University of Technology
الجامعة التكنولوجية

Computer Science Department
قسم علوم الحاسوب

Mobile and Network Security

أمنية الموبايل والشبكات

Lect. Wisam ALI Mahmood
م. وسام علي محمود

2025-2024

cs.uotechnology.edu.iq

References:

1. **Wireless and Mobile Network Security,Hakima Chuouchi, Maryline Laurent Makharicius, Wiley, 2009.**

2. **Mobile Computing Principles, Reza R. Far, 2005.**

Introduction to Mobile and Wireless Networks

Chapter 1

Introduction

Wireless networks in small or large coverage are increasingly popular. The first major success of wireless networks is rendered to Wi-Fi (IEEE 802.11), which opened a channel of fast and easy deployment of a local network. Other wireless technologies such as Bluetooth, also show a very promising future given the high demand of users in terms of mobility and flexibility to access all their services from anywhere.

Mobile cellular networks

1. Introduction

The first generation (1G) mobile network developed in the USA was the AMPS network (Advanced Mobile Phone System). It was based on FDM (Frequency Division Multiplexing). A data service was then added on the telephone network, which is the CDPD (Cellular Digital Packet Data) network. It uses TDM (Time Division Multiplexing). The network could offer a rate of 19.2 kbps and exploit periods of inactivity of traditional voice channels to carry data.

The second generation (2G) mobile network is mainly GSM (Global System for Mobile Communications). It was first introduced in Europe and then in the rest of the world. Another second-generation network is the PCS (Personal Communications Service) network or IS-136 and IS-95; PCS was developed in the USA. The IS-136 standard uses TDMA (Time Division Multiple Access) while the IS-95 standard uses CDMA (Code Division Multiple Access) in order to share the radio resource. The GSM and PCS IS-136 employ dedicated channels for data transmission.

The ITU (International Telecommunication Union) has developed a set of standards for a third generation (3G) mobile telecommunications system under the IMT-2000 (International Mobile Telecommunication-2000) in order to create a global network. They are scheduled to operate in the frequency band around 2 GHz and offer data transmission rates up to 2 Mbps. In Europe, the ETSI (European Telecommunications Standards Institute) has standardized UMTS (Universal Mobile Telecommunications Systems) as the 3G network.

The fourth generation of mobile networks based on both mechanisms of cellular networks and wireless networks of the IEEE or a combination of both. The ITU has stated the flow expected by this generation should be around 1 Gbps static and 100 Mbps on mobility regardless of the technology or mechanism adopted.
Despite their diversity, their goal has always been the same; to build a network capable of carrying both voice and data respecting the QoS (quality of service), security and above all reducing the cost for the user as well as for the operator.

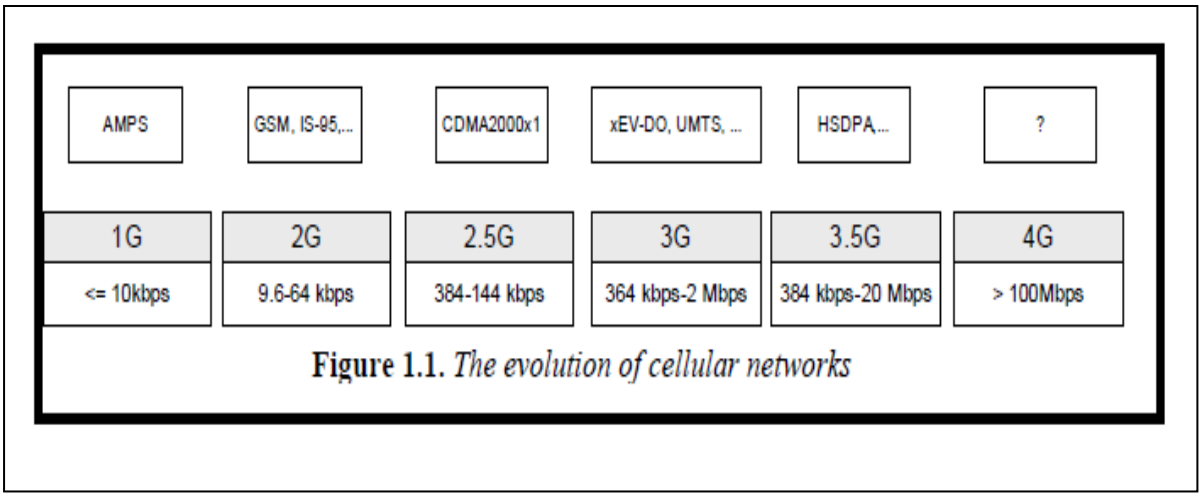


Figure 1.1. *The evolution of cellular networks*

2. *Cellular network basic concepts*

a) Radio resource
Radio communication faces several problems due to radio resource imperfection. In fact the radio resource is prone to errors and suffers from signal fading. Here are some problems related to the radio resource:

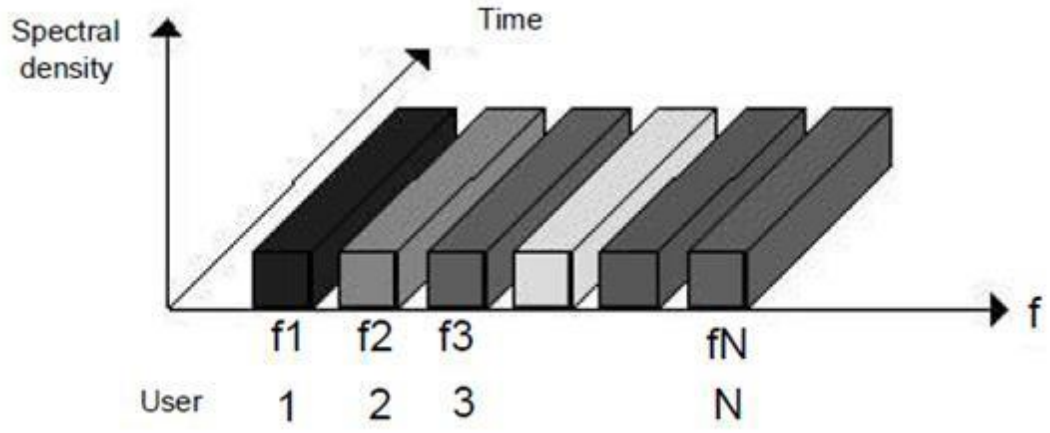
1. **Power signal**: the signal between the BS (base station) and the mobile station must be sufficiently high to maintain the communication. There are several factors that can influence the signal (the distance from the BS, disrupting signals, etc.).

2. **Fading**: different effects of propagation of the signal can cause disturbances and errors. It is important to consider these factors when building a cellular network. To ensure communication and to avoid interference, cellular networks use signal strength control techniques. Indeed, it is desirable that the signal received is sufficiently above the background noise. For example, when the mobile moves away from the BS, the signal received subsides. In contrast, because of the effects of reflection, diffraction and dispersion, it can change the signal even if the mobile is close to the BS. It is also important to reduce the power of the broadcast signal

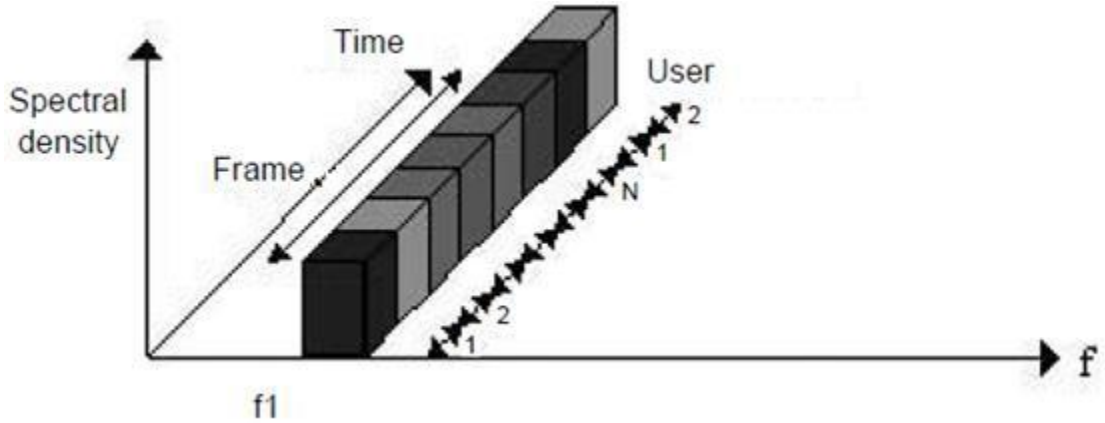
from the mobile not only to avoid interference with neighboring cells, but also for reasons of health and energy.

As the radio resource is rare, different methods of multiplexing user data have been used to optimize its use:

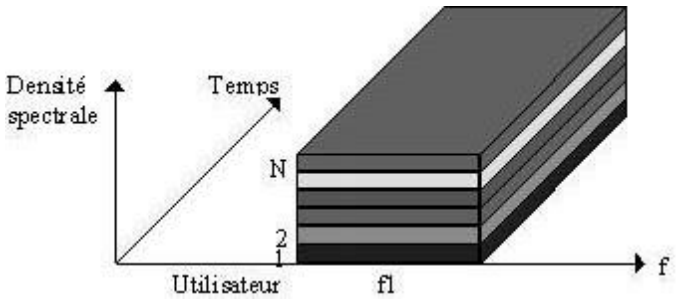
a. *FDMA* (Frequency Division Multiple Access) is the most frequently used method of radio multiple access. This technique is the oldest and it allows users to be differentiated by a simple frequency differentiation. Indeed, to listen to the user N, the receiver considers only the associated frequency fN. The implementation of this technology is fairly simple. In this case there is one user per frequency.



b. *TDMA* (Time Division Multiple Access) is an access method which is based on the distribution of the radio resource over time. Each frequency is then divided into intervals of time. Each user sends or transmits in a time interval from which the frequency is defined by the length of the frame. In this case, to listen to the user N, the receiver needs only to consider the time interval N for this user. Unlike FDMA, multiple users can transmit on the same frequency.



c. *CDMA* (Code Division Multiple Access) is based on the distribution code. It is spread by a code spectrum allocated to each communication. In fact, each user is differentiated from the rest of users with a code N allocated at the beginning of its communication and is orthogonal to the rest of the codes related to other users. In this case, to listen to the user N, the receiver has to multiply the signal received by the code N for this user.



b. Cell design

A cellular network is based on the use of a low-power transmitter (~100 W). The coverage of such a transmitter needs to be reduced, so that a geographic area is divided into small areas called cells. Each cell has its own transmitter-receiver (antenna) under the control of a BS. Each cell has a certain range of frequencies. To avoid interference, adjacent cells do not use the same frequencies, as opposed to two non-adjacent cells.

The cells are designed in a hexagonal form to facilitate the decision to change a cell for a mobile node. Indeed, if the distance between all transmitting cells is the same, then it is easy to harmonize the moment where a mobile node should change its cell. In practice, cells are not quite hexagonal because of different topography, propagation conditions, etc.
Another important choice in building a cellular network is the minimum distance between two cells that operate at the same frequency band in order to avoid interference. In order to do so, the cell's design could follow different schema. If the schema contains N cells, then each of them could use K/N frequencies where K is the number of frequencies allocated to the system.

The value of reusing frequencies is to increase the number of users in the system using the same frequency band which is very important to a network operator. In the case where the system is used at its maximum capacity, meaning that all frequencies are used, there are some techniques to enable new users in the system. For instance, adding new channels, borrowing frequency of neighboring cells, or cell division techniques are useful to increase system capacity. The general principle is to have micro and pico (very small) cells in areas of high density to allow a significant reuse of frequencies in a geographical area with high population.

c. Traffic engineering

Traffic engineering was first developed for the design of telephone circuit switching networks. In the context of cellular networks, it is also essential to know and plan to scale the network that is blocking the minimum mobile nodes, which means accepting a maximum of communication. When designing the cellular network, it is important to define the degree of blockage of the communications and also to manage incoming blocked calls. In other words, if a call is blocked, it will be put on hold, and then we will have to define what the average waiting time is. Knowing the system's ability to start (number of channels) will determine the probability of blocking and the average waiting time of blocked requests. What complicates this traffic engineering in cellular networks is the mobility of users. In fact, a cell will handle, in addition to new calls, calls transferred by neighboring cells. The traffic engineering model becomes more complex. Another parameter that is even more complicating for the model is that the system should accommodate both phone calls as data traffic, knowing that they have very different traffic characteristics.

d. Cellular system's elements

A cellular network is generally composed of the following:

– **BSs (Base Station):** situated at the heart of the cell, a BS includes an antenna, a controller and a number of transmitters and receivers. It allows communications on channels assigned to the cell. The controller allows the management of the call request process between a mobile and the rest of the network. The BS is connected to a mobile switching center (MTSO: Mobile Telephone Switching Office). Two types of channels are established between the mobile and the BS: the data channel and the traffic control channel. The control channels are used for associating the mobile node with the BS nearest to the exchange of information necessary to establish and maintain connections.

The traffic channels used to transport the user traffic (voice, data, etc.).

– **Mobile Switching Center (MTSO):** a MTSO manages several BSs generally bound by a wired network. It is responsible for making connections between mobiles. It is also connected to the wired telephone network and is thus able to establish connections between mobiles and fixed nodes. The MTSO is responsible for the allocation of channels for each call request and is also responsible for handover and recording the billing information of active call users.

The call process includes the following functions:

– **Initializing a mobile**: once the mobile node is turned on, it scans the frequency channels, and then it selects the strongest control call channel (setup). Each cell regularly controls the information on the band corresponding to its control channel. The mobile node selects the channel whose signal is the most important. Then the phone goes through a phase of identification with the cell (handshake). This phase occurs between the mobile and the MTSO. The mobile is identified following an authentication and its location is recorded. The mobile continues to regularly scan the frequency spectrum and decides to change the BS if it has a stronger signal than the previous cell phone. The mobile node also remains attentive to the call notification.

– **Call initiated by a mobile node:** the mobile node checks that the call channel is free by checking the information sent by the BS on the downlink control channel. The mobile may then issue the call number on the uplink control channel to the BS that transmits the request to MTSO.

– **Call notification**: the phone number is received, the switching center tries to connect to BSs concerned by the number and sends a call notification message to the called mobile node (paging). The call notification is retransmitted by BSs in the downlink control channel.

– **Acceptance of call:** the mobile recognizes its number in the call control channel and then responds to the BS to relay the message to the switch that will establish a circuit between the BSs of the calling and the called nodes. The switch will also select an available traffic channel in each of the two cells involved and sends the information related to that call to the BSs. The phones will then synchronize the traffic channels selected by the BS.

– **Active communication:** this is the process of exchanging data or voice traffic between the calling and called mobiles. This is assured by both BSs and the switching center.

– **Call blocking:** if all channels of traffic in a BS are occupied, the mobile will try a number of pre-configured times to repeat the call. In case of failure, an “occupied”

signal tone is returned to the user.

– **Call termination:** at the end of a communication, the switching center informs the BSs to free channels. This action is also important for billing.

– **Abandonment of call:** during a communication, if the BS fails to maintain a good level of signal (interference, low signal, etc.) it abandons the channel traffic of the mobile and notifies the switching center.

– **Call between a fixed terminal and a mobile node**: the switching center being connected to the landline or fixed network, it is then able to establish communication between these two networks. It can also join another mobile switching center through the fixed network.

– **Handover (Handoff):** when the mobile discovers a control channel where the signal is stronger than its current cell, the network will automatically change to the cell by transferring its mobile channel call to the new cell without the user noticing.

The main criterion used to take the decision to transfer the mobile is the measured signal power of the mobile node by the BS. In general, the station calculates an average over a time window to eliminate the rapid fluctuations resulting from multipath effects. Various techniques can be used to determine the moment of transfer of the mobile. In addition, this transfer can be controlled by either the network or the mobile. The simplest technique of handover decision is one that triggers the transfer as soon as the mobile detects a new signal stronger than the cell where it is connected.

Mobile generation

a. First generation (1G) mobile

First generation cellular networks such as CT0/1 (Cordless Telephone) for wireless and AMPS (Advanced Mobile Phone Service) for mobile communications were first characterized by analog communications. The first cellular networks are virtually non-existent today. The AMPS system was the 1st generation of the most widespread used network in the USA up to the 1980s. It has also been deployed in South America, Australia and China. In Northern Europe, the NMT (Nordic Mobile Telecommunications System) was developed. In the UK, the TACS (Total Access

b. Second generation (2G) mobile

Cellular networks such as second generation DECT for wireless and mobile phones for mobile were characterized by digital communications networks, unlike the first generation, which were analog. During the 1990s several digital technologies were developed:

– GSM (Global System for Mobile Communication), developed in Europe, operating at 900 MHz.

– DCS 1800 (Digital Cellular System) equivalent to GSM but operating at higher frequencies (1,800 MHz).
– PCS 1900 (Personal Communication System) and D-AMPS (Digital AMPS) developed in the USA.
– Finally, PDC (Pacific Digital Cellular) developed in Japan.

c. Third generation (3G) mobile

3G cellular networks operate around the frequency band of 2 GHz, providing a range of multimedia services to fixed and mobile users with a Quality of Service almost comparable to that of fixed networks. The International Telecommunications Union (ITU) has selected five standards for 3G mobile under the symbol IMT-2000 (International Mobile Telecommunications system for the year 2000). This is the WCDMA
(Wideband CDMA), TD-CDMA and TD-SCDMA standard used in the
European UMTS (Universal Mobile Telecommunication System) of CDMA2000,
EDGE (Enhanced Data rate for GSM Evolution) and the third generation of DECT.

1.3 IEEE wireless networks

1. Introduction

Many standards for wireless communication are being developed day after day and the price of their equipment becomes increasingly attractive. This will contribute to the success of these technologies. In this section, we introduce the standards that are the basis of many wireless networks.

2. *WLAN: IEEE 802.11*

The IEEE 802.11 standard describes the wireless area network characteristics.
Wi-Fi (Wireless Fidelity) corresponds initially to the name give to a certification delivered by the Wi-Fi Alliance which is a consortium of separate and independent companies that agrees on a set of common interoperable products based on the family of IEEE 802.11 standards.

The IEEE 802.11 can operate in two modes: infrastructure and ad-hoc. In the ad hoc mode or infrastuctureless mode, two WLAN stations can communicate directly with each other whenever they are in the same range spectrum without the intervention of the access point. Each WLAN station can be considered as an access point and a client station at the same time. However, in the infrastructure mode, the wireless network is controlled by the access point which is equipped with two interface networks:

– One wireless interface by which it receives all the exchanged frames in the cell and over which it retransmits the frames to the destination station in the cell.

– The second interface, which is ethernet, is used for communication with other access points or used for accessing the Internet. The set of all WLAN stations that can communicate with each other is called the basic service set (BSS). The distribution system (DS) connects more than one BSS and forms an extended service set. The concept of a DS is to increase network coverage through roaming between cells.

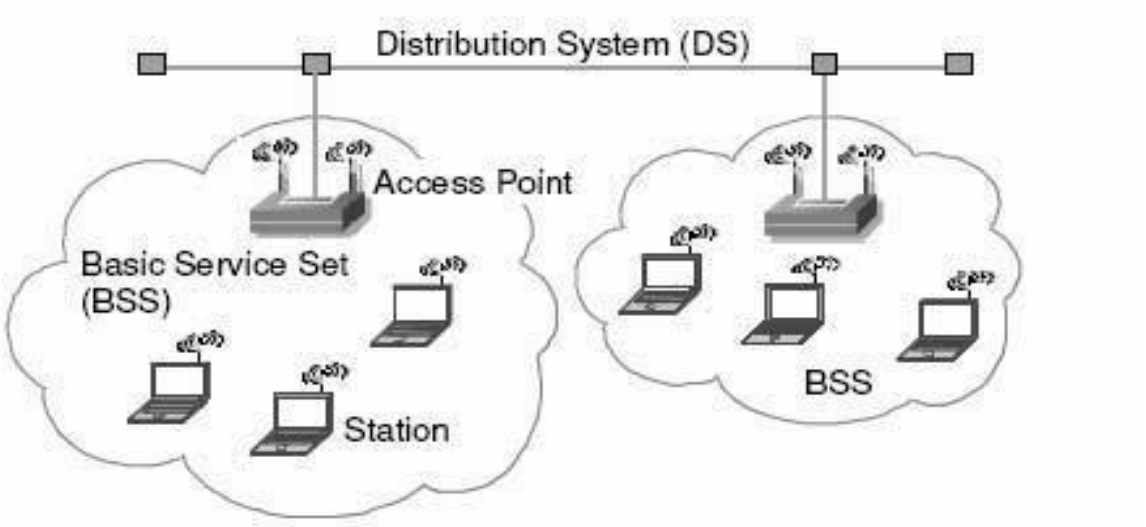


Figure 1.5 WLAN-infrastructure mode

a) Wi-Fi architecture

the IEEE 802.11 specifications address both the Physical (PHY) and Media Access Control (MAC) layers and are tailored to resolve compatibility issues between manufacturers of WLAN equipment. The MAC layer can be a common layer for the different types of physical layer adopted by this standard. This can be done without any modification to the MAC layer.

b) The PHY layer

Three PHY layers were defined initially for IEEE 802.11:

1) DSSS (Direct Sequence Spectrum): the principle of this is to spread a signal on a larger frequency band by multiplexing it with a signature or code to minimize localized interference and background noise. To spread the signal, each bit is modulated by a code. In the receiver, the original signal is recovered by receiving the whole spread channel and demodulating with the same code used by the transmitter. The 802.11 DSSS PHY also uses the 2.4 GHz radio frequency band.

2) FHSS (Frequency Hopping Spread Spectrum): this utilizes a set of narrow channels and "hops" through all of them in a predetermined sequence. For example, the 2.4 GHz frequency band is divided into 70 channels of 1 MHz each. Every 20 to 400 ms the system "hops" to a new channel following a predetermined cyclic pattern. The 802.11 FHSS PHY uses the 2.4 GHz radio frequency band, operating at a 1 or 2 Mbps data rate.

3) Infrared: the Infrared PHY utilizes infrared light to transmit binary data either at 1 Mbps (basic access rate) or 2 Mbps (enhanced access rate) using a specific modulation technique for each. For 1 Mbps, the infrared PHY uses a 16-pulse position modulation (PPM). The concept of PPM is to vary the position of a pulse to represent different binary symbols. Infrared transmission at 2 Mbps utilizes a 4 PPM modulation technique.

c) MAC layer and channel access method

The principal function of the MAC layer is to control the access to the medium. The IEEE 802.11 adopted two algorithms of controlling access to the channel: DCF (Distributed Coordination Function) and PCF (Point Coordination Function). The default method of access is DCF, which is designed to support asynchronous best effort data. Nowadays, the IEEE 802.11 works on this mode only. Fundamentally, the DCF deploys the CSMA/CA (Carrier Sense Multiple Access/Carrier Avoidance) algorithm. The most important part of this algorithm is the process of backoff which is applied before any frame transmission.

Whenever a WLAN station wants to sent data, it first senses the medium. If the later is idle, then the WLAN station will transmit its data, otherwise it changes its transmission. After detecting the medium being idle over a period of time DIFS (Distributed Interframe Spaces), the WLAN station will continue to listen to the medium during a supplementary random time called the backoff period. The frame then will be transmitted if the medium is idle after the expiration of the backoff period.

The duration of back off is determined by the CW (Contention Window) which has a value bounded by [CWmin, CWmax] maintained separately in each WLAN station in the BSS. A slotted back off time is generated randomly by each WLAN station in the interval of [0, CW]. If the medium is still idle, the back off time will be decremented slot by slot and this process will be continued as long as the medium is idle. When the back off time reaches 0, the WLAN station will transmit the frame. If the medium is occupied during the process of back off, the countdown to back off will be suspended. There it restarts with the residual values when the medium is idle for one consecutive DIFS. Whenever the frame received well by the recipient, the latter will send an acknowledgement (ACK) message to the sender. If the WLAN station does not receive the ACK, it deduces that there were a collision and in order to avoid

consecutive collisions, it will retransmit the same frame. The value of the CW will be doubled in the case of transmission failure.

The PCF method, also called the controlled access mode, is based on a polling method which is controlled by the access point. A WLAN station cannot transmit if it is not authorized and it cannot receive only if it is selected by the access point. This method is conceived for the real-time applications (voice and video) that demand delay management when transmitting data. This system is reservation-based access.

1.4 Mobile Internet networks

1. ***Introduction***

IP routing was designed without support for mobile nodes and was defined for fixed nodes. IP mobility has been made possible thanks to developments in wireless networks as well as developments in the miniaturization of portable and mobile terminals. IP mobility introduces new features in the network to ensure continuity of routing for mobile nodes on the move. ***These features are addressing, location management, re-routing and handover of the mobile's node***:

– ***Addressing*****:** in an IP network, support for mobile nodes requires two IP addresses: a fixed address of the mobile node, which is related to the home network that serves as an identification of the mobile node, and a temporary address that is related to the visited network. The temporary address changes as the mobile node moves from one temporary network to another. The temporary address is produced each time by a visited network.

– ***Location management***: a correspondence is maintained in the network between the fixed address and the temporary address of the mobile node. This correspondence is conducted by a new entity in the network, a mobility agent. The mobile node must securely send its new temporary address for the mobility agent to maintain the correspondence between the temporary address and the permanent address of the mobile node and thus can locate it in order to forward its traffic to its current location.

– ***Re-routing***: when the mobile node has an active session during its trip, it is the responsibility of the network to route the traffic to its new destination without interrupting the session.

– ***Handover*****:** the handover is the process of changing the point of attachment to the network. It contains the *discovery phase* of the new visited network and *attachment* to this new network. The handover is difficult when there is an ongoing session because the whole issue is to change the point of attachment without interrupting the session. From a performance protocol developed by the Internet Engineering Task Force (IETF) to support IP mobility (Mobile IP) creates too much latency

in micro mobility. In order to do this, supporting **IP mobility has been divided into two categories:** support for macro mobility and support for micro mobility.

Macro mobility happens when a mobile node moves between two different areas. Macro mobility can take place during an active session of a mobile user, or during a new session initiated by the user from a visited network in a new domain which is known as a nomadic or roaming user.

Micro mobility concerns a mobile node moving between two points of attachment belonging to the same area. Active research on this subject has raised several propositions; however, all lack efficient standards. The basic features of a mobility management are mainly management of the location and management of handovers.

Chapter 2
Vulnerabilities of Wired and Wireless Networks

2.1. Introduction

This section synthesizes the vulnerabilities common to modern telecommunications systems. The presented synthesis is general in the sense that it does not depend on particularities of any specific system like topology, form, used media, implementation, etc.

2.2.2. *Definition of security*

It is necessary to find a definition of security common to an asset, a service, an infrastructure and an info sphere for any concerned owner. We can find several definitions of security:

Security

noun (pl. securities) 1. the state of being or feeling secure. 2. the safety of a state or organization against criminal activity such as terrorism or espionage. 3. a thing deposited or pledged as a guarantee of the fulfillment of an undertaking or the repayment of a loan, to be forfeited in case of default. 4. a certificate attesting credit, the ownership of stocks or bonds, etc.

The owner wants to minimize risks and imposes counter-measures that he considers necessary to protect the asset (see Figure 2.2). He therefore describes the security objective.

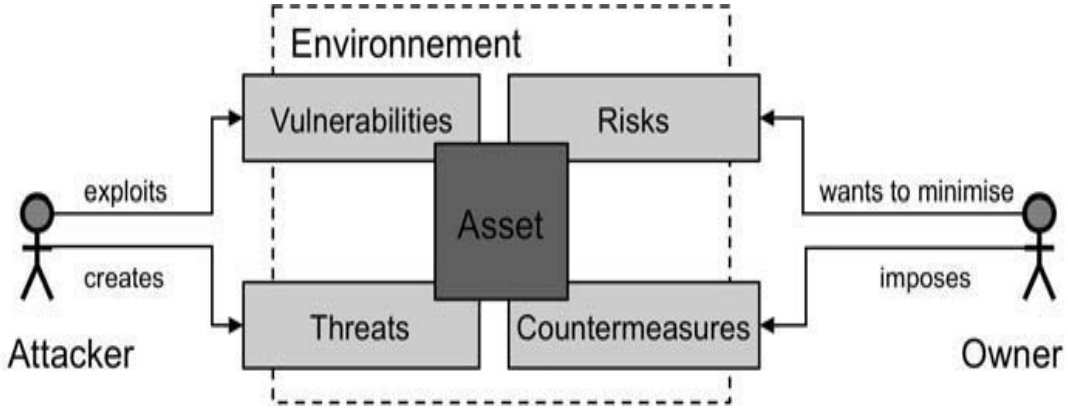


Figure 2.2. *Relationships between asset, attacker and owner*

2.3.2. *Threat models in telecommunications systems*

Threat models first describe the system, all actors in this system and their position in the system (for example, link, node). Then, the threat model introduces an attacker in the system and demonstrates the attacker's capacities, i.e. topological position in the system, resources, possible access, etc.

The traditional threat model to a communication channel is based on the minimal communication model minimalist involving two participants called Alice and Bob, and a communication channel (see Figure 2.3)

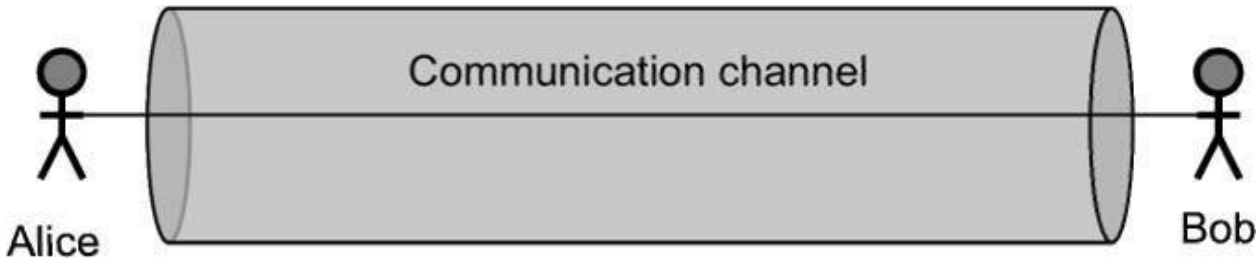


Figure 2.3. *A minimal communications model*

This model usually presumes an initial trust relationship between Alice and Bob. It is often used in cryptography, since it effectively limits possible attacks to the attacks against the communication channel between Alice and Bob. Yet, in the context of telecommunications systems, this model is not exhaustive, since other elements and vulnerabilities are present. Figure 2.4 introduces a more appropriate model, distinguishing between the two communicating parties (Alice and Bob) and at least one telecommunications infrastructure and its authority crossed by the communication channel. In general, this authority is neither Alice nor Bob but a real third party.

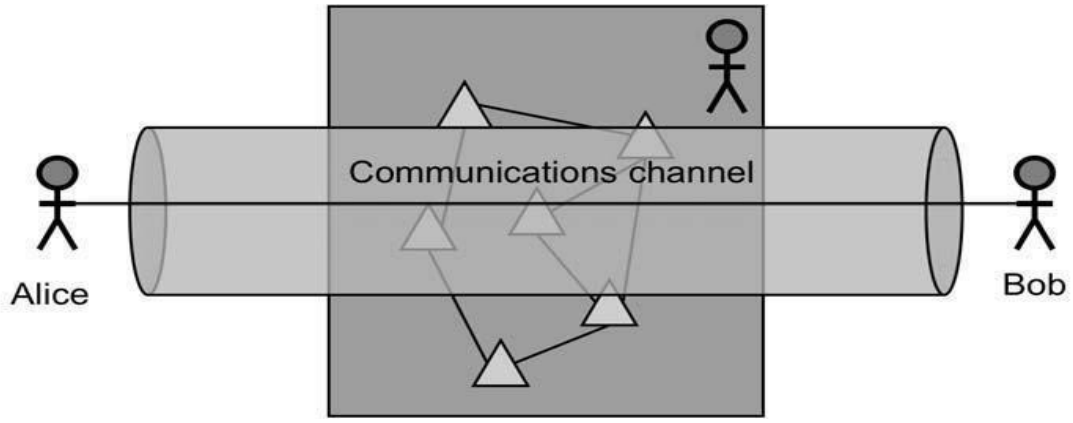


Figure 2.4. *Communications model with a telecommunications system*

The emergence of such a third party increases the complexity of the system, introduces new interfaces and vulnerabilities and may require a more complicated trust chain. It therefore widens the spectrum of possible threats. The trust model of Figure 2.4 may have very different forms, but in practice we assume one of the following:

– Alice and Bob trust each other in the sense of the intended communications, and they both trust the used telecommunications system to correctly provide the services (private network).

– Alice and Bob trust each other, but do not trust the crossed infrastructure (public network).

– Alice and Bob trust the telecommunications infrastructure but do not trust one another; they will use the infrastructure as a trusted third party (TTP) to establish a new trust relationship.

Figure 2.5 presents from left to right the typical threats against the actors and parts of this model. In the following, the attacker is denoted as Eve:

– Eve can attack one of communicating parties (Alice in the example) using vulnerabilities in the software and protective measures used by Alice. Strictly speaking, this threat is not related to the telecommunications system. However, a terminal with a connection interface to a telecommunications system is a more open entity and is thus more vulnerable. Often, attacks are possible because of vulnerabilities in the terminal and the visibility of the terminal involved in a telecommunications service. A typical example is the execution of malicious code on the platform used by Alice through a virus or by the overflow of reception buffers.

– Alternatively, Eve can attack the communication channel linking Alice to the telecommunications system. This attack may be non-intrusive (reading the

exchanged data) or intrusive (modification of exchanged data, injection of data, replay of old data). The possibility of such an attack depends on the channel. For example, a wireless channel is potentially more vulnerable against passive listening by a third person than a network cable, which normally at least requires physical access to the medium.

– Another possible attack against the channel exists within the telecommunications system. To do this, a form of access to the telecommunications system is normally required. If Eve is not the owner of the system, Eve may try to masquerade as a legitimate part of the infrastructure to attract Alice (or Bob) to use its services. In some cases, Eve can get physical access to communication channels forming part of the system or use system vulnerabilities to gain access to system components.

These forms of access can allow Eve to collect information on communications between Alice and Bob and to manipulate the data flow between the two.

– The intrusion into the infrastructure permits to mount "man in the middle" attacks. In this scenario, Eve positions as a junction point between Alice (or Bob) and the infrastructure such that all communications of Alice (or Bob) to the infrastructure traverse Eve. Without reliable and mutual authentication (i.e. identity verification) between Alice (or Bob) and the infrastructure, Alice and the infrastructure cannot find this kind of intrusion. However, "man in the middle" attacks are also possible if Eve can usurp the identity of the communicating opponent. To counter these attacks, mutual and reliable authentication between Alice and Bob becomes necessary.

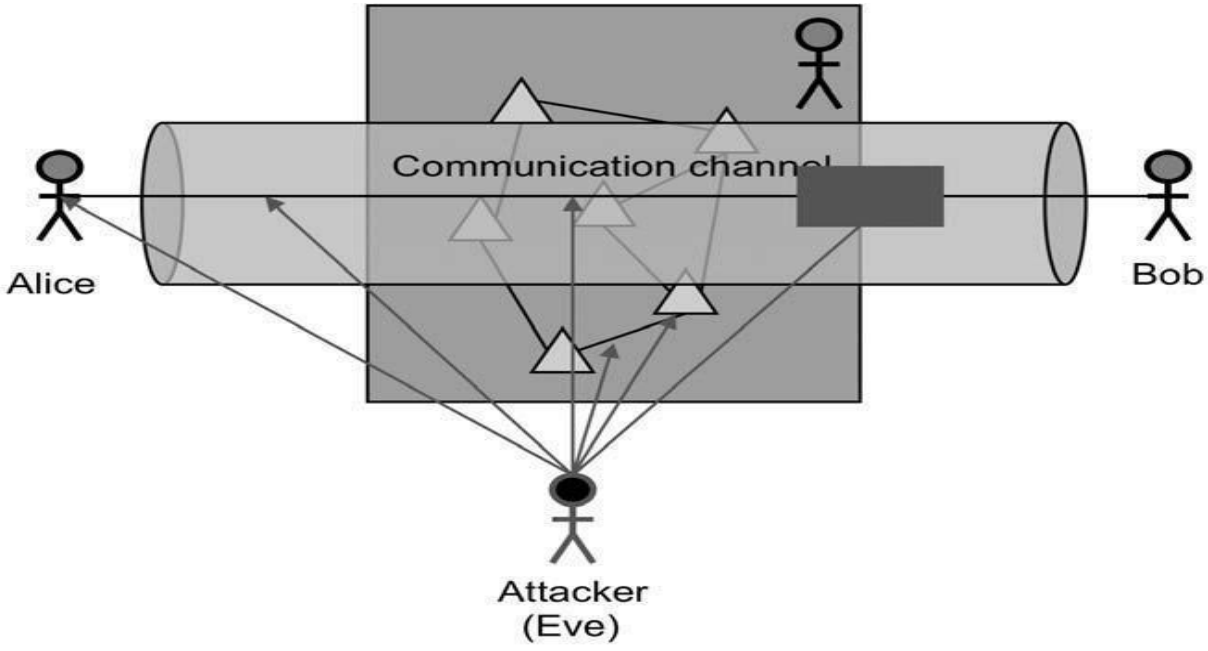


Figure 2.5. *Threat model for a telecommunications system and its participants*

From wireline vulnerabilities to vulnerabilities in wireless communications

We draw attention to the fact that we are talking about possible and not certain threats. Indeed, any technology can be used in a way such that a given vulnerability does not have any real impact. On the other hand, it is perilous to believe that we can impose constraints on the use of any technology in modern society.

1. *Changing the medium*

The wireless medium, in the sense of radio-based broadcast transmissions, is very vulnerable by nature, much more vulnerable than the wired medium. The medium allows wireless access for any actor: reading, injection, deletion and modification of data are possible for any actor in most configurations. In addition, all communications are purely virtual: generally, we cannot delimit the perimeter of the network (because of physical properties of the medium: the signal attenuation is strong, but the multi-path propagation, reflections/refractions, etc. often produce surprising results), nor can we distinguish different connected terminals. In other words, the medium does not limit the circle of actors involved in the processing of exchanged data. It does not detect whether access to the medium or to the transmitted data took place during the transmission.

For an attacker, such a wireless medium is often more attractive because it does not require the physical presence of the attacker. Well equipped, an attacker is able to mount attacks against natural medium vulnerabilities while remaining outside of the attacked area. In addition, attacks can be easily automated or at least semi-automated: the equipment can record all received frames for spying on the encountered wireless infrastructures , even without exploiting any particular flaws in the security measures usually implemented in such networks (mainly access control, confidentiality and integrity).

To overcome the transmission problems related to reliability and security of communications, management and control systems, finite state machines and protocol stacks used in these networks often exhibit elevated complexity

2. *Wireless terminals*

The terminals used in wireless networks are characterized by their portability. They are small, often equipped with a restrained human-machine interface (HMI), limited in terms of processing power and storage and powered by a battery. These

characteristics have a significant impact on the security of the terminal, and, by extension. The limited user interface often poses problems in pairing and access control phases (how to enter a password in a pair of headphones, how to establish a unique identity of a USB key, etc). Limited computing (CPU) and storage (memory, disk) capabilities introduce constraints with regard to possible calculations. For example, it is arduous to base the security of an embedded device, for example, of a sensor or of a mobile phone without a dedicated module, on public key cryptography, since the necessary private key computations take too much time and energy. Battery management implies several changes in the behavior (unexpected on/off, technically close to mobility) and requires additional management systems (standby management, mechanisms for paging, etc). In addition, the development of battery technology is constant but linear. This problem can be overcome through a high quality circuit design, sophisticated standby management and protocols and complex adaptive power management, complicating the terminal and making it potentially more vulnerable.

3. ***New services***

Beyond these aspects, networks based on radio transmissions add a degree of freedom to every transmission: the spatio-temporal context. With wireless communications, it is reasonable to talk about mobility, location of users connected through this medium. This new freedom justifies the implementation of new services for mobility or localization support (location-based services, etc). These new services are not reserved to wireless communications, but practice shows that with wireless technology they become truly interesting: mobility is not limited to wireless and wireless does not imply mobility, but in reality there is a considerable overlap.

Mobility represents a known problem for security considerations, not only because it introduces new mechanisms and subsystems and therefore results in a higher complexity, but before all through the presence of several potential authority domains. This complicates the chains of trust. It is often necessary to provide services to users from another authority domain, subject to a different security policy. Since security policies are not directly comparable, this often results in irresolvable requirements and cannot be realized. However, even the reception of mobile users from the same authority domain is complicated: the network has to verify that, after a period of absence, the configuration of the mobile user is still consistent with the security policy requirements of the domain. In practice, this often results in drastic measures regarding the access rights of mobile users, or mobile users have to pass through a quarantine period prior to regaining full user rights. As a result, mobile systems are normally more vulnerable, both from the

point of view of users and of operators. It is difficult to fulfill all security requirement in the CIA sense, but it is even more difficult to implement correct non-repudiation properties, a sufficient traceability (e.g. for billing) not leading to new abuse (anonymity, respect for the privacy requirements), to verify the lack of viruses, malware, etc., on the mobile terminal. The security of mobility must be treated with great caution. The problem is that the security mechanisms often become active simultaneously to the typical mobility mechanisms, like, handover treatment. These mechanisms interfere and extend handover delays;

Chapter 3
3. Fundamental Security Mechanisms

1. Basics on security

A) *Security services*

Security services refer to security concepts contrary to security mechanisms which include the set of cryptographic tools useful for implementing security services. The X.800 standard [X800] defines the security services (except for replay detection) as follows:

– **Availability:** "the property of being accessible and useable upon demand by an authorized entity";

– **Access Control**: "the prevention of unauthorized use of a resource, including the prevention of use of a resource in an unauthorized manner";

– **Data Integrity**: "the property that data have not been altered or destroyed in an unauthorized manner";

– **Data Origin Authentication**: "the corroboration that the source of data received is as claimed";

– **Peer Entity Authentication**: "the corroboration that a peer entity in an association is the one claimed". Note the clear distinction between "identification" and "authentication". Identification refers to an entity (user, equipment) claiming its identity by providing an identifier (name, pseudonym, email address, IP address, domain name), or the procedure to find the Identity of a user among N users known by the systems under several features. Authentication consists of proving the claimed identity by providing one or several authentication elements;

– **Confidentiality:** “the property that information is not made available or disclosed to unauthorized individuals, entities or processes”;

Encryption mechanisms enable the implementation of data confidentiality. Most of the time, the services of data integrity and data origin authentication are implemented by the same security mechanisms: hash function and MAC generation (see section 3.2.4).

b) *Symmetric and asymmetric cryptography*

Since the 1970s, two cryptography families emerged [SCH 96]. In symmetric cryptography, the enciphering and deciphering systems know the same cryptographic key, while asymmetric cryptography (known as public key cryptography) is based on two complementary keys – the public and private keys – one of them for encrypting and the other one for decrypting. Both families are hereafter described with a few examples of algorithms that are commonly used today, their advantages and drawbacks, as well as their complementarities.

Note that older cryptographic algorithms were based on the secret of the algorithm itself. This means that as soon as the algorithm was cracked, the cryptographs needed to invent a new one. The novelty of symmetric and asymmetric algorithms was to make public the whole enciphering/deciphering processing and to externalize the secret into a secret parameter also called the “cryptographic key”.

d. Symmetric cryptography

Symmetric cryptography is based on the usage of the same key to encrypt and decrypt data. These keys are called symmetric keys (sometimes secret keys). In the context of exchanges over a network, a transmitter encrypts data with a key and the destination entity decrypts the data with the same key. If the symmetric algorithms are efficient, and make it possible to reach a high data rate when encrypting/decrypting, they raise the problem of establishing the same key between the transmitter and the receiver, however. Sharing a key with each possible communicating entity, even in a closed group of entities, is a very high constraint, and rapidly leads to a big number of keys to be managed. Thus, it is better automating the establishment of these keys (see section 3.2.2.3).

The most well known symmetric algorithms are, in the chronological order of their definition: DES (Data Encryption Standard), 3DES (pronounced “Triple DES”), and AES (Advanced Encryption Standard). DES was invented in 1977 by IBM as the public encryption algorithm with secret keys of 56 bits and input of 64-bits data blocks. DES is based on permutation mechanisms and exclusive OR gates.
These fast operations make DES highly efficient, but brute force attacks are still able to crack the 56-bits keys by trying any combinations of keys. The 3DES algorithm was more robust and was successor of DES; 3DES applies DES three times, one after the other; the 3DES key is maximum 168 bits (3*56=168) and applies to the same input block size (64 bits). The 3DES algorithm is not always efficient from an encryption rate point of view, robustness to brute-force attacks, etc., so an international competition was launched in 1997 to elaborate a new algorithm to replace 3DES. After several selection steps, the Belgium Rinjdael algorithm was selected for its fast processing time, its portability on several platforms (hardware and software, 8 and 32 bits), several supported key lengths, etc. Thus, we talk indifferently about AES or Rinjdael. AES is the generic name of the algorithm winning the competition. It relies on inputs of 128-bit blocks and key lengths of 128, 192 or 256 bits.

Symmetric algorithms can work according to several modes. Usually we distinguish the ECB (Electronic Code Book) mode which consists of encrypting each of the blocks independently. Thus, the operation of encrypting a message consists of fragmenting a message into blocks of the expected size (dependent on the selected algorithm). Each of these blocks is then encrypted independently. The drawback of the ECB mode is that two similar blocks give similar encrypted blocks outputs. This makes ECB vulnerable, as a spy on the network can detect two similar encrypted blocks, attempt to guess their contents, and perform a brute-force attack by testing any combination of the keys until the assumed clear text message is obtained. The CBC (Cipher Block Chaining) mode, also known as the “chaining mode”, consists of processing a block by injecting the last computed ciphered block into the processing, such that encrypting two similar blocks leads to two different blocks. Attacks are more difficult to implement.

Asymmetric cryptography or public key cryptography

Asymmetric or public key cryptography relies on two encryption keys, called “asymmetric keys”. Both keys are generated at the same time and play a complementary role in that the encryption with one of the keys needs to be decrypted with the other key. Each key plays a specific role. The private key must

be known by only one entity and can be used for authenticating itself for instance. The public key can be largely published and it is better that public keys are largely published in order that any other entity can perform authentication. Obviously, knowledge of the public key does not enable us to deduce the complementary private key.

To authenticate the origin of message in a communication over a network, the transmitter must use its own private key, for instance to generate an electronic signature (see section 3.2.4) that it will append to the message before transmission. The receiver who knows the public key will be able to verify the validity of the signature and will have guarantees regarding the origin of the message. To ensure the confidentiality of a message, it is necessary to encrypt the transmitted message with the public key of the receiver. This public key is known by all entities and can be served to any entity to encrypt a message. However, the complementary private key is only known to the destination of the message; the receiver will be the only one able to decrypt the message.

RSA (Rivest, Shamir, Adleman) is the most well known asymmetric algorithm. It is based on the theory of prime numbers and encryption keys with classic key sizes of 512 classic, 1,024, 2,048 or 4,096 bits. Today a good level of security refers to keys of at least 2,048 bits. If the robustness of cryptographic algorithms is dependent on the length of the keys used, it is meaningless comparing the robustness of symmetric and asymmetric algorithms with respect to the same length of keys. Indeed, cracking an asymmetric algorithm (in order to discover the used key) does not require testing all the possibilities for keys like the symmetric algorithms;

Complementariness between the two cryptographic systems

Security protocols use both security protocols, each having a specific usage: – symmetric cryptography (or secret key cryptography) makes it possible to protect high bit rate data exchanged over a network; the processing speed of symmetric algorithms is used;

– Asymmetric (or public key) cryptography is used to initialize a secure connection between two entities of the network by enabling those entities to authenticate each other and to establish a symmetric key in a confidential way.

Hash functions

Hash functions aim to give a representative result of the message's content over a limited number of bytes. They are pretty like a more sophisticated CRC (Cyclic Redundancy Check).
The awaited properties of these hash functions are as follows:

– A result on a limited number of bytes (usually 16 or 20 bytes);
– Inability to recover the original message from the outcome of the function;
– two messages that differ by only 1 bit produce two results that differ by at least half a bit. Several terms for hash functions like irreversible functions or one-way functions are used indifferently.

Secure communication protocols and VPN implementation

Several security protocols are defined to protect communications going through a network. Generally, these protocols are based on **two successive phases**, namely:

– **The initialization phase** where two entities mainly authenticate and negotiate services and security mechanisms in order to protect their data, and agree on one or more symmetric key(s).

– **The data protection phase**: the services and security mechanisms and symmetric keys that were previously agreed on during the initialization are activated to protect data exchanges.

In this section, two popular security protocols, **IPsec and SSL,** are presented. For each of them, the two phases of operation are presented with the associated protocols, the security services supported to protect the exchanges during the two phases and the processing done over data for their protection. **A comparison of these two protocols and the possible usage of them in a VPN (virtual private network)**

1. *Secure Socket Layer (SSL) and Transport Layer Security (TLS)*

The original SSL [RES 01] was designed by Netscape Communications to protect e-commerce applications based on HTTP and became very popular because of its

systematic integration in Microsoft browsers Internet Explorer and Netscape Navigator. SSL is in the form of an additional protocol layer between the application and transport layers, Thanks to its position in the protocol stack, SSL can support protection for any TCP-based applications like ftp, telnet, smtp, etc. To enable SSL security, applications are needed to call a secure socket connection setup instead of a standard socket setup.

2. *IPsec protocol suite*

The IPsec protocol suite (IP security), In 1995, there were some IPsec products but they were not bought by industrials at that time. The IPsec market became increasingly larger after that time (after 1998). In 1998, the next series were strongly improved and were widely implemented in firewalls and other VPN equipment's /software. The latest series of 2005 introduced only a few improvements, thus avoiding the problems of compatibility between generations of IPsec equipment/software.

the IPsec protocol suite includes several protocols operating at various levels of the protocol stack:

– **Protocols protecting IP packets**. The two sub-protocols AH (Authentication Header) and ESP (Encapsulating Security Payload) make it possible to encrypt the contents of packets and/or to append a MAC. These sub-protocols are part of the IP layer.

– **Protocol implementing the initialization phase**. The IKE (Internet Key Exchange) protocol provides this role and takes the form of an application level module running over UDP port 500. It also negotiates all the security settings appropriate for the protection of data flows; this set of parameters is known as the "security association". Two types of protection are possible with IPsec. The tunnel mode makes it possible to define and manage an IPsec tunnel, while the transport mode provides direct protection of IP packets without additional encapsulation. The tunnel mode is mainly used in a VPN connecting two remote private networks. The transport mode is used to protect a simple connection

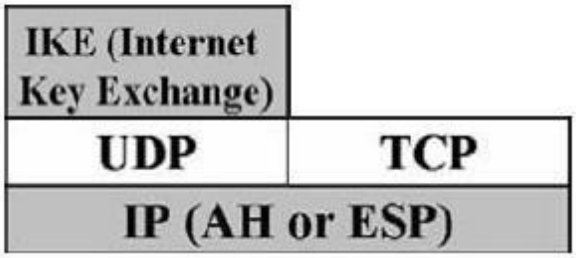


Figure 3.9. *Organization of the IPsec suite into layers*

Authentication mechanisms

Authentication mechanisms are becoming more and more sophisticated on the information system security market. They are designed to provide users and administrators with a certain ease of use (rapidity of authentication, simplicity of use), a minimal administration, great robustness (against possible intrusions), high reliability (to avoid authentication errors) and ubiquity of its usage. **These mechanisms, described in more detail in the following sections, can be divided into several categories according to:**

– what the entity knows, a password for example;
– what the entity owns, such as a smart card, a private key or a Kerberos ticket;
– what the entity is: this category covers the authentication techniques based on a user biometric features (fingerprint, iris, facial form, shape of hands, etc.);

Usually, the technical authentication solutions distinguish between a weak authentication and a strong authentication. For a weak authentication, an entity is authenticated with only one piece of authentication (e.g. password). Strong authentication plans to combine at least two elements of authentication, typically a password and a smart card. In order to diversify the authentication methods, there is a generic authentication protocol called EAP. This protocol is generic, in that it is independent of the authentication method. its role is limited to the transportation of authentication data between a client and a server. The content of these exchanges is not interpreted by the software layer EAP, but by the selected EAP method. As such, it brings the advantage that an EAP method suddenly detected as vulnerable can easily be changed to another more robust method while keeping the same EAP protocol. This makes the security equipment more flexible and able to evolve at low cost.

The EAP protocol is mainly operated in 802.11 (wireless) environments. Because of its limited role in encapsulation of authentication data, it is extremely simple and includes only four types of messages request, response, success and failure.

1. Password-based authentication

Passwords might be static or dynamic. Static passwords obey security policy in a company that can define a minimum number of characters and a lifetime (expressed in days or number of connections). These passwords can be cracked (i.e. discovered by a malicious person) or spied on (on a phone link, data network, etc.) and can lead to the disclosure of confidential information by an intrusive access to a computer account, for example.

To overcome these drawbacks, dynamic passwords, also called OTP (One-Time-Password), have been defined. At each new session, a different value of password must be provided. OTP techniques obviously require a perfect synchronization between the client wishing to authenticate and the authentication server. This synchronization of dynamic passwords can be based on a clock, a series of numbers, a sequence number, etc.
The PAP (Password Authentication Protocol) and CHAP (Challenge-Handshake Authentication Protocol) [RFC1994] were originally based on static passwords and supported authentication of remote users connected on the telephone network through the PPP (Point-to-Point Protocol) and a modem. The PAP requires the PPP client to send a login and password in clear text over the network. The CHAP is based on a random number provided by the network and the client has to send back a hash calculated over this random number and password.

The original PAP and CHAP used static passwords and were then improved by the use of dynamic passwords. This dynamic password is usually generated by a token owned by the user, i.e. a sort of calculator or electronic badge. Upon entering the PIN code, the token provides a dynamic password as a string. Of course, it is necessary that the token and the authentication server are fully synchronized for the server to successfully check the password. If the synchronization is based on a clock, then the risk is high that temporal drift occurs between clocks of the token and the server. Current techniques therefore cope with this possible temporal drift by making the authentication server adapt to the temporal drift according to the password returned by the token. As such, if the client regularly connects to the server, the password will always be accepted as ranging within the windows of acceptance.

2. Smart card-based authentication

This type of authentication is a direct result of authentication performed in the second-generation GSM (Global System for Mobile Communications) cellular network for which subscribers have a smart card in their mobile equipment provided by an operator.

3. Biometry authentication

The authentication of a user can rely on one or more of his biometric features called “biometric modalities”. The most common modalities are the fingerprint, iris, face, voice or handwriting signature. More and more commercial products using biometrics mainly to authenticate a user, for instance in order to limit the use of certain equipment (fingerprint readers to unlock a laptop), to control access to sensitive buildings, certain areas of an airport, etc.

To initialize a biometric system with biometric data, first some people need to be enrolled, i.e. sensors digitize their biometric data; then, an algorithmic treatment is made of them and a “template” is stored serving as a reference for future authentication of these persons. This template can be stored in a centralized server or into a smartcard, depending on the use case. During the authentication procedure, a sensor is again used to digitize a biometric modality. A comparison between these data and the template is made and the system can deduce whether it is the same person or not. Depending on the selected modality, but also the quality of the sensor, results may differ considerably.

In particular, errors can occur: a criminal can be accepted mistakenly and then impersonate a legitimate user, or a legitimate user can be denied. These two types of possible errors lead to define two rates for evaluating the reliability of a biometric system: the False Rejection Rate (FRR) (i.e. the rate of rejecting a legitimate user) and False Acceptance Rate (FAR) (i.e. the rate of accepting an impostor). The iris is the most reliable biometric method, but is applicable only to certain very strict applications because of the intrusive feeling that users experience with this method.

The fingerprint is unreliable, but is more naturally accepted by users. To ensure more reliable biometric systems, research is underway on a combination of biometric procedures (e.g. iris and fingerprint).

Today biometrics are very commonly used to control access to buildings. The user is provided with a smart card where a template of his fingerprint is registered. At the entrance of the building, a smart card reader and fingerprint reader enable the user to enter his smart card and to press his finger on the reader. The system then verifies that the given fingerprint is sufficiently close to the template to unlock the door. Let us note that biometric systems are not only limited to the authentication procedure. They can be used to identify a person among N entities. This function is useful for identifying a criminal from a list of known criminals in an airport or a football stadium.

Access control

Previous sections presented mechanisms aimed at providing confidentiality authentication and integrity services for communications. However, apart from targeting communications, attacks can also aim at other goals. Attacks against end systems can provide attackers with access to unauthorized resources. This can occur by taking advantage of weaknesses in authentication systems deciding whether communications should be established or not. This can also occur by exploiting software or hardware vulnerabilities in communication systems in order to bypass access control systems of existing resources. Finally, it can also be performed by taking advantage of the lack of proper separation between resources used by different users to monopolize resources and deny access to legitimate users or slow down their operations. In this section we consider network-based mechanisms aimed at protecting against these threats.

1. Firewalls

Firewalls appeared at the end of the 1980s. Their initial goal was to separate networks in order to protect insecure computers from attackers. Separation of networks is based on the amount of trust the security administrator can put in devices constituting them. The main task of the firewall is to control communications between networks with different trust levels in order prevent attacks from occurring. The notion of trust is often based on the level of control the security administrator has on operations performed by users and devices within a network. For instance, devices connected to the Internet through a network different from the protected network are often considered as unreliable since the security administrator has no way to limit operations executed by them. Within a network, devices and users are expected to be subject to the same security policy. We thus consider that they share the same level of trust which explains why

communications within a network are not controlled. The frontier between two networks of different trust levels is called the **security perimeter**. Two general characteristics are usually expected from a firewall:

– It must be incorruptible. An attacker must not be able to change the behavior of a firewall. It must moreover have a failsafe behavior, meaning that in the event of a failure, it must limit the ability of attackers to take advantage of it.

– It must control all communications. There must be no way for devices located across the security perimeter to communicate without their communications being controlled.

More precisely, firewalls are normally used to provide several types of services:

– reduce the ability of attackers to attack devices within the security perimeter by limiting the resources accessible to them by filtering the types of data units that can cross the security perimeter. This strategy is usually referred to as "attack surface reduction";

– prevent vulnerabilities in internal systems from being exploited by blocking or reformatting data units appearing as malicious;

– prevent obfuscation techniques from being used in order to prevent previously mentioned services from being correctly implemented. Obfuscation techniques are tricks used by attackers to hide their operations or to create a different understanding of communications between the firewall and the systems communicating through it.

The protocol level of the analysis performed by the firewall:

– At the network level, the content of the network level protocol (e.g. IP, ICMP) and transport level protocol (UDP, TCP) headers are used to decide whether packets should be accepted or not. This can for instance be used to limit the devices or services accessible to attackers.

– At the circuit level, firewalls take into account the notion of transport level connection. This allows them to check the link between packets belonging to the same communication. For instance, with TCP, when setting up a connection, a TCP connection setup segment from the source to the destination should be followed by a setup acknowledgement segment in the opposite direction.

– At the application level, firewalls require a filtering policy specific to the considered application. For instance, for the HTTP protocol, a firewall will usually provide the ability to decide which methods can be used on which objects for a given server. It will also usually reformat requests sent to a server in order to avoid ambiguous understanding of the requests between the server and the firewall.

The location. Two main classes of locations exist today:

– Network firewalls are located within a network. Therefore, they can protect devices belonging to a complete network. Their main advantage is their incorruptibility since they usually rely on specific hardware and/or software systems specifically tuned to increase their resilience to attacks. This resilience is further improved when the device is managed by a security professional.

– Personal firewalls are collocated with end systems. These tools have the ability to interact with the operating system as well as with the user in order to prevent attacks such as application hijacking that are more difficult to prevent using network firewalls. Moreover, these tools allow a security administrator to control every communication received or generated by the protected device. This type of tools tends to integrate with other host based security tools (anti-viruses, anti- spyware, etc.). Their main weakness is however their reliance on the security of the device they are expected to protect.

1. Intrusion detection

Firewalls are not always sufficient to prevent all kinds of attacks against end systems. For example, if we consider firewalls trying to block malicious traffic, when only considering attacks that can be detected, firewalls will usually focus on blocking a small portion of these attacks. This can be explained by several reasons:

– In order to block a communications, a firewall must have strong evidence that the communication constitutes an attack. In practice, attack detection tools can usually make mistakes and consider that a benign communication is an attack (this is usually referred to as “false positive”). It is usually impossible to block communications automatically for a given attack when the level of “false positive” detection events for this particular attack is too high since it would affect the traffic of legitimate users negatively.

– Most filtering systems operations can be detected by parties communicating through a firewall. Potential attackers can therefore try to adapt their behavior to the firewall in order to hide their operations. They can also try to attack a firewall to make it fail. It is therefore often interesting to have additional means to control communications in a more furtive way.

Chapter 9

Security in Mobile Telecommunication Networks

1. Introduction

Circuit-switched telecommunication networks were created at a time when there was a strong monopoly granted to government-owned corporations. Depending on the nation, the network operator was either a government-controlled company under government monopoly or a private company under a government-granted monopoly.

The principal objective was to guarantee the fulfillment of public service duties, i.e., the establishment of telecommunications over a national territory. The access to such a network structure was granted with an analog landline for which the user identifier, his localization and his billing address were identical. Transnational internetworking required for worldwide telecommunications was based on mutual agreements based on an operator's reputation. The first generation of cellular networks largely followed such a principle.

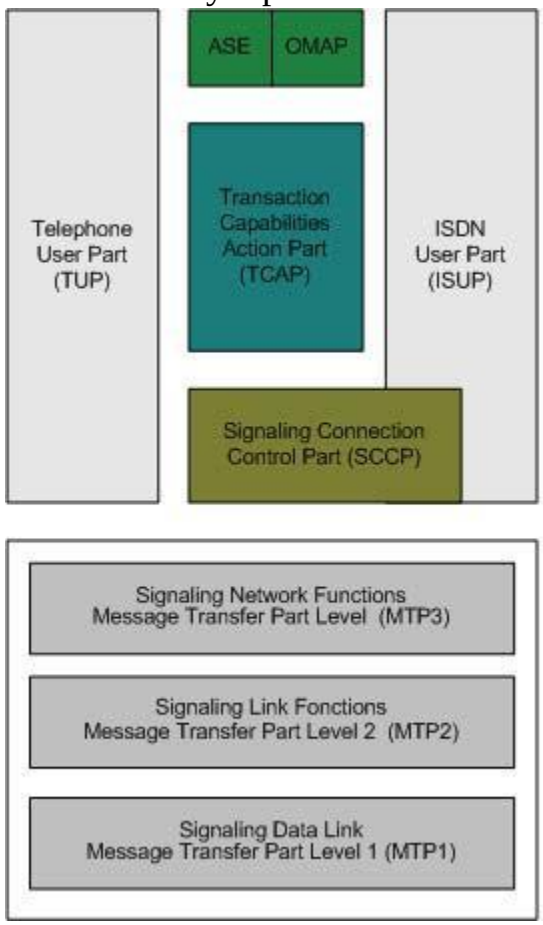
2. Signaling

Signaling in telecommunication networks has always been a problem for smooth network functionality. It is through robust signaling that calls are correctly routed to the correct destination and are specific subscribed services are billed to the correct person. The corollary is that signaling has already, since its beginnings, been a target for acts of sabotages from groups of persons aiming to illegally benefit from a telecommunications network or more dramatically seeking to hijack or to totally shatter it. It is therefore crucial to develop robust signaling protocols and if need be to identify its flaws and correct them.

The major signaling protocol used in telecommunication networks is the SS7 also referred to as *Common Channel Signaling System 7* (CCSS7) in North America. It

is used in public switched telephone networks, cellular networks and even in their interconnection with IP networks.

The SS7 protocol stack is composed of four layers. The first three are in charge of establishing point-to-point transfers while the fourth represents the application part of SS7. Figure 9.3 illustrates this 4-layer protocol stack.



The SS7 protocol stack consists of seven main functionalities:

– *Signaling Data Link (MTP1)*: this is the SS7 physical layer responsible of the interconnection.

– *Signaling Link Functions (MTP2)*: this is the link layer managing reliable transmissions (error detection, sequence checking).

– *Signaling Network Functions (MTP3)*: this layer routes signaling messages point-to-point through the SS7 network to a requested endpoint. It is also responsible for network management as it is in charge of homogenous traffic allocations or link redirections if the availability of a MTP2 datalink changes.

– *Signaling Connection Control Part* (*SCCP*): this extends the MTP layer by including advanced facilities such as global title translations (toll-free numbers or calling numbers for prepaid cards) into addresses and guarantees the transport of connectionless or connection-oriented services. Unlike MTP, SCCP establishes end-to-end connections.

– *Transaction Capabilities Application Part (TCAP*): this enables multiple and concurrent data exchanges between various applications through SS7 using the connectionless version of SCCP. The MAP messages that are exchanged between infrastructures and databases are also routed with TCAP.

– *Telephone User Part (TUP)*: this defines international signaling functions in order to establish communication. It does not allow the establishment of data links.

– *ISDN User Part (ISUP)*: this configures, manages and releases voice or data circuits

3. Security in the GSM

In 1982, the working group called in French "Groupe Spécial Mobile" (GSM). Its objective was the creation of a digital standard for 2G mobile telecommunication. This standard was been developed by the European Telecommunications Standards Institute (ETSI) in the 900 MHz and 1,800 MHz frequency bands. the GSM technology covers 100% of world nations and exceeds 2 billion users (including the recent extensions GPRS). The GSM network has been specifically created for voice communications. In order to join a GSM network, a potential customer has the choice of taking a subscription or buying a prepaid GSM calling card.

a) *GSM architecture*

A Public Land Mobile Network (PLMN) is a wireless communication system providing telecommunication services to mobile subscribers. The GSM network is the most popular example of a PLMN and each non-virtual GSM operator owns one. A GSM-PLMN network is composed of 4 major entities:

– The *Mobile Station (MS)* is usually a mobile telephone but more generally any device equipped with an adequate antenna and a SIM (*Subscriber Identity Module*) card may be considered to be a mobile station.

– The *Base Station Subsystem (BSS)* is composed of a network of radio relays called *Base Transceiver Stations (BTSs)* and of concentrators called *Base Station Controllers (BSCs).* BTSs transmit and receive signals from and to mobile stations and are the only radio interface of the total GSM system, as any other communication form between BTS and BSC as well as with the Core Network (CN) is performed by digital landline communication based on SS7 signaling.

– The *Network Switching Subsystem (NSS)* is in charge of a correct routing of voice calls between two GSM subscribers. It is composed of specialized interlined switches called *Mobile Switching Centers (MSCs)* that have several BSCs under their responsibility. Each MSC is associated with a specialized database called a *Visitor Location Register (VLR)* managing subscribers' information that is within the radio range covered by the MSC (or more precisely the summed radio coverage of all BTSs of the MSCs and BSCs). In order to manage subscribers' information for the complete PLMN, a unique database called the *Home Location Register (HLR)* also exists even though copies are appropriately located in the PLMN to mitigate the risks of failure. The HLR also contains the *Authentication Center (AuC)* which is in charge of the proper subscriber identification. The complementary information to the AuC is located into a subscriber's SIM card.

Local copies of the required information from the HLR are transferred to the respective VLRs in order to speed up the handling time of MSC requests. Finally, a specific MSC called a *Gateway MSC (GMSC)* is located at the entry point of a PLMN and therefore acts as a gateway .

– The *Operation and Maintenance Center (OMC)* is in charge of monitoring the correct functionality of all GSM systems and taking appropriate maintenance actions when necessary.

b. *Security mechanisms in GSM*

Most security protections provided by the GSM are located at the BSS and limited to access control and radio encryption. The emergence of application-layer security mechanisms securing the messages exchanged between SIM cards and an application server is more recent. GSM security is composed of three classes of protection:

– *subscriber identity protection*. For privacy issues, transmitting a subscriber identity in plain on a radio link must be avoided;

– *network access control* by means of SIM cards. The major functionality of the SIM is to securely hold and manage confidential information to allow the GSM network to formally identify a subscriber's identity.

– *radio communication encryption* between a MN (mobile node) and the BTS. Eavesdropping on radio communication being significantly easier than landline communication, it is absolutely vital to protect the radio link.

4. GPRS security

General Packet Radio Service (GPRS) is a mobile telecommunication standard derived from the GSM that theoretically promises a higher data throughput for sporadic traffic. While GSM is a 2G network, GPRS is in practice often described as a 2.5G network as it is technologically located between the GSM and the UMTS. The GPRS extends GSM by adding best-effort packet-switched communications for low latency data transmissions.

a. *GPRS architecture*

Unlike GSM, GPRS is able to provide a packet-based IP connectivity to a MN and also proposes a higher throughput by allocating radio resources as a function of the volume of information to be transferred. From an architectural point of view, a GPRS network exists in parallel with a GSM network benefiting from the later for voice communication but using its own infrastructure for data communication. The GPRS adds two new entities:

– *Serving GPRS Support Node (SGSN)*: this manages the attachments of the MN in the service zone and acts as a transit interface for packets on their way to a GGSN. The link between a SGSN and a GGSN is based on the IP, but user traffic is encapsulated in a proprietary protocol called the *GTP (GPRS Tunneling Protocol)*. Concerning security, the SGSN has the same role as a BSC as it is in charge of authentication, integrity and communication authorization.

– *Gateway GPRS Support Node (GGSN)*: this acts as an interconnection gateway between an operator's packet-oriented network and IP networks. The GGSN also runs a firewall in order to control access to its network.

b. GPRS security mechanisms

Due to the application-oriented GPRS development, its security shall be analyzed at the structural level as well as at the application level. Structural security may be separated into three parts: GPRS radio subsystem access control, GPRS session access control and GPRS network subsystem access control. We will illustrate below the various security mechanisms of the GPRS.

GPRS radio subsystem access control

The large majority of GPRS security mechanisms are identical to those of the GSM, notably authentication and access control. The novelty comes from packet oriented security instead of call-oriented security. This clearly impacts on the encryption that is now performed at the protocol level instead of the physical layer.

GPRS subscriber authentication

The GPRS subscriber authentication process is similar to that of the GSM. The major difference is that the authentication is not handled by a BSC but by a SGSN and uses a different and independent random number GPRS-RAND. Accordingly, the GPRS network provides a distinct challenge reply (GPRS-SRES) and a GPRS encryption key (GPRS-Kc) from the GSM network.

GPRS data encryption

GPRS data and signaling encryption, more commonly known as the *GPRS Encryption Algorithm (GEA)*. However, the encryption itself however varies from that of the GSM in several respects. First, the encryption is not only done between a MN and a BTS as for GSM but up to the SGSN. The GPRS-Kc key is therefore separately stored from the GSM Kc key. Then, unlike GSM, GPRS does not encrypt the physical channel itself but a logical channel at the LLC (Logical Link Control) layer as GPRS traffic is multiplexed on the exact same radio resource as that of the GSM. The encryption itself is therefore done at a higher protocol layer.

GPRS session access control

Unlike GSM, where a successful radio access also guarantees access to services (it is not necessary to re-authenticate to send a SMS) the GPRS must establish a

logical connection to make a MN reachable for a particular GPRS service. A specific mechanism called a PDP (Packet Data Protocol) context has been created in order to establish a logical link between a MN, a SGSN and a GGSN, possibly making the MN visible to external data networks (such as Internet or private networks). A PDP context allocates an IP address to the MN and defines the routing, billing. By analogy with mobile IP, a GGSN is similar to a Home Agent (HA) while a SGSN is similar to a Foreign Agent (FA) within a PDP context.

3G security

UMTS is one of the 3G mobile communication technologies. The objectives of UMTS are numerous and provide advantages relating to both voice and data communication. As this technology is based on a larger frequency band, a higher number of calls may be simultaneously serviced. Moreover, its throughput for data communication has been significantly increased. UMTS should theoretically mitigate the current quasi-constant saturation of existing GSM networks and offer higher quality services. In particular, the maximum throughput, which is theoretically five times higher, opens the door to multimedia applications.

1. UMTS infrastructure

The UMTS network has been compelled to guarantee a total interoperability with GSM/GPRS networks. Its infrastructure therefore includes GSM/GPRS-specific and UMTS-specific functionalities. The UMTS mostly reuses GSM and GPRS entities for voice calls or data transmissions. The major difference is located at the protocol layer for each interface and with respect to the radio technology.

UMTS security

The so-called 3G security systems define a higher security management for UMTS networks. New security provisions have been added such as the detection of rogue base stations, the strict control on the context for the transmission of secret keys, network mutual evaluation and identification, longer encryption keys, data integrity and subscriber identity protections. Moreover, a more powerful chip containing an elaborated *Universal Subscriber Identity Module* (USIM) replaces the GSM SIM card.

The novelty in 3G telephony mostly comes from the heterogenity of telecommunication operators. We not only face the interconnection of new cellular telephone operators but also the interconnection of new kinds of communication

operators such as Wi-Fi networks, corporate networks. Such configuration requires robust security management at the signaling and data planes in the UMTS core network. The innovation behind the UMTS is also on the radio part. Given that mobile terminals benefit from increased resources, it is now possible to use more powerful security mechanisms such as TLS or IPsec. A mutual authentication process has been added to the UMTS standard in order to solve some security flaws inherited from the GSM.

UMTS security is composed of five protection categories:

– *network access security*: mutual authentication between a MN and a UMTS network to mitigate attacks ;
– *network domain security*: protection of the signaling in the operator's NSS (network security service);
– *user domain security*: protection of the access to UMTS terminals;
– *application security*: secured data exchanges between UMTS terminals and UMTS networks at the application layer;
– *visibility*: visibility of the various security measures and the dependency of particular network services on specific security measures.

Network interconnection

The increasing use of packet-switched networks for real-time voice communications based on Voice-over-IP (VoIP) triggered an increased demand for access to SS7-based IN platforms, consequently requiring the interconnection of SS7 networks with the Internet or other data networks. Until recently, SS7 interconnections were limited for security reasons. Yet, in the light of this increasing demand, such a policy could be reconsidered. Indeed, IN services have already been successfully extended to cellular networks. Now, Local Exchange Carriers (LEC), competitive or not, as well as ISPs also request an access to the SS7 network of PSTNs. For instance, Internet Telephony Providers (ITP) would like to propose IN services such as number portability or free IP calls based on VoIP.

SS7 networks provide a very high stability and resilience but also contain connectivity and security issues. Data networks offer a simplified connectivity at the cost of reduced reliability. The interconnection of both worlds could be

beneficial by providing increased access to SS7 networks and a better resilience to data networks. It could however generate stability and security issues on SS7 networks. In order to connect the Internet to landline networks, it is necessary to make SS7 and IP networks transparently inter-operable. Several working groups have accordingly been created and have proposed four major standards: *H.323, SIP*. We give below a brief summary some of these protocols and refer

H.323

H.323 is a standard developed by the IUT and defines a multimedia communication protocol used for packet-switched networks.

SIP

The Session Initiation Protocol (SIP) is another standard, developed this time by the Internet Engineering Task Force (IETF). It is actually a signaling protocol managing video calls, telephony and instant messages where at least one participant belongs to a packet-switched network.

Megaco

The Megaco protocol, also called H.248, provides external control and management capabilities for data communications through a Media Gateway (MG) and is complementary to H.323 and SIP. Media Gateway Controllers (MGCs) are connected to and control MGs using H.248, whereas they communicate with each others using SIP or H.323.

Chapter 11

Security in Next Generation Mobile Networks

1.Introduction

The concept of next generation mobile networks appeared with the interconnection of telecommunication networks based on heterogenous telecommunication technologies and with specific value-added services proposed by different providers. Before such interconnection and independently of the type of technology used, there was in practice approximately one type of network per service. For example, a cell phone connected to the Internet could only access the limited Internet services proposed by its provider. Instead, the objective would be

to have a unique, possibly heterogenous, network for access to all telecommunication services. However, with heterogenous technologies and services, security concerns must be clearly addressed. Indeed, how can we guarantee data or network integrity or how can we control a correct billing for subscribed services when we have to deal with multiple intermediaries?

2. The SIP

The *SIP (Session Initiation Protocol)* has been created with this objective in mind. It not only makes it possible to establish multimedia sessions on the Internet but may also be used by any network connected to the Internet or having access to it. It works similarly to SS7 with respect to call establishments and is actually intended to replace it in the near future. The most prominent SIP application is *Voice-over-IP (VoIP)*. Unfortunately, SIP is not capable of managing user or network mobility by itself. The community therefore proposed an extension called *IMS (IP Multimedia Subsystem)* which significantly improves access control and subscriber management. IMS not only administers a controlled access for subscribers to networks, but also enables the interconnection of heterogenous networks. The objective of IMS is first to guarantee transparent access for subscribers to services and second to facilitate the establishment of new services proposed by different operators irrespective of to the telecommunication technologies employed or the exact location of the subscribers.

The SIP is an application-layer session initiation protocol standardized by the IETF. It is in charge of authenticating and locating the various actors of a SIP session. The SIP being independent of the type of data traffic, any type of communication protocol may be used. However, the Real-time Transfer Protocol (RTP) is the most widely used in practice for audio and video sessions. The SIP is also the open standard used by VoIP.

2.1. *SIP generalities*

SIP is a text protocol and shares similar response codes with HTTP. However, SIP differs from HTTP as a SIP agent is at the same time a client and a server. In general, SIP is composed of the following elements:

– *User Agent (UA)*: we may find it in all SIP phones or any other SIP-based applications. A communication between two SIP agents is established based on a *URI (Uniform Resource Identifier)* that is similar to an e-mail address.

– *Registar*: as we obviously need to know the IP address of the target SIP UA to establish a communication, the Registar is in charge of registering and maintaining this IP address into a database that will then link it with the target URI.

– *Proxy*: a SIP proxy has a middleman role between two SIP UAs in order to obtain their respective IP addresses. The SIP proxy retrieves the destination IP address from the database and then contacts the destination SIP UA. Data traffic never travels through a SIP Proxy but is directly exchanged between two SIP UAs.

– *Redirect Server*: a SIP redirect server receives requests from a SIP UA and is in charge of returning a redirection response indicating where the request should be retrieved.

– *Session Border Controller (SBC)*: this is a SIP-ready intelligent firewall. When a SIP UA initiates a SIP session, two connections are built, one for signaling and one for data transmission. Although this process does not pose any problem when both SIP UAs are located within the same subnetwork, firewalls or NAT (network address translation) separating different networks may not be aware of the relationship between these two connections. They could therefore reject traffic from a subscriber in its subnetwork even if signaling successfully established that connection. NATs further generate address translation problems between multiple temporary addresses established by ISPs and their visibility on the Internet.

2.2. *SIP security flaws*

Like SS7, SIP has not been conceived with default security mechanisms and like any textual protocol, it is very sensitive to attacks. We now provide some examples of typical attacks on SIP-based applications:

– *Registration hijacking*: a Registar evaluates the identity of a SIP UA based on the message header. The "FROM" field of the SIP header may yet be arbitrarily tampered with and opens the door to malicious (de-)registrations. By impersonating a SIP UA, a malicious user may request to replace URI contact addresses with its own contact information on the database. This demonstrates the requirement for authentication provisions between SIP UAs and SIP proxies.

– *Impersonating a proxy*: a SIP UA contacts a SIP proxy in order to correctly route its requests. The proxy may be impersonated by a malicious user and then perturb or even reroute requests to third parties. The mobility factor in a SIP network

further exacerbates such a security flaw. In order to combat possible security breaches, a mutual authentication process must therefore be established.

– *Tearing down sessions*: by passively listening to SIP call parameters and then by inserting a SIP control message "BYE", a malicious user may abruptly close a SIP session. By further inserting a SIP "RE-INVITE" message, it may then redirect a call to an arbitrary third party. In order to combat this kind of attack, SIP connection parameters must be hidden and the ID of the SIP UA must be authenticated.

– *Integrity*: it is unfortunately possible to arbitrarily tamper with the content of SIP messages with malicious data. A SIP proxy, even fully authenticated, should never have access to the content of a SIP message, especially during key agreement transactions.

– *Denial-of-Service (DoS)*: denial-of-service is an attack vector that aims at making a network element unreachable or unavailable. SIP proxies also have to be integrated into the Internet in order to be able to intercept legitimate requests from SIP UAs located around the world, SIP networks are therefore very vulnerable to a range of various DoS attacks. It should be noted that if SIP proxies are compromised or unavailable, the whole SIP network becomes non-operational considering that SIP UAs are not able to recognize each other and cannot access SIP databases. One vector to combat this security flaw is by controlling registration attempts.

3. VoIP

VoIP is a new technology that made it possible to federate the data and voice communication worlds. Before VoIP, the only solution to transmit voice communication was to establish a circuit between the caller and the callee, which had the advantage of guaranteeing very good communication quality unfortunately at an equivalently high price. With Internet communications, it became absurd to be able to transfer millions of data bits around the world at a very competitive price but still pay a high toll just to be able to talk. VoIP therefore equilibrated the equation by transmitting real-time voice mostly through the Internet at that time but, thanks to the ITU Next Generation Networks (NGNs), soon also through any packet-switched communication network.

VoIP contains a signaling layer and a media transport layer. The signaling protocol, principally the H.323 used by operators although SIP has recently showed an increasing popularity, handles subscriber localization, communication

setups and tear-downs. The media transport layer is principally the and is in charge of carrying media transmissions with real-time characteristics. IP eventually encapsulates the media packets and routes them through the network. VoIP has been designed with a full interoperability in mind. If voice calls are established within a same packet-switched network (wireless or IP), then no further structure is required. However, if voice calls are transmitted from or to circuit switched networks (PSTN or MPLN), then VoIP requires the following new elements:

– *Media gateway (MG)*: a media gateway interrupts a voice communication of a circuit-switched network, then samples and encodes the voice before eventually delivering it as voice packets to the IP network. The reverse operation is performed for a voice communication from an IP network.

– *Media Gateway Controller (MGC)*: also called "soft switch", this receives VoIP signaling information and assigns resources to MGs such as instructing them to send or receive voice packets.

– *Signaling Gateway (SG)*: this provides a transparent signaling interconnection between the SS7 network and the IP network. It is in charge of interrupting SS7 signaling if necessary or converting it to the IP format before directing it to the MGC. As such gateways are critical to VoIP networks, they are typically deployed in swarms.

– *IP-enabled Service Control Point (IP-SCP)*: is totally integrated into an IP network. It may also still be reached by SS7 networks. These elements do not have a unique denomination as they are developed by different standardization bodies or research groups. In H.323 for example, a MGC is called a Gatekeeper (GK), while SGs and MGs are simply both called a Gateway.

Several VoIP standards exist at the ITU or the IETF but the latter have recently appeared to take the lead over ITU. We now provide a brief description of the VoIP protocol stack proposed by the IETF:

– *Stream Control Transport Protocol (SCTP)*: this is actually the SIGTRAN protocol in charge of transporting SS7 signaling between a SG and a MGC or between a SG and an IP-SCP.

– *Megaco (H.248)*: this represents a control protocol between a MGC and several MGs.

– *Session Initiation Protocol (SIP)*: this manages calls between MGCs or between MGCs and SIP phones.

–*RTP*: this is a protocol based on UDP and transports voice packets with realtime constraints.

3.1. *VoIP security flaws*

The revolutionary aspect of VoIP is to be able to remove the complex proprietary structures of circuit-switched telecommunication operators. A VoIP may indeed avoid spending a fair amount of money in telecommunication switches and replace them with routers or soft switches at a more attractive price and accordingly transpose this financial saving on competitive prices for its services. The corollary is that the required architectures to build VoIP networks are basically "off the shelf", bringing more virtual operators into the arena and accordingly increasing the chances of intrusion or impersonation of VoIP networks. VoIP must therefore protect its signaling and data networks, as various attack vectors, which we briefly describe below, should not be ignored:

– *Confidentiality*: signaling is as important as communication considering that a compromised signaling may let a malicious user obtain sensitive information about a legitimate subscriber. For example, a compromised SG could be the source of eavesdropping attempts on VoIP calls or of the logging of calls passed by a legitimate subscriber.

– *Eavesdropping*: a conversation itself is also put at risk, a malicious user being able to intercept and tamper with VoIP packets in order to eavesdrop on it.

– *Man-in-the-Middle*: VoIP conversations are also vulnerable when it comes to Man-in-the-Middle attacks which could typically allow a malicious user to intercept a call and tamper with its parameters. Such an attack is also considered critical as it leads to identity thefts or call redirections that are totally transparent to the legitimate subscriber or the VoIP network.

– *DoS*: unlike circuit-switched networks, there is basically no guarantee on the available resources provided by VoIP networks. They may therefore be easy targets for DoS attacks rendering critical network elements totally inoperable and significantly reducing the Quality of Service (QoS) provided to the subscribers.

– *Non-repudiation*: once a destination accepts a call, it is important to have mechanisms guaranteeing that this destination cannot later deny having accepted it.

– *VoIP servers and terminals*: being computers in the first place, these are very vulnerable to attacks. It is indeed not trivial to compromise an analog telephone but the software contained in VoIP phones can be easily be tampered with.

4. IP Multimedia Subsystem (IMS)

The IP Multimedia Subsystem (IMS) is a new 3.5G to 4G standard and constitutes a further evolution compared to the SIP as it provides an intermediate layer in core networks to move from a classical call mode (circuit) to a session mode. IMS is based in part on SIP signaling but enhances it with its ability to open several sessions while on call. IMS may be considered as an intelligent SIP as it is able to open multimedia sessions and also to add intelligent routing rules in order to manage multimedia sessions considering new parameters such as localization, and the availability or the type of terminal. Initially created for cellular networks, IMS has been extended to wireless and landline networks in collaboration with TISPAN.

The IMS architecture therefore symbolizes the convergence between the worlds of mobile and fixed networks and the Internet.

4.1. *IMS architecture*

IMS includes a set of functions that are not specifically distributed per node. Different functions may exist on the same system or the same function may be distributed in different systems. We give below a summary of the various IMS entities:

– *Home Subscriber Server (HSS)*: this is a major database informing IMS core elements about call or session parameters.

– *Media Resource Function (MRF)*: this hosts multimedia resources in the subscriber's home network.

– *Application Server Function (AS)*: this hosts and executes telecommunication services such as MMS, SMS or Lawful Interception (LI). – *Serving Call Session Control Function (S-CSCF)*: this is a central node of an IMS network and is

located at the bottleneck of all signaling messages. It is actually a SIP server that is also in charge of controlling IMS sessions. It is always located in a subscriber's home network. The S-CSCF uses the DIAMETER protocol in order to securely contact the HSS to obtain information about a subscriber.

– *Interrogating Call Session Control Function (I-CSCF)*: this is a SIP proxy Server located at the edge of the IMS domain and acts as a gateway to a subscriber's home network. It also uses the DIAMETER protocol in order to question the HSS to obtain the location of a subscriber.

– *Proxy Call Sessions Control Function (P-CSCF)*: this is an IMS gateway and a SIP proxy server at the same time. It is in charge of authenticating a subscriber and initiating a transport mode IPsec ESP secured link with a subscriber. The P-CSCF accordingly protects information accessing an IMS network.

– *Breakout Gateway Control Function (BGCF)*: this is a SIP server that contains all routing functionalities for telecommunication networks. It is used when a subscriber calls a telephone number located in a circuit-switched network (PSTN or PLMN).

5. 4G security

The so-called fourth generation networks, also called next generation networks (NGN) by the ITU, are currently under development. They promise an unprecedented maximum wireless throughput of 100 Mb/s. While 3G networks witnessed the appearance of heterogenous networks with the transparent interconnection of IP, PSTN and PLMN, 4G networks will make provisions for a full heterogenity in the radio sub-system and the total superposition and cooperation of various radio technologies. For example, it is envisioned that a WLAN and a cellular network will be transparently connected without any communication or QoS interruption.

the 3GPP with IMS work together to define a network subsystem including the IMS as a core component which would be in charge of guaranteeing a total cooperation between the various fixed and mobile networks (PSTN, PLMN, WLAN) in an all IPv6 secured environment. Such 4G subsystems will be transparent to the subscriber, hiding their technical complexity and delegating to 4G User Agents (UAs) the choice and negotiation for the best communication technology to be used as a function of the requested multimedia application. A

larger cooperation between the various communication actors is therefore expected in order to provide high quality services to users in a secured environment.
The 4G networks composed of four key layers: *user*, *access*, *transport* and *service*, and a total transparency in the communication technologies and protocols used at each layer.