5G and Next-generation Mobile Computing

Prof. Carla Fabiana Chiasserini (carla.chiasserini@polito.it)
Dr. Corrado Puligheddu (corrado.puligheddu@polito.it)

Academic Year 2025-2026

Abstract

This course, "5G and Next-generation Mobile Computing," provides an in-depth exploration of the technologies and concepts shaping the future of mobile communications. Starting from the fundamental principles of wireless networks, the program covers the evolution of cellular generations from 4G to 6G, focusing on Radio Access Network (RAN) and Core architectures. Special attention is given to emerging technologies such as Edge Computing, connected autonomous vehicles, and the Internet of Things (IoT), highlighting how these innovations are transforming sectors like smart cities, telemedicine, and sustainable mobility through the integration of Artificial Intelligence and Machine Learning. The course also analyzes key concepts like virtualization, containerization, and Open-Radio Access Networks (O-RAN), offering both theoretical and practical insights (with lab activities) on how these technologies enable advanced and intelligent mobile services. The objective is to provide students with a solid understanding of the challenges and opportunities in the field of next-generation mobile computing, paving the way for innovative applications and high-performance, ubiquitous connectivity.

Contents

1 Fundamentals of Wireless Communications and Propagation

This lecture introduces the foundational concepts underpinning all wireless communications. We will explore how digital information is encoded onto analog radio waves through modulation, understand the critical role of radio frequencies as a finite resource, and analyze how signals propagate and degrade in real-world environments. Finally, we will cover the metrics used to assess signal quality and the fundamental techniques employed to combat errors.

1.1 The Principle of Modulation

At its core, communication involves transmitting information from a source to a destination. In wireless systems, this information must be impressed upon a radio wave that can travel through space. **Modulation** is the process of embedding information onto a sinusoidal carrier signal, which is characterized by a specific carrier frequency (f_c), by systematically altering one of its properties: amplitude, frequency, or phase.

1.1.1 Digital Modulation

Digital modulation translates digital information (bits) into variations of an analog carrier wave. The primary schemes are:

- **ASK (Amplitude Shift Keying):** The amplitude of the carrier is varied to represent different bits. For instance, a high amplitude could represent a '1', and a low (or zero) amplitude could represent a '0'.
- **FSK (Frequency Shift Keying):** The frequency of the carrier is shifted between predefined values. A common example is binary FSK (BFSK), where a frequency f_0 represents a '0' and a frequency f_1 represents a '1'.
- **PSK (Phase Shift Keying):** The phase of the carrier is shifted. For example, in 8-PSK, there are 8 distinct phase shifts, with each phase representing a unique 3-bit symbol.
- **QAM (Quadrature Amplitude Modulation):** A more complex scheme that combines both ASK and PSK. By varying both amplitude and phase, more bits can be encoded per symbol. For example, 16-QAM uses 16 unique amplitude/phase combinations to represent 4-bit symbols. These combinations are visually represented on a **signal constellation** diagram.

1.1.2 Analog Modulation

In analog modulation, a continuous baseband signal (like an audio signal) directly modifies the carrier. The main types are AM (Amplitude Modulation), FM (Frequency Modulation), and PM (Phase Modulation).

1.1.3 The Communication Chain

The modulation process fits into a larger communication chain. At the transmitter (TX), the information bits undergo **digital modulation** to produce a baseband digital signal. This signal then undergoes **analog modulation** to shift it to the carrier frequency before being sent out via the antenna. The receiver (RX) performs the reverse process: **analog demodulation** brings the signal back to baseband, and **digital demodulation** extracts the original information bits.

1.2 Wireless System Capacity and Radio Frequencies

The primary and most valuable resource in any wireless system is the radio frequency spectrum.

- **Value of Spectrum:** The ability to use more frequency bands directly translates to the ability to serve more users, thus having a significant economic value. The Italian 5G spectrum auction, where operators spent a total of €6.5 billion, is a testament to this.
- **Frequency Reuse:** Since frequencies are a costly and finite resource, they must be reused geographically to maximize the overall network capacity.
- **Frequency Trade-offs:** The choice of operating frequency involves critical trade-offs:
  - **Antenna Size:** Antenna size is proportional to the wavelength (λ). Since $\lambda = c/f$, higher frequencies (f) allow for smaller, more compact antennas.
  - **Attenuation:** Signal attenuation in free space is proportional to the square of the frequency (f^2). This means signals at lower frequencies travel farther and penetrate obstacles better.

  - **Network Capacity:** Given a fixed transmit power, higher frequencies lead to smaller coverage areas (cells). A network with smaller cells can reuse frequencies more aggressively, leading to higher overall capacity.

1.3 Radio Wave Propagation

The study of how radio waves travel from transmitter to receiver is fundamental to wireless network design.

1.3.1 Fundamental Definitions and Units

- **Energy per bit (E_b):** The amount of energy transmitted for each single bit of information.
- **Transmit Power (P_t):** The rate at which energy is transmitted, given by $P_t = E_b \times R_b$, where R_b is the bit rate.
- **Noise (N_0):** The noise power spectral density, representing the background noise level in the communication channel.
- **Decibels (dB and dBm):** To handle the vast range of signal power values, logarithmic units are used.
  - **dB:** A relative measure of power ratio: $dB = 10\log_{10}(\frac{P_1}{P_2})$.
  - **dBm:** An absolute measure of power relative to 1 milliwatt: $dBm = 10\log_{10}(\frac{P_1}{1\text{ mW}})$. For example, 10W is 40 dBm, and 1μW is -30 dBm.

1.3.2 Attenuation in Free Space (Path Loss)

In an ideal vacuum with no obstacles, the signal power attenuates as it spreads out. The received power (P_R) is described by the Friis transmission formula:

$$P_R = P_T G_T G_R \left(\frac{\lambda}{4\pi R}\right)^2 \tag{1}$$

where P_T is the transmitted power, G_T and G_R are the antenna gains, λ is the wavelength, and R is the distance. The Path Loss (L), which is the ratio P_T/P_R, shows that attenuation increases quadratically with both distance (R^2) and frequency (f^2).

1.3.3 Propagation in the Real World

Real-world propagation is far more complex due to various phenomena that alter the signal's path and strength:

- **Absorption:** Materials like rain, fog, and atmospheric gases (especially oxygen and water vapor at high frequencies) absorb energy from the radio wave, converting it to heat and weakening the signal.
- **Refraction:** Radio waves bend when passing through media of different densities, such as layers of the atmosphere. This can cause the signal path to curve, potentially extending the communication range beyond the geometric line of sight.
- **Diffraction:** Occurs when the radio wave path is obstructed by a sharp edge or surface (e.g., a building corner, a mountain peak). The wave bends around the obstacle, allowing reception even when there is no direct line of sight.
- **Reflection:** The wave bounces off a surface that is large relative to its wavelength, such as the ground, buildings, or walls.

1.4 Signal Impairments: Fading and Multipath

In most environments, the received signal is a sum of multiple copies of the transmitted signal that have traveled along different paths due to reflection, diffraction, and scattering. This is known as **multipath propagation**. These multiple copies (or "echoes") can interfere constructively or destructively at the receiver's antenna, causing fluctuations in the received signal amplitude, phase, and angle of arrival. This fluctuation is called **fading**.

- **Shadowing (Slow Fading):** Caused by large-scale obstacles like buildings or hills that obstruct the main signal path. This results in a gradual, slow variation in the average received signal power as the user moves over distances of tens or hundreds of meters.
- **Fast Fading:** Caused by the constructive and destructive interference of multipath components. It leads to rapid and deep fluctuations in signal strength over very short distances (on the order of half a wavelength). It can cause temporary signal loss of 20-30 dB.

1.5 Signal Quality and Performance Metrics

To quantify the quality of a received signal, we use the following metrics:

- **SNR (Signal-to-Noise Ratio):** In a simple point-to-point link, the quality is determined by the ratio of the received signal power to the noise power within the signal bandwidth. It is defined as:

$$SNR = \frac{P_{rx}}{N_0 B} \tag{2}$$

where P_{rx} is the received power, N_0 is the noise power spectral density, and B is the bandwidth.

- **SINR (Signal-to-Interference-plus-Noise Ratio):** In a cellular network with multiple users, the quality is affected not just by noise, but also by interference from other transmitters. SINR accounts for this:

$$SINR = \frac{P_{rx}}{I + N_0 B} \tag{3}$$

where I is the total interference power from other transmitters.

If the SNR or SINR falls below a certain threshold required by the modulation scheme, the receiver will be unable to correctly decode the data, leading to bit errors. The **Bit Error Rate (BER)** is a direct function of the SINR.

1.6 Error Detection and Recovery

To ensure reliable communication over an unreliable wireless channel, several error control techniques are employed.

1.6.1 Forward Error Correction (FEC)

FEC involves adding redundant information (parity bits) to the data stream at the transmitter. This allows the receiver to not only detect but also correct a certain number of errors without needing to ask for a retransmission.

- **Block Codes:** Divide the information into blocks of k bits and add $n - k$ redundant bits to form a larger block of n bits (a codeword). An example is a simple **repetition code**, where each bit is sent multiple times.

- **Convolutional Codes:** Process the information stream continuously, generating n encoded bits for every k input bits based on the current and previous input bits using a shift register.

1.6.2 Automatic Retransmission reQuest (ARQ)

In ARQ schemes, the transmitter adds information (like a CRC - Cyclic Redundancy Check) to detect errors. If the receiver detects an error in a packet, it discards it and sends a request to the transmitter to send it again.

- **Stop-and-Wait ARQ:** The simplest form. The sender transmits one packet and waits for an acknowledgment (ACK) before sending the next one.

- **Go-Back-N ARQ:** The sender can transmit several packets without waiting for an ACK. If the receiver detects an error in packet N, it requests a retransmission of packet N and all subsequent packets.

- **Selective Repeat ARQ:** A more efficient version where the receiver requests retransmission of only the specific packets that were received in error.

2 IEEE 802.11 (WiFi) - Architecture and Network Association

This lecture delves into the specifics of the most widespread Wireless Local Area Network (WLAN) technology: IEEE 802.11, universally known as WiFi. We will analyze its fundamental architecture, from the basic building blocks to complex network topologies, and detail the step-by-step process a device follows to connect to a WiFi network.

2.1 Introduction to the IEEE 802.11 Standard

The IEEE 802.11 is a family of standards that defines the specifications for the **Physical Layer (PHY)** and the **Medium Access Control (MAC) layer** of a wireless network. It governs the wireless interface between a client station and a base station (or Access Point), as well as peer-to-peer communication between clients.

As illustrated in Figure 1, the 802.11 standard fits into the lower two layers of the ISO/OSI model. The Data Link layer is subdivided into two sub-layers: the Logical Link Control (LLC), defined in the 802.2 standard, and the Medium Access Control (MAC) layer. Below them lies the Physical (PHY) layer.

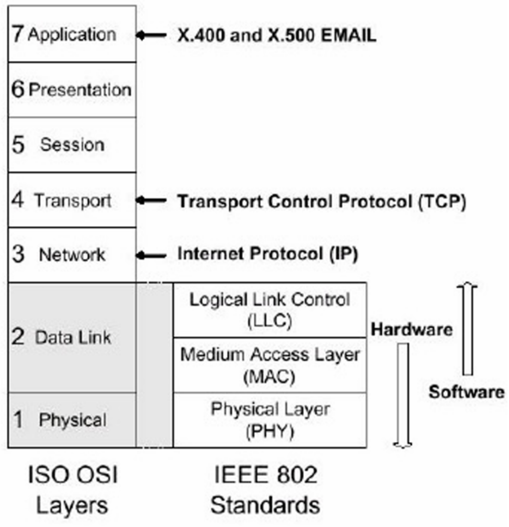


Figure 1: The IEEE 802.11 standard in the context of the ISO/OSI protocol stack.

The standardization process began in 1990, with the first official release published in 1997. Since then, the standard has been continuously evolving through numerous amendments to improve speed, reliability, and efficiency.

2.1.1 Standards Evolution and "WiFi" Denomination

To make the numerous amendments (e.g., 802.11b, 802.11g, 802.11ac) more accessible to consumers, the Wi-Fi Alliance introduced a simpler, generational naming convention. The IEEE 802.11-2016 release consolidated all commercially available versions of the standard up to that point. The new denominations are as follows:

- **WiFi 8** (under development, expected rel. 2028) to identify devices supporting **802.11bn**.
- **WiFi 7** (released 2024) to identify devices supporting **802.11be**.
- **WiFi 6** (released 2019) to identify devices supporting **802.11ax**.
- **WiFi 5** (released 2014) to identify devices supporting **802.11ac**.
- **WiFi 4** (released 2009) to identify devices supporting **802.11n**.
- **WiFi 3** (released 2003) to identify devices supporting **802.11g**.
- **WiFi 2** (released 1999) to identify devices supporting **802.11a**.
- **WiFi 1** (released 1999) to identify devices supporting **802.11b**.

Table 1: Evolution of Key IEEE 802.11 Radio Standards.

Parameter	**802.11 (Original)**	**802.11b**	**802.11a**	**802.11g**	**802.11n**
Approval Date	July 1997	Sep 1999	Sep 1999	June 2003	Sep 2009
Bandwidth	22 MHz	22 MHz	20 MHz	20 MHz	20 or 40 MHz
Operation Frequency	2.4 GHz	2.4 GHz	5 GHz	2.4 GHz	2.4 and/or 5 GHz
No. of non-overlapping channels	3	3	23	4	23 (at 20 MHz)
Data Rate / Channel	1, 2 Mbps	1, 2, 5.5, 11 Mbps	6, 9, ..., 54 Mbps	1, 2, ..., 54 Mbps	Up to 600 Mbps
PHY Layer Technology	FHSS, DSSS	DSSS	OFDM	DSSS / OFDM	OFDM+MIMO

2.1.2 Key Physical Layer Technologies Explained

The evolution of WiFi has been driven by advances in physical layer technologies. The terms in the table above refer to different methods of transmitting signals over the radio medium.

FHSS (Frequency-Hopping Spread Spectrum) A technique where the signal rapidly hops between many different frequencies in a pseudorandom, predetermined sequence known to both transmitter and receiver. This makes the transmission resistant to narrow-band interference and provides a basic level of security. It was used in the original 802.11 standard but is now obsolete in modern WiFi.

DSSS (Direct-Sequence Spread Spectrum) A technique where each information bit is replaced by a longer, higher-rate sequence of bits, known as a ”chipping code”. This spreads the signal's energy over a wider bandwidth, making it appear as low-power noise to unintended receivers. The primary benefit is its strong resistance to interference. DSSS was the technology behind the popular 802.11b standard.

OFDM (Orthogonal Frequency Division Multiplexing) A highly efficient modulation technique that is the foundation of almost all modern broadband systems (including WiFi 2/3/4/5/6 and LTE/5G). It works by splitting a single high-speed data stream into hundreds of slower, parallel sub-streams. Each sub-stream is then transmitted on a separate, closely spaced sub-carrier frequency. The ”orthogonal” nature of these sub-carriers allows them to be packed together without mutual interference, maximizing spectral efficiency. OFDM is particularly robust against multipath fading, a common problem in indoor environments.

MIMO (Multiple-Input Multiple-Output) A revolutionary antenna technology that uses multiple antennas at both the transmitter and receiver. MIMO can be used in two primary ways:

- *Spatial Multiplexing:* To increase data rate by transmitting multiple, independent data streams simultaneously over the same frequency channel. This effectively multiplies the throughput.
- *Transmit Diversity:* To increase reliability by transmitting the same data stream across different antennas in a way that makes the signal more robust to fading.

MIMO was the key innovation introduced in 802.11n (WiFi 4) and has been enhanced in all subsequent standards.

2.2 Fundamental 802.11 Architectures

The 802.11 standard defines two primary modes of operation, which form the basis of all WiFi network topologies.

2.2.1 Basic Service Set (BSS)

The **Basic Service Set (BSS)** is the fundamental building block of a WiFi network. It is defined as a set of nodes (stations, or STAs) that use the same coordination function to access the shared channel. The geographical area covered by a BSS is known as the **Basic Service Area (BSA)**, which effectively constitutes a single ”cell” in a WLAN. A BSS can operate in one of two configurations:

- **Infrastructure Mode:** This is the most common mode. The BSS contains wireless hosts (STAs) and a central base station known as an **Access Point (AP)**. All communication must pass through the AP; direct communication between STAs is disallowed. The AP acts as a gateway, connecting the wireless clients to a fixed network infrastructure.

Real-world example: Consider a typical **Home Wi-Fi** setup. Your router acts as the AP. Even if you want to send a file from your laptop to your printer (both connected to Wi-Fi), the data does not go directly from laptop to printer; it travels from the Laptop → Router (AP) → Printer.

- **Ad Hoc Mode (Independent BSS - IBSS):** In this mode, a group of 802.11 STAs can dynamically form a network *without* an AP. They communicate directly with each other in a peer-to-peer fashion. This is useful for creating temporary networks for specific purposes, such as a meeting in a conference room, interconnecting personal devices, or in battlefield scenarios. The IETF MANET (Mobile Ad hoc Networks) working group focuses on standards for this type of networking.

Real-world example: A **Disaster Relief operation** is a prime scenario. If an earthquake destroys cellular towers and power lines, rescue teams can create an Ad Hoc network between their tablets to share maps and medical data directly on the field, without needing an internet connection or a central router. Another conceptual example is Apple's **AirDrop** or handheld gaming consoles (like Nintendo DS) playing in local multiplayer.

2.2.2 Extended Service Set (ESS)

To provide broader coverage and enable user mobility, multiple infrastructure BSSs can be interconnected. This is achieved through a **Distribution System (DS)**, which is the backbone network that connects the Access Points. An **Extended Service Set (ESS)** is formed by one or more interconnected BSSs.

- Within an ESS, a STA can **seamlessly move** from one BSS to another. This transition requires cooperation between the access points.
- When a station moves, the new AP (e.g., AP2) must inform the original AP (AP1) that the station is now associated with it, ensuring that data frames are correctly routed.
- The original 802.11 standard did not specify the details of communication between APs during these transitions. Newer amendments like **802.11r (Fast BSS Transition)** have been introduced to standardize and speed up this process.

Real-world example: Think of a **University Campus (e.g., Eduroam) or a Corporate Office**. The entire campus constitutes the ESS. You might start a VoIP call in the library (connected to AP1). As you walk to the cafeteria, your phone disconnects from AP1 and connects to AP2. Thanks to the ESS architecture, your call does not drop, and your IP address typically remains the same.

2.3 The Process of Joining a BSS with an AP

For a station to become a functional member of an infrastructure BSS, it must complete a three-stage procedure. In an Independent BSS (ad hoc), neither authentication nor association procedures are required.

2.3.1 Stage 1: Scanning (Finding the Network)

The first step is for the station to discover available networks. This can happen in two ways:

1. **Passive Scanning:** The station listens on each channel for **Beacon frames**. These frames are transmitted periodically by the AP and contain synchronization information and network details, such as the network name (SSID).
2. **Active Scanning:** The station takes a more proactive approach by transmitting a **Probe Request** frame on the channels it wants to scan. Any AP within range that receives this request will respond with a **Probe Response** frame, providing the necessary network information.

2.3.2 Stage 2: Authentication (Verifying Identity)

Once an AP is found and selected, the station must authenticate with it.

- **Open System Authentication:** This is the default, two-step process. The station sends an authentication frame with its identity, and the AP replies with an acknowledgment (ack) or negative-acknowledgment (nack). It provides no real security.
- **Shared Key Authentication:** This method requires that both the station and the AP have been preconfigured with a shared secret key. It uses a challenge-response protocol to verify possession of the key, independent of the WiFi channel itself.

Challenge & Response Authentication Detail: This protocol works as follows, using a symmetric key cipher with a shared key K_{AB}:

1. *Request:* The prover (Alice) transmits her identity (ID_A) to the verifier (Bob) to request authentication.
2. *Challenge:* Bob sends a nonce (a random number, R_B) to Alice.
3. *Response:* Alice demonstrates that she shares the secret key K_{AB} by encrypting the nonce and transmitting the result, $K_{AB}(R_B)$, back to Bob. Bob can then verify this by encrypting the original nonce with its own key and comparing the results. The nonce is renewed in each session to prevent replay attacks.

2.3.3 Stage 3: Association (Connecting for Data Transfer)

Once a station is authenticated, it begins the final association process. This involves exchanging information about capabilities (e.g., supported data rates) and roaming information.

- The station sends an **Association Request** frame to the AP.
- The AP replies with an **Association Response** frame. In a roaming scenario, the new AP will also inform the old AP (via the Distribution System) of the station's new location.

Crucially, only after the association process is successfully completed can a station transmit and receive data frames on the network.

2.4 WiFi 1: Basic Physical-Layer Principles

While the WiFi standards have evolved significantly, the fundamental principles established in the early versions remain relevant across all generations. These principles govern how the radio spectrum is used and how devices adapt to changing channel conditions.

2.4.1 Frequency Bands and Channels

The original 802.11 and the popular 802.11b/g standards operate in the **2.4 GHz ISM (Industrial, Scientific, and Medical) band**. This band is unlicensed, meaning it can be used freely without requiring a license, which contributed significantly to WiFi's widespread adoption. However, it also means that WiFi devices must coexist with other technologies operating in the same band, such as Bluetooth, cordless phones, and microwave ovens.

To manage access, this frequency band is divided into a series of channels:

- The band is divided into 14 channels (though not all are available in every country).
- Each channel is **22 MHz** wide.
- The center frequencies of adjacent channels are spaced by only **5 MHz**.

This spacing creates a significant problem: **channel overlap**. As shown in Figure 2, a transmission on one channel will cause interference on several adjacent channels. To avoid this *Adjacent Channel Interference (ACI)*, network administrators should only use channels that are sufficiently spaced apart.

In the 2.4 GHz band, the only set of channels that do not overlap are **channels 1, 6, and 11**. These channels are spaced by 25 MHz, ensuring that their signals do not interfere. For this reason, it is a best practice to configure adjacent Access Points on these three channels. This also implies that in any given area, a maximum of three BSSs can operate simultaneously without causing interference to one another.

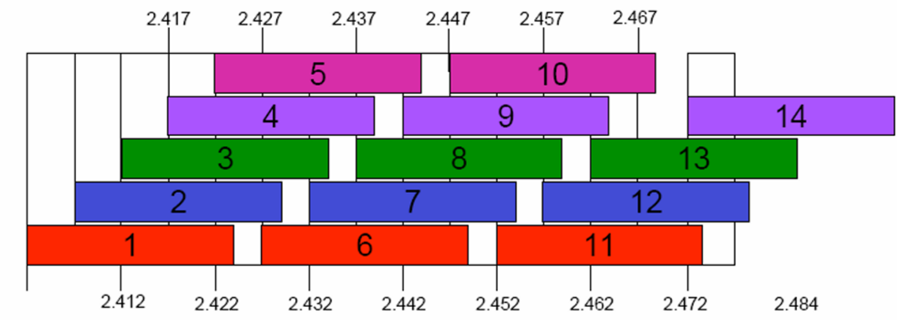


Figure 2: Overlapping Frequency Channels in the 2.4 GHz WiFi Band. Channels 1, 6, and 11 are non-overlapping.

2.4.2 Advantage of Multi-rate Transmission

A key feature of WiFi is its ability to operate at multiple data rates (e.g., 802.11b supports 1, 2, 5.5, and 11 Mbps). This capability is a direct consequence of the relationship between communication rate and channel quality.

- Higher data rates require more complex modulation and coding schemes, which in turn demand a higher minimum **Signal-to-Noise Ratio (SNR)** to be decoded correctly.

- As the distance from the transmitter increases, the signal strength decreases, leading to a lower channel quality (lower SNR).

This creates a fundamental **trade-off between communication range and link speed**. A station close to the AP with a strong signal can use a high data rate, while a station far away with a weak signal must fall back to a lower, more robust data rate. This multi-rate flexibility allows the network to meet diverse consumer demands and provide the best possible performance under varying coverage conditions. For example, a device might achieve 11 Mbps at 30 meters indoors but only 1 Mbps at 90 meters.

2.4.3 Rate Adaptation

To leverage the multi-rate capability, stations do not operate at a fixed speed. Instead, they employ a process called **Rate Adaptation** (or Rate Control).

- **Basic Rate:** Every BSS defines a *basic rate*, which is the lowest rate supported by all stations in the network (e.g., 1 Mbps). This rate is used for transmitting critical control and management packets (like ACKs and Beacons) to ensure that every station, regardless of its distance or channel quality, can correctly receive them.

- **Dynamic Data Rate Selection:** For data packets, stations constantly perform operations to sense the channel quality and automatically select the best (highest) possible rate that the current conditions can support.

- **Implementation:** The 802.11 standard does not specify a mandatory algorithm for rate adaptation, leaving it as a vendor-specific implementation choice. Common empirical approaches include:
  - Algorithms based on **SINR measurements** over a moving window. The station measures the signal quality and uses a lookup table to select the corresponding rate.
  - Algorithms based on **packet transmission success rate**, such as the popular Minstrel algorithm. The station probes different rates and statistically determines which one provides the best throughput by tracking the percentage of successfully acknowledged packets.

2.4.4 A Final Remark: The Half-Duplex Nature of WiFi and Its Alternatives

A fundamental constraint of all current 802.11 nodes is that they are **half-duplex** devices. This means:

1. They have only one transceiver (a combined transmitter and receiver).

2. At any given moment, a station can either transmit or receive (or sense the channel), but it **cannot do both simultaneously**.

This hardware limitation is the primary reason why WiFi cannot use Collision Detection (like wired Ethernet) and must instead rely on Collision Avoidance mechanisms, which will be the subject of the next part of our lecture.

It is useful to contrast this with the main alternative:

Full-Duplex: A full-duplex system allows a device to transmit and receive signals simultaneously. In traditional systems, this is achieved by using different physical resources for each direction, such as separate frequency bands (Frequency Division Duplex - FDD) or different physical wires (as in wired Ethernet).

In-Band Full-Duplex (IBFD): A more advanced and emerging technology aims to achieve simultaneous transmission and reception on the *same frequency channel at the same time.* This is extremely challenging because the device's own transmitted signal is immensely more powerful than the weak signal it is trying to receive, creating massive self-interference. IBFD systems require sophisticated self-interference cancellation (SIC) techniques to make this possible. While not yet a feature of standard WiFi, IBFD is an active area of research for future wireless systems (like 6G) as it holds the potential to theoretically double the spectral efficiency of a network.

For the scope of our analysis of the 802.11 MAC layer, however, we will assume the standard half-duplex model.

2.5 WiFi MAC Layer: Principles and Functions

The Medium Access Control (MAC) layer is a sub-layer of the Data Link Layer and represents the "brain" of a WiFi node. Its primary responsibility is to coordinate how and when multiple stations can access the shared wireless medium to transmit data. Given the half-duplex nature of WiFi and the inability to detect collisions, the MAC layer is built around a non-persistent Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) scheme.

2.5.1 Core MAC Functions

The 802.11 MAC layer performs several critical functions:

- **Resource Allocation:** Managing access to the shared wireless channel.
- **Data Segmentation and Reassembly:** Breaking down larger data packets from upper layers into smaller MAC frames for transmission, and reassembling them at the receiver.
- **Addressing:** Handling the MAC Protocol Data Unit (MPDU) address.
- **Frame Formatting:** Constructing the MPDU, which in 802.11 terminology is called a **frame**.
- **Error Control:** Providing mechanisms to check for transmission errors and manage retransmissions.

2.5.2 802.11 Frame Types

To perform these functions, the standard defines three distinct types of frames:

1. **Control Frames:** These are short frames used to manage the channel access and ensure reliability. Examples include acknowledgment frames (**ACK**) and the handshaking frames for channel reservation (**RTS** and **CTS**).
2. **Data Frames:** These frames carry the actual user information (the payload from upper layers).
3. **Management Frames:** These frames are used for managing the lifecycle of a connection. They handle tasks such as connection establishment and release (e.g., Beacon, Probe Request/Response, Authentication, and Association frames). Although they are exchanged as data frames, their content is not reported to the higher layers of the protocol stack.

2.5.3 Data Transfer Coordination Functions

The 802.11 standard defines two main coordination functions for data transfer:

DCF (Distributed Coordination Function): This is the fundamental and mandatory access mechanism. It is designed for asynchronous, delay-tolerant traffic such as file transfers or web browsing. DCF is a fully distributed, contention-based scheme where all stations compete for channel access. This will be the main focus of our analysis.

PCF (Point Coordination Function): This is an optional, centralized access mechanism designed for real-time, synchronous traffic like audio and video. It operates on top of DCF and is based on a polling scheme controlled by a central coordinator (the Access Point, which acts as a Point Coordinator or PC). In this mode, the AP polls stations to grant them contention-free transmission opportunities. Due to its optional nature and implementation complexity, PCF is rarely used in commercial products.

2.6 Timing and Synchronization in DCF

The DCF mechanism relies on precise timing to function correctly. This is achieved through the concepts of time slots and interframe spaces.

2.6.1 Time Slots

Time in the network is divided into fixed-duration intervals called **slots**. The slot is the fundamental system unit of time, and its duration is dependent on the specific physical layer being used. The slot time must account for factors like the TX/RX turnaround time (the time it takes for a radio to switch from transmitting to receiving) and the power detection time (the time needed to reliably sense if the channel is busy). For example, in early WiFi standards, this resulted in a slot time of 20 µs. All stations in a BSS are synchronized on a slot basis, either with the AP in infrastructure mode or among each other in ad-hoc mode, making the system synchronous at this micro-level. This synchronization is maintained through the periodic reception of Beacon frames.

2.6.2 Interframe Spaces (IFS)

To establish a priority system for accessing the channel, DCF defines several fixed-duration time intervals known as **Interframe Spaces (IFS)**. A station must wait for a specific IFS period before performing an action. A shorter IFS interval corresponds to a higher priority. The primary types are:

- **SIFS (Short IFS):** The shortest and highest-priority IFS. It is used for immediate response actions that must happen without contention, such as sending an ACK frame after receiving data, or a CTS frame after an RTS.
- **PIFS (Point coordination IFS):** A medium-priority IFS, longer than SIFS but shorter than DIFS. It is used by the Point Coordinator (in PCF mode) to gain priority access to the channel to start a contention-free period.
- **DIFS (Distributed IFS):** The longest of the main IFS types. A station with new data to send must first sense the channel as idle for a continuous period of DIFS before it can attempt to transmit. This ensures that it does not interfere with a higher-priority SIFS or PIFS response.
- **EIFS (Extended IFS):** A very long IFS used as a recovery mechanism. If a station detects a frame that it cannot decode (i.e., a transmission error), it must wait for EIFS before attempting to access the channel. This longer delay ensures that the station does not accidentally interfere with a pending ACK that might be sent between the two original communicating parties.

The exact duration of these IFS periods depends on the physical layer implementation. The priority hierarchy is:

$$\textbf{SIFS} < \textbf{PIFS} < \textbf{DIFS} < \textbf{EIFS}$$

These precisely defined time intervals are the fundamental tools that allow the distributed CSMA/CA mechanism to operate in a coordinated and fair manner.

2.7 The DCF Access Scheme: CSMA/CA

The Distributed Coordination Function (DCF) is the mandatory and most fundamental MAC mechanism in IEEE 802.11. It is a contention-based protocol based on the principle of **Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA)**.

The "Carrier Sense" part means that a station must listen to the medium before transmitting to check if it is busy. The "Collision Avoidance" part signifies that, unlike wired Ethernet's Collision Detection, the protocol's main goal is to proactively reduce the probability of collisions, since they cannot be detected in real-time. Every station with a new data frame to transmit must repeat the contention procedure. This is a distributed scheme, meaning there is no central coordinator; all stations run the same algorithm and compete for access.

2.7.1 The Two Basic Rules of DCF

The entire DCF procedure is governed by two fundamental rules that dictate when a station is allowed to begin a transmission.

1. **Rule 1: Sense Before Transmitting.** A station can only begin a data transmission if the medium has been idle for a sufficient amount of time.
   - If the previous frame was received correctly (or the channel has been idle for a long time), the medium must be sensed as idle for at least one **DIFS** (Distributed Interframe Space) period.
   - If the station detected a transmission error in the previous frame, it must wait for a longer period, the **EIFS** (Extended Interframe Space), to avoid interfering with a potential ACK that it was unable to hear.
2. **Rule 2: Defer and Backoff if Busy.** If a station senses the medium as busy, it must wait for the channel to become idle. This waiting is referred to as **access deferral**. Once the channel has been continuously idle for a DIFS period, the station does not transmit immediately. Instead, it must execute an **exponential backoff procedure** to resolve contention with other stations that might also be waiting.

2.7.2 The Basic Unicast Transmission Sequence

In the simplest case of a successful unicast (one-to-one) transmission without contention, the sequence of events is as follows:

Sender: Senses the channel is idle, waits for a DIFS period, and then transmits its entire data frame. Since this is wireless, there is no collision detection during the transmission.

Receiver: Upon correctly receiving the data frame, it waits for a SIFS period (which is shorter than DIFS, giving it priority) and then sends back an **ACK** (Acknowledgment) frame to confirm the successful reception.

Other Stations: Any other station that hears either the data frame or the ACK frame understands that the medium is busy and will defer its access accordingly.

If the sender does not receive an ACK within a specified timeout period, it assumes the transmission failed (due to a collision or channel errors) and will schedule a retransmission.

2.7.3 The Hidden Terminal Problem: A Fundamental Challenge

A fundamental impediment in wireless networks is the **hidden terminal problem**. This scenario occurs when a station is visible to the Access Point (or intended receiver), but not to other stations that are also communicating with that same AP.

Consider three stations, A, B, and C, arranged linearly. Station B can hear both A and C, but A and C are out of range of each other.

1. Station A senses the medium. It cannot hear C, so it assumes the channel is free and starts transmitting to B.
2. At the same time, Station C senses the medium. It cannot hear A, so it also assumes the channel is free and starts transmitting to B.

As a result, station B receives two frames simultaneously, causing a **collision**. Neither A nor C is aware of this collision. They will only realize the transmission failed when they do not receive an ACK from B. This problem significantly degrades network performance, especially in highly populated networks.

2.8 Solving the Hidden Terminal Problem with RTS/CTS Handshaking

To mitigate the hidden terminal problem, DCF provides an optional mechanism for **explicit channel reservation**. This is a four-way handshaking procedure using two small control frames: **RTS (Request to Send)** and **CTS (Clear to Send)**. This mechanism is typically used for frames that are larger than a certain configurable size (the $RTS_{Threshold}$).

2.8.1 The RTS/CTS Procedure and Virtual Carrier Sensing

The complete transmission sequence involves four frames, as illustrated in Figure 3.

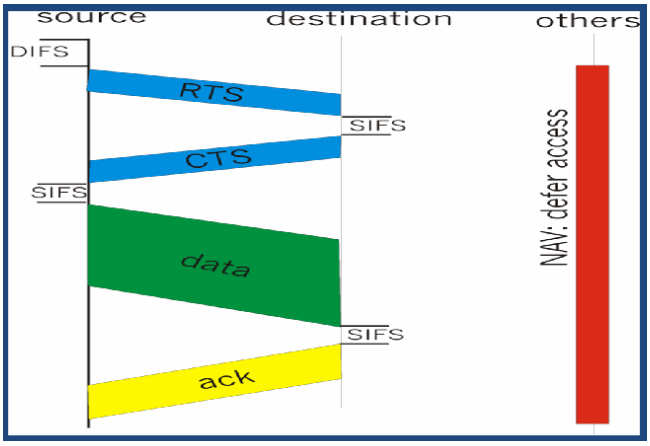


Figure 3: The four-way handshaking sequence using RTS/CTS. The timeline shows the frame exchange between the source and destination, and the resulting channel reservation (NAV) set by other stations.

The procedure unfolds as follows:

1. The **source** station, after sensing the channel is idle for DIFS and completing its backoff countdown, sends a short **RTS** frame to the destination. This frame essentially asks for permission to send a data frame.
2. The **destination**, upon receiving the RTS, waits for a SIFS period and replies with a short **CTS** frame. This frame grants permission to the source.
3. The source receives the CTS, waits for a SIFS period, and transmits its **data** frame.
4. The destination correctly receives the data, waits for a SIFS period, and sends the final **ACK** frame to confirm successful reception.

2.8.2 How Handshaking Solves the Problem

The effectiveness of this mechanism lies in the fact that the RTS and CTS frames are short and broadcast to all nearby stations. Both frames contain a **Duration** field, which specifies the total time required for the rest of the exchange (e.g., Data + ACK). This enables *Virtual Carrier Sensing*.

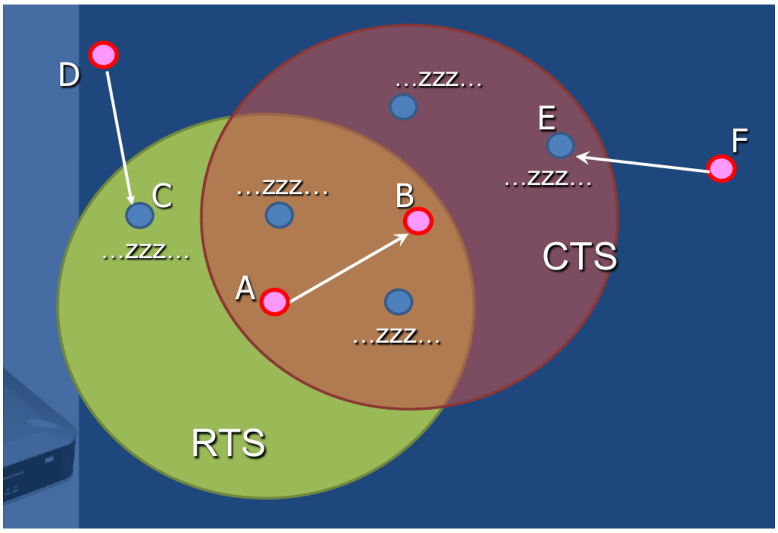


Figure 4: Visualizing how RTS/CTS solves the hidden terminal problem.

Let's analyze the scenario from Figure 4, where station A wants to transmit to station B, and D is a hidden terminal for A.

- Station A sends an RTS to B. Any station that can hear A (like station C) receives the RTS, reads the Duration field, and sets its **Network Allocation Vector (NAV)**. The NAV acts as a reservation timer; the station will not attempt to access the channel until this timer expires. These stations are now "silenced".
- Station B replies with a CTS. Any station that can hear B (including the hidden terminal D) receives the CTS. It also reads the Duration field and sets its own NAV, effectively deferring any transmission.

In this way, the CTS frame silences the hidden terminal D, clearing the channel for A's data frame to be received by B without collision. Because the RTS and CTS frames are very short, the probability of them colliding is much lower. Even if they do collide, the time wasted is minimal compared to a collision involving a large data frame. This makes the RTS/CTS handshake an effective, albeit overhead-intensive, solution for improving reliability in the presence of hidden terminals.

2.9 The DCF Basic Access Mode

Having introduced the general principles of CSMA/CA, we now delve into the detailed rules and mechanisms of the DCF Basic Access Mode, which does not use the RTS/CTS handshake. This is the default mode for transmitting all frames.

2.9.1 Carrier Sensing: Physical and Virtual

A station's decision to transmit is based on its assessment of the channel's state (busy or idle). This is achieved through two complementary carrier sensing mechanisms:

- **Physical Carrier Sensing:** Performed at the physical layer, this is the direct detection of radio energy on the channel. The station listens to the medium to determine if another station is transmitting.
- **Virtual Carrier Sensing:** Performed at the MAC layer, this is a mechanism that uses information from frame headers to predict how long the channel will be busy. The frame header of 802.11 frames contains a *Duration* field that indicates the time, in microseconds, required to transmit the current frame and its subsequent acknowledgment (if any). This information is used to set the Network Allocation Vector.

2.9.2 Network Allocation Vector (NAV)

The **Network Allocation Vector (NAV)** is the concrete implementation of virtual carrier sensing. It is a timer maintained by each station that indicates the predicted duration for which the medium will remain busy.

- When a station (that is not the intended recipient) receives a frame, it reads the Duration value and sets its NAV timer accordingly.
- The station considers the medium to be busy for the entire duration of the NAV.
- If a station receives a frame from an upper layer while its NAV is active, it behaves as if it had physically sensed the channel and found it busy, thus deferring its access.
- The channel is considered idle at the MAC layer only when the NAV timer expires.

The NAV is a crucial tool for preventing collisions, especially in the context of the RTS/CTS mechanism.

2.9.3 Unicast Traffic: Transmission Procedure

The procedure for transmitting a standard unicast frame follows a precise sequence for the transmitter, receiver, and neighboring stations.

Transmitter's Role:

1. It **senses** the channel.
2. If the channel is idle, it waits for a time equal to **DIFS**.
3. If the channel remains idle for the entire DIFS period, it transmits its MPDU (MAC Protocol Data Unit).

Receiver's Role:

1. It computes the checksum on the received frame to verify its integrity.
2. If the transmission is correct, it waits for a time equal to **SIFS** (which gives it priority over any station waiting for DIFS).
3. It sends an **ACK** frame back to the transmitter. The ACK should be transmitted at a rate less than or equal to the rate used by the transmitter for the data frame, ensuring the transmitter can receive it.

Neighbors' Role: Any neighboring station that overhears the data frame will:

1. Set their NAV to the value indicated in the transmitted MPDU.
2. The NAV is set to cover the duration of the data frame transmission plus one SIFS period plus the time for the subsequent ACK: $NAV = MPDU\_tx\_time + 1\ SIFS + ACK\_time$.

This sequence is visualized in Figure 5.

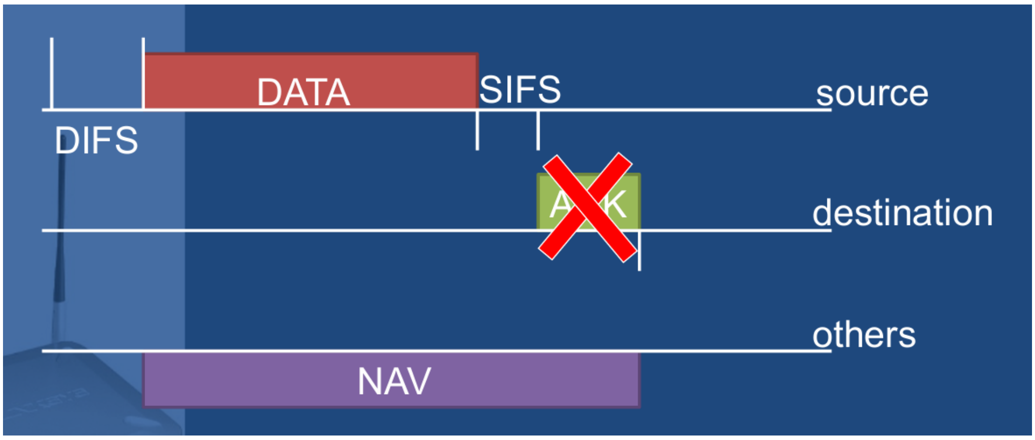


Figure 5: The timeline of a basic unicast MPDU transmission and the NAV set by neighboring stations.

2.9.4 Unicast Retransmissions

A frame transmission may fail due to a collision on the radio channel or due to errors induced by noise and fading.

- A failed transmission is inferred when the transmitter does not receive an ACK within a specified timeout.
- The MAC layer uses a simple **ARQ (Stop&Wait)** scheme: a failed transmission is re-attempted until a maximum number of retransmissions is reached.
- After a failed attempt, the station must **re-contend** for the channel by performing the backoff procedure.

2.9.5 The Backoff Procedure in Detail

A station must perform the backoff procedure (for collision avoidance) before accessing the channel in three main cases:

1. When it has a new packet to send, but it senses the channel as busy.
2. When it has just transmitted a frame unsuccessfully and needs to retransmit it.
3. When it has just transmitted a frame successfully (this is called a **post-backoff** and ensures fairness).

The backoff procedure follows these steps:

1. If a station senses the channel as busy (physically or virtually via NAV), it waits for the channel to become idle.
2. As soon as the channel is sensed as idle for a continuous **DIFS** period, the station:
   (a) Computes the backoff time interval.
   (b) Sets its internal backoff counter to a random value chosen from this interval.
   (c) Decrements the counter with time, but only while the channel remains idle.
3. The station is allowed to transmit its frame only when its backoff counter reaches 0.

Backoff Value Calculation (Exponential Backoff): The backoff time is not a fixed value. It is an integer corresponding to a number of time slots, chosen randomly from a uniformly distributed interval $[0, CW]$.

- **CW (Contention Window):** This is the range from which the random backoff value is chosen. It is updated at each transmission attempt using a **binary exponential backoff** algorithm: $CW = 2^{BE} - 1$, where BE is the Backoff Exponent.
- **First Attempt:** For the first attempt ($i = 1$), $BE_1 = BE_{min}$, so $CW_1 = CW_{min} = 2^{BE_{min}} - 1$.
- **Subsequent Attempts:** For each subsequent retransmission attempt ($i > 1$), the exponent is incremented: $BE_i = BE_{i-1} + 1$. This doubles the Contention Window: $CW_i = 2^{BE_i} - 1$.
- The Contention Window is capped at a maximum value: for any i, $CW_i \leq CW_{max}$.

This exponential increase in the contention window size dramatically reduces the probability of repeated collisions, as colliding stations will likely choose very different backoff values in their next attempt.

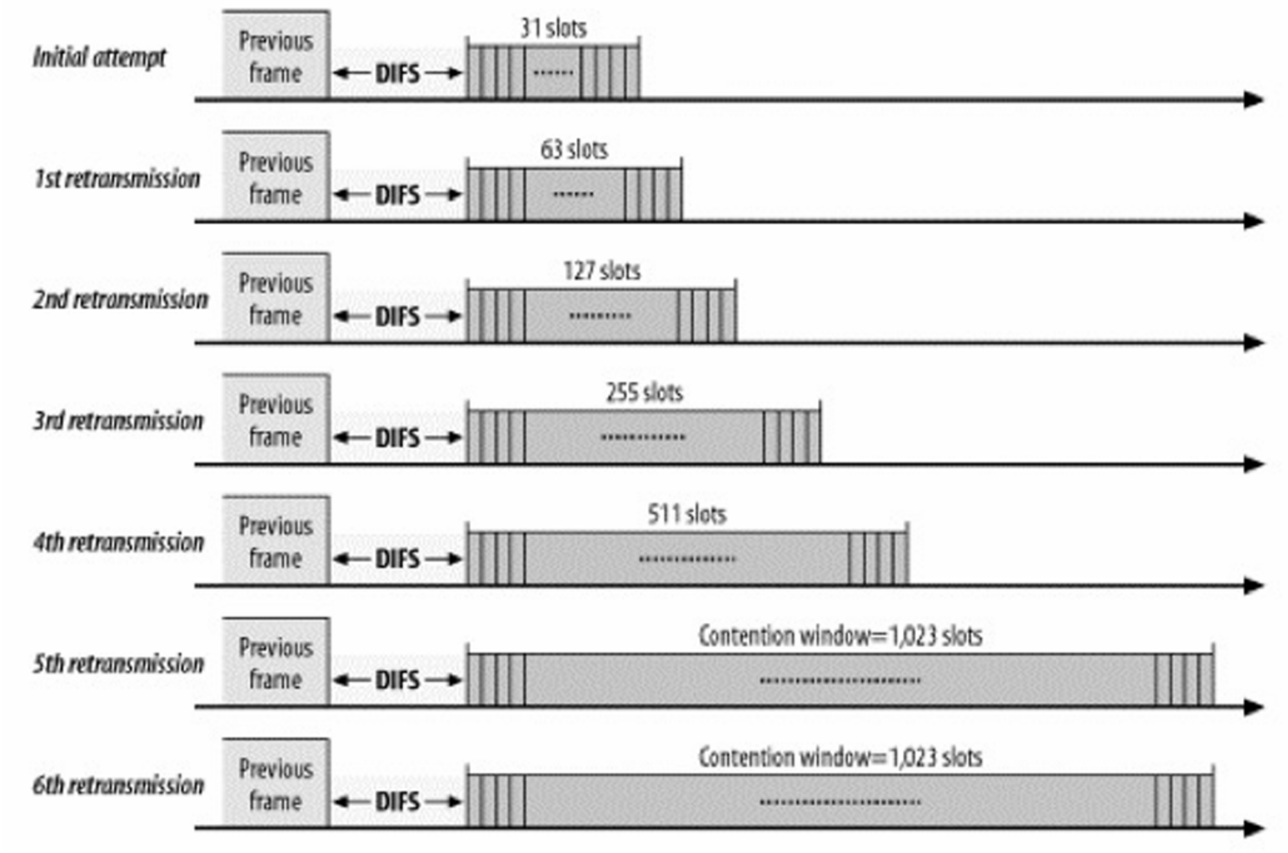


Figure 6: An example of exponential backoff, showing how the Contention Window (and thus the range of possible backoff slots) grows with each retransmission attempt.

Backoff Counter Decrease: The backoff counter does not decrease continuously.

- While the channel is sensed as busy, the backoff counter is **frozen**.
- The decrease can only be resumed after the channel has been sensed as idle for a continuous **DIFS** period.
- While the channel is idle, the station decrements the backoff value until either the channel becomes busy again (freezing the counter) or the counter reaches 0 (allowing transmission).

Post-Backoff: A station performs the backoff procedure not only on failures but also after completing a successful transmission. This *post-backoff* is crucial for fairness, preventing a station that just captured the channel from immediately grabbing it again. A station is only allowed to transmit after waiting just DIFS (without backoff) in two specific cases:

- The station has just entered the BSS.
- Its transmission queue has been idle, its post-backoff time has already passed, and no error was detected right before.

2.9.6 Special Cases: Broadcast/Multicast and EIFS

Broadcast/Multicast Traffic: These frames are handled differently from unicast frames.

- They are always transmitted at the **basic rate** to ensure all stations in the BSS can receive them.
- They are **never retransmitted** because there is no ACK mechanism for broadcast/multicast traffic (as this would cause an ACK storm). This makes them inherently unreliable.

EIFS (Extended Interframe Space): The EIFS is a special recovery mechanism.

- It is used by a station (let's say A) when its PHY layer notifies the MAC layer that a received frame was in error (i.e., failed the checksum).

- Station A must then wait for an EIFS period (which is longer than DIFS) after the corrupted frame ends before it can contend for the channel.
- The EIFS must be long enough to allow another station (C) to send an ACK to the original transmitter (B). This prevents station A, which did not correctly hear the frame from B, from interfering with the ACK from C to B.
- If station A receives an error-free frame during the EIFS period, the EIFS is terminated, and the station resynchronizes to the actual busy/idle state of the channel, resuming normal medium access (DIFS and backoff if necessary).

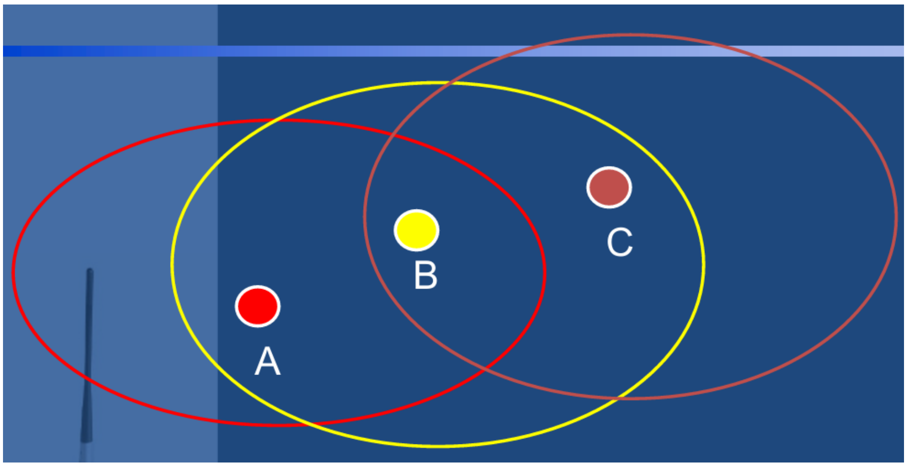


Figure 7: Example of EIFS usage. B transmits to C, but A receives an error. A must wait for EIFS, allowing C's ACK to be sent back to B without interference from A.

2.9.7 Summary of DCF Basic Access

- **Sender:** Transmits a frame after sensing the channel idle for DIFS + a random backoff.
- **Receiver:** If the frame is received correctly, it sends an ACK after a SIFS period.
- **Neighbors:** Remain silent if they heard the NAV. Otherwise, they detect the channel is busy via physical sensing and freeze their backoff counters.
- **Collisions:** There is no real-time collision detection. A collision is inferred when an expected ACK is not received. The sender then initiates an exponential backoff procedure to avoid subsequent collisions.

2.10 DCF Access with Handshaking (RTS/CTS)

While the basic access mode is sufficient for many scenarios, it is vulnerable to the hidden terminal problem. To address this, the 802.11 standard defines an optional four-way handshaking mechanism using Request to Send (RTS) and Clear to Send (CTS) frames.

2.10.1 Motivation for the Handshake

The primary reasons for using the RTS/CTS handshake are:

- **To reserve the channel explicitly**, which is the most effective way to combat the hidden terminal problem.
- **To avoid costly collisions with large frames.** When two large data frames collide, the bandwidth and time spent transmitting them are completely wasted. If, instead, two short RTS frames collide, the loss is minimal, and the subsequent data transmission is more likely to succeed.

This mechanism is particularly useful in environments with a large number of stations contending for the channel, where the probability of collisions is higher. The handshake is typically enabled only for frames larger than a configurable $RTS_{Threshold}$. RTS/CTS frames themselves are always transmitted at a robust *basic rate* to ensure all stations can decode them.

2.10.2 The Handshake Procedure in Detail

The roles of the transmitter, receiver, and neighbors in the full handshake are as follows:

Transmitter:

1. After successfully contending for the channel, sends an **RTS** (20 bytes long) to the destination.
2. Waits for the CTS response.
3. Upon receiving the CTS, waits for SIFS and then starts transmitting the data frame.

Receiver:

1. Upon receiving the RTS, waits for SIFS and acknowledges the request by sending a **CTS** (14 bytes long).

Neighbors:

- **Neighbors of the transmitter:** Read the duration field in the RTS and set their NAV accordingly.
- **Neighbors of the receiver (including hidden terminals):** Read the duration field in the CTS and update their NAV.

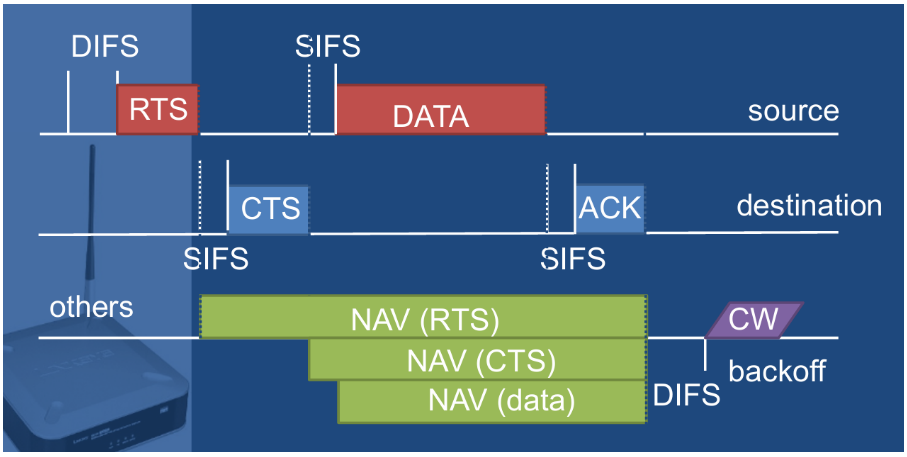


Figure 8: Detailed timeline of an MPDU transmission with RTS/CTS handshake and the corresponding NAV settings.

Figure 8 illustrates how the NAV is used to reserve the medium. The NAV set by the RTS frame ('NAV(RTS)') covers the entire transaction (CTS + SIFS + Data + SIFS + ACK). The NAV set by the CTS frame ('NAV(CTS)') is shorter, as it only needs to cover the remaining time.

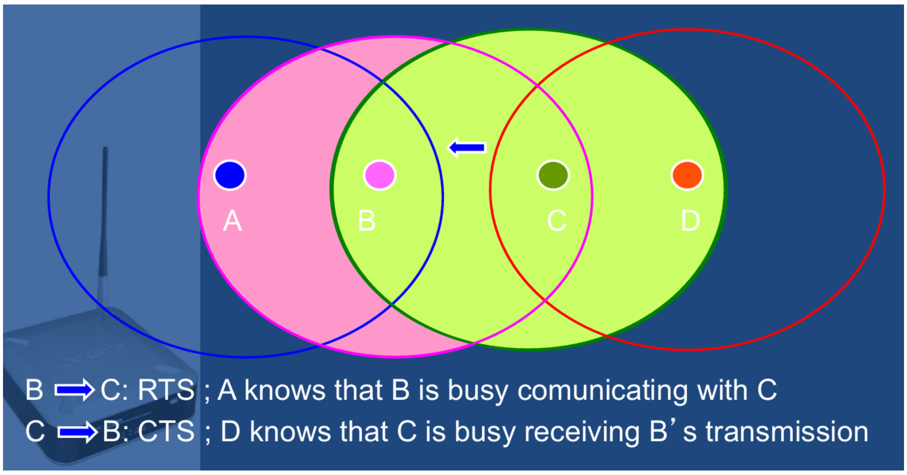

A
B
C
D
B ⇨C: RTS ; A knows that B is busy comunicating with C
C ⇨B: CTS ; D knows that C is busy receiving B' s transmission

Figure 9: Conceptual diagram of how RTS/CTS solves the hidden terminal problem. Station A is silenced by the RTS from B to C, while the hidden station D is silenced by the CTS from C to B.

2.10.3 Important Operational Details

- A station that detects only the RTS message will set its NAV accordingly and remain silent, even if it does not hear the corresponding CTS.
- If the station originating the RTS does not receive a CTS within a certain timeout, it assumes the destination is unreachable or the CTS was lost. It will not transmit its data frame and will start a new backoff procedure to re-attempt access later.
- A station that receives only the CTS will set its NAV accordingly, even if it did not hear the original RTS. This is the key mechanism that solves the hidden terminal problem, as shown in Figure 9 where station D is silenced.
- **Failure Case:** The mechanism is not foolproof. If a station close to the receiver does not hear the CTS (perhaps due to a collision with another transmission), it may wrongly assume the channel is free and attempt to transmit, causing a collision at the receiver during the data frame reception.

2.11 The Point Coordination Function (PCF) for QoS

While DCF is the mandatory default, the 802.11 standard also defines an optional, centralized access scheme called the **Point Coordination Function (PCF)**.

2.11.1 Basic Characteristics

- PCF is designed to support services with **Quality of Service (QoS)** requirements, such as real-time traffic, by providing a **contention-free access** to the channel.
- It requires a central **Point Coordinator (PC)**, which is always the Access Point. Therefore, PCF can only be implemented in infrastructure mode networks.
- The mechanism is based on a **polling** scheme, where the PC grants stations permission to transmit one by one.
- Stations enabled to operate under the PCF mode are known as **CF-aware** (Contention-Free aware).

2.11.2 PCF Operation and Coexistence with DCF

PCF is designed to coexist with DCF. Time is divided into periods where DCF is active and periods where PCF is active.

- The **Superframe** (or CFP Repetition Interval) defines the overall structure. Each Superframe contains a **Collision Free Period (CFP)**, where PCF is active, and a **Collision Period (CP)**, where DCF is active.
- The CFP is initiated by the AP sending a special **Beacon signal**. This Beacon frame contains a duration value that all stations use to set their NAV, effectively silencing all DCF-based contention for the entire duration of the CFP.
- The CFP terminates when the AP transmits a **CF_end** frame, at which point the CP begins and stations can start contending for the channel again using DCF.

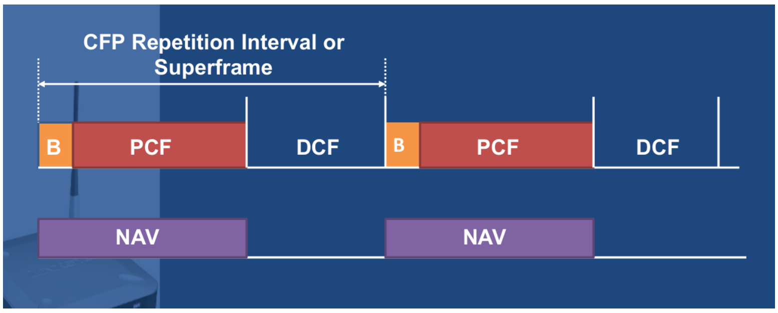
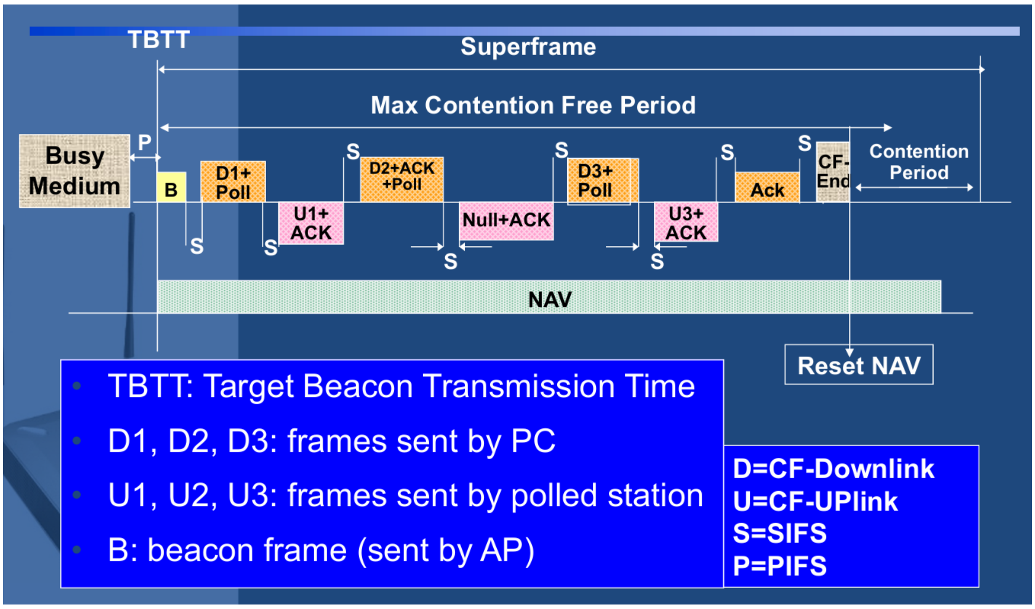


Figure 10: Coexistence between DCF and PCF within a Superframe structure. The NAV is used to suspend DCF operation during the CFP.

2.11.3 The PCF Protocol within a Superframe

During the CFP, the PC controls the medium by polling stations from a static polling list, which is built based on stations declaring their wish to participate during the association process.



- TBTT: Target Beacon Transmission Time
- D1, D2, D3: frames sent by PC
- U1, U2, U3: frames sent by polled station
- B: beacon frame (sent by AP)

D=CF-Downlink
U=CF-UPlink
S=SIFS
P=PIFS

Figure 11: Detailed protocol exchange within a PCF Superframe. (D=CF-Downlink, U=CF-Uplink, S=SIFS, P=PIFS).

As shown in Figure 11, the sequence is highly structured:

1. The period starts at the **TBTT (Target Beacon Transmission Time)**. The AP sends a Beacon to start the CFP.
2. The PC can then send a downlink data frame to a station, which may also contain a poll (**D1+Poll**).
3. The polled station responds after SIFS with an uplink frame, which may also contain an acknowledgment (**U1+ACK**).
4. The PC can also "piggyback" an ACK for the previous uplink frame with the next downlink frame and poll (**D2+ACK+Poll**).

2.12 Challenges for QoS in Early WLANs

Despite being designed for QoS, the PCF mechanism in early WiFi standards (WiFi 1-3) had several practical problems that limited its adoption:

1. **Unpredictable Beacon Delay:** A station in the DCF period does not have to stop its transmission just because a TBTT is reached. It continues its ongoing transmission. This means the Beacon frame, and thus the start of the CFP, can be delayed (up to 4.9 ms with large frames), making real-time guarantees difficult.
2. **Unknown Transmission Duration:** The PC does not know in advance how long a polled station will transmit, making it hard to schedule the CFP efficiently.
3. **Static Polling List Overhead:** The polling list is static and does not adapt to traffic needs. The PC must poll every station on the list, even if they have no data to send, leading to significant polling overhead and wasted time.

These limitations led to the development of a much more flexible and widely adopted QoS mechanism in later standards: **802.11e (Enhanced Distributed Channel Access - EDCA)**.

Part I
5G commnication

3 Introduction

3.1 TIM Group Overview

TIM is the main telecommunication and ICT operator in Italy and one of the most important in Latin America. Key figures (referred to Y2024):

- **Employees**: 17,751
- **Mobile Customers**: 91,639,000
- **Revenues** (€): 14,442,000,000
- **Sectors**: Fixed and mobile communications, internet and media, systems and solutions for business, R&D.
- **Strategic Markets**: Italy, Brasil.

3.2 TIM Technology Innovation – Wireless Access Activities

The department focuses on:

- R&D activities, scouting, testing, and performance evaluation of wireless systems (e.g., OFDMA).
- Study of innovative radio access technologies and antenna systems.
- Technological trials (e.g., 5G) and support to the engineering department.
- Involvement in standardization bodies and fora (3GPP, ETSI, NGMN, O-RAN, etc.).
- Recent participation in EU and national projects (e.g., 6G SNS, HEXA-X-II, CORENEXT, GOALS, RESTART).

3.3 Mobile Cellular System Evolution Path

Mobile networks have evolved through multiple generations, moving from analog voice to ultra-broadband and digital services:

- **1G (1980s, TACS)**: Voice Only, Analogic (CS voice).
- **2G (1990s, GSM)**: Voice and text, Digital (CS voice, PS data).
- **2.5G (GPRS)**: Introduced Packet Switched (PS) data services.
- **3G (2000s, HSPA)**: Voice, text, and Mobile Broadband (Mobile BB), supporting Video and Digital services.
- **4G (2010s, LTE/LTE-A)**: Mobile (Ultra) BB. The focus is on this generation and its evolution.
- **5G (2020s)**: Enhanced Mobile Broadband (eMBB), enhanced Machine Type Communication (eMTC), ultra Reliable and Low Latency Communications (uRLLC).
- **Towards 6G**: A tiny bit of 6G is included in the discussion.

4 The 4G Standard (a.k.a. LTE)

4.1 Standardization Activity: From 3G/HSPA to 4G/LTE

- **3GPP (3rd Generation Partnership Project)**: A global industry collaboration managing the standards and development of mobile communication technologies from 2G to 5G.
- **HSPA Evolution (3G/3.5G)**:

  - **Rel-5 (HSDPA)**: DL up to 14.4 Mbps.
  - **Rel-7 (HSPA+)**: DL up to 21 Mbps, UL up to 11 Mbps (64QAM/16QAM).
  - **Rel-8 (HSPA+ DC-HSDPA)**: DL up to 42 Mbps (Dual Cell).
  - **Rel-9 (HSPA+ Dual Band)**: DL up to 84 Mbps (Dual Band, 64QAM, MIMO).
  - **Rel-10 (HSPA Advanced)**: DL up to 168 Mbps (4C-HSDPA, MIMO, 64QAM).
- **LTE Introduction (4G)**:
  - **Rel-8 (LTE)**: DL up to 300 Mbps, UL up to 75 Mbps.
  - **Rel-10 (LTE-Advanced)**: DL up to 3 Gbps, UL up to 1.5 Gbps (Target).

4.2 LTE System Characteristics and Requirements

The original system requirements for LTE (defined in 3GPP TR 25.913) included:

- **Peak Data Rate**: 100 Mbit/s DL and 50 Mbit/s UL with 20 MHz bandwidth.
- **Spectral Efficiency**: Target 3÷4 times Rel-6 HSDPA (DL) and 2÷3 times Rel-6 HSUPA (UL) in a loaded network.
- **Mobility**: Optimized for low speed (0÷15 km/h), high performance for high speed (15÷120 km/h), and support for even higher speeds (120÷500 km/h).
- **Coverage**: Previous performance granted in cells up to 5 km, decreasing performance up to 30 km, and support up to 100 km.
- **Capacity (Control Plane)**: At least 200 users per cell in active state (5 MHz BW).
- **Latency**:
  - **User Plane**: $<$ 5 ms (single user with a single data stream).
  - **Control Plane**: $<$ 100 ms (camped state → active state).
- **Spectrum Flexibility**: Supports different allocations (1.25÷20 MHz), FDD and TDD.
- **Architecture**: Packet switched only, supporting end-to-end Quality of Service (QoS).
- **Interworking**: Granted with existing 3GPP systems and non-3GPP systems.

4.3 LTE Network Architecture: Evolved Packet System (EPS)

The EPS, also known as SAE (Service Architecture Evolution), is the evolution of the 3G system and includes:

- A new radio access, OFDM based, named **E-UTRAN** (Evolved Universal Terrestrial Radio Access Network).
- A new core network, completely IP based, named **Evolved Packet Core (EPC)**.

Main Innovations in EPS: Use of OFDMA, Packet-based service only, Always-on service, Flat architecture in the E-UTRAN.

Key Components (Functional Definitions):

- **UE (User Equipment)**: User device.
- **eNB (evolved Node B)**: The base station in E-UTRAN.
- **MME (Mobility Management Entity)**: Control plane (similar to SGSN in 3G).
- **Serving GW (SGW)**: Anchor point for inter-3GPP RAN mobility (user plane similar to SGSN in 3G).
- **PDN GW (PGW)**: Gateway to the Packet Data Network (similar to GGSN in 3G).
- **PCRF (Policy Control and Charging Rules Function)**: Manages policy control and charging.

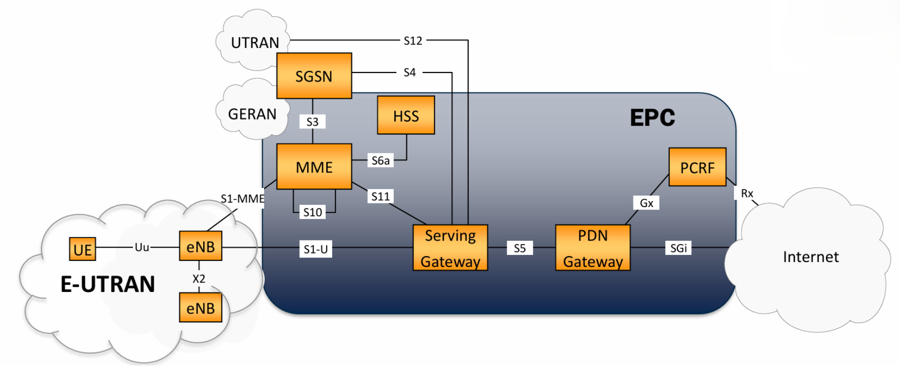


Figure 12: Evolved Packet System (EPS) Architecture

5 LTE Radio Access – Key Enablers

5.1 Key Ingredients in LTE

The Long Term Evolution (LTE) is based on several key enabling elements:

- Orthogonal Frequency Division Multiplexing (OFDM).
- Multiple Input Multiple Output (MIMO) / Multiple Input Multiple Output (MIMO).
- Higher order modulations & Adaptive Modulation and Coding (AMC).
- Scalable bandwidth (1.4 MHz to 20 MHz).
- Flat Architecture.

5.2 Orthogonal Frequency Division Multiplexing (OFDM)

5.2.1 Advantages of OFDM

OFDM offers three main advantages over previous techniques:

- **High robustness to channel variations** (multipath fading).
- **High spectral efficiency**.
- **Efficient Hardware implementation**.

5.2.2 OFDM vs. WCDMA in Fading Channels

- **Fading**: In wireless channels, multiple reflected paths cause frequency-selective fading (non-flat channel response).
- **WCDMA**: A wideband signal sees a frequency-selective channel, causing high distortion and requiring complex channel equalization.
- **OFDM**: Converts the frequency selective channel into N_c ”flat” fading channels, one for each sub-carrier. This provides **high resistance to multipath fading** and **”built-in” frequency domain equalization**.

5.2.3 Orthogonality in OFDM

- The basic principle is similar to Frequency Division Multiplexing (FDM) but with higher spectral efficiency due to **partial overlapping** of subcarriers.
- **Orthogonality** is achieved because each subcarrier spectrum reaches its peak where the other subcarriers are null.
- This property allows for easy generation using the **Inverse Fast Fourier Transform (IFFT)** at the transmitter and decoding using the **Fast Fourier Transform (FFT)** at the receiver.

[h!]

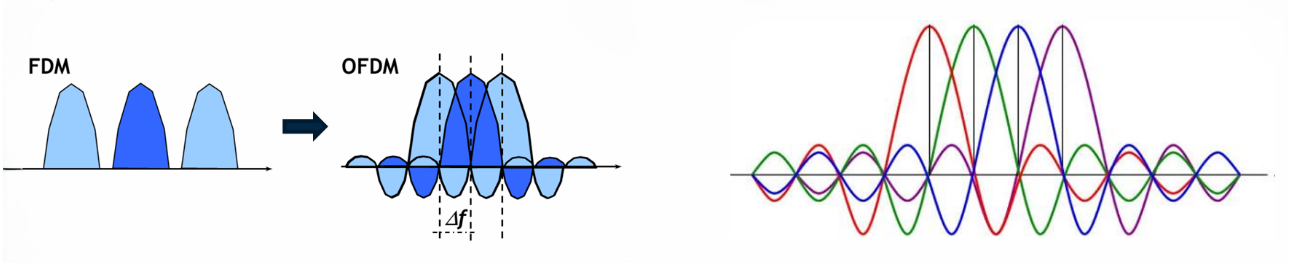


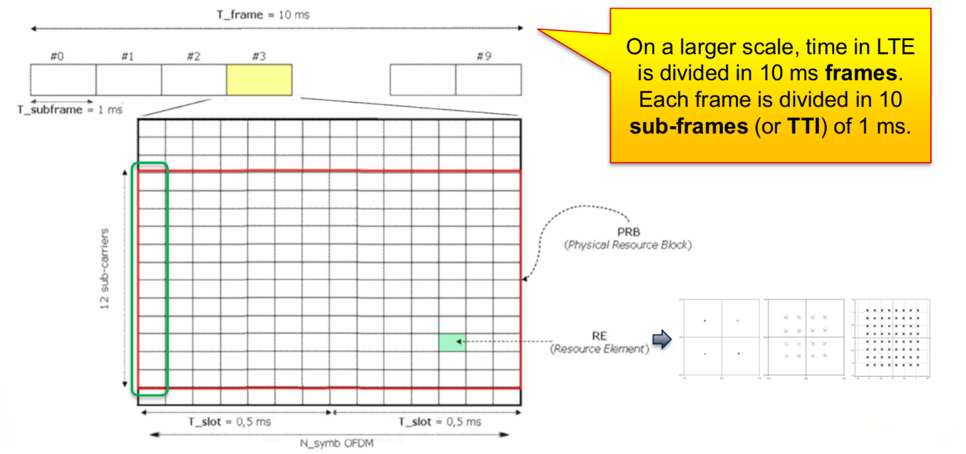
Figure 13: FDM evolution.

5.3 OFDM and Fast Fourier Transform (FFT)

- A single OFDM symbol signal $x(t)$ is a sum of N_c modulated subcarriers, each multiplied by a pulse $u(t)$ of duration $T_u = 1/\Delta f$ (where Δf is the subcarrier spacing).
- $x(t)$ in the time interval $mT_u \leq t < (m+1)T_u$ represents one OFDM symbol.
- The complex modulation symbols $a_k^{(m)}$ applied to the k-th subcarrier are input to an **Inverse Discrete Fourier Transform (IDFT)** to generate the time-discrete OFDM signal samples.
- When $N = 2^m$ (where N is the total number of samples and N_c is the number of subcarriers), the IDFT is implemented efficiently using the **Inverse Fast Fourier Transform (IFFT)**.

5.4 LTE Downlink (DL) Radio Access

5.4.1 The OFDM Time-Frequency Grid



On a larger scale, time in LTE is divided in 10 ms **frames**. Each frame is divided in 10 **sub-frames** (or **TTI**) of 1 ms.

Figure 14: OFDM time-frequency grid.

- **Resource Element (RE)**: One sub-carrier in one OFDM symbol. Carries a QPSK, 16QAM, or 64QAM symbol (2, 4, or 6 bits).
- **Resource Block (RB)**: A frequency-time grid of 12 sub-carriers (180 KHz) and 14 OFDM symbols.
- **Time Transmission Interval (TTI)**: The time domain is divided into 10 ms frames, each divided into 10 sub-frames (or TTI) of 1 ms. 1 TTI contains two 0.5 ms slots (14 OFDM symbols total).

5.4.2 Multiple Access in Downlink: OFDM-A

- OFDM simplifies multi-user allocation (multiple access) by assigning different subcarriers to different users.
- Resources are scheduled to users in terms of multiple RBs on each TTI (1ms).

5.4.3 Downlink Subframe Structure and Physical Channels

Every subframe carries both user data and control information.

Control and Data Allocation:

- **Control Information**: Mapped to control channels (PDCCH, PCFICH, PHICH) allocated on the first $n \leq 3$ OFDM symbols of each TTI.
- **Reference Signals (RS)**: Known symbols used by the receiver to estimate the channel.
- **Users Data**: Mapped on the **PDSCH** (Physical Downlink Shared Channel), covering the remaining resources.

Physical Channels in Downlink:

- **PDSCH**: Carries user data, paging request, SIBs.
- **PDCCH** (Control): Provides DL Control Information (DCI) including PDSCH/PUSCH allocation.
- **PCFICH** (Control): Indicates the PDCCH format (number of used symbols).
- **PHICH** (Control): H-ARQ Ack/Nack feedback for UL transmission.
- **PBCH** (Broadcast): Carries the MIB (Master Information Block).
- **PMCH** (Multicast): Used for eMBMS (multicast traffic).

5.4.4 Downlink Reference Signals (RS) and Cell Id

- **Reference Signals (RSs)** are pseudo-random sequences inserted in the first and third-to-last OFDM symbols of every time slot.
- **RS Sequence Dependency**: Depends on the time slot number and the **Physical Cell Identifier (PCI)**.
- **RS Usage**: Used by the receiver to: 1) Compensate channel effects during decoding; 2) Estimate overall channel quality (e.g., CQI calculation, MIMO).
- **PCI**: Every LTE cell is identified with a specific PCI. 504 cell IDs are available, divided into 168 groups of 3 specific IDs.
- **Channel Estimation**: By observing the received signal $\mathbf{y}$ and knowing the transmitted RS $\mathbf{r}$, the channel effect $\mathbf{h}$ can be estimated as $\hat{\mathbf{h}} = \mathbf{y} \cdot \mathbf{r}^*/||\mathbf{r}||^2$.

5.5 Adaptive Modulation and Coding (AMC)

Adaptive Modulation and Coding (AMC) is a key technique to maximize spectral efficiency by adjusting the modulation and coding scheme (MCS) based on the channel quality perceived by the UE.

- **Feedback Mechanism**: LTE terminals provide a feedback to the eNB called the **Channel Quality Indicator (CQI)**.
- **eNB Action**: The eNB uses the CQI to select the user, the modulation scheme, and the coding protection robustness (MCS).
- **CQI Definition**: On every TTI (Time Transmission Interval), the UE estimates the channel quality using **Reference Signals (RS)** and sends the CQI to the eNB.
- **Transport Format Suggestion**: The CQI suggests a transport format (modulation and coding rate) suitable to receive the next packet with a **Block Error Rate (BLER) of 10%**.
- **Modulation Schemes**: LTE uses **QPSK** (2 bits/symbol), **16QAM** (4 bits/symbol), and **64QAM** (6 bits/symbol) in both DL and UL. Higher order modulations offer higher spectral efficiency but lower protection (i.e., less robustness to interference).
- **Throughput Trade-off**:
  - **Low SINR (High Interference)**: A robust transport format is used (low modulation order, high coding protection) to achieve low error probability, but this results in lower data rates.
  - **High SINR (Low Interference)**: A less robust transport format is used (high modulation order, low coding protection) to achieve low error probability, resulting in higher data rates.

5.6 Multiple Input Multiple Output (MIMO)

MIMO is one of the main enablers for high LTE speeds, increasing the spectral efficiency of the system.

- **Mechanism**: A signal processing technique that exploits several antennas at the transmitter and receiver, operating simultaneously on the same frequencies.
- **Channel Requirement**: MIMO works better with OFDM and a **multipath-rich channel** (where transmitted signals reach the receiver via different propagation paths and delays).
- **Modes of Operation**:
  1. **Spatial Multiplexing (SM)**: To **increase peak throughput**. Different data flows are transmitted, which are orthogonal "in space." Multiple antennas are needed at the receiver to separate the flows (requires advanced space-time equalization algorithms).
  2. **Space Frequency Coding (SFC) / Transmit Diversity (TxD)**: To **increase coverage** and robustness. Different copies of the same data flow are transmitted with different encoding to exploit spatial diversity and improve received signal quality at the cell edge.
- **MIMO in LTE (Downlink only)**: MIMO uses a **fixed set of precoding matrices (codebook)** since the eNB does not know the full channel matrix.
- **Precoding Usage**:
  - **Open Loop**: Predefined precoders are cyclically allocated. The UE reports a **Rank Indicator (RI)** to signal the number of parallel streams.
  - **Closed Loop (CL)**: The UE identifies and reports the best precoding matrix in the available set using the **Precoding Matrix Indicator (PMI)**, along with the RI.
- **MIMO Modes (Rel-8)**:
  - **Mode 3 (Open-loop Spatial Multiplexing)**: Dynamically adapts the transmission rank (RI). Falls back to **Mode 2 (Transmit Diversity)** if channel conditions worsen.
  - **Mode 4 (Closed-loop Spatial Multiplexing)**: Dynamically adapts the rank (RI) and uses the best precoding matrix suggested by the UE (PMI). Falls back to **Mode 6 (Closed-loop rank 1 precoding)**.
  - **Throughput Evolution**: MIMO 2x2 on 10MHz (75 Mbps) → MIMO 4x4 on 10MHz (150 Mbps) → MIMO 8x8 on 10MHz (300 Mbps).

5.7 Scalable Bandwidth

LTE uses a scalable bandwidth design based on the OFDM structure, allowing flexible deployment.

- **Flexibility**: Unlike UMTS/WCDMA (fixed at 5 MHz), the OFDM structure allows the system to easily adapt the bandwidth simply by changing the number of available sub-carriers.
- **Supported Sizes**: The LTE standard supports bandwidths of **1.4MHz, 3MHz, 5 MHz, 10MHz, 15MHz, and 20 MHz**.
- **Italian Deployment Example**: In Italy, LTE frequencies (auctioned in 2011) are deployed in different 5 MHz blocks at 800, 1800, and 2600 MHz.
- **Multilayer Approach (TIM)**:
  - **2600 MHz**: Used to **maximize capacity** (Metropolitan Area).
  - **1800 MHz**: Band with **high capacity** and **limited interference**.
  - **800 MHz**: Band with **high coverage** and **limited interference** (Rural Area).

5.8 LTE Release 8 Peak Performance

Theoretical values achievable according to Release 8 3GPP specifications (20 MHz BW):

Factors Affecting Data Rate Performance: Data rate per user depends on:

1. **User position** (SINR): Users near the antenna (low interference) can use less robust transport formats (high data rate).
2. **Amount of traffic in the cell**: High traffic leads to high interference and lower SINR, potentially forcing the use of more robust (lower data rate) transport formats.

Table 2: LTE Release 8 Peak Performance (Theoretical)

Metric	Downlink (DL)	Uplink (UL)
Peak Data Rate	300 Mbps (4x4 MIMO) 150 Mbps (2x2 MIMO)	75 Mbps (1x2 SIMO)
Peak Spectral Efficiency	$\approx$ 16.3 bit/s/Hz	$\approx$ 4.3 bit/s/Hz (1x2 SIMO)
Data Plane Latency Control Plane Latency	$>$ 10 ms (round trip delay) 100 ms (idle to active state)	

6 LTE Advanced (Rel-10 and beyond)

LTE Advanced defines the evolution path from standard LTE (Rel-8/9) towards 5G, primarily focusing on improving throughput, capacity, and supporting new services.

Main Radio Features (Timeline):

- **Rel-10 (2011)**: Introduction of **Carrier Aggregation (CA)** and **Enhanced MIMO** (up to 8 DL and 4 UL layers).
- **Rel-11 (2013)**: Introduction of **DL and UL CoMP** (Coordinated Multi-Point).
- **Rel-12/13 (2015/2016)**: Introduction of **NB-IoT** and **LTE-M** for Massive Machine Type Communications.

6.1 Carrier Aggregation (CA)

Carrier Aggregation, also known as **4GPLUS**, is the key feature in Rel-10 to significantly increase peak data rates by combining multiple component carriers (CCs).

- **Mechanism**: Aggregates multiple separate frequency blocks (carriers) from different bands (e.g., 2600 MHz, 1800 MHz, 800 MHz) into a single, wider virtual channel.
- **Throughput Gain**:
  - **Single Carrier (e.g., 20 MHz total)**: $\approx$ 110 Mbps.
  - **2-Carrier Aggregation (e.g., 2CC × 5 MHz + 2CC × 5 MHz)**: $\approx$ 225 Mbps (Example: 2600 MHz + 1800 MHz).
  - **3-Carrier Aggregation (3C, e.g., 3CC × 5 MHz + 3CC × 5 MHz + 2CC × 5 MHz)**: $\approx$ 300 Mbps (Example: 2600 MHz + 1800 MHz + 800 MHz).

6.2 LTE Broadcast (eMBMS)

Evolved Multimedia Broadcast Multicast Services (eMBMS) is an LTE feature designed for efficient delivery of the same content to a large number of users simultaneously.

- **Unicast vs. Broadcast**:
  - **Unicast**: One data channel per device, limited on the maximum number of users.
  - **Broadcast (eMBMS)**: One data channel for the content, **unlimited number of users**, providing high quality for everybody.
- **Use Case**: Ideal for high-demand content delivery such as live TV streaming to a large audience in a localized area (e.g., stadium, city center).

6.3 Multi Antenna Techniques: Higher Order and Multi-User MIMO

LTE Advanced introduced higher-order MIMO configurations and Multi-User MIMO to further enhance throughput.

- **Configuration Evolution**: LTE MIMO evolves from 2x2 to higher orders like **MIMO 4x4** and **MIMO 8x8**.
- **Challenge**: Higher order configurations require radiating systems with increasing dimension and complexity on both the UE side (device engineering) and the Base Station (BS) side (site design).

- **Single User MIMO (SU-MIMO)**: Dedicated beams/layers for a single UE (e.g., 4x4 or 8x8).
- **Multi-User MIMO (MU-MIMO)**: Multiple UEs are served simultaneously by the same eNB resources, with beams spatially separated to avoid interference. This increases cell capacity (multiple users served in parallel).

6.4 Coordinated Multi-Point (CoMP) Transmission/Reception

CoMP is an LTE-Advanced Rel. 11 feature designed to improve the coverage of high data rates, enhance cell-edge throughput, and increase overall system throughput.

- **Mechanism**: Different eNBs **coordinate their transmissions** to reduce interference for users, or **cooperate** to increase the received signal quality.
- **Benefit**: Particularly beneficial for **cell-edge users** whose performance is typically degraded by interference from neighboring cells.
- **Result**: The cell-edge user experience is significantly improved compared to a conventional LTE system without CoMP.

7 Massive Machine Type Communication (MTC)

7.1 Narrowband Internet of Things (NB-IoT)

NB-IoT was introduced in **3GPP Release 13** to support Massive MTC.

Main Capabilities:

- **Bandwidth**: Very small bandwidth deployment (**200 kHz**).
- **Coverage**: **Extended coverage** ($\geq$ 20 dB enhancement) compared with existing cellular technologies (e.g., GPRS). Deep coverage: 164 dB MCL.
- **Battery Life**: Optimized for very long terminal battery life ($\geq$ 10 years). Uses eDRX and PSM (Power Saving Mode).
- **Connections**: Support for **massive connections** ($\geq$ 50K devices/cell).
- **Cost**: Optimized for ultra-low terminal cost ($<$ 5\$, relative BOM cost $<$ 12%).
- **Mobility**: Nomadic Only (cell reselection supported, but not full mobility).
- **Peak Data Rate**: $\approx$ 200 kbps DL/UL ($\approx$ 20/60 kbps sustainable).
- **Latency**: $<$ 10 sec at deep coverage (164 dB).

Coverage Extension Techniques (20 dB gain):

- **Increased Power Spectral Density (PSD)**: In the uplink, single tone transmission concentrates the UE power (e.g., 23 dBm) in a small bandwidth (3.75 or 15 kHz).
- **Burst Repetition**: Soft combining of multiple repeated bursts at the receiver.
- **Power Efficient Modulation**: Use of schemes with low PAPR (e.g., $\pi/2$-BPSK) to avoid complex and expensive power amplifiers.

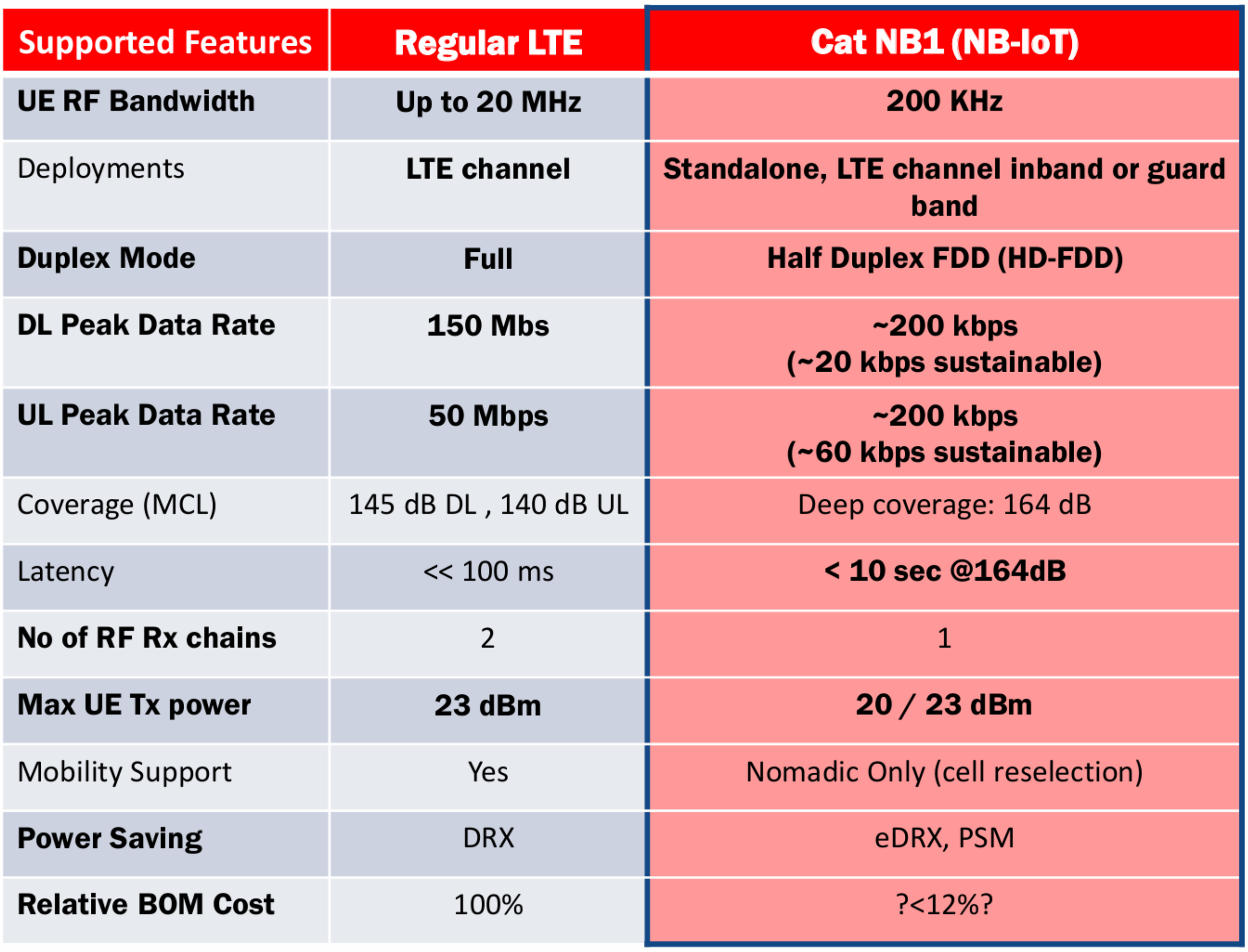
Flexible Deployment Modes: NB-IoT can reuse existing GSM or LTE infrastructure.

1. **Stand-alone operation**: Utilizes GSM or scattered spectrum.
2. **Guard band operation**: Utilizes unused Resource Blocks (RBs) within an LTE carrier's guard-band.
3. **In-band operation**: Utilizes RBs within a normal LTE carrier.

7.2 LTE-M (a.k.a. eMTC or Cat M1)

LTE-M was developed in parallel with NB-IoT, serving as an evolution of the previous UE Category 0 activity.

- **Bandwidth**: Wider bandwidth (**1.4 MHz**).
- **Duplex Mode**: Half Duplex (optional).
- **Peak Data Rate**: Higher data rate (DL $\approx$ 800 Kbps, UL $\approx$ 1 Mbps).
- **Receiver**: Traditional LTE receiver, stripped down of some functions, with additional features to improve coverage and reduce power consumption (BOM cost $\approx$ 20% of LTE).
- **Mobility**: Full Mobility Support (not explicitly listed, but implied by the traditional LTE receiver design compared to NB-IoT's Nomadic).



Supported Features	Regular LTE	Cat NB1 (NB-IoT)
UE RF Bandwidth	**Up to 20 MHz**	**200 KHz**
Deployments	**LTE channel**	**Standalone, LTE channel inband or guard band**
Duplex Mode	**Full**	**Half Duplex FDD (HD-FDD)**
DL Peak Data Rate	**150 Mbs**	**~200 kbps (~20 kbps sustainable)**
UL Peak Data Rate	**50 Mbps**	**~200 kbps (~60 kbps sustainable)**
Coverage (MCL)	145 dB DL , 140 dB UL	Deep coverage: 164 dB
Latency	<< 100 ms	**< 10 sec @164dB**
No of RF Rx chains	2	1
Max UE Tx power	**23 dBm**	**20 / 23 dBm**
Mobility Support	Yes	Nomadic Only (cell reselection)
Power Saving	DRX	eDRX, PSM
Relative BOM Cost	100%	?<12%?

Figure 15: LTE and NB-IoT Comparison.

Part II
The 5G Core Network and Enabling Technologies

The transition from 4G to 5G represents a fundamental paradigm shift in the architecture of mobile networks. While previous generations focused primarily on higher data rates and hardware evolution, 5G aims to support a heterogeneous ecosystem of services through a complete redesign of the Core Network (CN). This transformation is driven by software, moving away from the "one size fits all" approach of LTE towards a flexible, programmable, and service-oriented architecture. To understand 5G, we must first analyze the two key technologies that enable this flexibility: **Software Defined Networking (SDN)** and **Network Function Virtualization (NFV)**.

8 The Need for a New Architecture: Verticals and KPIs

5G is not just about faster smartphones; it is designed to serve "Vertical Industries" (e.g., automotive, industry 4.0, e-health), each with conflicting performance requirements (KPIs):

- **Enhanced Mobile Broadband (eMBB):** Services like Virtual Reality (VR) and UHD streaming require huge throughput (100 Mbps to 20 Gbps) and moderate latency.
- **Ultra-Reliable Low Latency Communications (URLLC):** Mission-critical applications such as autonomous driving (V2X) or remote surgery require ultra-low latency (< 1 ms) and virtually zero packet loss (99.999% reliability), but not necessarily high data rates.
- **Massive Machine Type Communications (mMTC):** Scenarios like Smart Cities or agriculture require connecting millions of sensors (10^6 devices/km^2) with low data rates and extremely long battery life.

The traditional LTE architecture, built on specialized hardware nodes, cannot simultaneously satisfy these diverse requirements efficiently. A rigid network cannot be optimized for both high-throughput video and low-power sensors at the same time. This necessitates the concept of **Network Slicing**, enabled by SDN and NFV.

9 Software Defined Networking (SDN)

Software Defined Networking (SDN) addresses the rigidity of traditional telecommunications infrastructure. In legacy networks, the "brain" (control logic) and the "muscle" (packet forwarding) are bundled together inside proprietary hardware appliances (routers, switches, firewalls). This vertical integration creates "closed platforms" where innovation is slow, and configuration is complex and vendor-specific.

9.1 The SDN Paradigm Shift

SDN proposes a radical separation of concerns by decoupling the **Control Plane** from the **Data Plane**.

- **The Data Plane (The Muscle):** The network devices (switches) become simple, fast, and "dumb" forwarding elements. They no longer run complex routing algorithms (like OSPF or BGP). Their only job is to match packets against a table and perform simple actions (forward, drop, modify). They are implemented in highly efficient hardware (Merchant Silicon) but perform no "thinking".
- **The Control Plane (The Brain):** The intelligence is removed from the switches and centralized in a logically central software entity called the **SDN Controller**. The controller has a global, abstract view of the entire network topology and state. It computes the optimal paths and instructs the switches on how to handle traffic.

Analogy: This evolution is similar to the computing industry's shift from Mainframes (vertically integrated, proprietary hardware/OS/apps) to the PC era (horizontal architecture with open interfaces between standard hardware, OS, and applications). SDN brings this "horizontal" revolution to networking.

9.2 Interfaces and Programmability

The power of SDN lies in its open interfaces:

1. **Southbound Interface (SBI):** This connects the Controller to the physical switches. The de-facto standard protocol is **OpenFlow**. Through OpenFlow, the controller pushes ”Flow Tables” to the switches. A flow table entry consists of a *Rule* (matching header bits like IP, MAC, Port), an *Action* (what to do), and *Statistics* (counters).

2. **Northbound Interface (NBI):** This exposes the network's capabilities to applications ”on top” of the controller via programmable APIs (usually RESTful). This allows developers to write network applications (e.g., Load Balancers, Firewalls, Traffic Engineering apps) that program the network behavior dynamically, treating the network like a software platform.

9.3 SDN in Cellular Networks

Applying SDN to the mobile Core Network solves several LTE limitations. In LTE, the Packet Gateway (P-GW) is a centralized bottleneck that handles everything: routing, billing, QoS, and legal interception. These P-GWs are expensive, proprietary hardware. With SDN, we can:

- **Offload the Data Plane:** Expensive tasks can be offloaded to simple OpenFlow switches, reducing the cost per bit.

- **Optimize Routing:** Traffic can be steered dynamically. For instance, local traffic (e.g., Content Delivery Network caches) can be routed locally without traversing the entire core network, reducing latency.

- **Unified Management:** A single controller can manage the forwarding across different technologies (LTE, 5G, Wi-Fi), enabling seamless vertical handovers.

10 Network Function Virtualization (NFV)

While SDN separates control from data, **Network Function Virtualization (NFV)** decouples the network software from the proprietary hardware. Traditionally, a Mobile Network Operator (MNO) needing a new function (e.g., a Multimedia Messaging Service Center) had to buy a specific physical ”box” (appliance) from a vendor. This led to high CAPEX (Capital Expenditure), high OPEX (Operating Expenditure), and long Time-to-Market.

10.1 The NFV Concept

NFV proposes to virtualize these functions. Instead of physical appliances, we use **Virtual Network Functions (VNFs)**. A VNF is simply software implementing a network function (like a Router, a Firewall, or an LTE EPC) running on standard, high-volume servers (Commercial Off-The-Shelf - COTS) located in data centers. This allows operators to:

- Consolidate equipment: Use the same physical servers for different functions (e.g., running a firewall during the day and a video transcoder at night).

- Scale elastically: If traffic spikes, the operator simply spins up more Virtual Machines (VMs) or containers, rather than buying new hardware.

- Speed up innovation: Deploying a new service becomes as simple as installing software.

10.2 Service Chaining

In the NFV world, a ”Network Service” is defined as a chain of VNFs connected in a specific order, known as a **Forwarding Graph**. For example, a ”Secure Internet Access” service might consist of a chain: *Firewall* → *Deep Packet Inspection* → *NAT*. The NFV infrastructure manages the logical connections between these virtual functions, regardless of which physical server they reside on.

11 Network Slicing

Network Slicing is the culmination of SDN and NFV. It allows an operator to partition a single physical network infrastructure into multiple, isolated **logical networks**, called ”slices”. Each slice is an independent, self-contained network tailored to a specific use case:

- An **Automotive Slice** might be configured for ultra-low latency, with VNFs deployed at the network edge (MEC) and strict resource reservation.
- An **IoT Slice** might be optimized for massive signaling handling but low bandwidth, using lightweight control plane functions.
- A **Mobile Broadband Slice** might focus on high-throughput caching and video optimization.

Thanks to SDN, the traffic for each slice is isolated and routed independently. Thanks to NFV, the resources (computing, storage) are allocated dynamically to each slice. This creates a multi-tenant environment where ”Verticals” (e.g., a car manufacturer or a hospital) can buy a ”slice” of the 5G network that behaves exactly like their own private physical network.

11.1 The 5G Business Ecosystem

This architecture introduces new business roles:

1. **Infrastructure Provider (InP):** Owns the physical data centers, antennas, and cables.
2. **Mobile Network Operator (MNO):** Leases raw resources from the InP to instantiate and manage network slices.
3. **Tenant:** The third-party customer (e.g., Netflix, BMW) that purchases a specific slice to deliver services to its end users.

Part III

From Cloud to Edge and MEC

This section of the course analyzes the evolution of computing architectures supporting mobile networks, starting from the Cloud Computing paradigm up to Edge Computing and the ETSI MEC (Multi-access Edge Computing) standard. The goal is to understand how to reduce latency and improve efficiency for next-generation applications.

12 Evolution towards Mobile Cloud Computing

The current Internet faces challenges it was not originally designed for, specifically **mobility** (not natively supported), massive data access, and security. The architectural evolution followed a specific path:

1. **PC-based computing:** Local computing on fixed machines.
2. **Mobile computing:** The advent of smart devices.
3. **Cloud Computing:** Centralization of storage and computation in large remote data centers.

12.1 Cloud Computing

According to the NIST definition, Cloud Computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort.

The Cloud is based on **5 essential characteristics**:

- **On-demand self-service:** A consumer can use computing capabilities automatically without requiring human interaction with each service provider.
- **Broad network access:** Capabilities are available over the network and accessed through standard mechanisms by heterogeneous client platforms (e.g., mobile phones, laptops).
- **Resource pooling:** The provider's resources are pooled to serve multiple consumers using a multi-tenant model, dynamically assigned according to demand. The user generally has no control over the exact physical location of the resources.
- **Rapid elasticity:** Capabilities can be elastically provisioned and released to scale rapidly outward and inward commensurate with demand.
- **Measured service:** Resource usage is automatically monitored, controlled, and reported, providing transparency for both the provider and consumer.

Cloud services are delivered according to three main models:

- **SaaS (Software as a Service):** The user uses the provider's applications running on a cloud infrastructure (e.g., Gmail, Dropbox). The consumer does not manage the underlying infrastructure.
- **PaaS (Platform as a Service):** The user deploys their own applications created using programming languages and tools supported by the provider. The user controls the deployed applications but not the underlying infrastructure (servers, OS).
- **IaaS (Infrastructure as a Service):** The user provisions fundamental resources (processing, storage, networks) where they can deploy and run arbitrary software, including operating systems. The user has control over the OS and deployed applications.

12.2 Mobile Cloud Computing (MCC)

Mobile Cloud Computing (MCC) integrates cloud computing into the mobile environment to overcome the intrinsic limitations of mobile devices (battery, storage, processing power).

- **Advantages:**
  - *Extending battery lifetime:* Offloading computation to the cloud saves energy on the device.
  - *Improving storage and processing:* Access to virtually unlimited resources.

- *Improving reliability:* Data is backed up on multiple remote computers.

- **Main Issues:**
  - *Low Bandwidth:* Radio resources are scarce compared to wired networks.
  - *Availability:* Congestion or coverage issues may cause cloud services to become unavailable.
  - *Privacy & Security:* Sensitive data leaves the device to be processed by third parties; mobile devices are also exposed to threats like malicious codes.

13 Edge Computing

To address the latency and bandwidth limitations of centralized Cloud, the paradigm shifts to **Edge Computing**. The fundamental idea is to distribute computing, storage, and networking closer to users and end devices.

The key features of Edge Computing are:

- **Proximity (Lower Latency):** Being close to the source of information reduces latency, which is crucial for applications like AR/VR or autonomous driving.
- **On-premises:** It can run isolated from the rest of the network, ensuring local reliability, resilience, and privacy.
- **Network Context Information:** Real-time network data (e.g., radio conditions, cell load) can be used by applications to offer context-aware services.

Fog Computing vs Edge Computing: While often used interchangeably, *Fog Computing* typically refers to a continuum of distribution from the Cloud down to the Things, whereas Edge Computing focuses specifically on the edge of the network.

14 Multi-access Edge Computing (MEC)

MEC is the standard defined by ETSI (European Telecommunications Standards Institute) to implement Edge Computing in mobile networks. Originally "Mobile Edge Computing", it was renamed "Multi-access" to include non-cellular technologies (like Wi-Fi). MEC transforms the base station into an intelligent programmable platform.

14.1 ETSI MEC Reference Architecture

The ETSI MEC architecture is modular and consists of three main levels:

14.1.1 Mobile Edge Host Level (The "Body")

This is the physical infrastructure at the network edge. It includes:

- **MEC Host:** Contains the virtualization infrastructure (data plane) to execute applications.
- **MEC Platform:** Provides essential functionalities to run MEC apps. It includes the *Service Registry* (to discover available services), *DNS handling*, and *Traffic Rules Control.*
- **MEC Apps:** Virtualized applications (VMs) running on top of the infrastructure, consuming and providing MEC services.

14.1.2 MEC Host Level Management (Local Management)

- **MEC Platform Manager:** Manages the lifecycle of applications on a specific host and informs the orchestrator of events.
- **Virtualization Infrastructure Manager (VIM):** Manages the virtualized resources (compute, storage, network) of the host.

14.1.3 MEC System Level Management (The "Brain")

- **MEC Orchestrator:** The core component with a global view of the system. It receives requests from users/operators, selects the most appropriate MEC host (based on latency, resources, and requirements), and triggers application instantiation.

14.2 Key MEC Services

MEC applications can access contextual services via standardized APIs:

- **RNIS (Radio Network Information Service):** Provides real-time information regarding radio network conditions and user statistics (e.g., cell load, signal quality).
- **LS (Location Service):** Provides authorized applications with location-related information about UEs (e.g., via Cell ID or geolocation).

14.3 MEC Host Migration

Since MEC servers are distributed, user mobility poses a challenge. If a user moves to a new cell served by a different MEC host, the system must decide whether to migrate the application state to the new server or keep serving the user from the original one (accepting higher latency).

15 Computation Offloading

Offloading is the procedure of transferring the execution of a computing task from the mobile device to the edge or remote cloud. The decision is a trade-off between **Energy** and **Latency**.

Let us define:

- T_{local}, E_{local}: Time and energy to execute the task locally.
- $T_{offload}, E_{offload}$: Time and energy to execute the task remotely.

Local execution depends on the device's CPU:

$$T_{local} = \frac{\text{Task Size}}{\text{CPU Speed}_{local}}$$

Cloud/Edge execution introduces transmission delays:

$$T_{offload} = T_{uplink} + T_{processing\_remote} + T_{downlink}$$

Decision Rule: Offloading is beneficial if and only if:

1. It saves time: $T_{offload} < T_{local}$ (critical for low-latency apps).
2. It saves energy: $E_{offload} < E_{local}$ (critical for battery life).

However, in dynamic network environments, transmission time (T_{uplink}) can vary significantly. If bandwidth is low, offloading might take longer than local execution, even if the remote CPU is much faster.