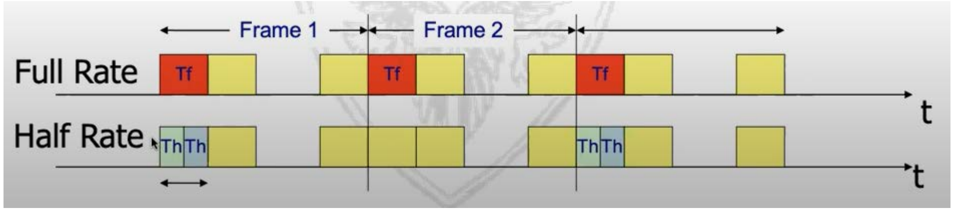


Figure 14: How full & half rate use frame

Interleaving is used in transmission of traffic and it works like this: let's assume i have to transmit some information and i split them in 4 different blocks; there can be issues when transmitting to the physical channel that creates a high BET (bit error rate). Many blocks can be corrupted, and even with FEC (forward error correction), if a high percentage of information are lost, then the message can't be reconstructed. But if the error can be distributed across all the packet, then the FEC would work. A technique used to do this is interleaving, which takes information from different blocks and send packages of these mixed information; the destination is able to reconstruct the original information and apply FEC.

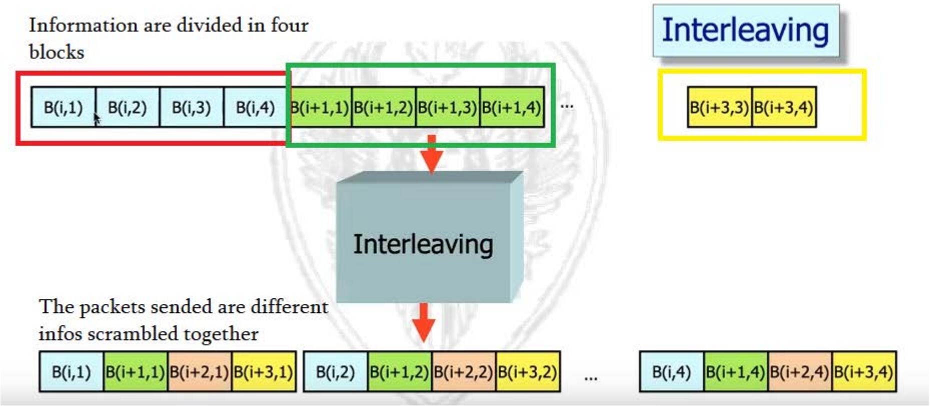


Figure 15: Interleaving process

The information that is cut in 4 blocks is long 114 bits and is produced in 20ms of calling.

If a particular frequency is prone to high BER, in order to not penalize a single MS, frequency hopping (implemented with a pseudo random algorithm) is adopted; this technique is used with interleaving to minimize the information loss.

6.1.5 Mapping between physical and logical channel

Signaling requires low bit rate compared to user's transmission; for this reason a whole slot for each frame for a logical channel is a waste. So the transmission of logical channel is done with multiframing: each slot of a frame is reserved to signaling (but not to a single logical channel), and every x frames the same allocation will be replicated.

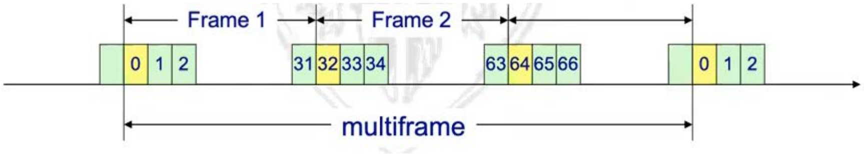


Figure 16: Transmission of signals over different frames

Now let's look at a typical data burst (which is the name of packets in GSM) It carries 114 bits of information, where the coded bits are the actual

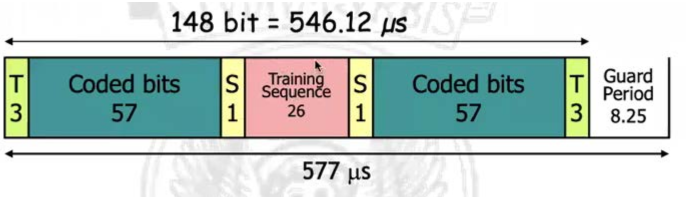


Figure 17: Data burst

data (traffic, but also control information, FEC bits, etc.), the T_3 are bits that signal the start and the end of data, S_1 are flags that define the start and the end of the training sequence, the training sequence, that is a predefined pattern of bits to perform equalization (the string of bits is already known, so it can be seen how it's distorted, and based on that perform equalization), and the guard period [explained in 2.2, radio interface: synchronization]. A channel that uses one slot per frame has a frame rate of 114bit (the length of data burst) / 4.6ms (the length of a time slot) = 24.7 kb/s; however, in full rate encoding the data is transmitted at a rate of 22.8 kb/s, meaning that 1.9 kb/s (24.7 - 22.8) is not used, and it's equivalent to one slot every 13 frames. For example, SACCH is transmitted on one slot every 26 frame, corresponding to a rate of 950bit/s.

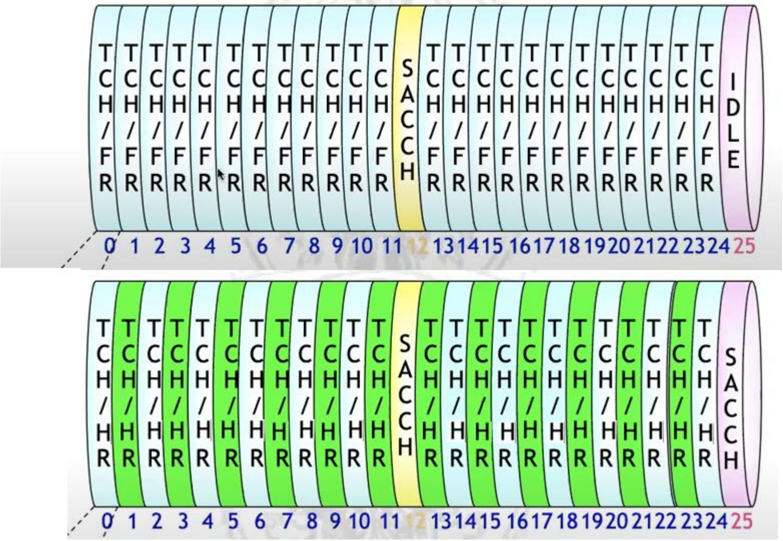
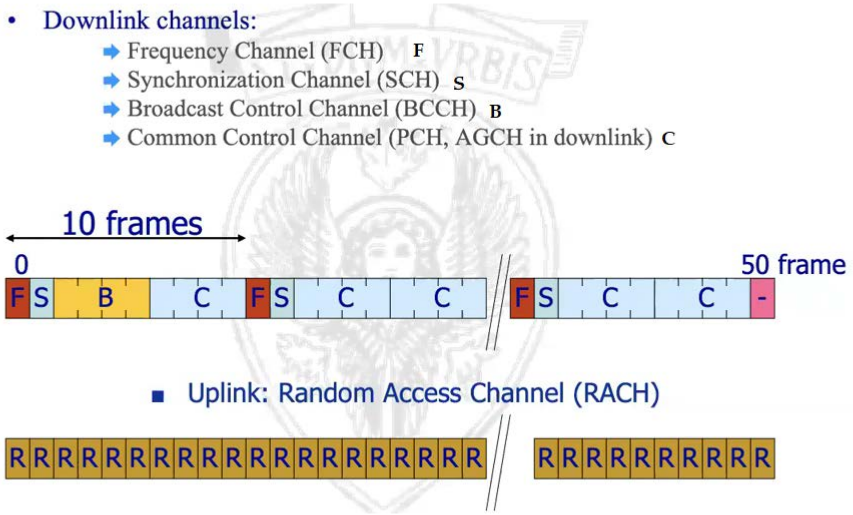


Figure 18: A multiframe architecture

In the image above, 26 frames can be seen (from 0 to 25) both in full rate and half rate style. The idle channel is the result of the 1.9kb/s unused transmission of SACCH, and it's used by mobile station to check if adjacent cells are transmitting BCCH. In half rate mode, there are two user (the blue one and the green one), and a SACCH per user (yellow and red).
Remember that FACCH channel opportunistically exploits TCH (traffic control channel, and in particular the "coded bits" part of the data burst) to send its own data, so in this model FACCH transmits too.
A given frequency called main carrier (or "C0"), where frequency hopping isn't performed, is used to transmit broadcast and common control channels (and in particular they're transmitted over the slot 0); the channels are multiplexed in 51 frames (235ms), meaning that a channel will transmit in the slot 0 every 51 frames.



- Downlink channels:
  - Frequency Channel (FCH) F
  - Synchronization Channel (SCH) S
  - Broadcast Control Channel (BCCH) B
  - Common Control Channel (PCH, AGCH in downlink) C

- Uplink: Random Access Channel (RACH)

Figure 19: Transmission over the main carrier frequency

Keep in mind that what you see in figure 18 is the slot 0 of every frame. Other slot of "C0" are used for 8 SDCCH (so 8 users), and each one uses 3 slots with a superframe of 26 ($8 \times 3 + 1idle + 1A$).

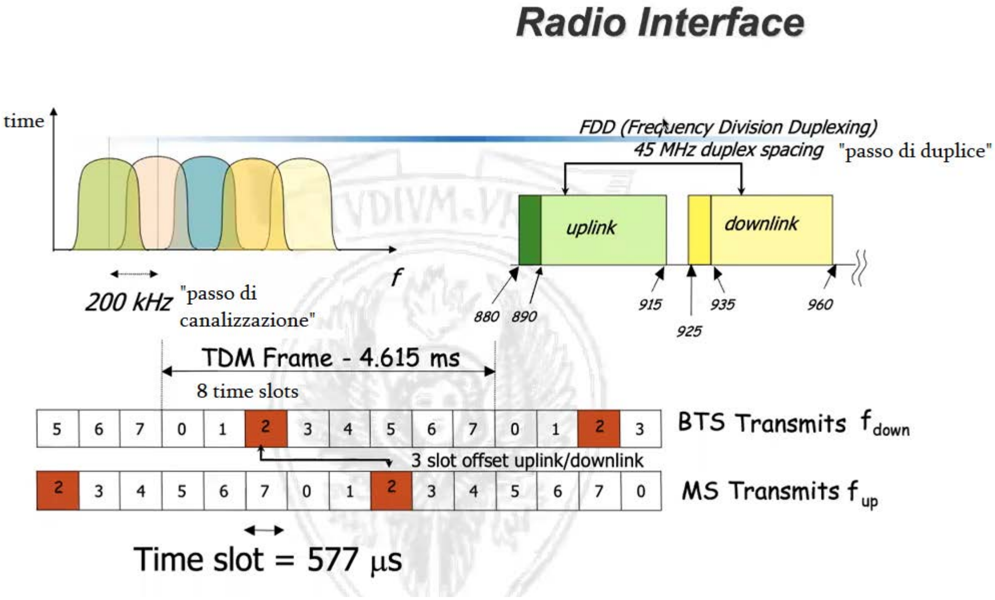
Data burst classification there different types of data burst: normal burst, used to transmit user's data over traffic channels, and access burst; the latter is used to transmit over the RACH channel and for the first time access on the system. It also have a longer time guard, both because in the first access MSs don't have an assigned one [explained in section 2.2 radio interface: synchronization], and also to avoid overlapping in RACH requests. Other bursts are: the frequency correction burst, which consist in a sequence that corresponds to a sinusoidal pattern and it's used one the FCCH (frequency correction channel), the synchronization burst, who carries information to synchronize to the SCH channel (so some bits to calculate the frame number, and the BSIC); notice that these information are critical, so there are a lot of bit of redundancy to be sure to apply the FEC correctly, and a longer training sequence in the center. Last, there's the dummy burst, and it's used because it's important that the user recognize the main carrier C0 (because broadcast and common control channels are transmitted here), so when there are no information to transmit over the main carrier, this dummy burst (consisting of random bits) is sent, so the user is still able to perform measurement of power [the dummy burst guarantees that there's more power on the BCCH than in other channels, but i really don't understand how; anyway, every timeslot in the C0 frequency needs to be used, and the dummy is used to fill them if no infos need to be transmitted].

6.2 Radio interface

The following image will show a little recap on the radio part of GSM; in this section we will go in-depth on some topics. Now let's go in-depth on some topics.

Power control has the goal of minimizing the power usage of battery powered devices and reduce the interference in the system. To do so, the BTS controls the power output of MSs; in sends commands to decrease/increase the power at steps of 2dB. The objective is to bring the power received the BTS just above what is needed for reception; the measurement on signal is sent downlink on the SACCH channel.

Synchronization covers different aspects: frequency synchronization, where the MS must retrieve precisely the frequency of the radio carrier, slot synchronization, where the MS must have infos about the current time slot, frame synchronization, MS must know the current frame number, and, optionally, base station synchronization, where BTS have a syncrhonous clock . With these information (frequency, time slot, frame number), MSs can known which logical channel is being transmitted (that is, if the physical-logical channel mapping is known). And how are these information obtained? In the case of frequency

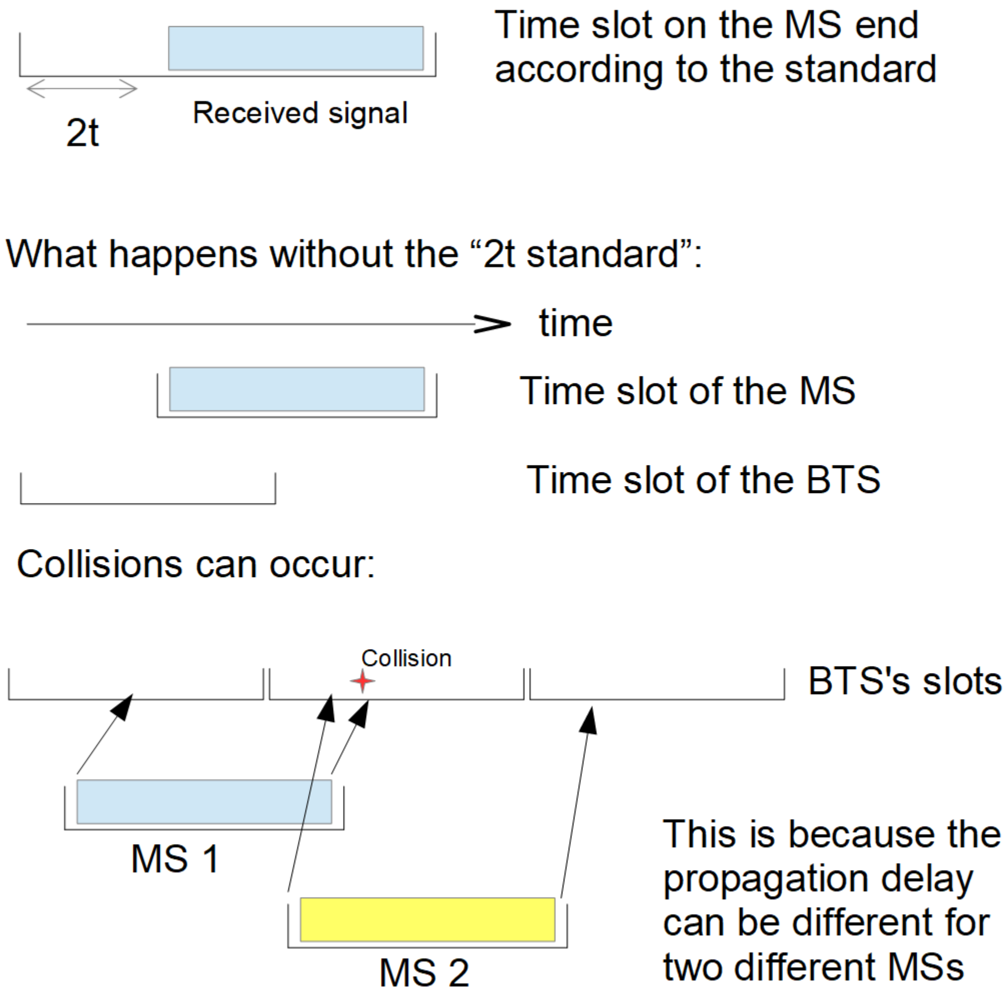


5 6 7 0 1 2 3 4 5 6 7 0 1 2 3

2 3 4 5 6 7 0 1 2 3 4 5 6 7 0

Figure 20: Recap of GSM radio's features

synchronization, the MS can obtain that by listening to the broadcast common control channel, where the BTS transmits, at regular intervals, a sequence of bits that corresponds to a sinusoidal pattern that is used to synchronize the receiver of the MS with the transmitter of the BTS. To synchronize time slots and frame, the BTS transmits information on the broadcast channel that allow the MSs to reconstruct them; the logical SCG channel is used to transmit the frame number. Slot synchronization isn't an easy task; a MS can't just assume that when it receive the signal, then that's the start of the frame slot. The amount of time that the signal takes to travel from the BTS to the MS must be taken into account (distance from BTS to MSs can be as long as 35km); the time slot is defined as: an initial time 2τ where the MS doesn't transmit, and the rest is as long as the received signal. In this way, overlapping between different time slots on the BTS end is avoided. Notice that τ is defined as the between the BTS and the MS times the speed of light (which transmission speed of the signal); it's multiplied by two to have the worst possible case covered. See the figure below for a visual explanation. In the worst possible case where the distance is at maximum, 68bits per second are wasted in time (233ms) where the MS does nothing; to avoid this, in the GSM system the BTS estimates this delay and send this information to the MS, who will transmit in advance; in this way, the guard time (our previous 2τ) is reduced to 33ms and only 9 bits are lost. These information are exchanged trough the SACCH channel, and they keep exchange them as the MS moves.



What happens without the “2t standard”:

Collisions can occur:

This is because the propagation delay can be different for two different MSs

Figure 21: About time slot synchronization

7 19/03

7.1 Mobility management

Radio coverage in a GSM system is guaranteed by BTSs, who transmit in a range (cell) and together can cover large areas; every BTS is capable to cover 35km, but depending on the terrain and the number of users, this range can be smaller. Cells can also be divided into smaller cells, and in theory the coverage can be seen as a circle or different geometric structure (triangles, hexagons, etc.). Frequency reuse is widely used (see section 1.2).
Mobility management guarantees the service in our GSM system, where MS moves around, but it's not possible to always know the exact location of the user (there would be too many control message to keep track of the position); tasks of mobility management differs based on the state of the serviced MS.

Idle Mobile Station if the MS is idle (not in conversation, so a circuit is not established to it), then it must be traced to be able to send paging messages and to be able to receive calls request (the MS wants to start a call), so procedure of cell selection (discover in which cell the MS is), cell re-selection (the MS has moved and is in another cell) and location updating (the MS has moved into a different location area) must be applied.
Now let's see in detail what happens when a MS is switched on (cell selection procedure):

1. The MS scan frequencies (it knows the frequency to scan because they're described by standards -primary GSM or DCS, European or American) to search for the main carrier (C0), where common and broadcast control messages are transmitted (remember that the main carrier can be easily identified because it transmits constantly and at higher power compared to other carries; also frequency hopping isn't applied).

2. The MS will synchronize itself frequency-wise with the BTS trough the Frequency Correction CHannel

3. The MS will synchronize itself time-wise (so it will receive the TDMA's frame number) trough the Synchronization CHannel, who also transmits the BISC (identification code of the BTS)

Now the MS is correctly synchronized with the BTS of the residing cell, and can receive the information transmitted over the BCCH channel (LAC, CGI, MCC, MNC). If the MS stays idle, these steps are repeated, so if a better signal is received from adjacent BTS, a cell re-selection will be performed.
What a MS must do after the cell selection (or re-selection but IF AND ONLY IF the new cell is in a different location area) procedure is the location updating (basically, the MS must inform the MSC/VLR that it's residing in (one of) his location area(s)). Now, the location updating is performed every time the MS moves from a LA to another, but there are different cases: the first time the MS

register on a LA, the MS moves from a LA to another but the MSC/VLR is the same, and when it moves from a LA to another that is managed by a different MSC/VLR [the professor doesn't cover all these cases, but they're covered more in depth in 7.3: GSM procedures]. In the first and third case, when the MS is switched on, it sends a registration request to the VLR, who in turn will inform the HLR; the HLR will ack and send some information about the user (such as the profile). The VLR will inform the user that its registration has been successful, and (if we're in the third case) the HLR will send a de-registration message to the old VLR.
Now that the MS is correctly synchronized and its location known, it can initiate calls and receive them, so it can be paged. The paging message is sent to notify the user of the incoming call (so he can decide whether to respond or not), and also to know the cell ID (CGI, all the acronyms can be consulted in section 1.5), so that a circuit can be established if the call is accepted. The MSC/VLR starts the paging procedure (how we've arrived at the MSC/VLR from a calling user is explained in section 1.4.2 where the calling procedure is shown) sending a ”page” message to the BSC of that location area, who in turn will tell the BTS to transmit over the PCH (Paging CHannel) the paging message (it contains the TMSI, so the MS with the corresponding one can answer).

Active Mobile Station If the MS is active (so is in a conversation), then the handover service must be guaranteed; it is the procedure bu which a MS in conversation changes the associated BTS. The key points of the handover procedure are that is the network that decides whether to perform an handover or not based on measurements made by BTSs and MSs, and that it must be fast and efficient; MSs measure the strength of signal from the channel BCCH from adjacent cells, and the strength and quality of the signal on the traffic chanel TCH, while BTSs monitor the strength and qualith of the signal on the MS → BTS link. But exactly when an handover must be performed? If it's when the signal it's very low, then the risk is of losing signal; if it's done in too much advantage, there will be too many handover requests. In the situation A, is possible that we continuously perform handovers, because the signal-to-noise

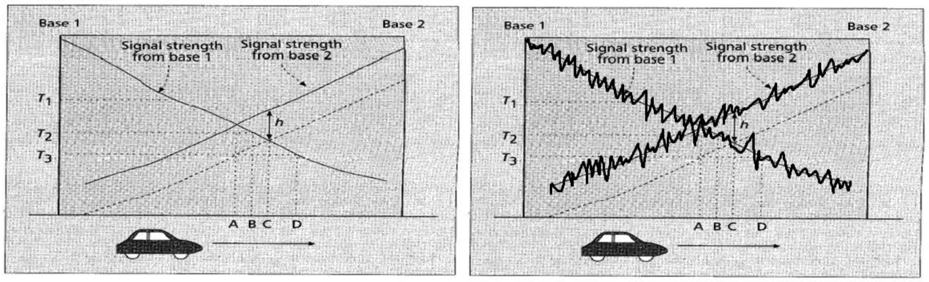


Figure 22: Various scenarios of when an handover is performed

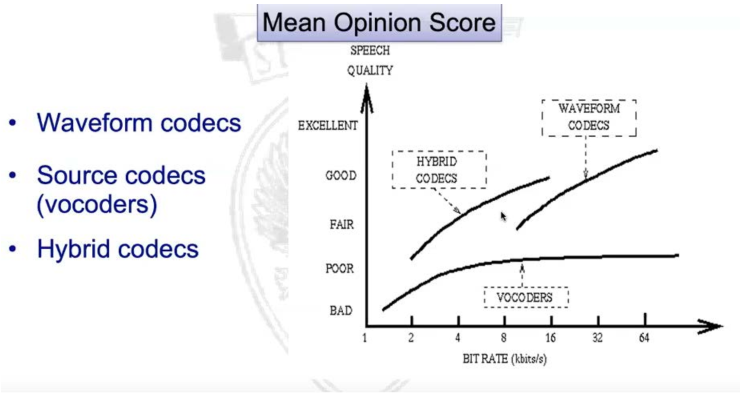
line isn't perfectly straight but has up and downs, so if a MS sit still in the central point, it will perform handovers when the signal goes down whether it's connected on Base 1 or Base 2. In situation B, the handover is performed only if the signal is under a certain threshold; this is a solid approach. In situation C, the handover is performed if the signal-to-noise ratio of the new BTS is higher than the one of the previous by at least h. [situation D is probably situation B but further].

Notice that the decision of when performing an handover is not standard and it's left to decide by cellular operators. It's also possible that the new cell can't accept the handover because it has no channels left; in this case it's better to not accept the call at all, since being able to maintain the call it's perceived more positively by users than be able to call a bit and then suddenly lose the signal. It's possible to reserve some channel to manage handovers, but it will decrease the general capacity of the system, so it's a process that requires an accurate estimation about the traffic dynamics. There are also two more possibilities: the handover can be queued, so the MS will remain connected to the old BTS until there's room in the new, or it can be done in a subrating scheme, where a channel in the new BTS is split two channels that browse at half the original speed.

There are two types of handover: hard, where the radio link is removed and re-established, and the MS is connected to a single BTS (used in GSM-2G), and soft, where the MS is connected to different BTSs (used in UMTS-3G).

7.2 Voice coding

To encode the voice and thus transform it in bits, various technique exists; their trade-off is between the quality perceived and the bitrate needed.



Mean Opinion Score

- Waveform codecs
- Source codecs (vocoders)
- Hybrid codecs

Figure 23: Various types of voice encoding

7.2.1 Waveform codex

These codex digitize the waveform of the ingress signal, so they create a description that can be used to reconstruct it with sufficient accuracy; no knowledge on how the signal was generated is needed, meaning that this codex can digitize any signal. The information needed are the signal bandwidth (in case of voice it's < 4KHz), and the maximum tolerable quantization noise. Thanks to the Nyquist theorem, knowing the bandwidth allows to know how frequently the signal must be sampled to capture all the information. There are two standard waveform codex that differs in speed of emission / transmission, the PCM and the ADPCM.

PCM Pulse Code Modulation is a standard that assume that the bandwidth B of the signal is 4kHz (the voice), and thus the sampling bandwidth is $B_c = 8$kHz (for the Nyquist theorem, $B_c = 2B$ is sufficient); 8 bits are used per sample, so the rate is 64kb/s ($8kHz \times 8b$). What does it mean to use 8 bits to encode the signal? It means that the values that the signal can assume are 2^8, so each value can be represented by 8 bits. The quantization error is fixed (step/2) and it represent the fact that the real value of the signal at a certain point has been encoded in a imprecise way (e.g, with only 8 bits maybe a point has been encoded as 11001101, but if more bits had been used, it would receive a different and more accurate string); in fact, the higher the quantity of bits, the smaller the step and the subsequent error (but the data rate will increase); when considering error, the important thing is the proportion between the value of an encoding and the applied error (the step is on the right-most bit, so mistaking 11110111 for 11110110 it's not so important, but mistaking 0000001 with 0000000 is, because in proportion the error is bigger). It can be proved that to minimize the error, 12 bits are sufficient $^{\text{obligatory "the proof is left to the reader"}}$, but this would increase the rate at 96kb/s; so the solution is to divide the signal in a non-uniform quantization: this quantization considers that bigger samples can stand bigger errors, while little samples must be encoded with little errors (because weak signal get more distorted), possibly having the same encoding as if 12 bits are used. Figure 24 illustrates this concept.

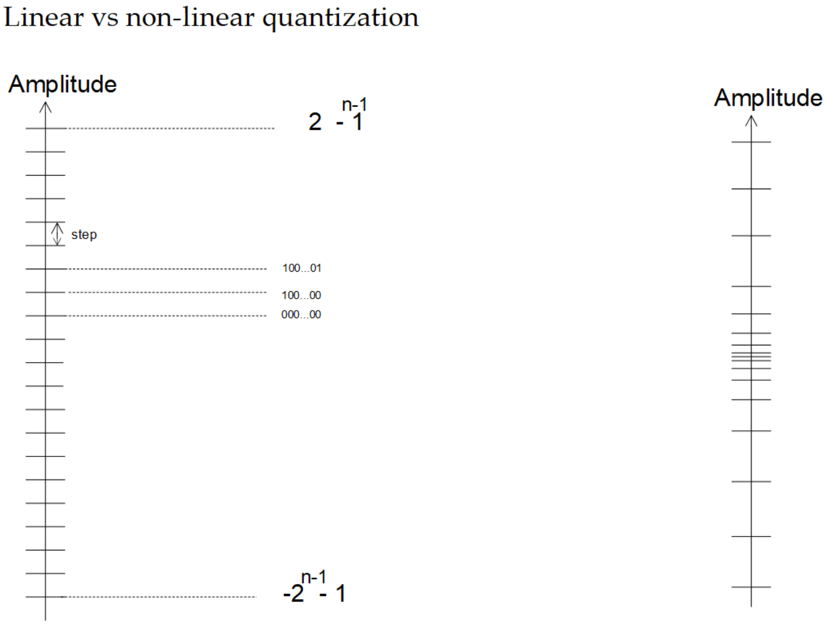


Figure 24: Types of quantization

The process so is to use a uniform quantizer of 12 bits and then compress it introducing a non uniform quantizer, reducing the bits to 8 per sample.

ADPCM Adaptive Differential Pulse Code Modulation exploits the idea of using a predictor to estimate the value of a sampling in a certain instant based on the history; in this way, with the predictor operating both on the transmitter and the receiver on the quantized signal (so both can apply it on the same signal), the only information needed to be transmitted are the difference between the predicted value and the real one, and this is encoded using a smaller number of bits, thus reducing the bit rate at 32kb/s (standard, can operate also at 40, 32, 24, 16 kb/s) . Making the predictor and the quantizer sensible to the strength of the signal (so the step can be adjusted as the amplitude of the signal) improves the performance, and it's why is called "adaptive".

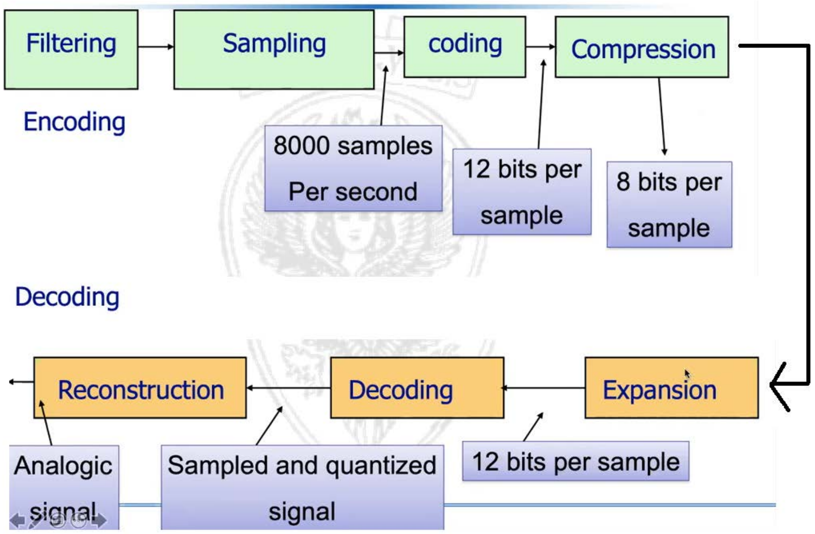


Figure 25: Phases of the encoding-decoding process

7.2.2 Source codex (vocoders)

These codex are made to encode the human voice specifically (so they aren't a general signal codex), having the idea to reconstruct the right signal using less information (because the general behaviour is known). They have high complexity, high delay, and are especially sensible to errors, noise and non-human sounds.
The goal is to transmit only some information to reconstruct the signal, and the process can be described in 3 phases:

1. The codex analyses a piece of vocal signal using an analysis filter, extracting the parameters of interest

2. These parameters (and a excitation signal [it's basically a random signal]) are fed into a filter synthesis that is exactly the same as the one on the receiving end; this filter will try to reconstruct the original signal using the parameters and the excitation signal

3. The difference between the constructed signal and the original one must are minimized using the Mean Squared Error criteria

This process is based on the assumption that the vocal apparatus can be described with a transfer function with the form:

$$H(z) = \frac{A(z)}{B(z)} \tag{3}$$

Where $A(z)$ and $B(z)$ are polynomials, and z a discrete step in time; the parameters we search for in the above process are some coefficient and/or degree of $A(z)$ and $B(z)$ polynomials.
The delay these coders experience are caused by the analysis filter, where a part of a voice signal is analyzed (typically between 5-30ms), and the optimization part, where the error is reduced; notice that the process is iterative.

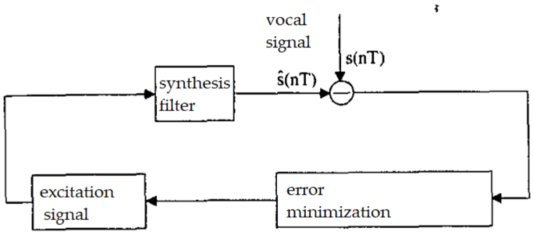


Figure 26: Vocoders process to create the parameters to dispatch

Linear vocoder - LPC the voice sample is approximated by a linear combination of past samples:

$$\sum_{k=1}^{p} a_k s(n-k) \tag{4}$$

The bit rate is very low: <2.4 kb/s.
The corresponding transfer function is:

$$H(z) = \frac{1}{\sum_{k=1}^{p} a_k s(n-k)} \tag{5}$$

Codebook Excited Linear Prediction - CELP to overcome the synthetic voice of vocoders, the excitation signal can be chosen from a variety in the CELP notebook (which is a set that contains various signals) to better approximate the given voice. The signal chosen is the one that produce the minimum error; at this point, the index of the chosen signal is sent (this practice is called vector quantization, because once the parameters are identified, is sufficient to sent an index to have them on the other end) with the parameters. This method allows bit rates as low as 4.8kb/s.

Hybrid codex they don't start to code the voice from a pseudo random sequence but from one that is representative of the real signal.

7.3 GSM Procedures

IMSI attach When a MS is switched on, it must perform:

1. Cell selection: it decides which BTS to tune to
2. Registration (location updating): the MS notifies the MSC about its presence in the location area

[Cell selection procedure is already explained in detail in 7.1 Mobility management: idle mobile station]

Registration there are two possible cases based on the received LAI:

1. It is the same that is stored in the SIM: then the IMSI attach procedure is invoked, and it will activate the IMSI in the MSC/VLR (it was previously stored [this is known because the LAI is stored in the SIM] but with a "detached" flag, indicating that the MS was turned off, and thus to not send paging messages to it)
2. LAI is different from the stored one, or no LAI is stored: the location update is invoked

Location updating is performed when: the LAI recevied by a MS isn't the same as stored in the SIM (or when no LAi is stored at all), when the location area is changed by the MS (roaming), and periodically every 30 minutes; in this last case if the update is not perceived, the user will be flagged as detached. There are two cases [there are more but she didn't explain them]:

1. The new LA is under the same MSC/VLR jurisdiction: in this case the VLR updates the LAC associated with the IMSI, a new TMSI is generated, and the MS will be authenticated again; the HLR isn't notified of these changes. The location update request includes: the request, old TMSI and old LAI; the response gives a new TMSI and ACK. As for the logical channel, first of all the MS will ask for a channel on the RACH, then the BTS will respond and assign a channel on the AGCH; all the other messages (the actual location update request and subsequent messages) will be exchanged on the assigned SDCCH.

2. The new LA is under the jurisdiction of a different MSC/VLR: after receiving the location update request, the VLR must exchange some data with the old VLR: he sends the TMSI and obtains the IMSI; afterward, the HLR is notified that the MS has changed location (and he will memorize the new VLR associated with the MS). The HLR orders the old VLR to delete the data of the user, and the MS is notified that the location update was a success.

Call set-up: PSTN to MS

1. A PTSN/ISDN digit the MSISDN number of the user it wants to call

2. The PTSN network route the call to the GMSC of the PLMN of the user using the NDC (national destination code)

3. The GMSC receive the call set up request on the CCSS7 network (it's the signaling system used to communicate with other network elements) who contains the MSISDN number of the called user

4. The GMSC identifies the HLR that contains the data of the user

5. The GMSC request routing information from the HLR

6. The HLR identifies the VLR associated with the MS

7. The HLR requests a roaming number from the MS' associated MSC/VLR

8. The MSC/VLR allocates a MSRN (roaming number) to be used for the call

9. The MSC/VLR starts the paging procedure: it identifies the LA thanks to the IMSI, and sends a paging command to every BSC in the network

10. The BSC make the BTSs send the paging message (containing the TMSI) over the paging channel PCH
11. The MS responds on the RACH channel with a request of a dedicated channel (SDCCH)
12. The MSC initiate the ciphering and authentication procedures
13. A traffic channel TCH is allocated for the call
14. The MSC/VLR notifies the caller that the called's MS is ringing
15. The called accepts the call
16. The connection between the two is established

Call set-up: MS originated call

1. The MS dial the number of the desired user
2. The MSC analyses the caller data and can either authorize or refuse the call; if is authorized, then the call routing procedure is started
3. If the called user is in the same GSM network, a send routing info procedure is started to obtain the MSRN (and from now on, the same steps as PTSN to MS applies, starting from 4 [but the protagonist won't be the GMSC, the originating MSC will handle these requests])
4. If the user is on another GSM network, then the call is routed to the GMSC

Handover this procedure is started by the network based on measurements provided by the MS and BTSs; when the MS connects to a cell, the BSC will send to it a list of alternative BCCH frequencies (in total 6 and they are the one from adjacent cells), whose signal must be monitored by the MS [this is mentioned in figure 13: parameters transmitted in uplink, and in this image all the parameters transmitted by the MS and used in the handover procedure are present]. These measurements are sent to the BSC on the SACCH every 480ms, and an handover may be started because of them. The handover procedure has a set of requirements, such as the criteria to be used to decide whether an handover is necessary [discussed in 7.1 mobility management: active mobile station], and the procedures to commute the communication from one channel to another; this should happen fast (under 100 ms) and should not be perceived by the user; the handover parameters provided by a BTS are: the signal strength from the MS on TCH, the quality of signal on the TCH from the MS, and the distance of the MS (timing advance). Now let's see the decisions that can trigger an handover:

1. The quality and/or the strength of the signal are below a threshold (low quality transmission)

2. The distance between the BTS and the MS is above a given threshold (timing advance)

3. Excessive traffic load on a cell

4. Control and maintenance

There are 4 types of handover:

1. In the same cell and the same BSC: this simple handover is decided by the BSC only, and happens when the TCH is low quality but still high strength (or when the TCH isn't available anymore for some reason), and no adjacent BTS can provide better quality; in this case both the frequency and the channel are changed.

2. In a different cell but the same BSC: this handover is decided by the BSC only (but the MSC/VLR is notified), who will search for the best BTS and TCH based on the measurements of both MS and BTSs. The BSC connects to the new BTS and allocate a channel for the MS, and order the MS to connect on the new frequency and channel (these communications occur on the FACCH channel, who subtract resources from the TCH). The MS then move on the new TCH and can continue the communication.

3. On another cell and another BSC (same MSC/VLR): the initial BSC decides to do an handover and based on the measurements of both MS and BTS, locates a new BTS and TCH; since the new BTS is under the control of another BSC, the initial BSC relay the handover to the MSC/VLR, who establish a connection to the new BSC. Similarly as before, the new BSC connects to the BTS and reserve the chosen TCH. On the FACCH, the MS is told the new TCH. Once the MS is tuned on the new TCH, the MSC/VLR route the call onto the new BSC, and the communication can continue; in the end, the MSC/VLR releases the resources used on the initial BSC. If the LA changed, a location update must be started by the MS.

4. On a different MSC: the initial BSC decides to do an handover over the cell of another MSC/VLR; he will send this request to the initial MSC/VLR, who will in turn send this request to the new MSC/VLR. The new MSC/VLR allocates an handover number and sends it to the initial MSC/VLR, so the latter can route the call towards it. The new MSC/VLR orders prepares a connection to the BSC and the BTS, and order them to allocate the new TCH; the new frequency and TCH are sent to the MS using the FACCH by the initial MSC/VLR. The MS tunes on the new TCH, and the old resources are released. Now the MS will certainly perform a location updating.

8 26/03

8.1 GPRS

GPRS also called 2.5G and 2G+ introduces packet switching adding network elements (that are IP routers associated with an IP) that forward packet to external IP networks such as internet; at the BSC level, a PSU (packet switching unit) is implemented to route the packages not to the MSC/VLR but to other IP network elements, and also to deal with dynamic resource management. Common Control channels are added as well as dedicated ones.

8.1.1 Other GPRS innovations

1. Enhanced full-reate codec (ACELP): it's a linear predictive speech coding (LPC) based on CELP with an algebraic structure [see 7.2.2 for speech coding techniques] that has a better performance and compression

2. Adaptive multi rate (AMR): a new codec that can adapt to the available channel and changes rate depending on the channel propagation conditions; this is possible because AMR decides the type of encoding depending on the available resources.

3. Tandem Free Operation (TFO): limits the use of transcoding (the act of converting a coded info to another different coded format), who degrades the voice quality, in MS to MS communication. [insights on TFO are included in the last section]

4. Enhanced data rate by an improvement of the physical layer and the allocation of multiple slots for the same MS.

5. Localization services: the position of the MS can be obtained triangulating BTSs, and this is accurate around 100m (not goo as satellites, but good enough to provide some services).

6. Number portability: a user can now change operator maintaining the same number

7. Change of paradigm: now the SIM can ask the MS to perform operations, such as: set up a call to a number stored in the SIM, send telephone number to the SIM to decide if a call can be made, pass information, execute commands, launch a browser that redirects to a web address, etc.

8. introduce CAMEL: an environment for applications & services support, customization and customization.

9. cordless telephony system: the MS can be used as a cordless phone connecting to the Home Base Station and from there connect to the cellular operator network [it seems it has not been implemented; what a shame]

8.2 EDGE

Edge is a radio access technology that improved the data rate without changing the cellular mobile system architecture; it has a data rate of 59 kb/s (vs 22.8 of GSM), and the data burst can carry more data bits (384 vs 114 of GSM). This is possibly because while the symbol rate (it's the rate at which different symbols are transmitted per unit of time) is the same, each symbol can carry 3 bits instead of 1 (this is called phase modulation, and is also used in Wi-Fi and satellite television). EDGE also introduces adaptive modulation: based on the channel condition, the data rate can be modified (in case the channel has BER, a high data rate can be switched to a low, more robust data rate); this is also implemented in later generation cellular systems.
EDGE is also called 2.75G.

8.3 3G

This technology is born with the objective of overcoming the limit of bandwidth and data rate of 2G/2G+ generation, as well as supporting more services (like different classes of traffic), and integrating communication with satellites.
To improve the data rate, the capacity is increased (creating smaller cells, macro micro and pico, so that the same channel serve less people but grant more bandwidth) and also the total bandwidth is different with 155mhz for terrestrial communication and 75mhz for satellites; aggressive modulation (modify the data rate based on the channel condition) is another approach applied to increase data rate.
The MAC is CDMA, used only in 3G, and there's a clear division between the radio access network and the core network, enabling support for hybrid networks; both data and voice over an ip network. Since CDMA allows multiple users to communicate using the same physical resource, the MS can be attached to more than one BTS, so the same data stream can be sent over different physical channels; also, the handover can be soft.
CMDA works in this way: each user has a unique code called chipping sequence, and communicates over the same frequencies as the others; since the chipping sequences are orthogonal with respect to one another, there's minimal interference (if the users are perfectly synchronized, then there's no interference at all; this is impossible due to the complexity of the network: a little delay somewhere and the users are no more in perfect synchro). To encode data, they are multiplied with the chipping sequence, and to decode them, the inner product between the decoded data and the chipping sequence is done.

8.4 4G

Goals: lower mobile energy consumption , higher data rate, uniformity of service (even at the edge of cell), reduced delays, simplified network architecture, spectral efficiency (the information rate that can be transmitted over a given bandwidth, formed by technologies and standards), and more.

8.5 LTE

The goals are: to operate in a wide range of frequencies and sizes of spectrum allocations, fast connection time, support mobility up to 500km/h with cell radius varying between 5-100km, flexible inter-operation with other radio access technologies (such as 3G and Wi-Fi; this is useful so the service is guaranteed in the migration phase), low complexity and power consumption, and cost effective deployment.
The MAC is 0FDMA in downlink and SC-FDMA in uplink (multicarrier, and thanks to 0FDMA, intersymbol interference is avoided); multi antenna technology (Multiple Input Multiple Output, MIMO, is a method for multiplying the capacity of a radio link using multiple transmission and receiving antennas to exploit multipath propagation) is also used. A very aggressive resource dynamic allocation is used, and has been improved since 3G. The concept of bearer is introduced: the set up of a tunnel or virtual circuit IP where data and voice flows that can connect the MS to external network; a MS can have different bearers active at the same time with different QoS demands.

9 23/04

[for this section i highly recommend the slides of the autonomous networking course @ https://twiki.di.uniroma1.it/twiki/view/AN/WebHome]

9.1 Ad hoc wireless networking

An ad hoc wireless network is a wireless multi-hop infrastructure-less network whose devices act as a source / destination of messages and act as relay for packets; it self organizes, self configure and self maintain. It has various application scenarios: disaster recovery applications, military networks, personal area networking, wireless sensor network and IoT, inter-vehicular communication, etc.; it's deployed principally when an infrastructure cannot be used.
Features:

1. Highly dynamic
2. Simple protocols with low energy consumption and low overhead (ideally)
3. Number of nodes depends on the application (typically 10-100)
4. Traffic can be low or high

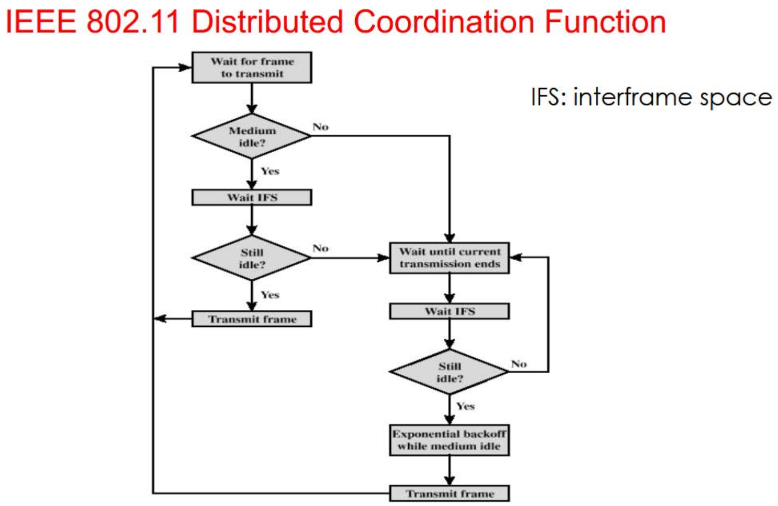
9.1.1 Wireless sensor network (WSN)

WSN networks have a mostly static topology (e.g. networks to monitor the environment), but it's also possible that the nodes move; ideally, the number of nodes must not degrade the performance. Nodes are typically sensor with limited battery capacity and low processing power, so protocols must be simple, low energy consuming, and light, since the memory is also limited; the traffic flows from sensors to one ore more sink(s) (the data gathered by sensors are analyzed/processed by sinks).

9.2 MAC in Ad hoc networks

9.2.1 802.11

Is a protocol based on CSMA/CA (collision avoidance, since collisions can't be detected in wireless networks). It uses ACKs and the NAV to deal with the hidden/exposed terminal problem. For a complete explanation of this protocol, click here to see the slide of professor Maselli on this topic. [i can't explain it better or faster]



IEEE 802.11 Distributed Coordination Function

IFS: interframe space

Wait for frame to transmit
Medium idle?
No
Yes
Wait IFS
Still idle?
No
Yes
Transmit frame
Wait until current transmission ends
Wait IFS
Still idle?
No
Yes
Exponential backoff while medium idle
Transmit frame

Figure 27: How 802.11 works

9.2.2 Other approaches

A TDMA approach would require a synchronized environment; it's not a good idea: nodes can be a lot, can move, and the overhead would significantly increase. Hybrid solutions have been proposed, such as: ATLAS: Adaptive Topology- and Load-Aware Scheduling.

9.3 Routing in non ad hoc networks

Let's see the classical routing algorithms.
Two approaches for intra-AS (autonomous systems, a set of routers whose prefixes and routing policies are under common administrative control, e.g. a uni-

versity's network) routing: link state, where each switching node (notice that in ad hoc wireless networks, every node is a switching node [because every node can route packets]) keeps information on the topology and then applies shortest path algorithms (e.g. Dijkstra) to construct its own routing table, and distance vector, where switching nodes exchange information (such as the routing table and hop count) and choose the best route based only on the distance; this approach uses the Bellman-Ford algorithm to determinate the best route. Now, ad hoc wireless networks have different needs from classical networks, a routing protocol should have: minimal control overhead, low processing overhead, no loops, multi-hop path capability, dynamic topology maintenance, and it must be self-starting. In Bellman-Ford, switching nodes only exchange information about the length of paths, without further context; for this reason is not a good idea in networks where the topology is dynamic.

9.4 Routing in ad hoc networks: proactive protocols

These protocols are called proactive because they create and maintain routing tables before it's necessary to use them; they are updated upon a change and also periodically (so they maintain a consistent view of the network). High overhead and route convergence time.
Also, since traffic is generally low in wireless network, is not worth to update routes constantly.

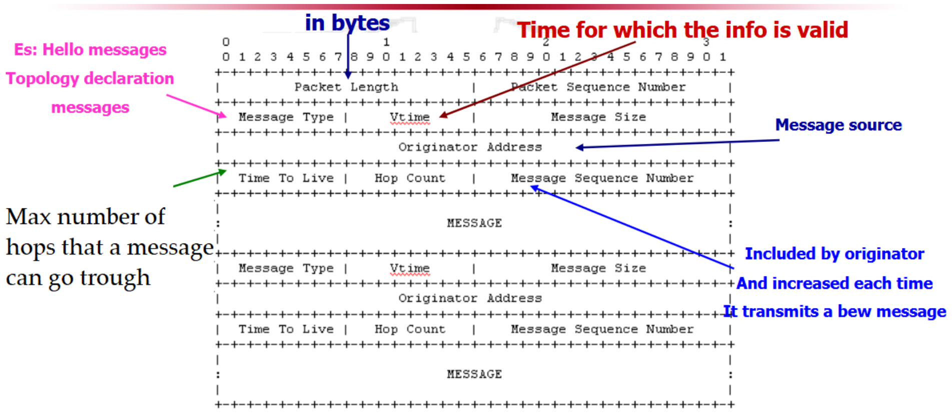
9.4.1 Dynamic Destination-Sequenced Distance-Vector (DSDV)

In DSDV routing each entry in the routing table is tagged with a sequence number; periodically, destination nodes send updated entries with a new sequence number that describes the cost (number of hops) and how recent is the route; these updates get propagated by the other nodes. Notice that the sequence number is decided by the transmitter, and is always preferable to chose a route with the newest sequence number. The data broadcasted includes: the destination address, the number of hops to reach the destination, the sequence number (the one created by the destination; it is not changed by relay nodes); in the header is stored the address of the sender (HW/NET). There are two type of updates: full dump or incremental; this is done to decrease bandwidth consumption.
A broken link has infinite cost, and can be detected in the following ways: an exceeded threshold of the MAC in re-transmitting a frame, no HELLO messages has been received in a while, or the periodic update hasn't reached its neighbors. Updates signaling broken links (or ex broken links that now work) are transmitted immediately. To decrease overhead, new route are transmitted after a time that depends on the history.

Correctness of DSDV a distance vector algorithm is correct if no loop can occur. Assume that $G(X)$ denotes the route graph from sources to a destination x and has no loops (assumption); a change occurs in i in two cases: the link from i to its parent $P(i)$ broke and must be put to inf (a loop can never occur in this setting), or if i receives from a neighbor k a new route with a higher sequence number or equal sequence number but a smaller number of hops. If the new route has a higher sequence number and is thus chosen, it can't contain a loop, because if it does, the (new route) sequence number would be smaller than the one already stored in i [why? i don't know and the professor's explanation is a riddle]. Meanwhile if the new route has a smaller number of hops, it can't contain loop because distance vector algorithms always maintain loop-free paths (and this is true with the assumption that link weights are static or decreasing).

9.4.2 Optimized Link State Routing (OLSR)

Is a link-state protocol designed for MANET networks; the basic idea is to reduce the overhead of flooding by identifying a set of nodes (multi-point relays) in charge of forwarding the information during the flooding process. A node Y is called multi-point relay if at least one of its one hop neighbors (e.g. X) has chosen him to relay all the valid (not duplicate nor expired) broadcast information it sends to him; nodes that receive information from X but aren't multi-point relays simply store the information without re-transmitting it. The set of multi-point relays chosen by a node X is written as $MPR(X)$, and X is called the selector of these multi-point relays; to ensure that each route to a destination can be found, is sufficient to declare the links between MPRs and their selectors. The time between updates can be tuned to be reactive to topological changes, and it doesn't require a reliable connection: the messages are not required to be sent in sequence, and they can be reconstructed because they use sequence number; the needed information are periodically transmitted in UDP datagrams. This protocols supports both multicast and sleep mode operation, thus supporting most of MANET related issues.

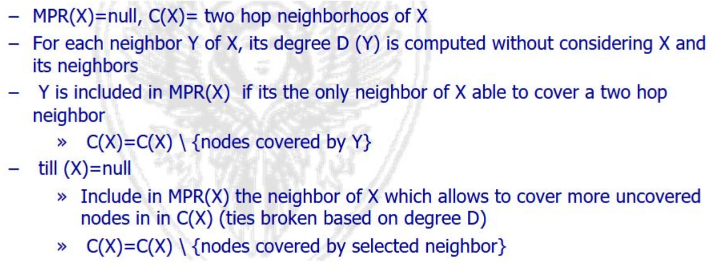


```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Packet Length         |    Packet Sequence Number     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Message Type |     Vtime     |         Message Size          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Originator Address                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Time To Live |   Hop Count   |    Message Sequence Number    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
:                            MESSAGE                            :
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Message Type |     Vtime     |         Message Size          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Originator Address                       |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Time To Live |   Hop Count   |    Message Sequence Number    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
:                            MESSAGE                            :
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 28: Packet format of OLSR

Hello messages are used to verify that links are up and running (link sensing) and also to to exchange neighborhood information to compute two hop neighborhoods (the usefulness of two hop neighborhoods is seen in the next paragraph).

MPR selection each node Y in a MPR of a node X is selected to cover the 2-hop neighborhood of node X; a greedy algorithm can be used.



- MPR(X)=null, C(X)= two hop neighborhoos of X
- For each neighbor Y of X, its degree D (Y) is computed without considering X and its neighbors
- Y is included in MPR(X) if its the only neighbor of X able to cover a two hop neighbor
  - » C(X)=C(X) \ {nodes covered by Y}
- till (X)=null
  - » Include in MPR(X) the neighbor of X which allows to cover more uncovered nodes in in C(X) (ties broken based on degree D)
  - » C(X)=C(X) \ {nodes covered by selected neighbor}

Figure 29: A greedy algorithm to compute MRPs

Control and data packet flow upon receiving a packet, if the receiver is a relay node, it will: reduce the message TTL, increase the number of hops by 1, and broadcast it on all node interfaces of his selector; this is done assuming the message is valid: non-zero TTL, and not a duplicate (multi-point relays keep a set of duplicate messages for this purpose). Information about topology are as limited as possible, and are exchanged to all the nodes; each node can then run locally a shortest path algorithm to determinate the route to the destination.

9.5 Routing in ad hoc networks: reactive protocols

These protocols discover routes on demand when they must be used; the latency in acquiring routes is increased but working routes can be temporally stored, and it scales better producing less overhead. The major drawbacks of this approach are the longer delays, the energy consumption operating in promiscuous mode, and the fact that route discovery and maintenance is very sensible to node mobility.

9.5.1 Ad-hoc On-demand Distance Vector (AODV)

In this protocol, a node does not have to maintain or discover routing to a destination until he's on the path of a message or need to send one; it's similar to DSDV, but there are no periodic updates, only on-demand. When a source node desires to send a message to some destination node and does not already have a valid route to that destination (cache memory is used), it initiates a

path discovery process to locate the other node; source node broadcasts a route request (RREQ) packet to its neighbors, who then forward the request to their neighbors, and so on, until either the destination or an intermediate node with a “fresh enough” route to the destination is located. During the process of forwarding the RREQ, intermediate nodes record in their route tables the address of the neighbor from which the first copy of the broadcast packet is received, thereby establishing a reverse path; once the RREQ reaches the destination or an intermediate node with a fresh enough route, the destination intermediate node responds by unicasting a route reply (RREP) packet back to the neighbor from which it first received the RREQ; as the RREP is routed back along the reverse path, nodes along this path set up forward route entries in their route tables which point to the node from which the RREP came. If a node receives further RREP for the same request, it will keep the one with a higher sequence number, or the one with the less hop count.
If a link goes down, the one-hop neighbors send a Route Error (RERR) message upstream to its own neighbors; neighbors upon receiving RERR, will adjust their route excluding broken links, and will send RERR to the upstream neighbors.

Optimizations - Expanding Ring Search prevents flooding of the network during route discovery, controlling TTL of RREQ to search incrementally larger areas of network. It produces less overhead when successful but produces long delay if the route is not immediately discovered.

Optimizations - Local Repair Repairs breaks (in active routes) locally instead of notifying the source, using small TTL (the destination hasn’t moved far probably); if the first attempt to repair is unsuccessful, RERR is sent to the source as normally. If successful, less overhead is produced, and packet delay and loss are decreased; if unsuccessful, it creates a longer delays and packet loss.

9.6 Dynamic Source Routing (DSR)

This protocols use source routing: instead of saving the next hop and the cost to the destination into routing tables, the entire route is inside the header of the packet (the route is built typically by the source node). To discover a route to a destination, the source node send a RREQ packet that contains: a pair ”initiator, request id” that uniquely identify the request, the target node, and a route record where the traversed route is accumulated. Nodes that receive the RREQ and are not the destination rebroadcast it, appending their address into the route record; once the packet has reached the destination, this latter unicasts a RREP packet back to the initiator, who can now use the route to send data.
Notice that when a node receive a RREQ and the pair ”initiator, request id” has been seen recently, then it is discarded. To avoid loops, if a node sees that his address is already appended in the route, then it will discards the packet.
As for broken links, a RRER packet is sent containing the two ends of the broken

links, and route containing that link are invalidated; if an invalidated route is needed, then a route discovery is initiated.

Optimizations

1. Path shortening: if a packet receive a RREQ and knows a shorter route, then it will signal it sending an unsolicited RREP
2. Exponential back-off when the network is disconnected to reduce the number of RREQs
3. Promiscuous mode: a node discovers new routes because neighbors nodes transmit them

9.7 AODV vs DSR

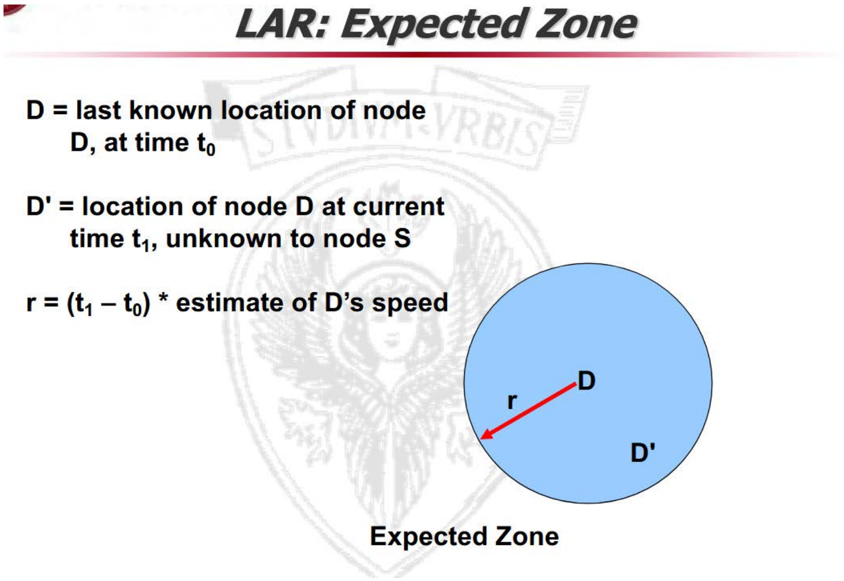
Both are reactive protocols, but: AODV uses routing tables and next-hop entries, while DSR uses route cache and source routing; DSR cache entries doesn't have a lifetime, while AODV entries does.

9.8 Geographically enabled routing

If the position of the nodes in known, it can be exploited to find routes. This requires additional hardware, and may be a waste if the network doesn't need localization for its task.

9.8.1 Location Aided Routing (LAR)

Exploits location to limit the flood of RREQ requests. Two zones: expected zone: where the destination is expected to be, and request zone, an area that contains both the expected zone as well as the sender node.
LARs protocols are greedy forwarding protocols that exploits the geographical position.



LAR: Expected Zone

D = last known location of node D, at time t_0

D' = location of node D at current time t_1, unknown to node S

r = $(t_1 - t_0)$ * estimate of D's speed

Expected Zone

Figure 30: The expected zone

9.8.2 LAR - 1

In this version the request zone is the smallest rectangle that fits both the sender and the expected zone; only nodes in this zone can forward route requests. Clearly, every node must know its own location to determine whether it is inside the zone or not. The request zone is stated in the RREQ.
If route discovery using the smaller request zone fails to find a route, the sender initiates another route discovery (after a timeout) using a larger request zone (up to the entire network). The rest of the route discovery is similar to DSR. Initially, location information for node X becomes known to Y only during a route discovery; this location information is used for future route discoveries. Destination may also proactively share its location, but it would increase the overhead (solved later in DREAM).

9.8.3 LAR - 2

Each node relays RREQ if it is closer to the destination than the source (greedy forwarding).

9.9 Distance routing effect algorithm for mobility (DREAM)

It's a proactive protocol; the initial node can determinate the route to the destination looking at its location table. It does what is called directional routing: based on the current time, the sender determinate the area where the destination can be found (its direction); it then transmit the packet to all the neighbors in that direction, and it continues until the destination is reached.
The locations of node must be updated, deciding how often and making it efficient; the more a node move, most often it has to send updates. Also, closer nodes appear to move faster, so neighbors must exchange location messages often. Location messages include an age used to determinate how far that packet must travel.
The weakness of this protocol are the fact that is flooding (it's directional, but still flooding), and that it's not suitable for network of nodes that move continuously (too much overhead).

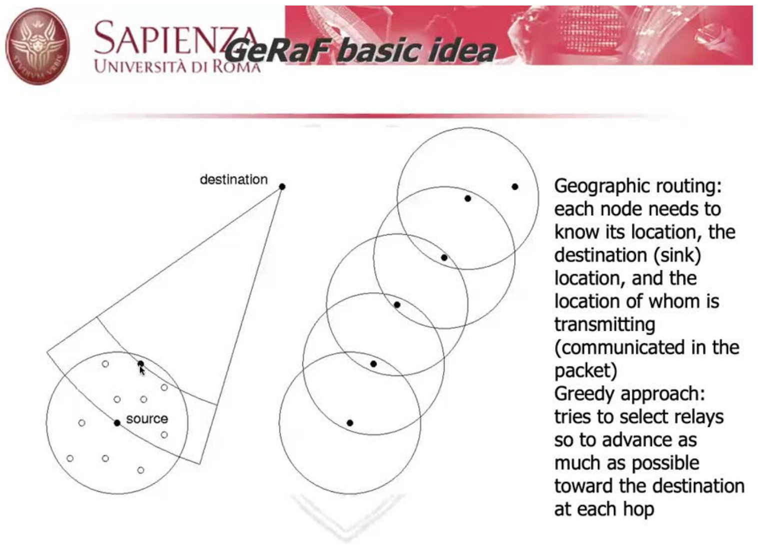
10 05/05

10.1 Practical geographical routing in IoT devices

These protocols keep in mind that IoT devices are low-powered and need to follow a duty cycle (the duty cycle describes the rate between the on time and the on + off time of the transceiver).

10.1.1 GeRaF

It integrates the geographical routing, MAC and sleep/awake schedule.



GeRaF basic idea

Geographic routing: each node needs to know its location, the destination (sink) location, and the location of whom is transmitting (communicated in the packet)
Greedy approach: tries to select relays so to advance as much as possible toward the destination at each hop

Figure 31: GeRaF basics

Notice that in the above figure, the white nodes are asleep and black ones are awake.
This protocol is cross layer, meaning that information from different layer of the stack are used simultaneously; in this case, MAC and routing are done together: RTS messages are also used to locate active neighbor, and a policy is applied to decide the best relay for the message (e.g. the node that is the nearest to the destination). This works because the position of both the nodes and the sink is known; low overhead, no routing tables.

Next hop selection the node that wants to transmit to the sink send an RTS to all the nodes within tx range; the tx range is subdivided in (typically) four or more regions and the awake nodes that are in the best region (e.g. closer to the destination) can respond with a CTS. Collisions can occur, and it will be the MAC the one that sorts them out. If no one answers in the selected region, then ”worst” regions are polled. If there’s no collision and a node responds, the data will be sent to him (along with the various ACK etc. of the MAC protocol).

Dealing with collisions when a collision occurs, the sender sends a collision packet to every affected node. When a node receives the collision packet and its previous CTS resulted in a collision, then it will resend the CTS with a probability p. At this point the cycle repeats: if only one responded to the sender, then it will transmit the data; if no one responded, then they will be asked to re-transmit (with the same probability p), and if a collision occurs, then the interested nodes will have to send the CTS with the same probability p (this step is different from the previous because only the nodes that experienced a

collision will try to re-transmit).
In IoT systems, introducing a little jitter (a variation from the mean of delays in the network) is enough to avoid collisions, because it will de-synchronize the communications; practical implementations of GeRaF exploits this method making best relays respond first.

Can the destination be always reached? If no area responds, a backoff time will be waited and then the procedure is redone; but this protocols presents the same problems as other geographical routing protocols: the nearest node to the sink may be unable to actually reach the sink. In this case a backtracking mechanism has to be added to try another route: if a node can't find a new relay, then it will decrease its probability to be chosen as the next hop (decreasing the duty cycle or the probability to transmit a CTS) until dead ends are practically blacklisted. Notice that the dead end problem is particularly severe in networks with a low densiy of nodes; increasing the number of nodes will instead decrease the probability of stumbling into dead ends since the connectivity will be greater. With only one hundred nodes, less than 40% of packets will arrive to the sink successfully. Another solution would be create a topological map of the graph and with some assumption find the best route; the problem is that its practical implementation doesn't provide good results, since the initial assumptions (such as "if a node is within Tx range of another one, then there's a link between them) aren't precise in practice.

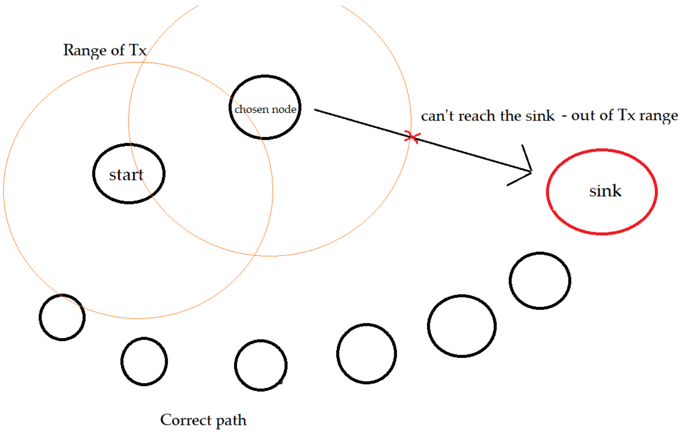
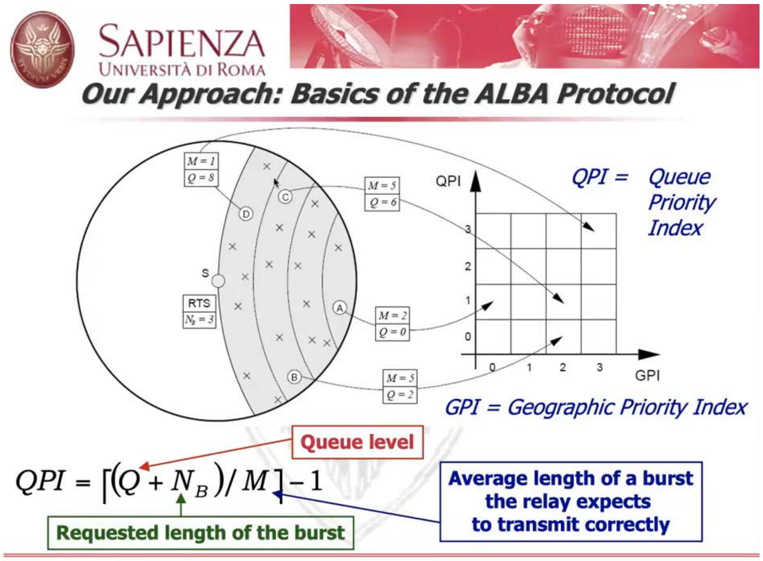


Figure 32: A situation where the wrong node is chosen to send infos to the sink

10.2 Adaptive Long-Balancing Algorithm (ALBA)

Is a cross-layer integrated protocol stack for medium-large sensor networks; as GeRaF, it is a geographical routing protocol and is cross layered, but implements a coloring node system to deal with dead ends. When ALBA selects the best relay, it will send to it a burst of packets (instead of one); the length of the burst is adapted depending on the BER of the selected link. It also tries to load-balance the network choosing the next hop based on its queue (influences the QPI, queue priority index), its proximity to the sink (represents the GPI, geographical priority index), the history, and the number of packets that the sender must transmit. Notice that the lower the indexes, the best are the candidates, and the queue is referred to the number of packets that are (already) waiting to be sent from that relay. In the QPI equation, if $(Q + N_b)/M$ is equal to 1, then it means that the packets in queue and the extra one the node want to relay to that node can be correctly sent with a single burst, meaning that no extra time for MAC is needed. To chose a relay, first the nodes with the lower QPI are selected, and then the one with the lower GPI is chosen; this helps in load balancing because nodes with a low queue are preferred. Notice that to save energy, nodes employ a duty-cycle routine, and thus only active nodes will contend to be the next relay; awaking nodes can participate in the QPI selection, but not on the GPI one (so the process is not slowed down). This approach doesn't solve the dead end problem.



SAPIENZA
UNIVERSITÀ DI ROMA

Our Approach: Basics of the ALBA Protocol

QPI = Queue Priority Index

GPI = Geographic Priority Index

$$QPI = \lceil (Q + N_B)/M \rceil - 1$$

Queue level

Requested length of the burst

Average length of a burst the relay expects to transmit correctly

Figure 33: Alba basics; the squares represent regions

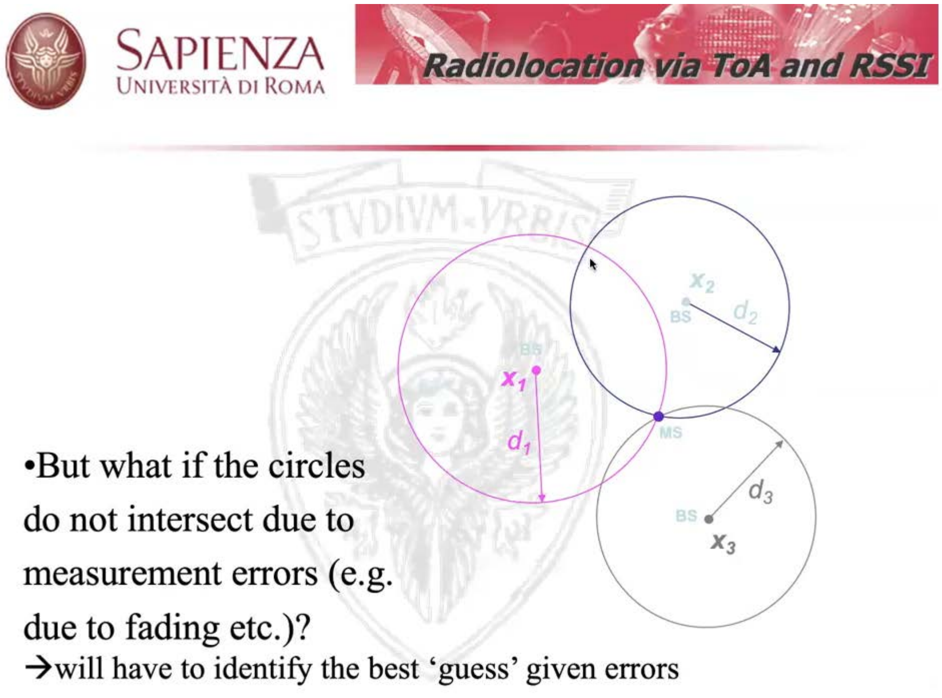
Rainbow to solve the dead end problem, a color code is applied to nodes. Yellow nodes represents the one that exhibit a greedy path to the sink; initially, all the nodes are yellow. If a node recognizes himself as a dead end, then it will color himself with red, and will send back (negative advancement to the sink) its packet to red or yellow nodes. If red nodes can't advance nodes, they will color themselves as blue, and will search for blue and red relays to send packed forward (positive advancement to the sink). If even this way they can't find a route, then they will color themselves as violet and relay packets backward (in a negative advancement to the sink) to blue and violet nodes. There can be many colors, e.g. h, and each color has a label C_i; yellow nodes have always the C_1 label, and in general: odd-labeled nodes relay forward, even labeled nodes relay backward, and both relay to neighbors with the same label C_k or the previous one C_{k-1}, with the exception of yellow nodes, who relays only to other yellow nodes labeled C_1.

10.3 Localization in sensor networks

Why a GPS isn't always used? Because it has high power consumption, requires additional hardware, works only with LOS (line of sight) satellites, and it gives too much information: the exact position isn't required, the relative one can suffice; global coordinates can be inferred centering the network around the sink. Localization of nodes can be achieved by some specialized hardware (Angle of Arrival, analysis of Received Signal Strength, Time of Arrival [basically a ping]), or it can be estimated using the number of hops (range free approach).

Angle Of Arrival - AoA the position of a node can be estimated calculating the angle of (at least) two signals that intersect said node; high directional antennas are used of this method

Time of Arrival, Received Signal Strength both gives a circle where the device must lie; to estimate the position, at least three circle must that intersect with each others must be seen.



SAPIENZA UNIVERSITÀ DI ROMA

Radiolocation via ToA and RSSI

- But what if the circles do not intersect due to measurement errors (e.g. due to fading etc.)?

→will have to identify the best ‘guess’ given errors

Figure 34: Functioning of the radiolocation with RSS or ToA

For a 3D localization, an additional anchor will be needed: two sphere intersecting will form a circle; a third one intersecting with the circle will give two points of uncertainty (like two circles intersecting together), so a fourth one is required.
MSE between the estimated position and the anchor is the metric used to minimize the error when the position is evaluated. The practice where beacon nodes helps the others to estimate their position is called multilateration; it is a simple solution that works for small networks, but suffers if errors begin to accumulate.

10.3.1 Range free approach

In this approach some anchors calculate their relative positions between each others using the number of hops; the localization error is high (because the number of hops is used to estimate the geographical position).

10.4 Energy-efficient MAC protocols

The sources of energy waste are: collisions, overhearing, control packet overhead, idle listening; a MAC protocol for sensors network must be energy efficient, and guarantee fairness, low end to end latency, and scalability. For these reasons, an awake/asleep schedule with a given duty cycle is widely adopted; in some protocols like S-MAC, these schedules are synchronized among nodes (periodical synchronizing packet are exchanged; they can get desynchronized because of clock drift), and every node transmits over the same slot if possible, resulting in a waste of radio resources.

10.4.1 S-MAC

Selecting a schedule First of all, a node must decide a schedule and broadcast it to its neighbors: at startup, the node listen for a sync packet; if such packet is received, then the original node become a *follower*, setting this schedule as its own, and then it will broadcast it after waiting a random time t_d. Now all the nodes with that schedule are synchronized and cannot collide. If the sync packet isn't received, the node will sleep for a random time T; when he wakes up, it will broadcast this time (that is the new sleep time), becoming a *synchronizer*. If a node receive a sync packet that is followed by more than one node, but has already its own schedule, it will adopt both and broadcast the new one; this happens to border nodes, and makes them able to synchronize to two or more groups of nodes, consuming more energy. Each node maintains a schedule table that keeps schedules of known neighbors.

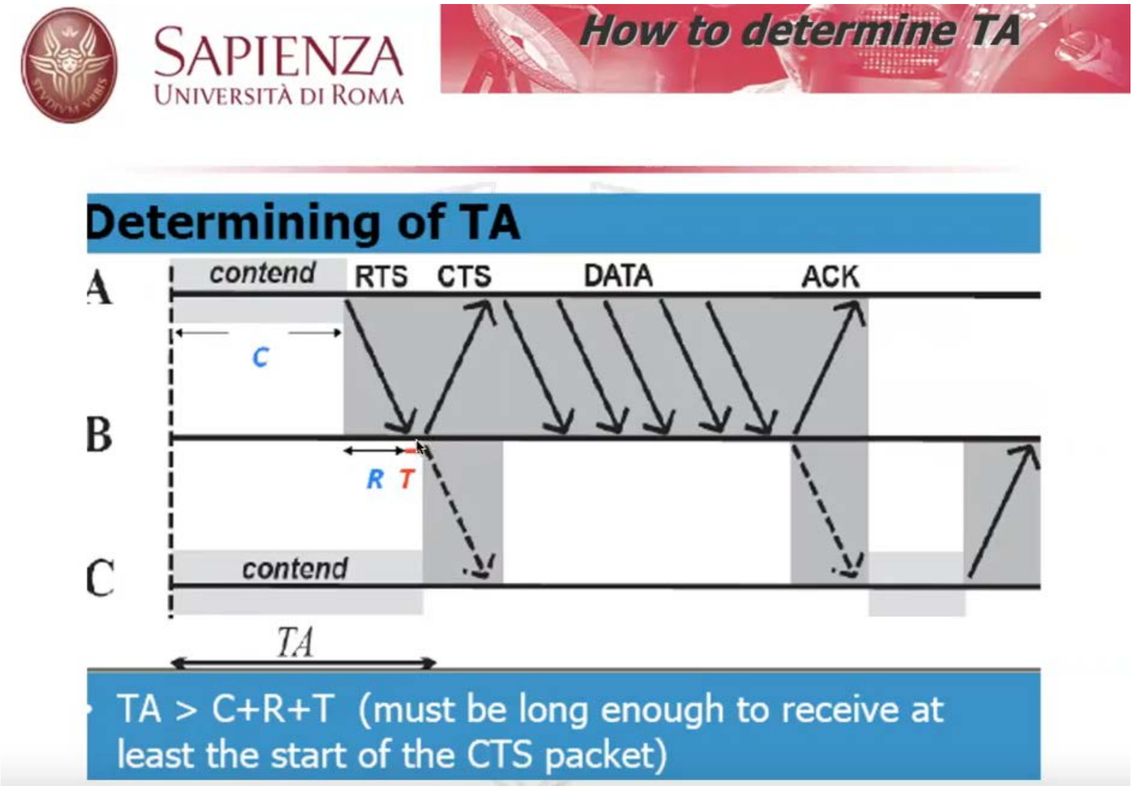
Sending a packet before sending a packet, the source will wait for the destination to be awake, and then it will behave like in CSMA/CA: it will perform carrier sense for a random period of time, and if the medium remains idle, it will begin the RTS/CTS exchange; if this latter has been successful, the data is sent, followed by an ACK. NAVs are used to decide for how long go to sleep before trying to re-access the medium, and bursts of packets are sent to optimize the time used for the RTS/CTS handshake. The transmission of ACKs are sufficient to deal with the hidden terminal problem.

Problems As always, synchronization is a pain. There's the phenomena of clock drift and SYNC packets may be lost, so they are exchanged frequently (and this means control overhead). Throughput is decreased because only the active part of a frame is used to transmit; moreover, it is further decreased because transmission and reception occur only during the ON time. Delays are experienced since the ON time of each relay must be waited.

10.4.2 T-MAC

To find a balance between ON time (who potentially waste energy in idle listening) and the repetition of the duty cycle (potentially increasing the latency, because packets may have to wait for a node to wake up again), in T-MAC the active time is changed dynamically adapting on the traffic level of the network: higher the traffic, longer the active time.
It function very similarly to S-MAC: same exchange of sync packets and CSMA/CA usage, with the data burst. However, if a node doesn't receives transmissions from neighbors for a period of time TA, the active time is aborted and the node will go to sleep; this timer will reset in the following cases: if any data is received on the radio channel, communication is sensed, data are being transmitted, or RTS/CTS are exchanged by neighbors (it's important to remain active when neighbors communicate because it means that the node may be the next relay).

How to determinate TA it must be long enough for the node to contend the channel, and to sent RTS and receive CTS.



How to determine TA

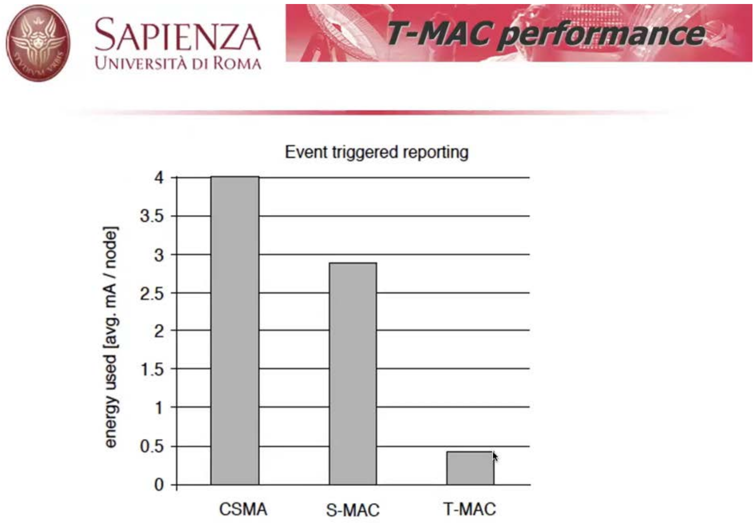
Determining of TA

- TA > C+R+T (must be long enough to receive at least the start of the CTS packet)

Figure 35: Equation to determinate TA

Optimizations there are a number of new changes from S-MAC for optimization: when a node sends the RTS but doesn't receive the CTS, it must be because: the RTS collided, the receiving node cannot answer due to RTS/CTS overhead, or the receiving node is asleep; in the first two cases, reducing the *AT* timer would be wrong, so a node will send RTS two times before going asleep. Also, mechanism are introduced to notify nodes (at the beginning of their awake time) that there will be traffic for them, so they shouldn't go to sleep; this problem is called early sleep, and there are two solutions [seen in this course]: the future RTS, FRTS, that notifies a node to remain awake because it will receive traffic soon, and the full buffer priority: when a node has an almost full buffer and hears a RTS, instead of responding with a CTS, he will respond with a RTS to the destination of his queued packets, and (after hearing the CTS) send him a burst, emptying its queue.

10.4.3 Performance



SAPIENZA
UNIVERSITÀ DI ROMA

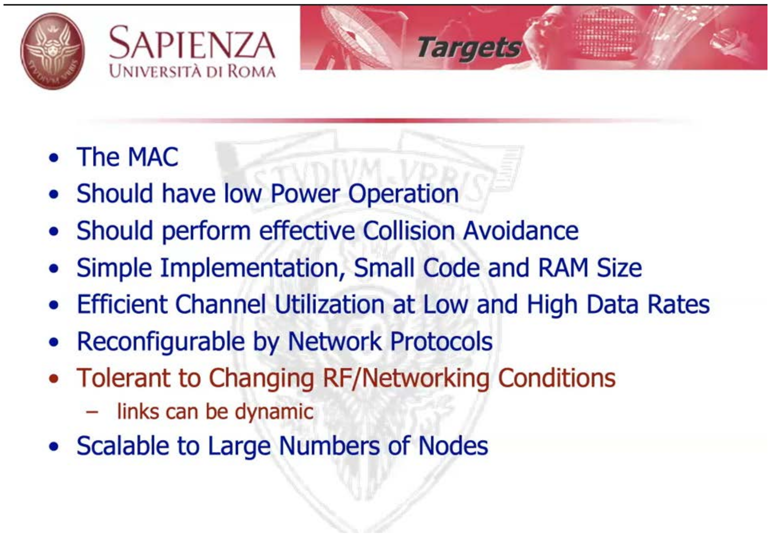
T-MAC performance

Figure 36: A comparison between the different MACs

11 12/05

11.1 Asynchronous MAC protocols

These protocols don't exchange the asleep/awake schedules, because a fixed schedule isn't the best idea: maybe in a period of time the traffic will be higher and would require more awake time, or maybe the traffic will be low and the awake time is wasted power. T-MAC implements this idea, but the problem of low throughput is still present: all the channels are active only for a brief period of time; also synchronization is very inefficient because of clock drifting: even if every node is perfectly synchronized, they will still require periodic synchronization. Other sources of energy consumption in synchronous MAC are: idle listening (to listen to schedules), and large control overhead. [a paper confronting the following protocols and other ones can be found here; there are also experiments]



SAPIENZA
UNIVERSITÀ DI ROMA

Targets

- The MAC
- Should have low Power Operation
- Should perform effective Collision Avoidance
- Simple Implementation, Small Code and RAM Size
- Efficient Channel Utilization at Low and High Data Rates
- Reconfigurable by Network Protocols
- Tolerant to Changing RF/Networking Conditions
  - links can be dynamic
- Scalable to Large Numbers of Nodes

Figure 37: A recap of MAC's objectives in wireless systems

11.1.1 B-MAC

Berkley-MAC [the professor advises to read the paper to prepare for the exam].

Collision avoidance for collision avoidance, B-MAC estimates the channel noise and determine if the channel is free by taking samples and checking if they are under the average noise level; if in 5 samples there's no outlier (a sample below the noise level), then the channel is assumed busy.

Transmitter and receiver sides when transmitting, a node will attach at the start of the packet a preamble; receiving nodes wake up and listen for activity for an interval of time equal to the length of the preamble; if they sense activity, then they'll remain awake to receive the packet. After reception (or a timeout), they'll return to sleep.

Downsides the energy consumed in transmitting will be high because of the long preamble; moreover, every node that overhear the preamble will stay awake until the end, just to discover they weren't the target. The latency caused by preambles increases hop by hop.

11.1.2 X-MAC

It's meant to take the features of B-MAc and make them better: in the preamble is embedded the address of the destination, so nodes can return to sleep without having to listen to all the preamble. The preamble is a series of short preambles divided by pauses; these latter allow the destination to send an ACK to interrupt the preamble and to start the transmission of data.

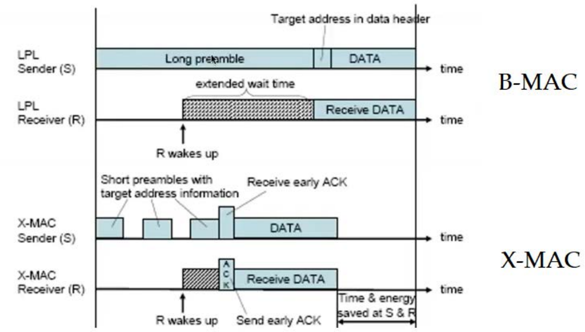


Figure 38: A comparison between B-MAC and X-MAC transmission

11.1.3 WiseMAC

WiseMAC tries to use minimum sized wake preamble; all sensor nodes sense medium at the same constant period T_w independently (meaning that the sampling period is the same but maybe in different times). The main idea is to learn the sampling schedules of direct neighbours of a node; these schedules are used to minimize size of preambles. To recover packet losses, a link level acknowledgement is used: the WiseMAC ACK packets are not only used to carry acknowledgement information but also to inform other nodes (including sender) the remaining time of next sampling. These other nodes store this time in their tables. Using this information, a node transmits a packet with minimized size of preamble; the duration of the wake-up preamble covers the small potential clock drift between the clock at source and destination.
The minimum preamble is calculated by: $min(4\theta L, T_w)$, where θ is the frequency tolerance (the deviation from the nominal frequency) and L the interval between communications (updated at every ACK received by neighbors).
In high traffic, the packet overhead is low, but in low traffic is high; however, the power consumption resulting from this high overhead is low.

11.2 6LoWPAN

[The two paper i used to write this part: one and two].
This protocols exploit ipv6 for IoT devices.

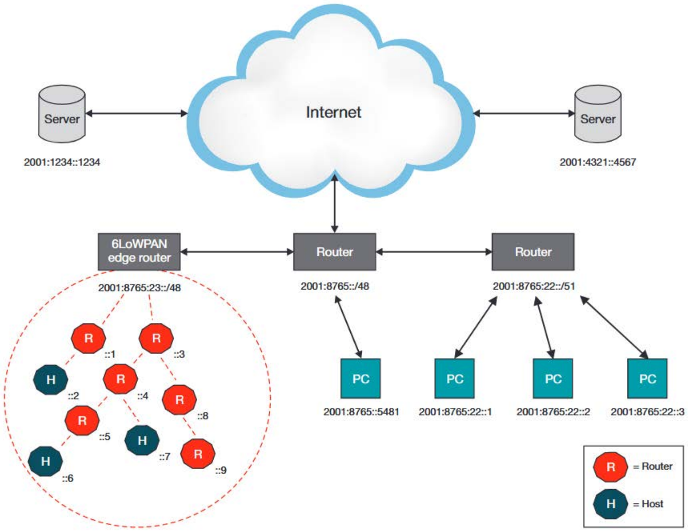


Figure 39: An ipv6 network including a 6LowPAN network

Network architecture As can be seen in the figure, the 6LowPAN network is connected to another ipv6 network with an edge router; this latter does three main things: communicates with other ipv6 networks (including internet), manages the local data exchange between the devices in the 6LoWPAN network, and the generation and maintenance of the 6LoWPAN network.[6LoWPAN networks are defined as "stub" networks, and is because incoming data is tipically directed to one node inside the network]. IP routers are sufficient mean to communicate with other networks since 6LoWPAN uses ipv6; and they can be connected using WiFi, 3G, Ethernet etc.. A 6LoWPAN network isn't divided in different subnetworks, meaning that all the nodes share the same ipv6 prefix; this is because to reduce overhead, the size of the addresses is reduced to 16 or 64 bits (this latter used only by edge routers to communicate with ipv6 to external networks), and if the prefix is always the same, it can be ignored and not included into addresses. Addresses are configured using ipv6 neighbor discovery (a modified version, since the classical one have some wrong assumptions, such as that all the neighbors are always awake).

Protocol stack Observe the picture confronting TCP/IP and 6LoWPAN stacks; ICMP is used for neighbors discovery, and between the datalink and network layers there's this adaption layer that enables transmission of ipv6 datagrams over 802.15.4 radio links. Both UDP and IP headers must be compressed to match the size of 802.15.4 frame: 127 Byte. Let's look at the physical layer: with 802.15.4, it's possible to use the ISM bandwidth, so either a channel of 868 Mhz, or the 16 channels of the 2.4Ghz band; this means that the lowest data rate obtained by the 868 mhz channel is 20kbps, and the highest is 250kbps.

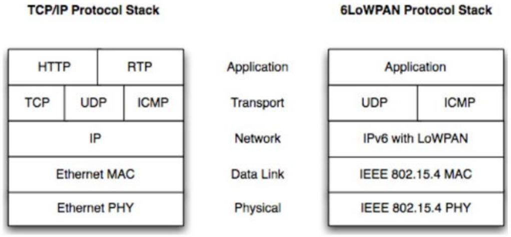


Figure 40: A comparison between TCP/IP and 6LowPAN stacks

Packets, headers Thanks to the compressed addressing, 6LoWPAN frames can use more bit per payload without modifications regarding the size of the frame.

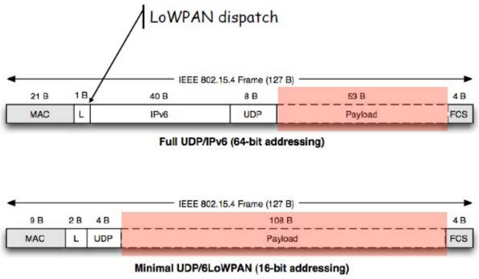

LoWPAN dispatch
IEEE 802.15.4 Frame (127 B)
21 B
1 B
40 B
8 B
53 B
4 B
MAC
L
IPv6
UDP
Payload
FCS
Full UDP/IPv6 (64-bit addressing)
IEEE 802.15.4 Frame (127 B)
9 B
2 B
4 B
108 B
4 B
MAC
L
UDP
Payload
FCS
Minimal UDP/6LoWPAN (16-bit addressing)

Figure 41: A comparison between 802.15.4 and 6LowPAN frame

There are three main situations: when the topology is a mesh network (so there are intermediate coordinators, see the next section) and the payload must be fragmented (fragmentation is no longer used in ipv6 in general, but there is needed due to the little size of packets); when the payload must be fragmented but there are no intermediate hops, and when the payload doesn't need to be fragmented. The datagram size describes the original size of the un-fragmented datagram; the tag is used to identify sements that belong on the same datagram, and the offset describes the position of this fragment in the un-segmented datagram. The reassembly is done at link level. To further optimize the size, if a segment is sent and it is the first one (so it's offset is 0), then the *datagram offset* part of the message is omitted, and instead the third bit in the header is set to 1.

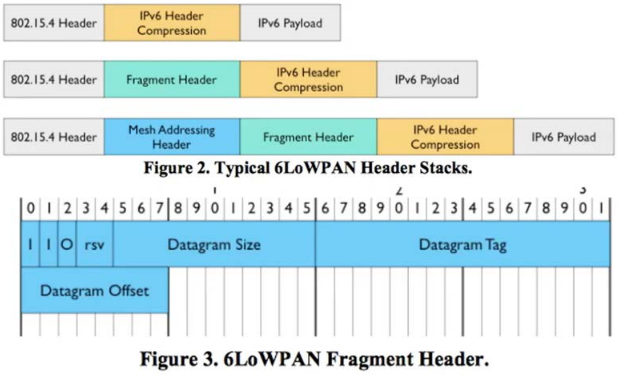


Figure 2. Typical 6LoWPAN Header Stacks.

Figure 3. 6LoWPAN Fragment Header.

Figure 42: Different situations and fragmentation specifics

Topology - 802.15.4 The following paragraph shows topology specifications from the 802.15.4 protocol. First of all, independently of topology, remember that the nodes are low powered, and that they can do duty cycles. There are three main type of nodes: simple ones (called end devices), that can't relay information for others and are very limited, a middle type one (called coordinator), that can relay information but still doesn't have high capabilities, and PAN coordinators that are the same type as before, but have been elected to coordinate the whole network. There are also three types of topology (see figure 43). Now let's discuss about roles: PAN coordinators pick the frequency (or frequencies) to use in the network, along with the ID; it also manages requests to join, will coordinate communications, disappeared nodes and can relay packet. When a new network is started, the PAN coordinator gets elected; it will then assign itself a 16 bit address and pick a frequency. New nodes entering the network will scan for the coordinator, send him a request to join (it will be ACKed back), and the coordinator may assign them 16 bit addresses. After this process and after the coordinator has decided how to route in case of a mesh topology, the data exchange can begin. As of the MAC, it can operate in two modes: bea-

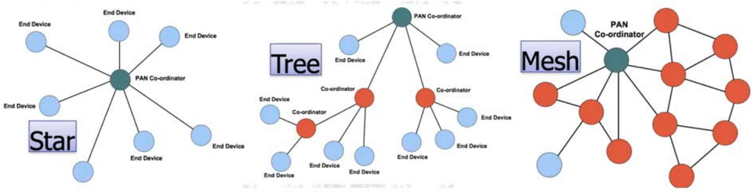


Figure 43: The three types of topology in the 802.15.4 protocol

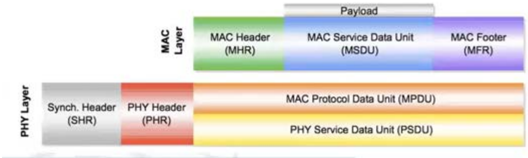


Figure 44: The two packets of 802.15.4, both the data-link and physical one

conless, and in this case it will use CSMA/CA, and beacon. In beacon mode, the beacon is a PAN coordinator, and this frame is utilized: in red, the time used by the beacon to broadcast network information, frame structure, and notifications of pending node messages; in orange, a contention frame that can be used by anyone using CSMA/CA, and in green the contention-free period, long 7 time slots, reserved for nodes that require a guaranteed bandwidth.

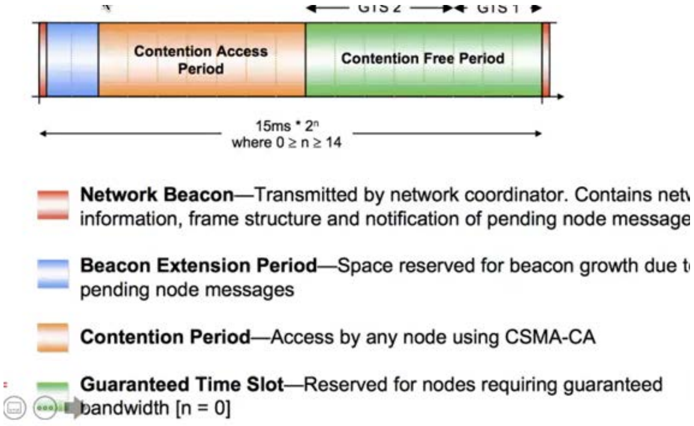

GTS 2
GTS 1
Contention Access Period
Contention Free Period
15ms * 2ⁿ
where 0 ≥ n ≥ 14
Network Beacon—Transmitted by network coordinator. Contains netv
information, frame structure and notification of pending node message
Beacon Extension Period—Space reserved for beacon growth due t
pending node messages
Contention Period—Access by any node using CSMA-CA
Guaranteed Time Slot—Reserved for nodes requiring guaranteed
bandwidth [n = 0]

Figure 45: The beacon frame

12 14/03

12.1 LPWAN

Low power WAN is a type of IoT technology. A single base station covers 10km and services millions of devices; these sensor nodes are low cost and low power, and can communicate in very long distances.

12.1.1 LoRa

LoRa is a physical layer that provide connectivity to low powered smart objects, and LoRaWAN is a complete stack that allow to build (WA)networks on top of LoRa links. In LoRa, the physical layer is patented and cannot be changed, but the MAC can be chosen (LoRaWAN is not a mandatory choice). To reach far distances, bits are coded using a spread spectrum; in this way, the signal is very robust to interference, multipath and fading.

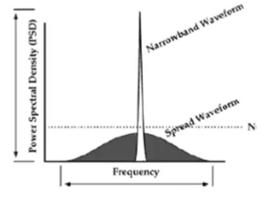


Figure 46: Spread vs narrow waveform

Chirps, symbols and spreading factors [For easy insights on this section (they're in italian) - here, and the paper from the lesson is here] a chirp is a sinusoidal signal whose frequency increases or decreases over time. The usage of chirp signals with a spread spectrum means that the chirp signal will use all the bandwidth to transmit a message, thus making it robust against noise. For this protocol, **SYMBOLS ARE CALLED CHIRPS**, because they are transmitted as such. [Recall from the previous lessons about 2G that symbols encode one or more bit]. LoRa uses spread spectrum and encodes information with multiple chirps; let *symbol rate* represent the rate at which the information are sent, and *spread factor* (SF) the duration of a chirp. Then, there are 2^{SF} possible waveforms (different symbols); LoRa can trade off data rate for sensitivity by selecting the amount of spread to use (7 - 12): lower SF means more chirps are sent per second, so more information can be encoded per second, resulting in an increased data rate; higher SF means less chirps are sent per second (because the duration of a single one is increased), so less information encoded per second, so that to send the same amount of information, more airtime is needed, resulting in a higher consumption of power. However, it gives the opportunity to the receiver more opportunities to sample the signal which results in better sensitivity. Each transmitted symbol encodes SF bits, and (again) there are 2^{SF} different symbols.

Top graph: SF modulation can change in time, unlike other protocols that

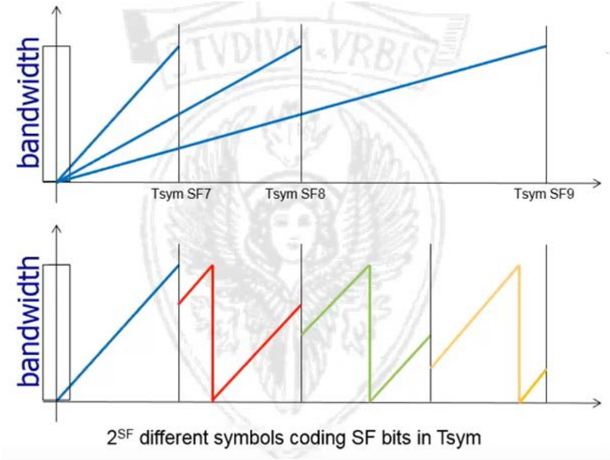


Figure 47: Example of different SFs and the form of the subsequent symbols

use a carrier frequency that never changes. Bottom graph: some of all the 2^{SF} possible symbols (that are derived by cycling the original frequency) of one of the SF modulations (again, they can be between 6 and 12). Theoretically, every different SF modulations are orthogonal between each other, but in reality this is true only if they are perfectly synchronized.

Demodulation to demodulate the signal, the chirp and the inverse chirp are complex multiplied, and the fast Fourier transform is applied on the result. With the fast Fourier transform, the energy becomes concentrated on single tones (frequencies); basically with the two operation (complex multiplication and fft), a peak is obtained, and it indicates the time shift of the original chirp. How is this information useful to extract information? Watch figure 47: a frequency is chosen and various symbols encoding bits are obtained by shifting it; there are 2^{SF} symbols and each one encodes a sequence of SF bits; knowing how much the original frequency has been shifted, let us know which sequence of bits that chirp was encoding.

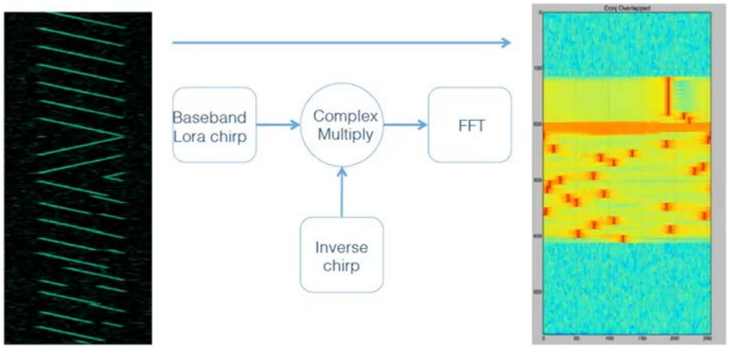


Figure 48: LoRa demodulation various chirps

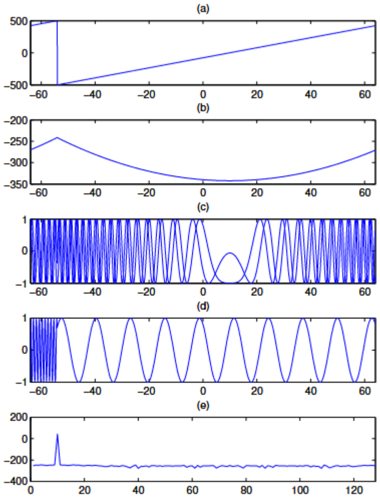
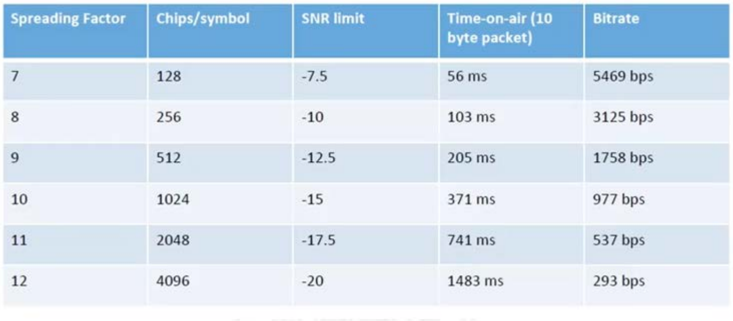


Figure 49: LoRa demodulation of a single chirp

On figure 49, you can see: a) the original chirp, b) the inverse chirp c) the complex multiplication of them d) the latter signal sampled at chirp rate e) the peak obtained by the Fast Fourier Transform.

ISM frequency bands as many other protocols, LoRa uses the ISM free-to-use bands. For each band, we can choose 7-12 spreading factors (SF); there are three main bands used: 125kHz, 250kHz and 500kHz. As you can observe in Figure 50, the higher spreading factor, the high number of chirps, time on air and lower bitrate. Notice that the SNR (signal-to-noise ratio) is negative, and this is because even if the received strength of the signal is below the noise floor, is still possible to extract information; with a spreading factor of 12, even if the signal is 1000 lower than the noise floor, it's still usable.

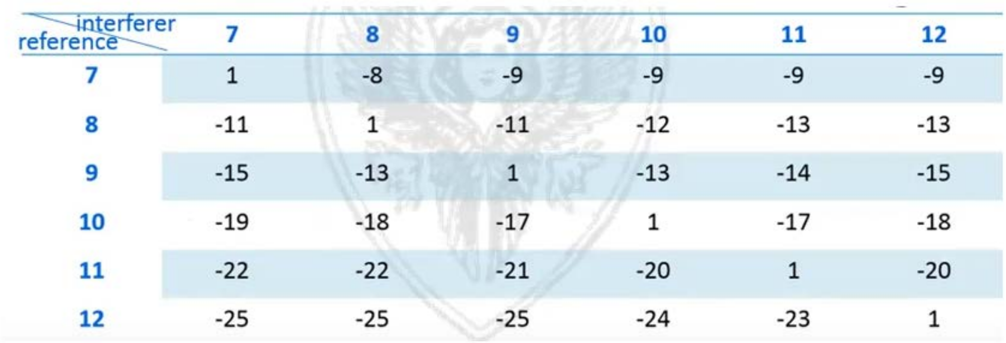


Spreading Factor	Chips/symbol	SNR limit	Time-on-air (10 byte packet)	Bitrate
7	128	-7.5	56 ms	5469 bps
8	256	-10	103 ms	3125 bps
9	512	-12.5	205 ms	1758 bps
10	1024	-15	371 ms	977 bps
11	2048	-17.5	741 ms	537 bps
12	4096	-20	1483 ms	293 bps

Figure 50: Various SF in the 125kHz band

Coverage as stated before, higher spreading factor means higher sensitivity; changing the spreading factor from 6 to 12 will increase the range by a factor of 2.5. The range is typically of a few km (in an experiment, up to 7km) in an urban setting with $\eta = 4$. In LOS (line of sight), the range is up to hundreds of km.

Inter SF interference since different signals can't be perfectly synchronized, they won't be perfectly orthogonal too, so they can interfere with each other (even with different SF). If one of the signal is at least some decibel above the other, that transmission will be successful; this amount changes with the amount of SF, since higher SF is more robust to noise.



interferer reference	7	8	9	10	11	12
7	1	-8	-9	-9	-9	-9
8	-11	1	-11	-12	-13	-13
9	-15	-13	1	-13	-14	-15
10	-19	-18	-17	1	-17	-18
11	-22	-22	-21	-20	1	-20
12	-25	-25	-25	-24	-23	1

Figure 51: How higher in decibel a signal must be to be correctly demodulated

Notice that for two signal with the same SF, one of the two must be 1dB above the other to be correctly demodulated.

12.1.2 LoRaWAN

Is a communication protocol and architecture that uses LoRa physical layer; it supports: secure bi-directional communications, mobility and localization, and defines how node should communicate. In a LoRaWAN network there are end nodes connected to a gateway via LoRa links, and these gateways can be connected to other network via LoRa links or other links such as 3G, ethernet, etc.. Is possible that gateways receive and relay the same information, and if this happens, the duplicates are filtered at the network server. Gateways only relay information.

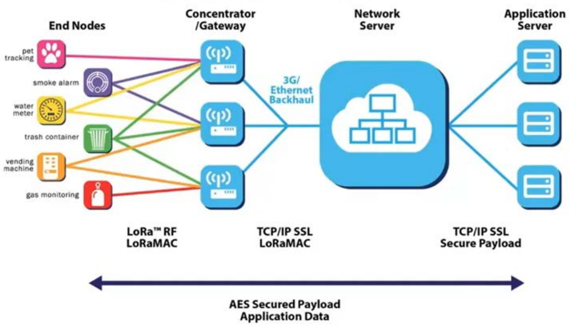


Figure 52: The architecture of a LoRaWAN network

Devices in LoRaWAN there are three types of devices: Class A, where each uplink communication is followed by two short downlink receive windows; class B, is like A but there are extra receive windows at scheduled times (called beacons). And class C, that has continuous receive windows except when it transmits. There are different types of node to balance the battery life and the downlink communication delay; class A is the most energy efficient, followed by B and then C.

Gateway they receive on all the channel for all the time (for all the spreading factors); there's no network controller, no frequency re-use and such. Each the sensor can communicate to every gateway, and every correctly demodulated packet is forwarded to the network server. At this level there are no filters.

Network server is responsible to listen for every packet from gateways, identify duplicates and filter them; they must validate data (security stuff) and correctly send them to the correct application server. Since there's a central time reference for every gateway, the network server is capable of localization. Since every decision on the network is made by the network server, the gateways are very cheap to implement.

LoRaWAN frame it starts with a preamble of known chirps; then an header, a CRC to control the accuracy of the header, and the payload. [Really, no interesting things are being said about LoRaWAN frames]

Security there are two layers of security: network and application. The first authenticates users, adds message integrity checks and encrypt the whole packet; the later encrypts its payload. These mechanism can be statically configured (pre-configured) or over the air; both of them uses AES128 for encryption.

Over the air activation - OTAA first of all, LoRaWAN describes a join procedure prior participating in data exchange with the network server; every time a node is disconnected or lose the session context information, it must re-do the procedure. The requirements to participate in the procedure are: two ID, device and application, the first identifies the device, the second the join server and is used for session key derivation; and also two keys: application and network; both are for a specific end-device, but the later is given by the network. If the application key is compromised, only that node will be compromised. When a node wants to join communication, it will send a join request to the network server, with a specific join ID and the device ID. The network server replies with a message that is encrypted with the network key.

12.1.3 Performance of LoRa and LoRaWAN

In LoRa, when a node needs to communicate, it will immediately send the packet; there's no synchronization, and in case of collision the packet can be resent or cancelled; usually they are cancelled, since ACKs from gateways are rarely used. It's similar to the ALOHA, so it's very limited in terms of performance: there's only the 18% throughput over the network capability.
Depending on the SIR (signal to interference ratio), different things can happen in case of collision (refer to Figure 51). Different SF can be received correctly even in case of collision if the SIR is sufficiently low, thus the throughput can be much higher than 18%. If devices are ”competing” to send packet with the same SF (Intra-SF), the interference will arrive from devices that are closer to the gateway than that node.

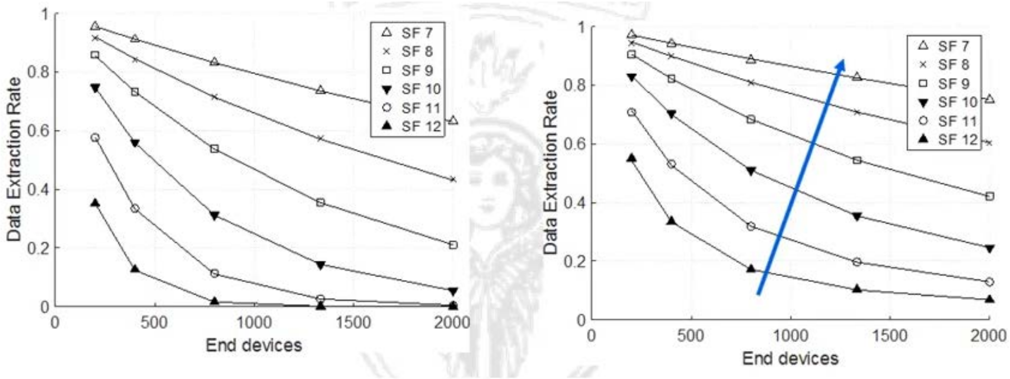


Figure 53: Performance comparison ALOHA vs ALOHA with capture in intra-SF

On the other hand, when devices are competing using different SF, the interference will arrive from nodes that are closer to that node than the ones near the gateway. In this case, there's degradation in performances, because one packet will be transmitted more powerfully than the other, so the other one will be lost (and this is because they aren't perfectly orthogonal).

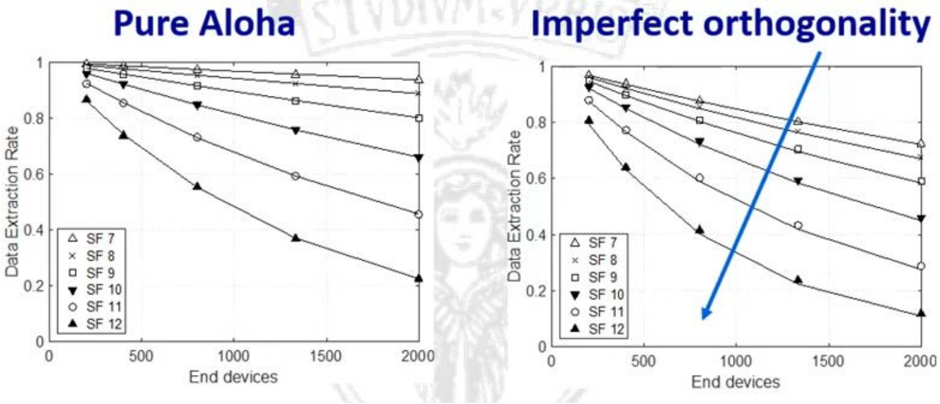


Figure 54: Performance comparison ALOHA vs ALOHA with capture in inter-SF

Scalability LoRa cells can't sustain high loads, so the duty cycle must be under 10% of the network capacity per channel. However, gateways can work on 8 channel at the same time, and can opportunistically manage SFs and transmission power. If the density of devices increases, more gateways can be deployed. Having different ; gateways (almost) means to have M different systems.

12.1.4 LoRa SF allocations

Adaptive Data Rate - ADR given the SNR and RSS (received signal strength), select the lower possible SF; this is done to have the highest possible data rate. This is a sub-optimal solution; since devices can have very different positions, it would be better to use different SFs to obtain the maximum orthogonization. This can be done using different frequencies or different SFs (even if they aren't perfectly orthogonal, the interference will be low).

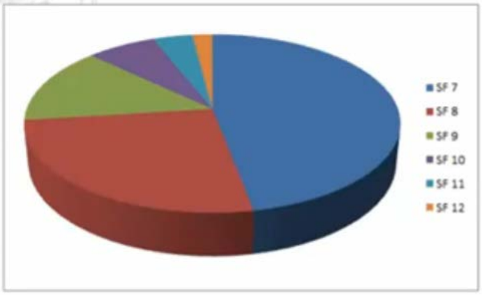


Figure 55: Distribution of different SF over the network to balance load

Of course more node will have the highest data rate, and only few the SF 12, because of the high airtime (SF 12 airtime is 32 times longer than SF 7). And how these nodes must be geographically distributed over the network? Circular (e.g. highest SFs far away and lower near the gateway) or random? Random is better because it's more probable to do the capture (understanding symbols in a collision), and also because it doesn't disadvantage far users, who can suffer from high inter-SF interference.
Does power control help? Absolutely no, it will destroy the packet capture opportunities so much that is equivalent to put every node the furthest from the gateway.

Sequential waterfilling - EXPLoRa users are ordered as a function of their RSSI from the gateway (basically the distance); SF are allocated in proportion starting from the 7, but sometime a perfect proportion isn't possible because some nodes may be too far away from the gateway to use lower SFs.

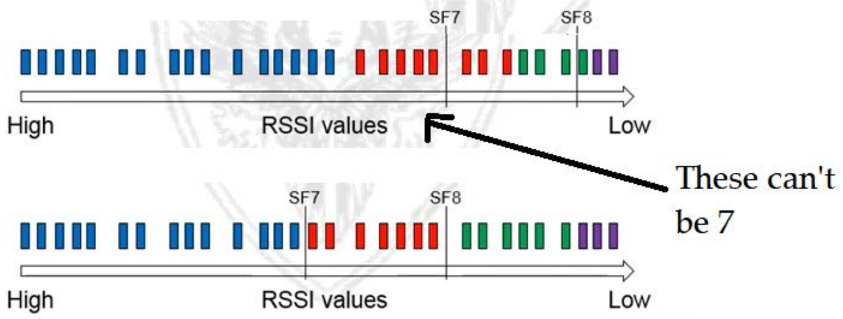


Figure 56: Proportionality of SFs

It's possible to optimize more taking into account the different RSSI of consecutive nodes: if the RSSI differs from the previous node by more than 3dB, then the same SF is kept; otherwise, use another one compatible (again, can't use 7 if a node is too far away from the gateway).

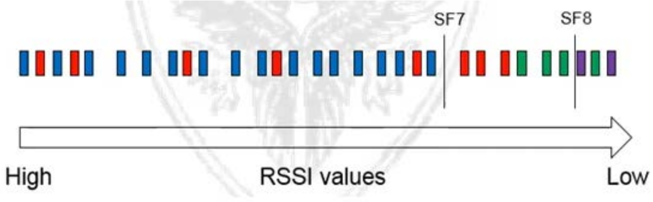


Figure 57: Waterfilling taking into account RSSI and othogonality

When dealing with multiple cells, nodes are divided in sets; their distance is generalized taking into account the visible gateways: if a pair of nodes see the same gateway(s), they will be closer. It can also deal with nodes that refer to different ISPs [but the professor didn't go much in detail]. Actually there's a fairness problem: with packet capture, closer nodes are almost always successful in transmission, but far away nodes are disadvantaged. Also, SFs can be distributed at random, but they perform worst in multi cell and multi ISPs.

13 19/05

13.1 Collection Tree Protocol - CTP

The base idea is to send beacon messages to construct a tree structure, and nodes can relay their information to their best parent (the one that can send data to the sink faster). Desired properties:

1. Reliability: if there's a route, 90% of the packets must be delivered successfully

2. Robustness: it must work without tuning or configuration in a wide range of network conditions

3. Energy efficiency

4. Hardware independent

Another challenge is that link quality changes fast, even every 0.5 seconds; however, the topology is static.
To select the best parent, the metric used is the ETX; it represents the expected number of transmission to reach the sink. So the one with the lower ETX is selected; since there's an exchange of these metrics along neighbors to construct routing tables, it's basically a distance vector approach.

There are two types of information: control and data. The first are the beacon messages, and they are exchanged regularly when the sink starts the beaconing process. These beacon messages are used to exchange information about ETXs; then the node will use them to compute the best route. In the data packet is embedded the estimated ETX, and sequence numbers that allow to know the percentage of packet that has been received, allowing to estimate the link quality.

To compute the best route, the algorithm is basically Bellman-Ford but instead of hops, the metric is the ETX: first of all the node calculate the ETX to arrive on each of his 1-hop neighbors, and then this will be summed to the ETX declared by those neighbors; the lower sum will be chosen. This computed ETX will be included in beacon messages sent by this node, allowing lower neighbors to know it and possibly choose him as their relay. To compute the ETX, the operation is a weighted sum of past transmission and current samples. If a node is congested, it can communicate it and nodes won't choose him as a relay; it is communicated setting a congested flag equal to 1 either in a data packet or a beacon packet. To avoid loops and count to infinite problem, there's a pull flag that can be set to 1 (either in beacon or data messages).

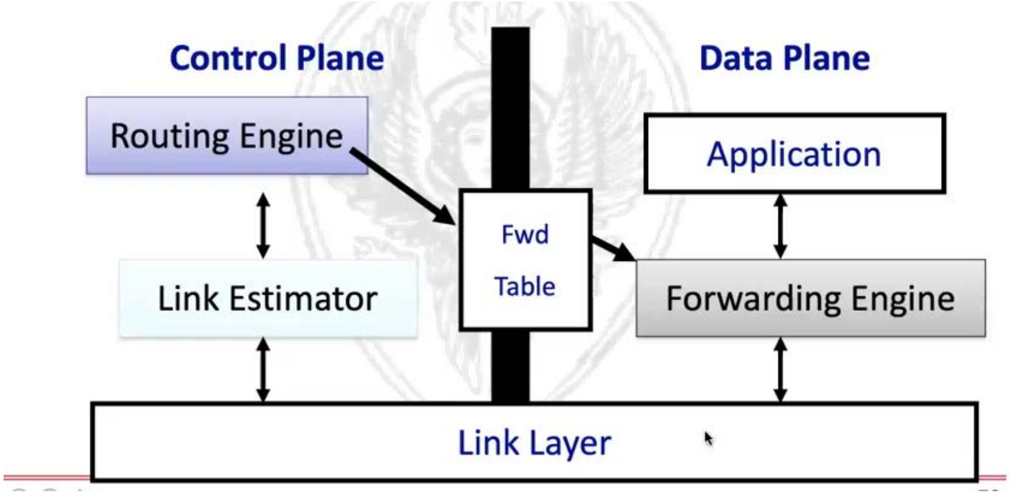


Figure 58: Architecture of nodes in CTP

As seen in the picture, there are two main parts in a node: a control plane, that manages routing tables exchanging beacon messages (both receiving them and sending them) and estimating the ETX (the link estimator is the one that computes the ETX on neighbor's link); and a data plane, that performs forwarding of information and performs some check, like filtering duplicates, detecting loops (and repairs them), etc.. Of course to forward it consults the forwarding table created by the control plane.
How to detect a loop? When a node receive a packet, it will check that the ETX embedded in it has a higher value than itself; if it's true, the ETX will be changed as the one on the current node, and then the packet is forwarded. If it's not true, the pull bit is flagged as 1, and the route must be recomputed starting a beaconing process (exchanging beacon messages to recompute routes between neighbors).How can this happen? for example when a link suddenly broke down, and the ETX cost of a node will change suddenly.

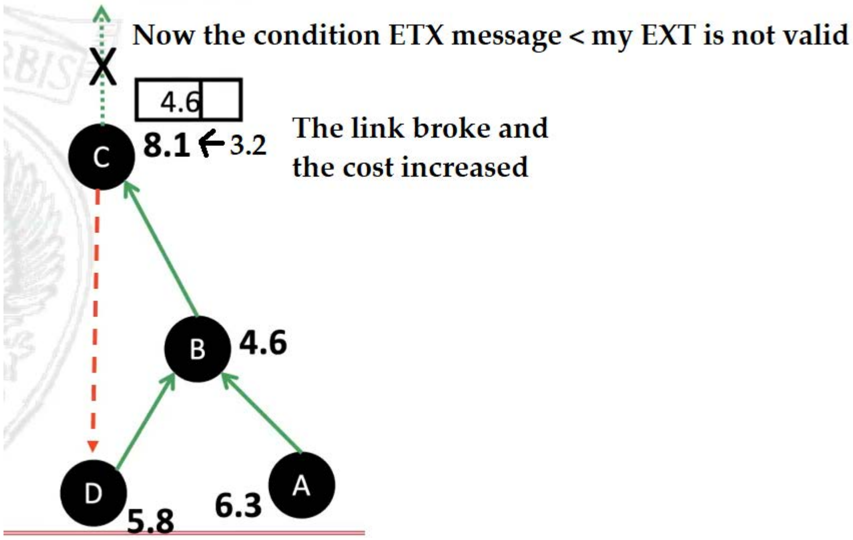


Figure 59: Principle of loop in CTP exchange

Beaconing process is started by the sink periodically, or when something went wrong (e.g. pull = 1). If everything is stable, the time between periodical beaconing processes doubles (up to a maximum); this time is reset if some inconsistency is detected. At the beginning, this interval will be short.

Performance actually very good, 98% of packets are correctly delivered.

13.2 Extensions of 802.15.4

802.15.4 is a standard physical/mac for IoT. As stated in 6LoWPAN, it operates in two modes: beaconless, and in this case the MAC is CSMA/CA, and beacon mode, where beacons define a frame where: in the first and last part the beacon transmits information about the network (e.g. the length of the frame), some time slots were for contention with CSMA/CA, and 7 time slots were for reserved communication for node that asked for a "secure" window of time to transmit. The limit of this protocol are: no delay guarantees, no resilience to interference (because frequency hopping is not adopted), not ideal in high traffic, and there's no standard energy otpimization in tree topologies (thought optimizations exists - they just need to be specified).

13.2.1 802.15.4e

This variant encapsulate some added features.

Radio Frequency Identification Blink - BLINK mode supports effective ID exchange for identification, location and tracking.

Asynchronous multi-channel adaption - AMCA improves the performance of networks that use 802.15.4 in beaconless mode (e.g. mesh networks); it increases the capacity of the network supporting multiple channels.

Deterministic and Synchronous Multi-channel Extension - DSME enables support for time-critical application for large networks in beacon mode. The limit was that only 7 slots are reserved and multiple channels are not supported. In this variation, there are multiple super-frames that cycle, and only the first is like the normal frame: the others have the time slots that were for the contention window occupied by other guaranteed time slots.

Low Latency Deterministic Network - LLDN supports applications that need low and deterministic latency (the delays are known). This is achieved by multi channel extension (PAN coordinators how have multiple transceiver and can transmit simultaneously over multiple channels), slotted beacon enabled frames, and shorter slots, packets and addresses; there are also optimization so PAN can now ACK to let nodes know if information have been received.

Time Slotted Channel Hopping - TSCH supports some of the aspects already stated in other variants: slotted access, multi channel and frequency hopping. It's topology independent, supports high network capacity and has high reliability and predictable latency, while allowing low duty cycles. IT works in this way: nodes synchronize on a cycling frame (that consists in n timeslots), and every time slot allow a node to send data and receive ACK. There are also multiple frequencies identifying multiple channels. To exchange data, a two dimension table is used for scheduling: on one axis there are the time slots, and on the other the channels; communications are scheduled using this table.

13.3 ROLL

Must build a DODAG (destination oriented DAG) whose root is the sink or edge router. The router starts the procedure by sending DIO messages; if a node is also capable of routing, it will also forward the DIO. When the DIO message is received, a node must tell the root that he is joining the network sending a DAO message that will travel up to the root; the root can solicite nodes to send DAO using DIS messages.
As for the routing, there are two modes: storing and non-storing. In storing mode, routers know routes and packets will need to travel only to most common ancestor of the destination (because that ancestor will know how to route to the destination). Otherwise, only the root knows routes and packets must travel back to him to be routed.
This protocol is used in ipv6, and DIS,DAO,DIO are ICMPv6 messages; if 6loWPAN uses ipv6 for routing (so it communicates to external networks), then ROLL is employed.

14 Insights

Warning: this section is optional and just for your personal knowledge; the professor either didn't explain these topics in depth or doesn't require them in the exam.

14.1 Tandem Free Operation

In GSM networks, a call between two MSs involve a dual encoding/decoding process; speech signals are first encoded in the originating MS, converted to PCM in the local transcoder (PCM coding is required for communications that go in networks different from GSM ones, such as PSTN), converted back to a GSM codec (e.g. AMR) in the remote transcoder and finally converted back to speech at the terminating MS. In this configuration the two transcoders are operating in tandem introducing a voice quality degradation. It is possible to eliminate this problem by removing the two transcoding operations in the voice path if the two MS are using the same codec: TFO steals the least significant bit of PCM encoding to send the GSM coded information, and the receiving MS will just read these less significant bit to have the GSM better quality information. Notice that stealing the less significant bits means that if they're lost for some reason, the most significant bits can still be used to carry the PCM signal. Why the GSM encoding has a better quality of voice if PCM has higher bit rate? Because PCM encodes an encoding, and this degrades the quality. How is it possible that the least significant bits can be used for GSM codecs? It's possible thanks to the high data rate of PCM, 64kb/s: GSM codecs works at a data rate as low as 8 kb/s and at a maximum of 16 kb/s, so these information can be included.