Wireless Networking

Mohammad M. Banat
Department of Electrical Engineering
Jordan University of Science and Technology
Irbid 22110, Jordan
Email: banat@just.edu.jo

March 15, 2026

Contents

Preface

List of Figures

List of Tables

List of Acronyms

2G	Second Generation	12
4G	Fourth Generation	12
5G	Fifth Generation	12
AM	Amplitude Modulation	12
ASCII	American Standard Code for Information Interchange	9
ATM	Asynchronous Transfer Mode	19, 29, 31
AWGN	Additive White Gaussian Noise	23
BEP	Bit Error Probability	23
CCDF	Complementary Cumulative Distribution Function	45
CDF	Cumulative Distribution Function	40, 41, 45
CDMA	Code Division Multiple Access	11, 13
CRC	Cyclic Redundancy Check	33
CSMA-CD	Carrier Sense Multiple Access with Collision Detection	vii
DSL	Digital Subscriber Line	19
e-Mail	Electronic Mail	19, 35
FDMA	Frequency Division Multiple Access	11, 12, 13
FM	Frequency Modulation	12
GSM	Global System for Mobile Communications	12
I/O	Input/output	15, 16
IDC	Index of Dispersion for Counts	44, 45
IID	Independent and Identically Distributed	42
IP	Internet Protocol	vii, 32, 33, 34

IPv4	Internet Protocol Version 4	29
IPv6	Internet Protocol Version 6	30
ISO	International Organization for Standardization	30, 31
ISP	Internet Service Provider	19, 20
LAN	Local Area Network	vii, 18, 19, 20
MAC	Medium Access Control	33
MAN	Metropolitan Area Network	18, 19
MS	Mean Square	43
NAP	Network Access Point	20
OFDMA	Orthogonal Frequency Division Multiple Access	12
OSI	Open Systems Interconnection	30, 31, 32, 33, 34, 35, 36
PC	Personal Computer	18
PDF	Probability Density Function	40, 41, 45
PDU	Protocol Data Unit	32, 33, 35, 36, 37
PGF	Probability Generating Function	41, 43, 46, 47
PMF	Probability Mass Function	41
RS-232	Recommended Standard 232	23, 24
SAP	Service Access Point	35, 36, 37
SDU	Service Data Unit	35, 36
TCP	Transmission Control Protocol	vii, 33, 34
TDMA	Time Division Multiple Access	11, 12
TTL	Time to Live	33
TV	Television	19
USB	Universal Serial Bus	24
WAN	Wide Area Network	18, 19, 20
WAP	Wireless Application Protocol	vii
Wi-Fi	Wireless Fidility	33
WLAN	Wireless Local Area Network	vii, 18
WSS	Wide Sense Stationary	43, 44

Course Overview

0.A Course in Brief

0.A.1 Course Catalog

3 Credit hours (3 h lectures, R1). Introduction to wireless and mobile networks. Types of networks, performance criteria. Queuing analysis of networks. Transmission control protocol (TCP)/internet protocol (IP). Local area networks (LANs), wireless local area networks (WLANs), protocols and performance analysis of carrier sense multiple access with collision detection (CSMA-CD). Mobile networks, cellular wireless networks, ad hoc networks and their routing protocols. Sensor networks, Bluetooth networks. Transport protocols for wireless networks, wireless application protocol (WAP), wireless network security.

0.A.2 Textbook

No single textbook. Students will be referred to several recent books and journal articles (mainly survey papers).

0.A.3 Recommended Readings

- Communication Networks: A Concise Introduction [1].
- Data Networks [2].
- Networks [3].

0.A.4 Instructor

Dr. Mohammad M. Banat (banat@just.edu.jo).

0.B Technical Content

0.B.1 Prerequisites

Level	Subjects
Background	Probability and Random Variables
	Communication Systems
Advanced	Digital Communications
	Random Processes

0.B.2 Topics

Weeks	Topics
1-2	Introduction to Communication Networks
3-5	Introduction to Queuing Theory
6-9	Physical, Data, Link and Network Layers
10	Network Topologies
11	Network Performance
12	Circuit and Packet Switching
13	Local Area Networks
14	Optical Networks
15-16	Wireless Networks

0.C Evaluation

Assessment Tool	Due Week	Weight %
Mid-Term Exam	9	25
Term Project Report	13	15
Term Project Presentation	14	10
Final Exam	16	50

0.D Mathematical Notations

Italic symbols like a and M denote scalar quantities and functions. Bold italic symbols like $\boldsymbol{a}$ and $\boldsymbol{M}$ denote vector quantities and functions. Bold non-italic upper case symbols like $\mathbf{M}$ denote matrices. $\mathbb{C}$ denotes the set of complex numbers. $\mathbb{R}$ denotes the set of real numbers. $\mathbb{I}$ denotes the set of integer numbers. $\mathbb{E}$ denotes the set of even integer numbers. $\mathbb{O}$ denotes the set of odd integer numbers. A set with an over-circle $\mathring{\mathbb{S}}$ denotes the non-negative subset of set $\mathbb{S}$. A set with an over-plus $\overset{+}{\mathbb{S}}$ denotes the positive subset of set $\mathbb{S}$. Set $2^{\mathbb{S}}$ is composed of the elements of $\mathbb{S}$ as powers of two. $\mathbb{I}_{k,l}$, where $k \leq l$ and $k, l \in \mathbb{I}$, denotes the set of integer numbers from k to l. $\mathbb{I}_N$, where $N \in \overset{+}{\mathbb{I}}$, denotes the set of integer numbers from 0 to $N - 1$. Note that $\mathbb{I}_N = \mathbb{I}_{0,N-1}$. Closed real interval $[a, b]$ denotes the set of all values $x \in \mathbb{R}$ such that $a \leq x \leq b$. Open real interval (a, b) denotes the set of all values $x \in \mathbb{R}$ such that $a < x < b$.

Part I

Background and Preliminaries

Chapter 1

Review of Signals and Linear Systems

Chapter 2

Review of Random Processes

Chapter 3

Review of Communication Systems

Part II

Networking Foundations

Chapter 4

Introduction to Communication Networks

4.A Signals

In communication systems, electrical signals are represented as functions of the time variable t, for example, $x(t)$, $g(t)$, $s(t)$, etc. These signals may represent information-carrying data such as voice, text, video, or any other media. A signal may also be a carrier that is modulated to carry the data or information signals. An important measure of a signal $x(t)$ is its energy $\mathcal{E}_x$, given mathematically as

$$\mathcal{E}_x = \int_{-\infty}^{\infty} |x(t)|^2 \, dt. \tag{4.1}$$

When $\mathcal{E}_x$ has a finite value, $x(t)$ is called an energy signal. The average power P_x of a signal $x(t)$ is the time average of its energy, and is defined mathematically as

$$P_x = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} |x(t)|^2 \, dt. \tag{4.2}$$

Often, in practical systems, there is a finite time interval of interest. In this case, the parameter T covers this time interval only. One example is periodic signals of period T since the average power in any period is the same. Energy signals have zero average power. Conversely, signals with nonzero average power are called power signals; such signals have infinite energy. Since all real-world signals have finite duration, theoretically, all real-world signals are energy signals. However, in many cases, signals of interest last a very long time and so are better treated as power signals. In general, we will treat signals of short durations as energy signals and long durations as power signals.

Wireless communication signal powers usually vary over a very wide range of values, typically from below 1 pW to tens of kW. Given such a wide range, comparisons of power levels can be challenging. For this reason, it is more convenient to compress the range of values using logarithmic compression. Using a logarithmic scale also makes the modeling of signal attenuation easier. The logarithmic scale of power levels is called the decibel or dB scale. Power level P is represented in the dB scale as

$$P_{\text{dB}} = 10 \log_{10} P. \tag{4.3}$$

The dB scale can also be represented relative to mW values as

$$P_{\text{dBm}} = 10 \log_{10} \left(\frac{P}{1 \text{ mW}} \right). \tag{4.4}$$

Note that

$$P_{\text{dBm}} = P_{\text{dB}} + 30. \tag{4.5}$$

4.B Data

Data communications and networking are changing the way we do business and the way we live. Business decisions have to be made ever more quickly, and the decision makers require immediate access to accurate information. Why wait a week for that report from far away to arrive by mail when it could appear almost instantaneously through communication networks? It is very important to know how networks operate, what types of technologies are available, and which design best fills which set of needs.

The development of the personal computer brought about tremendous changes for business, industry, science, and education. A similar revolution is occurring in data communications and networking. Technological advances are making it possible for communications links to carry more and faster signals. As a result, services are evolving to allow use of this expanded capacity. Research in data communications and networking has resulted in new technologies. One goal is to be able to exchange data such as text, audio, and video from all points in the world. We want to access the Internet to download and upload information quickly and accurately and at any time.

When we communicate, we are sharing information. This sharing can be local or remote. Between individuals, local communication usually occurs face to face, while remote communication takes place over distance. The term telecommunication, which includes telephony, telegraphy, and television, means communication at a distance (tele is Greek for “far”).

Data communication is the exchange of data between two devices via some form of transmission medium such as a wire cable. For data communication to occur, the communicating devices must be part of a communication system made up of a combination of hardware (physical equipment) and software (programs).

4.B.1 Effectiveness of data communication

The effectiveness of a data communication system depends on four fundamental characteristics: delivery, accuracy, timeliness, and jitter.

- **Delivery:** The system must deliver data to the correct destination. Data must be received by the intended device or user and only by that device or user.
- **Accuracy:** The system must deliver the data accurately. Data that have been altered in transmission and left uncorrected are unusable.
- **Timeliness:** The system must deliver data in a timely manner. Data delivered late are useless. In the case of video and audio, timely delivery means delivering data as they are produced, in the same order that they are produced, and without significant delay. This kind of delivery is called real-time transmission.
- **Jitter:** Jitter refers to the variation in the packet arrival time. It is the uneven delay in the delivery of audio or video packets. For example, let us assume that video packets are sent every 30 ms. If some of the packets arrive with 30 ms delay and others with 40 ms delay, an uneven quality in the video is the result.

4.B.2 Data Communications System Components

A data communications system has five components, as shown in Fig. 4.1.

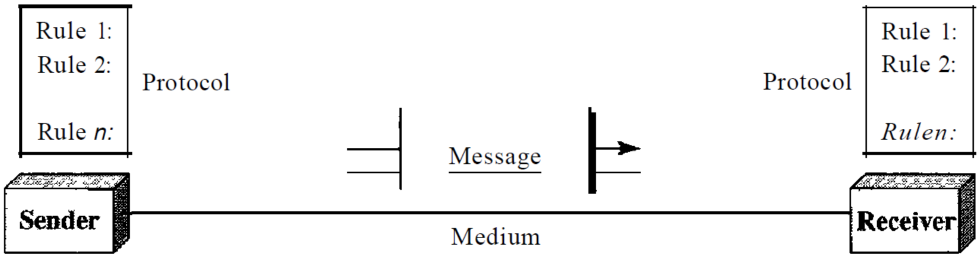

Fig. 4.1: Data communication system components

- **Message:** The message is the data to be communicated. Popular forms of information include text, numbers, pictures, audio, and video.
- **Sender:** The sender is the device that sends the data message. It can be a computer, telephone handset, video camera, and so on.
- **Receiver:** The receiver is the device that receives the data message. It can be a computer, telephone handset, television, and so on.
- **Transmission Medium:** The transmission medium is the physical path by which a message travels from sender to receiver. Some examples of transmission media include twisted-pair wire, coaxial cable, fiber-optic cable, and radio waves.
- **Protocol:** A protocol is a set of rules that govern data communications. It represents an agreement between the communicating devices. Without a protocol, two devices may be connected but not communicating, just as a person speaking French cannot be understood by a person who speaks only Japanese.

4.B.3 Data Representations

Data can have different forms such as text, numbers, images, audio, and video.

- ***Text:*** In data communications, text is represented as a bit pattern, a sequence of bits (Os or Is). Different sets of bit patterns have been designed to represent text symbols. Each set is called a code, and the process of representing symbols is called coding. Today, the prevalent coding system is called Unicode, which uses 32 bits to represent a symbol or character used in any language in the world. The American Standard Code for Information Interchange (ASCII), developed some decades ago in the United States, now constitutes the first 127 characters in Unicode and is also referred to as Basic Latin.

- ***Numbers:*** Numbers are also represented by bit patterns. However, a code such as ASCII is not used to represent numbers; the number is directly converted to a binary number to simplify mathematical operations.

- ***Images:*** Images are also represented by bit patterns. In its simplest form, an image is composed of a matrix of pixels (picture elements), where each pixel is a small dot. The size of the pixel depends on the resolution. For example, an image can be divided into 1000 pixels or 10,000 pixels. In the second case, there is a better representation of the image (better resolution), but more memory is needed to store the image. After an image is divided into pixels, each pixel is assigned a bit pattern. The size and the value of the pattern depend on the image. For an image made of only black and white dots (e.g., a chessboard), a 1-bit pattern is enough to represent a pixel. If an image is not made of pure white and pure black pixels, you can increase the size of the bit pattern to include gray scale. For example, to show four levels of gray scale, you can use 2-bit patterns. There are several methods to represent color images. One method uses a 24-bit color code, and is called RGB, where each color is made of a combination of three primary colors: red, green, and blue. The intensity of each color is measured, and an 8-bit pattern is assigned to it. The values for red, green, and blue are each specified on a scale from 0–255 (decimal) or 00–FF (hex). The number of RGB colors is equal to

$$\begin{aligned} N_{RGB} &= 2^{24} \\ &= 256^{3} \\ &= 16,777,216. \end{aligned} \tag{4.6}$$

  Although these colors are more than our eyes can distinguish, this is an easy representation for computers to handle, and is a great way for rendering images and videos. Higher numbers mean lighter, lower numbers mean darker. Another method is called YCM, in which a color is made of a combination of three other primary colors: yellow, cyan, and magenta.

- ***Audio:*** Audio refers to the recording or broadcasting of sound or music. Audio is by nature different from text, numbers, or images. It is continuous, not discrete. Even when we use a microphone to change voice or music to an electric signal, we create a continuous signal.

- ***Video:*** Video refers to the recording or broadcasting of a picture or movie. Video can either be produced as a continuous entity (e.g., by a TV camera), or it can be a combination of images, each a discrete entity, arranged to convey the idea of motion.

4.B.4 Data Flow

Data flow between two devices can be simplex, half-duplex, or full-duplex.

- ***Simplex:*** In simplex mode, the communication is unidirectional, as on a one-way street. Only one of the two devices on a link can transmit; the other can only receive. Keyboards and traditional monitors are examples of simplex devices. The keyboard can only introduce input; the monitor can only accept output. The simplex mode can use the entire capacity of the channel to send data in one direction.

- ***Half-Duplex:*** In half-duplex mode, each station can both transmit and receive, but not at the same time. When one device is sending, the other can only receive, and vice versa.

  The half-duplex mode is like a one-lane road with traffic allowed in both directions. When cars are traveling in one direction, cars going the other way must wait. In a half-duplex transmission, the entire capacity of a channel is taken over by whichever of the two devices is transmitting at the time.

  The half-duplex mode is used in cases where there is no need for communication in both directions at the same time; the entire capacity of the channel can be utilized for each direction.

- ***Full-Duplex:*** In full-duplex mode (also called duplex), both stations can transmit and receive simultaneously.

  The full-duplex mode is like a two-way street with traffic flowing in both directions at the same time. In full-duplex mode, signals going in one direction share the capacity of the link with signals going in the other direction. This sharing can occur in two ways: either the link must contain two physically separate transmission paths, one for sending and the other for receiving; or the capacity of the channel is divided between signals traveling in both directions.

  One common example of full-duplex communication is the telephone network. When two people are communicating by a telephone line, both can talk and listen at the same time.

  The full-duplex mode is used when communication in both directions is required all the time. The capacity of the channel, however, must be divided between the two directions.

4.C Multiple Access

The wireless communication infrastructure is designed to service many users simultaneously. The available radio resources must, therefore, be shared between different users, allowing for multiple access.

- ✷ ***FDMA***: Frequency division multiple access (FDMA).
- ✷ ***TDMA***: Time division multiple access (TDMA).
- ✷ ***CDMA***: Code division multiple access (CDMA).

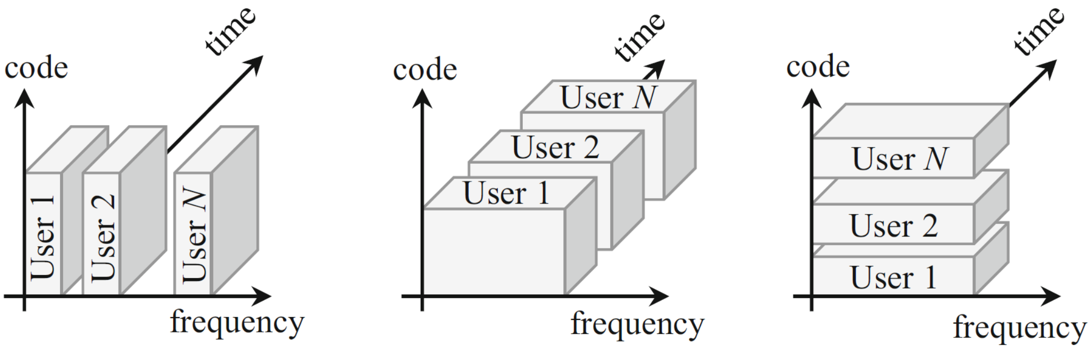

Fig. 4.2: Multiple access schemes

4.C.1 FDMA

In FDMA schemes, N data streams, belonging to N different users, are transmitted at different carrier frequencies, as illustrated in Fig. 4.3.

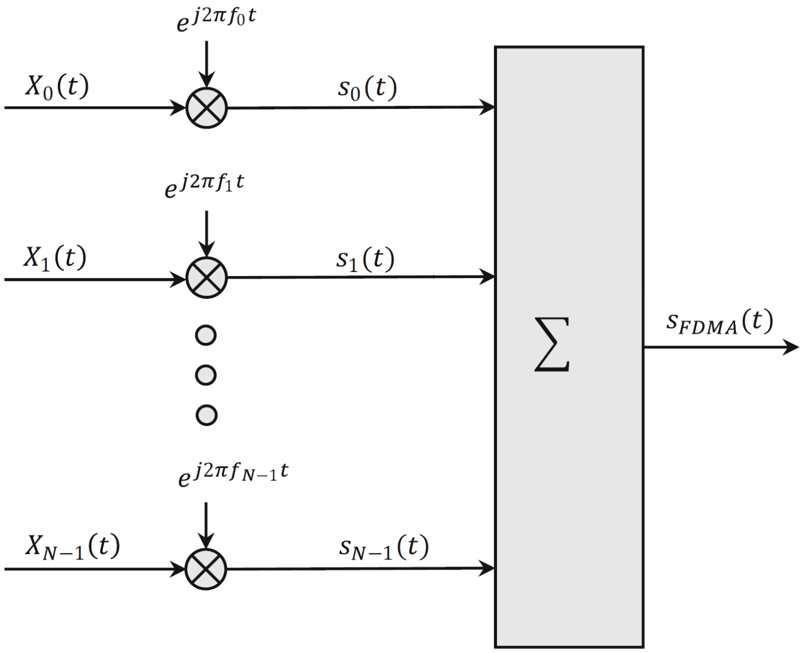

Fig. 4.3: Frequency division multiple access

The combined transmit signal can be expressed as

$$s_{\text{FDMA}}(t) = \sum_{n=0}^{N-1} X_n(t) e^{j2\pi f_n t}. \tag{4.7}$$

Typically, the spacing between the carrier frequencies is kept constant, i.e., for $1 \leq n \leq N-1$ we have

$$f_n - f_{n-1} = \Delta f, \tag{4.8}$$

where Δf is known as the carrier spacing. The receiver structure for this type of transmission is shown in Fig. 4.4.

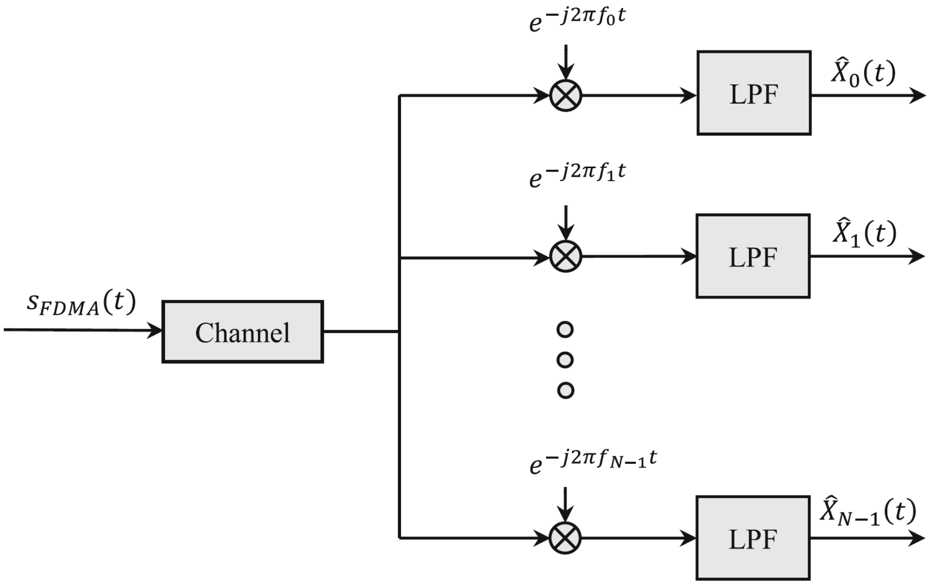

Fig. 4.4: Frequency division multiple access receiver

FDMA was typical for the first generation of cellular systems and is similar to how the analog amplitude modulation (AM) and frequency modulation (FM) radio station operate. Fourth generation (4G) and fifth generation (5G) systems also use a form of FDMA, but a far more efficient one called orthogonal frequency division multiple access (OFDMA).

4.C.2 TDMA

In TDMA, the data stream is divided into frames, which are further subdivided into time slots. Each user is allocated one slot. The global system for mobile communications (GSM), a second generation (2G) cellular system, is TDMA-based. There are eight slots shared dynamically among different users. The carrier frequency is common for all users. Slot assignment on the transmit and receive side is synchronized to avoid inter-user interference and can be dynamically assigned for increased capacity. The TDMA frame structure is shown in Fig. 4.5.

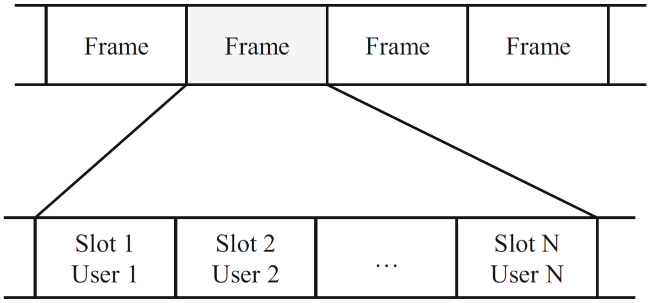

Fig. 4.5: TDMA Frame Structure

4.C.3 CDMA

CDMA was introduced in the second generation of cellular systems. In CDMA-based schemes, N data streams, belonging to N different users, are multiplied by different mutually orthogonal codes $c_n(t)$, where $n \in \mathbb{I}_N$, as illustrated in Fig. 4.6.

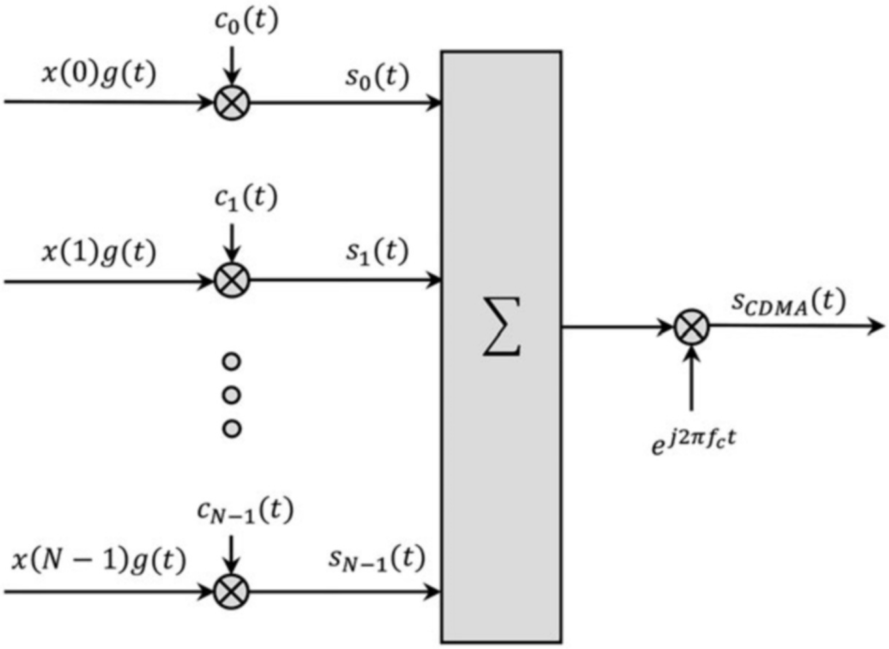

Fig. 4.6: Code division multiple access

The orthogonality condition that allows detection of the signals at the receiver is given by

$$\int_0^{T_s} c_k(t)c_l(t)dt = \begin{cases} 1, & k = l \\ 0, & k \neq l \end{cases}. \tag{4.9}$$

We note that this requires the length of the code to be at least N which, in turn, means the bandwidth of the transmission expands by a factor of N (at least). The receiver structure is similar to that of the FDMA system.

4.D Networks

A network is a set of devices (often referred to as nodes) connected by communication links. A node can be a computer, printer, or any other device capable of sending and/or receiving data generated by other nodes on the network.

Most networks use distributed (as opposed to centralized) processing, in which a task is divided among multiple computers. Instead of one single large machine being responsible for all aspects of a process, separate computers (usually a personal computer or workstation) handle a subset.

4.D.1 Network Criteria

A network must be able to meet a certain number of criteria. The most important of these are performance, reliability, and security.

- ***Performance:*** Performance can be measured in many ways, including transit time and response time. Transit time is the amount of time required for a message to travel from one device to another. Response time is the elapsed time between an inquiry and a response. The performance of a network depends on a number of factors, including the number of users, the type of transmission medium, the capabilities of the connected hardware, and the efficiency of the software. Performance is often evaluated by two networking metrics: throughput and delay. We often need more throughput and less delay. However, these two criteria are often contradictory. If we try to send more data to the network, we may increase throughput but we increase the delay because of traffic congestion in the network.
- ***Reliability:*** In addition to accuracy of delivery, network reliability is measured by the frequency of failure, the time it takes a link to recover from a failure, and the network robustness in a catastrophe.
- ***Security:*** Network security issues include protecting data from unauthorized access, protecting data from damage and theft, and implementing policies and procedures for recovery from breaches and data losses.

4.D.2 Network Attributes

- ***Type of Connection:*** A network is two or more devices connected through links. A link is a communications pathway that transfers data from one device to another. For visualization purposes, it is simplest to imagine any link as a line drawn between two points. For communication to occur, two devices must be connected in some way to the same link at the same time. There are two possible types of connections:
    - ***Point-to-Point:*** A point-to-point connection provides a dedicated link between two devices, as shown in Fig. 4.7. The entire capacity of the link is reserved for transmission between those two devices. Most point-to-point connections use a wire or cable to connect the two ends, but other options, such as microwave

or satellite links, are also possible. When you change television channels by infrared remote control, you are establishing a point-to-point connection between the remote control and the television control system.

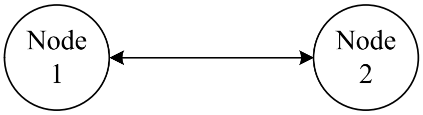

Fig. 4.7: Ponit-to-point link

- ***Multipoint:*** A multipoint connection is one in which more than two devices share a single link. In a multipoint environment, the capacity of the channel is shared, either spatially or temporally. If several devices can use the link simultaneously, it is a spatially shared connection. If users must take turns, it is a time-shared connection.

Physical Topology:

The term physical Topology refers to the way in which a network is laid out physically. Two or more devices connect to a link; two or more links form a topology. The topology of a network is the geometric representation of the relationship of all the links and linking devices (usually called nodes) to one another. There are four basic topologies possible: mesh, star, bus, and ring.

- ***Mesh:*** In a mesh topology, every device has a dedicated point-to-point link to every other device. The term dedicated means that the link carries traffic only between the two devices it connects. To find the number of physical links in a fully connected mesh network with n nodes, we first consider that each node must be connected to every other node. Node 1 must be connected to $n-1$ nodes, node 2 must be connected to $n-1$ nodes, and finally node n must be connected to $n-1$ nodes. The number of needed physical links in a mesh topology is equal to

$$N_{\text{mesh}} = n\left(n-1\right). \tag{4.10}$$

However, if each physical link allows communication in both directions (full duplex mode), we can divide the number of links by 2. In other words, the number of full duplex mesh topology physical links is equal to

$$N_{\text{fd,mesh}} = \frac{n\left(n-1\right)}{2}. \tag{4.11}$$

To accommodate that many links, every device on the network must have $n-1$ input/output (I/O) ports to the other $n-1$ stations.

A mesh offers several advantages over other network topologies.

- 1- The use of dedicated links guarantees that each connection can carry its own data load, thus eliminating the traffic problems that can occur when links must be shared by multiple devices.

- ➤ 2- A mesh topology is robust. If one link becomes unusable, it does not incapacitate the entire system.
- ➤ 3- There is the advantage of privacy or security. When every message travels along a dedicated line, only the intended recipient sees it. Physical boundaries prevent other users from gaining access to messages.
- ➤ 4- Point-to-point links make fault identification and fault isolation easy. Traffic can be routed to avoid links with suspected problems. This facility enables the network manager to discover the precise location of the fault and aids in finding its cause and solution.

The main disadvantages of a mesh are related to the amount of cabling and the number of I/O ports required.

- ➤ 1- Because every device must be connected to every other device, installation and reconnection are difficult.
- ➤ 2- The sheer bulk of the wiring can be greater than the available space (in walls, ceilings, or floors) can accommodate.
- ➤ 3- The hardware required to connect each link can be prohibitively expensive.

For these reasons a mesh topology is usually implemented in a limited fashion, for example, as a backbone connecting the main computers of a hybrid network that can include several other topologies. One practical example of a mesh topology is the connection of telephone regional offices in which each regional office needs to be connected to every other regional office.

✦ ***Star:*** In a star topology, each device has a dedicated point-to-point link only to a central controller, usually called a hub. The devices are not directly linked to one another. Unlike a mesh topology, a star topology does not allow direct traffic between devices. The controller acts as an exchange. If one device wants to send data to another, it sends the data to the controller, which then relays the data to the other connected device. Following are some advantages of the star topology.

- ➤ 1- A star topology is less expensive than a mesh topology. In a star, each device needs only one link and one I/O port to connect it to any number of others. This factor also makes it easy to install and reconfigure. Far less cabling needs to be housed, and additions, moves, and deletions involve only one connection: between that device and the hub.
- ➤ 2- A star topology is robust. If one link fails, only that link is affected. All other links remain active. This factor also lends itself to easy fault identification and fault isolation. As long as the hub is working, it can be used to monitor link problems and bypass defective links.

Following are some disadvantages of the star topology.

- ➤ 1- A star topology is strongly dependent of the whole topology on the hub. If the hub goes down, the whole system is dead.

- ➤ 2- Although a star requires far less cable than a mesh, each node must be linked to a central hub. For this reason, often more cabling is required in a star than in some other topologies.

✦ ***Bus:*** The preceding examples all describe point-to-point connections. A bus topology, on the other hand, is multipoint. One long cable acts as a backbone to link all the devices in a network.

Nodes are connected to the bus cable by drop lines and taps. A drop line is a connection running between the device and the main cable. A tap is a connector that either splices into the main cable or punctures the sheathing of a cable to create a contact with the metallic core. As a signal travels along the backbone, some of its energy is transformed into heat. Therefore, it becomes weaker and weaker as it travels farther and farther. For this reason there is a limit on the number of taps a bus can support and on the distance between those taps.

Advantages of a bus topology include ease of installation. Backbone cable can be laid along the most efficient path, then connected to the nodes by drop lines of various lengths. In this way, a bus uses less cabling than mesh or star topologies. In a star, for example, four network devices in the same room require four lengths of cable reaching all the way to the hub. In a bus, this redundancy is eliminated. Only the backbone cable stretches through the entire facility. Each drop line has to reach only as far as the nearest point on the backbone.

Disadvantages include difficult reconnection and fault isolation. A bus is usually designed to be optimally efficient at installation. It can therefore be difficult to add new devices. Signal reflection at the taps can cause degradation in quality. This degradation can be controlled by limiting the number and spacing of devices connected to a given length of cable. Adding new devices may therefore require modification or replacement of the backbone.

✦ ***Ring:*** In a ring topology, each device has a dedicated point-to-point connection with only the two devices on either side of it. A signal is passed along the ring in one direction, from device to device, until it reaches its destination. Each device in the ring incorporates a repeater. When a device receives a signal intended for another device, its repeater regenerates the bits and passes them along.

A ring is relatively easy to install and reconfigure. Each device is linked to only its immediate neighbors (either physically or logically). To add or delete a device requires changing only two connections. The only constraints are media and traffic considerations (maximum ring length and number of devices). In addition, fault isolation is simplified. Generally in a ring, a signal is circulating at all times. If one device does not receive a signal within a specified period, it can issue an alarm. The alarm alerts the network operator to the problem and its location.

However, unidirectional traffic can be a disadvantage. In a simple ring, a break in the ring (such as a disabled station) can disable the entire network. This weakness can be solved by using a dual ring or a switch capable of closing off the break.

- ✦ ***Hybrid:*** A network can be hybrid. For example, we can have a main star topology with each branch connecting several stations in a bus topology.

4.E Network Categories

Today when we speak of networks, we are generally referring to two primary categories: LANs and wide area networks (WANs). The category into which a network falls is determined by its size. A LAN normally covers an area less than two squared miles. A WAN can be worldwide. Networks of a size in between are normally referred to as metropolitan area networks (MANs) and span tens of squared miles.

- ✷ ***LAN:*** A LAN is usually privately owned and links the devices in a single office, building, or campus. Depending on the needs of an organization and the type of technology used, a LAN can be as simple as two personal computers (PCs) and a printer in someone's home office; or it can extend throughout a company and include audio and video peripherals. Currently, a LAN area is limited to a few squared kilometers.

  LANs are designed to allow resources to be shared between PCs or workstations. The resources to be shared can include hardware (e.g., a printer), software (e.g., an application program), or data. A common example of a LAN, found in many business environments, links a workgroup of task-related computers, for example, engineering workstations or accounting PCs. One of the computers may be given a large-capacity disk drive and may become a server to clients. Software can be stored on this central server and used as needed by the whole group. In this example, the size of the LAN may be determined by licensing restrictions on the number of users per copy of software, or by restrictions on the number of users licensed to access the operating system.

  In addition to size, LANs are distinguished from other types of networks by their transmission media and topology. In general, a given LAN will use only one type of transmission medium, but this is not always the case. The most common LAN topologies are bus, ring, and star.

  Early LANs had data rates in the 4 to 16 Mbps range. Today, however, speeds are normally 100 or 1000 Mbps.

  WLANs are the newest evolution in LAN technology.

- ✷ ***WAN:*** A WAN provides long-distance transmission of data, image, audio, and video information over large geographic areas that may comprise a country, a continent, or even the whole world. A WAN can be as complex as the backbones that connect the Internet or as simple as a dial-up line that connects a home computer to the Internet. We normally refer to the first as a switched WAN and to the second as a point-to-point WAN.

  - ✦ The switched WAN connects the end systems, which usually comprise a router (internetworking connecting device) that connects to another LAN or WAN. An early example of a switched WAN is X.25, a network designed to provide connectivity between end users.

- ✦ The point-to-point WAN is normally a line leased from a telephone or cable television (TV) provider that connects a home computer or a small LAN to an internet service provider (ISP). This type of WAN is often used to provide Internet access. A good example of a switched WAN is the asynchronous transfer mode (ATM) network, which is a network with fixed-size data unit packets called cells.
- ✦ Another example of WANs is the wireless WAN that is becoming more and more popular.

✷ ***MAN:*** A MAN is a network with a size between a LAN and a WAN. It normally covers the area inside a town or a city. It is designed for customers who need a high-speed connectivity, normally to the Internet, and have endpoints spread over a city or part of city. A good example of a MAN is the part of the telephone company network that can provide a high-speed digital subscriber line (DSL) to the customer. Another example is the cable TV network that originally was designed for cable TV, but today can also be used for high-speed data connection to the Internet.

4.F The Internet

Today, it is very rare to see a LAN, a MAN, or a WAN in isolation; they are connected to one another. When two or more networks are connected, they become an internetwork, or internet.

As an example, assume that an organization has two offices, one on the east coast and the other on the west coast. The established office on the west coast has a bus topology LAN; the newly opened office on the east coast has a star topology LAN. The president of the company lives somewhere in the middle and needs to have control over the company from home. To create a backbone WAN for connecting these three entities (two LANs and the president's computer), a switched WAN (operated by a service provider such as a telecommunications company) has been leased. To connect the LANs to this switched WAN, however, three point-to-point WANs are required. These point-to-point WANs can be a high-speed DSL line offered by a telephone company or a cable offered by a cable TV provider.

The Internet has revolutionized many aspects of our daily lives. It has affected the way we do business as well as the way we spend our leisure time. Count the ways you've used the Internet recently. Perhaps you've sent an electronic mail (e-Mail) message to a business associate, paid a utility bill, read a newspaper from a distant city, or looked up a local movie schedule. Or maybe you researched a scientific topic, booked a hotel reservation, chatted with a fellow, or shopped for a car or some other product. The Internet is a communication system that has brought a wealth of information to our fingertips and organized it for our use. The Internet is a structured, organized system.

4.F.1 Internet History

A network is a group of connected communicating devices such as computers and printers. An internet (note the lowercase letter i) is two or more networks that can communicate with each other. The most notable internet is called the Internet (uppercase letter I), a collaboration of more than hundreds of thousands of interconnected networks. Private individuals as well as various organizations such as government agencies, schools, research facilities, corporations, and libraries in most countries of the world use the Internet. Millions of people are users. Yet this extraordinary communication system only came into being in 1969.

In the mid-1960s, mainframe computers in research organizations were standalone devices. Computers from different manufacturers were unable to communicate with one another. The United States Department of Defense was interested in finding a way to connect computers so that the researchers they funded could share their findings, thereby reducing costs and eliminating duplication of effort.

The Internet has come a long way since the 1960s. The Internet today is not a simple hierarchical structure. It is made up of many WANs and LANs joined by connecting devices and switching stations. It is difficult to give an accurate representation of the Internet because it is continually changing: new networks are being added, existing networks are adding addresses, and networks of defunct companies are being removed. Today most end users who want Internet connection use the services of ISPs. There are international service providers, national service providers, regional service providers, and local service providers. The Internet today is run by private companies, not the governments.

- ***International ISPs:*** At the top of the hierarchy are the international service providers that connect nations together.

- ***National ISPs:*** The national ISPs are backbone networks created and maintained by specialized companies. There are many national ISPs operating in North America; some of the most well known are SprintLink, PSINet, UUNet Technology, AGIS, and internet Mel. To provide connectivity between the end users, these backbone networks are connected by complex switching stations (normally run by a third party) called network access points (NAPs). Some national ISP networks are also connected to one another by private switching stations called peering points. These normally operate at a high data rate (up to 600 Mbps).

- ***Regional ISPs:*** Regional ISPs are smaller ISPs that are connected to one or more national ISPs. They are at the third level of the hierarchy with a lower data rate.

- ***Local ISPs:*** Local ISPs provide direct service to the end users. The local ISPs can be connected to regional ISPs or directly to national ISPs. Most end users are connected to the local ISPs. Note that in this sense, a local ISP can be a company that just provides Internet services, a corporation with a network that supplies services to its own employees, or a nonprofit organization, such as a college or a university, that runs its own network. Each of these local ISPs can be connected to a regional or national service provider.

4.G Network Protocols and Standards

4.G.1 Network Protocols

In computer networks, communication occurs between entities in different systems. An entity is anything capable of sending or receiving information. However, two entities cannot simply send bit streams to each other and expect to be understood. For communication to occur, the entities must agree on a protocol. A protocol is a set of rules that govern data communications. A protocol defines what is communicated, how it is communicated, and when it is communicated. The key elements of a protocol are syntax, semantics, and timing.

- ***Syntax:*** The term syntax refers to the structure or format of the data, meaning the order in which they are presented. For example, a simple protocol might expect the first 8 bits of data to be the address of the sender, the second 8 bits to be the address of the receiver, and the rest of the stream to be the message itself.
- ***Semantics:*** The word semantics refers to the meaning of each section of bits. How is a particular pattern to be interpreted, and what action is to be taken based on that interpretation? For example, does an address identify the route to be taken or the final destination of the message?
- ***Timing:*** The term timing refers to two characteristics: when data should be sent and how fast they can be sent. For example, if a sender produces data at 100 Mbps but the receiver can process data at only 1 Mbps, the transmission will overload the receiver and some data will be lost.

4.G.2 Network Standards

Standards are essential in creating and maintaining an open and competitive market for equipment manufacturers and in guaranteeing national and international interoperability of data and telecommunications technology and processes. Standards provide guidelines to manufacturers, vendors, government agencies, and other service providers to ensure the kind of interconnectivity necessary in today's marketplace and in international communications. Data communication standards fall into two categories: de facto (meaning "by fact" or "by convention") and de jure (meaning "by law" or "by regulation").

- ***De Facto:*** Standards that have not been approved by an organized body but have been adopted as standards through widespread use are de facto standards. De facto standards are often established originally by manufacturers who seek to define the functionality of a new product or technology.
- ***De Jure:*** Those standards that have been legislated by an officially recognized body are de jure standards.

4.H Data Transmission

Telecommunication networks can be roughly distinguished between broadcast networks and switched networks. In the first case, all the nodes receive the same information transmitted by a source node. This is the case of radio and television networks. In the second case, the transfer of information (voice, data, etc.) requires routing or switching operations at the different network nodes, which are encountered along the path from the source to the destination.

The information sent from source to destination along the network can be identified with the generic term of "traffic". Each link along the source-to-destination path in the network conveys traffic that is typically the aggregate contribution of many users. A generic definition of traffic should entail the notion of random variables and stochastic processes. For the sake of simplicity and referring to the transmissions on a link, we can find that a generic traffic is characterized by the mean frequency of information arrival and the mean duration of transmission.

- ***Mean Frequency of Information Arrival:*** The mean frequency of information arrival is measured in calls per second in a telephone network or packets per second in a packet data network, and will be denoted as λ.
- ***Mean Duration of Transmission:*** The mean duration of transmission $\mathrm{E}[X]$ of each arrival (e.g., referring to the length of a call or to the transmission time of a packet) on a link. The product of the mean arrival frequency and the mean transmission time yields the traffic intensity, ρ, which is a dimensionless quantity, measured in Erlangs, and defined as

$$\rho = \lambda \mathrm{E}[X] . \tag{4.12}$$

In particular, in an old telephone network, the traffic is analog and its intensity is measured as the product of the mean call arrival rate and the mean call duration. The traffic intensity at a local exchange represents the mean number of simultaneously active phone calls. In a data network, the traffic is digital; the traffic intensity at a node can be obtained as the product of the mean packet (or message) arrival rate and the mean packet (or message) transmission time. When different and independent traffic flows combine at the entrance of a node, the resulting total traffic intensity is equal to the sum of the traffic intensities of the single flows.

Referring to a generic link (i.e., a transmission line), the traffic intensity expresses the percentage of time that the input traffic occupies the link. Hence, the maximum (limit) load condition for a single communication line is represented by a traffic intensity of one Erlang ($\rho = 1$). Access links in the network are typically characterized by time-varying traffic conditions with low-intensity values (e.g., $\rho < 0.6$ Erlangs). Instead, transit links in the network have more regular traffic with medium–high intensity values (e.g., $\rho \approx 0.8$ Erlangs).

As it is evident from these initial considerations, two nodes not only exchange information generated by traffic sources but also need to exchange signaling (i.e., control) messages, which are necessary for the appropriate management of the network. Signaling can be required to establish an end-to-end path in the network for the exchange of information between source and destination. Moreover, signaling may be needed to provide acknowledgments of received data or to request retransmissions.

Each link in the network is characterized by the transmission of signals from the transmitter to the receiver through the channel. Due to the disturbances and distortions introduced on the signal by the communication channel, a modulator can be used at the transmitter in order "to transpose" the frequency spectrum of the signal in a band suitable to traverse the channel; correspondingly, a demodulator is necessary at the receiver. However, baseband transmissions (i.e., non-modulated) are also possible, for instance, in the case of transmissions on cables.

There are two generic forms of signals evolving in time, which can be transmitted in telecommunication systems, i.e., analog signals and digital signals. In the first case, we have a continuously varying signal that represents the electrical transduction of physical data. In the second case, only a few signal levels are possible (e.g., two values corresponding to the representation of bits "0" and "1," but there could also be more than two symbols). Digital signals have the advantage that, since only a few levels are possible, additive noise can be quickly canceled at the receiver using a simple threshold detector (let us refer here to a baseband signal). Finally, digital signals provide a common language, which permits to integrate different media, such as audio, video, and data.

Let us focus on digital transmissions.We refer to the well-known Shannon theorem: in a communication channel, it is possible to transmit up to a maximum bit-rate C (i.e., channel capacity measured in bits per second), guaranteeing that, with both suitable coding and digital modulation, the bit error probability (BEP) can be made as small as needed. In particular, for a bandlimited waveform channel with zero-mean additive white Gaussian noise (AWGN), where the one-sided power spectral density of the noise is equal to N_0, the channel capacity can be expressed as [4]

$$C = W \log_2 \left(1 + \frac{P}{N}\right), \tag{4.13}$$

where W is the channel bandwidth, P is the received signal power, and N is the received noise power given by

$$N = W N_0. \tag{4.14}$$

From (4.13), we can see that generically there is an important relationship between the available bandwidth of the transmission medium W and the bit-rate that can be achieved with a certain quality in terms of BEP. The capacity formula depends on the channel; for instance, a different capacity expression is obtained for the classical binary symmetric channel. The main characteristics of digital transmissions are detailed below.

- ***Serial or Parallel Transmissions:*** Serial transmissions involve sending data bit by bit over a single communication line. In contrast, parallel communications require at least as many lines as the number of bits in a word being transmitted (for an 8-bit word, at least 8 lines are needed). Serial transmissions are beneficial for long-distance communications, whereas parallel transmissions are suitable for short distances (cabling is limited to 5–10 m) or when very high transmission rates are required. The Recommended Standard 232 (RS-232) standard is the classical serial interface for the exchange of information between data terminal equipment and data communications equipment. This standard is characterized by the typical 25-pin D-shaped connectors. It allows transmission speeds from 110 bit/s to 19.2 kbit/s for a distance up to 15 m. Serial ports can be used in personal computers to connect mouse, modem, or special peripherals. Today, RS-232 has been superseded by the universal serial bus (USB) port that is much faster and has connectors that are easier to use. RS-232 ports are still used on programmable boards to upload the operating system on the local memory.

Serial transmissions can be of two different types: synchronous or asynchronous. We refer below to baseband transmissions. Data transmitted between nodes are organized into bits, bytes, and group of bytes, named packets. Synchronization involves delimiting and recovering bits, bytes, and packets. The synchronization type depends on the clocks used by the sender and the receiver.

- ***Synchronous Transmission:*** In synchronous transmissions, there is a global clock or synchronized clocks used in transmission and reception. The transmission unit is a packet of bits, sent together in a stream. The packet contains overhead bits (they are typically concentrated in a header, but some of them could also be in a trailer) and a data payload, as shown in Fig. 4.8.

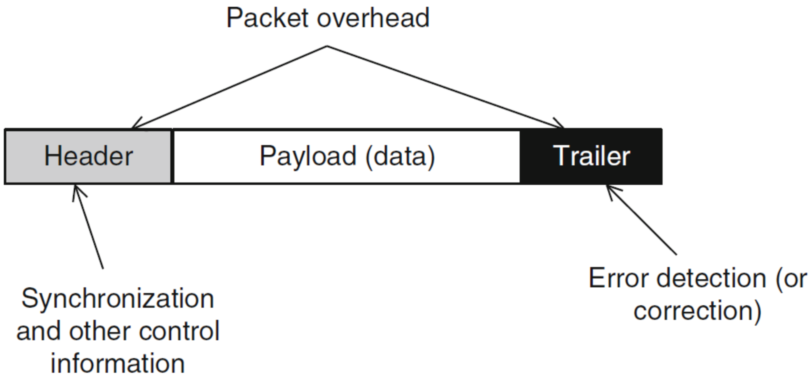
Packet overhead
Header
Payload (data)
Trailer
Synchronization and other control information
Error detection (or correction)

Fig. 4.8: Generic packet format for synchronous transmission

  The receiver must resynchronize at each new packet. Suitable bit sequences are at the beginning of a packet so that the receiver can acquire the right synchronism at the packet level (moreover, bits have adequate representation to ease the bit synchronization; this is typically accomplished by a suitable line code). Typically, 1–2 bytes are needed for packet synchronization. Since the packet can be sufficiently long, synchronous transmissions allow us to achieve higher efficiency than asynchronous ones. Synchronous communications are well suited to high bit-rate transmissions.

- ***Asynchronous Transmission:*** In asynchronous transmissions, transmitter and receiver clocks are independent. Asynchronous transmission is useful for human input/output data (e.g., a keyboard input) with random arrival times and transmission lines characterized by long idle phases. Let us refer to the transmission of a character of one byte (7-bit ASCII code plus a parity bit) at once. Since there is no direct clock information exchanged between the receiver and the transmitter, the receiver must explicitly resynchronize at the first bit of each byte. To achieve such synchronization, additional start and stop bits must be used for sending each byte. Subsequent bits are recovered by estimating bit boundaries. Let us consider the example shown in Fig. 4.9 for the asynchronous transmission of a character (i.e., one byte).

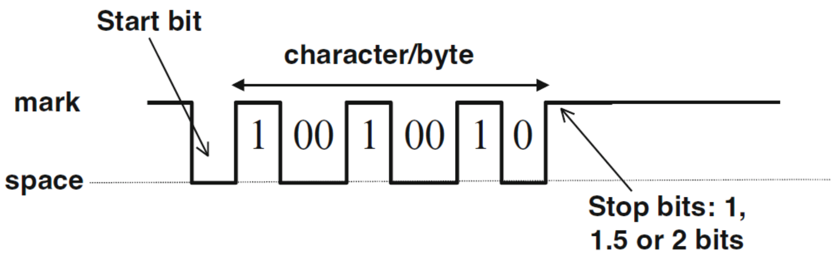

Fig. 4.9: Example of asynchronous transmission

The transmission of bit "1" is characterized by a high signal level, whereas the transmission of bit "0" corresponds to a low level. The start bit is a "0" and the end bit is (or bits are) "1" just to be sure that there is at least one transition in the character. Of course, the extra bits to manage the asynchronous transmission reduce the efficiency: 10–11 bits are needed to transmit a character of 8 bits; hence, 27.2% of the link capacity is lost due to the asynchronous protocol.

- ***Duplexing:***
  - ***Simplex.***
  - ***Half Duplex.***
  - ***Full Duplex.***

In half duplex and full duplex data exchanges, traffic is bidirectional. A data exchange is symmetric if both parties send a similar traffic load. This is the typical case of phone conversations. Otherwise, we have an asymmetrical situation. A typical example for computer networks is when a client connects to a remote server: the amount of data sent by the client is much lower than that provided by the server (typically, a 1:10 ratio can be considered).

- ***Bit Rate:***
  - ***Fixed Bit Rate:*** In this case the bit rate is equal to a fixed quantity R.
  - ***Variable Bit Rate:*** When the bit rate is variable, it can be described as an evolving-in-time quantity $R(t)$ as shown in Fig. 4.10.

    $R(t)$ can be modeled as a stochastic process. Digital traffic flows can be roughly distinguished into two broad families:

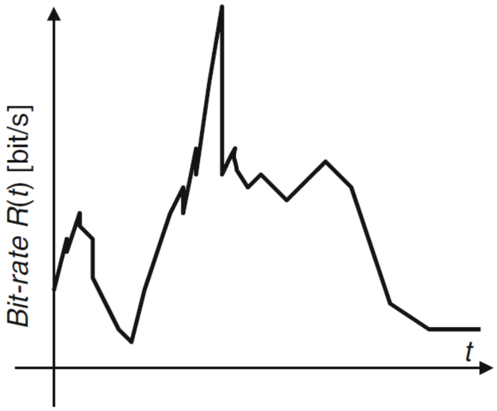

Fig. 4.10: Variable bit rate

- **Elastic Traffic** - typically referred to as data traffic, which can tolerate throughput variations, depending on network conditions.
- **Inelastic Traffic** - typically referred to as real time traffic for which the rate cannot be adjusted depending on network congestion.

Let us refer to real-time traffic and, in particular, to voice or video traffic sources. In both cases, we can consider variable bit-rate traffic sources. In the voice case, we have a constant bit-rate generation during a talking flow and a negligible traffic generation during a silent pause (ON–OFF voice traffic source). In the video case, bit-rate variations can be obtained since the images to be coded vary so that different compression values are achieved. Very bursty data traffic sources are those related to Internet traffic, where the bit-rate generated has very low values for long time intervals, but high and sudden peaks are possible. A fixed link capacity assigned to a bursty traffic source based on its peak traffic value can represent a wastage of resources. This is a crucial aspect to take into account when designing a network. If we aggregate the variable bit-rates generated by bursty traffic sources, we obtain a more smoothed traffic (i.e., traffic with lower variations) for which it is easier to predict the required capacity needs.

Referring to a data traffic source, we can define the burstiness β as the ratio between the maximum bit-rate $R_{\max}$ and the mean bit-rate $\mathrm{E}[R]$, i.e.,

$$\beta = \frac{R_{\max}}{\mathrm{E}[R]}. \tag{4.15}$$

For an ON–OFF voice traffic source, bit-rate $R(t)$ is equal to $R_{\max}$ in the on phase and equal to 0 in the off phase. Hence,

$$\mathrm{E}[R] = P_{\text{on}} R_{\max}, \tag{4.16}$$

where P_{on} denotes the percentage of the time spent by the source in the ON phase (i.e., activity factor). In conclusion, the ON–OFF traffic source has a burstiness degree given as

$$\beta_{\text{ON-OFF}} = \frac{1}{P_{\text{on}}} \\ > 1. \tag{4.17}$$

Assuming that the voice source traffic is transmitted over a digital line of capacity $R_{\max}$ bit/s, the burstiness degree represents the maximum (ideal) number of different ON–OFF voice sources that can be multiplexed onto the digital line. If the various voice sources would be ideally coordinated in their ON and OFF phases, we could have exactly $1/P_{\text{on}}$ voice sources sharing the use of the same line where they transmit alternately.

4.I Switching Techniques

Historically, three different types of switched networks can be distinguished: circuit switching, message switching, and packet switching networks. Each of these switching methods is suitable for a specific traffic type, whereas it could not be used (or it could be not efficient to use) for the transfer of other traffic classes. In general, circuit switching is well suited to traffic, which is regular (almost constant) for a sufficiently long time with respect to the procedures to set up the circuit. In contrast, message and packet switching are more appropriate for data traffic and, in particular, for variable bit rate and bursty traffic.

- ***Circuit Switching:*** Circuit switching is the solution adopted in old telephone networks. When a user makes a phone call towards another user, the network establishes an end-to-end physical (i.e., electrical) connection for the whole duration of their conversation. The following subsequent phases characterize a circuit-switched connection and the related service.

    - ***Circuit Setup:*** In the case of a phone call, this phase starts when the originating user dials the phone number of the destination and ends when the originating user receives a tone, indicating whether the destination is available or not. In this phase, an end-to-end circuit is built and resources are reserved on the links and at the nodes along the path.

    - ***Information Transfer:*** Information transfer from a user to the other. In the case of the telephone service, this phase corresponds to the phone conversation between the two users. During this phase, an end-to-end physical connection is available and no network procedure is involved. Voice is transparently conveyed at the destination by the network.

    - ***Circuit Release:*** When the phone call is over (one of the two users closes the connection), the network operates a series of operations to release the resources reserved along the path. These resources can be made available to other users.

- ***Message Switching:*** Message switching technology was born in the 1960s. In this case, each message represents an autonomous information unit, typically composed of a variable number of bits. Subsequent messages for the same source–destination pair follow a path decided based on the dynamic state of the network. A network resource (i.e., a link) is used just for the time necessary to transmit a message; soon after it is available to serve other messages. To explain the message switching technique, let us refer to the example in Fig. 4.11, where terminal A sends a set of messages (i.e., messages M1, M2, and M3) to terminal B.

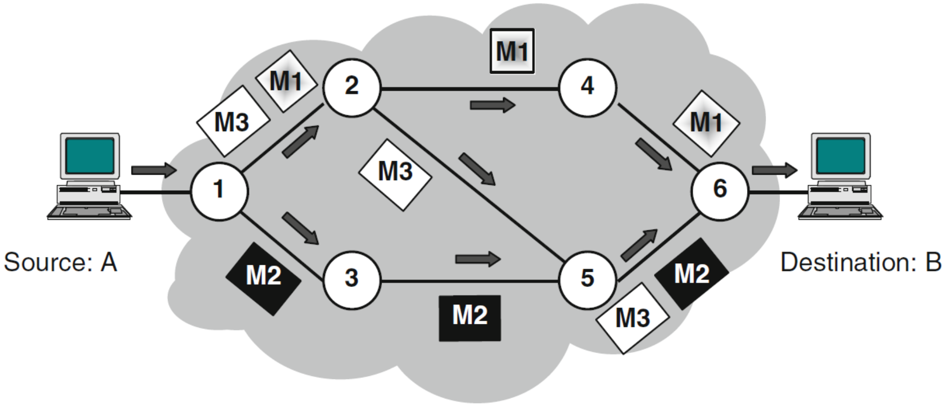

Fig. 4.11: Message switching

Each message is simply composed of a header and a payload. The header contains the address of source A and the address of destination B. Each message is autonomous, since it contains all the information for routing it to the destination. Each message crosses several nodes and links. When a message reaches a node (i.e., switching element), it is stored in a buffer and its header is processed to obtain the destination address. Based on this information, the node determines to which output link (and related node) the message has to be forwarded to reach its destination. Each node is of the "store-and-forward" type.

The telegram network technology was based on message switching. Message switching is the right solution for data traffic networks, characterized by bursty traffic. However, this technology has been overtaken by packet switching, which can achieve better performance in terms of fast switching at nodes and lower transmission delays on links.

✷ ***Packet Switching:*** Packet switching can be considered as an evolution of message switching. In particular, a message is segmented in packets of reduced length, each having a header (control information) and a payload carrying a fragment of the message. The header contains many control fields to manage the transmission of data on the links from source to destination. There should also be a counter to determine the number of payload fragments needed to reassemble the original message. Each packet is an autonomous entity.

Packet-switched transmissions may occur according to two different methods: virtual circuit and datagram. In both cases, buffers are needed at the different network nodes to store the packets to be transmitted on the various output links.

✦ ***Virtual Circuit:*** In the virtual circuit mode, a "logical" path is established in the network from source to destination: there is a setup phase similar to that described for circuit-switched networks. Once the path has been defined in the network, the packet forwarding is very fast from node to node (nodes have not to determine a new route at each new packet, since the flow has a well-defined path). All the packets of a traffic flow have the same route from source to destination, as shown in Fig. 4.12.

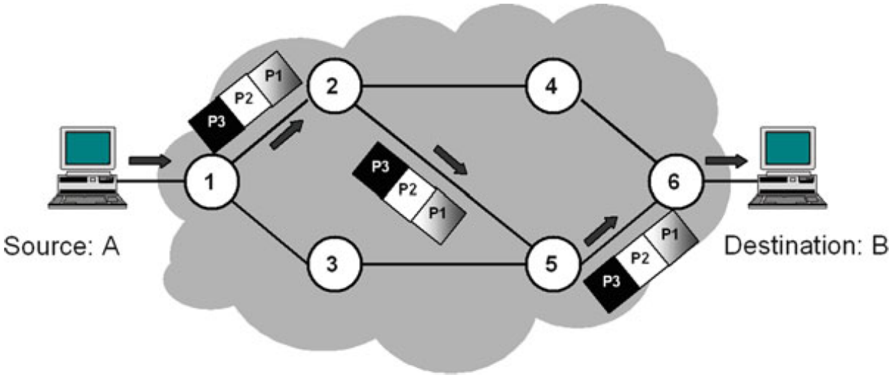

Fig. 4.12: Packet switching based on virtual circuits

Therefore, packets are received in the same order of generation; no reordering is needed at the destination. The virtual circuit mode is quite common in telecommunication networks (e.g., ATM networks).

- ***Datagram Mode:*** In the datagram mode, each packet is independently routed through the network towards its destination. Hence, packets generated from the same message may have different paths along the network from source to destination. Consequently, packets may arrive at the destination in a different order with respect to that of their generation. The destination node has to reorder the packets using a sequence number contained in the packet header. This transmission mode is similar to message-switching. The datagram transmission mode is employed in the Internet since it allows some advantages as follows:
  - No circuit must be created before the exchange of data between source and destination.
  - This switching mode is more robust to network faults, malfunctioning, and congestion. The route of packets can be dynamically adapted in response to changing network conditions. On the other hand, in the virtual circuit mode, after a node fault/congestion, all the virtual circuits crossing that node are interrupted/affected.

The datagram transmission mode requires that each packet contains the geographical address of the destination that must be processed at each node to find the appropriate output port. In the Internet Protocol version 4 (IPv4), the address field requires 32 bits. An IPv4 address is expressed as a series of four decimal numbers separated by periods, such as 192.168.0.1. Each of the four numbers can range from 0 to 255. In the Internet Protocol version 6 (IPv6), the address field requires 128 bits. An IPv6 address is represented in hexadecimal format, divided into eight groups of four hexadecimal digits separated by colons (e.g., 2001:0db8:85a3:0000:0000:8a2e:0370:7334). An Internet packet contains the addresses of both source and destination in the header.

With packet-switching, the message is fragmented into many packets, each with header information. Let's assume that the message originates three packets. These packets are sent in sequence. Each packet is queued and processed independently at each node. In the virtual circuit mode there is an initial setup phase for establishing the end-to-end path, similarly to circuit switched calls. After this phase, packets are quickly forwarded to the next node without requiring a heavy processing load. The processing of packets in each node is heavier in the datagram mode. Hence, the virtual circuit mode is convenient if a more regular and sufficiently heavy traffic load is sent from node A to node B.

4.J The ISO/OSI Reference Model

A suite of protocols must be used to properly exchange data at each interface along a path between two network nodes. These protocols are organized according to a stack. This is the layering approach, namely, dividing a task into smaller pieces and then solving each of them independently. Each protocol layer has to perform a suitable function, which permits the above layers to address other aspects. Each layer provides communication services to the layer above. The protocol stack architecture was standardized in the 1970s by the International Organization for Standardization (ISO) with the famous name of Open Systems Interconnection (OSI) reference model. The target was to define an "open system," meaning that different network elements can interwork independently of the manufacturers. The ISO/OSI protocol stack entails seven protocol layers, as shown in Fig. 4.13.

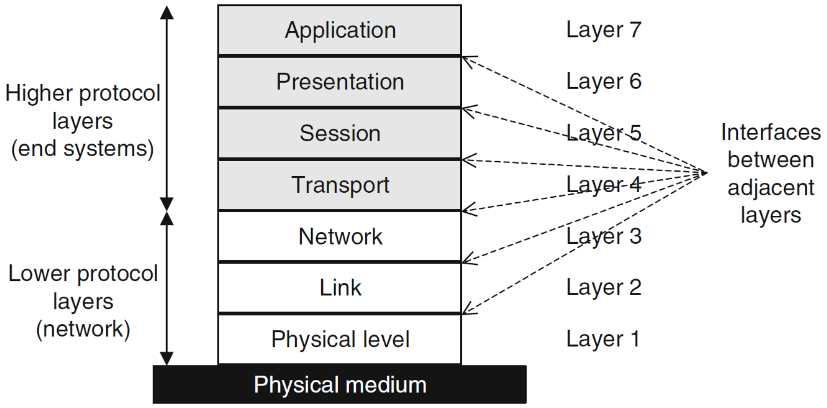

Fig. 1.15: OSI reference model for the protocol stack

Fig. 4.13: OSI reference model

Lower protocol layers (i.e., physical, link, and network layers) are present in every node of the network, including source and destination, which are called "End Systems." Instead, higher protocol layers (i.e., transport, session, presentation, and application) are present only at the source and destination.

Note that current trends in the design of the protocol stack also envisage interfaces between non-adjacent layers, thus violating the classical ISO/OSI classical structure. This is the cross-layer design, recently conceived for wireless networks, where a direct dialogue is also possible between protocols at non-adjacent layers. Fig. 4.14 shows the dialogue between user A and user B; these are the "End Systems," implementing the OSI protocol stack from layer 7 to layer 1.

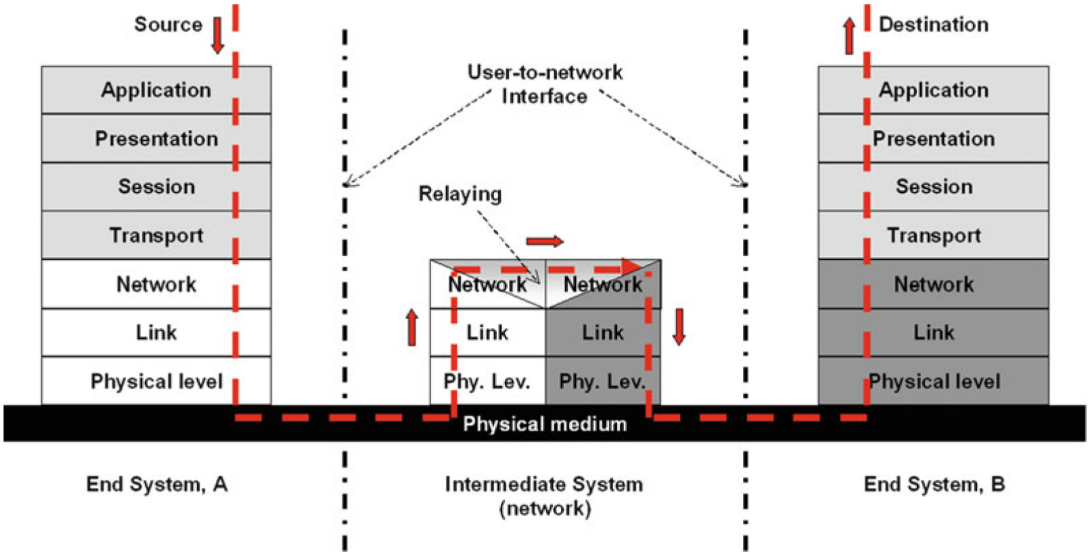

Fig. 4.14: OSI data path

A and B exchange data through a telecommunication network, which is denoted as "Intermediate System." Each network node in the intermediate system supports a reduced protocol stack; typically, only a few layers are implemented (in Fig. 4.14, only layers 1, 2, and 3 are adopted). Starting from source A, data are forwarded progressively from layer 7 to layer 1 and, then, transmitted. Data propagate through the physical medium, thus reaching the next node in the network (i.e., intermediate system). At this node, the information is reprocessed from layer 1 up to layer 3, assuming a network layer switching like in the Internet. When layer 3 is reached, data are not passed to upper layers but are managed at layer 3 to be passed again to the appropriate output link, thus going to layer 2 and physical layer, where transmission is performed. The function performed by layer 3 in the intermediate system in Fig. 4.14 is named "relaying." The protocol stacks on the left and right sides of the node of the intermediate system may be different. Note that the intermediate system can also implement the relaying function at different layers, depending on the network technology. In particular, the relaying function is at layer 1 in circuit-switched networks, at layer 2 in ATM networks, and at layer 3 in X.25 networks and the Internet.

4.J.1 Functions of the OSI Layers

Following is an outline of the main functions performed by the different layers of the OSI reference model.

- ***Layer 1:*** Physical layer, which directly carries out the transmission of bits through the physical medium. It is the layer that actually interacts with the transmission media, the physical part of the network that connects network components together. This layer is involved in physically carrying information from one node in the network to the next.

  The physical layer has complex tasks to perform. One major task is to provide services for the data link layer. The data in the data link layer consists of 0s and 1s organized into frames that are ready to be sent across the transmission medium. This stream of 0s and 1s must first be converted into another entity: signals. One of the services provided by the physical layer is to create a signal that represents this stream of bits.

  The physical layer must also take care of the physical network, the transmission medium. The transmission medium is a passive entity; it has no internal program or logic for control like other layers. The transmission medium must be controlled by the physical layer. The physical layer decides on the directions of data flow. The physical layer decides on the number of logical channels for transporting data coming from different sources.

- ***Layer 2:*** Data link layer has the primary function to regulate the access to physical layer resources and to recover errors through retransmission techniques. The data link layer transforms the physical layer, a raw transmission facility, to a link responsible for node-to-node (hop-to-hop) communication. Specific responsibilities of the data link layer include framing, addressing, flow control, error control, and media access control.

  The data link layer divides the stream of bits received from the network layer into manageable data units called frames. The data link layer adds a header to the frame to define the addresses of the sender and receiver of the frame. If the rate at which the data are absorbed by the receiver is less than the rate at which data are produced in the sender, the data link layer imposes a flow control mechanism to avoid overwhelming the receiver.

  The data link layer also adds reliability to the physical layer by adding mechanisms to detect and retransmit damaged, duplicate, or lost frames. When two or more devices are connected to the same link, data link layer protocols are necessary to determine which device has control over the link at any given time.

- ***Layer 3:*** Network layer has the task to route the traffic in the network from source to destination. The network layer is responsible for the source-to-destination delivery of a packet, possibly across multiple networks (links).

  A packet is a protocol data unit (PDU) at the network layer (Layer 3). It contains data along with a header that includes IP addresses (source and destination) and other routing information. Packets are primarily responsible for routing data across different networks. For example, an IP packet contains fields like the source IP address, destination IP address, and the time to live (TTL), which ensures the packet doesn't circulate indefinitely.

A frame, on the other hand, is the PDU at the data link layer (Layer 2). It encapsulates the packet with additional information required for local delivery between directly connected devices. A frame includes medium access control (MAC) addresses (source and destination), a type/length field, and often an error-checking mechanism like a cyclic redundancy check (CRC). Frames are used for physical transmission of data over a specific medium, such as Ethernet or wireless fidility (Wi-Fi).

In essence, packets are used for routing data between networks, while frames handle the actual transmission of data between devices on the same network. For example, when sending data over the Internet, a packet is encapsulated within a frame for local delivery. Once it reaches the next network, the frame is stripped, and the packet is forwarded to its destination. Whereas the data link layer oversees the delivery of the packet between two systems on the same network (links), the network layer ensures that each packet gets from its point of origin to its final destination.

✷ ***Layer 4:*** Transport layer performs the end-to-end control of the traffic flow from source to destination. Specific tasks are flow control (to avoid overwhelming the destination with too much traffic that it cannot handle) and congestion control (to avoid injecting too much traffic in the network, thus causing congestion at an intermediate node, also called “bottleneck”).

The transport layer is responsible for process-to-process delivery of the entire message. A process is an application program running on a host. Whereas the network layer oversees source-to-destination delivery of individual packets, it does not recognize any relationship between those packets. It treats each one independently, as though each piece belonged to a separate message, whether or not it does. The transport layer, on the other hand, ensures that the whole message arrives intact and in order, overseeing both error control and flow control at the source-to-destination level.

Computers often run several programs at the same time. For this reason, source-to-destination delivery means delivery not only from one computer to the next but also from a specific process on one computer to a specific process on the other. The transport layer header must therefore include a type of address called a service-point address in the OSI model and port number or port addresses in the TCP/IP protocol suite.

A transport layer protocol can be either connectionless or connection-oriented. A connectionless transport layer treats each segment as an independent packet and delivers it to the transport layer at the destination machine. A connection-oriented transport layer makes a connection with the transport layer at the destination machine first before delivering the packets. After all the data is transferred, the connection is terminated.

✷ ***Layer 5:*** Session layer manages the dialogue between the two end application processes. The session layer is responsible for managing and controlling the dialogues (sessions) between applications running on different devices. This layer ensures that communication sessions are established, maintained, and terminated in an orderly and reliable manner.

The session layer handles several critical tasks to facilitate communication between systems.

- ***Session Establishment, Maintenance, and Termination:*** It sets up, manages, and gracefully closes communication sessions between applications. This includes negotiating session parameters like authentication and communication direction (e.g., full-duplex or half-duplex).
- ***Synchronization:*** It uses checkpoints or markers in the data stream to ensure recovery in case of disruptions, such as network failures. This prevents data loss and ensures continuity.
- ***Dialog Management:*** It determines which device can send or receive data at a given time, supporting both half-duplex (one-way communication at a time) and full-duplex (simultaneous communication) modes.
- ***Activity Management:*** It divides data into logical units called activities, allowing independent processing of each activity.
- ***Resynchronization:*** In case of disruptions, it restores the session to a known state using predefined synchronization points.

Devices that interact with the session layer include firewalls, proxy servers, and application servers that manage session creation and control. While the session layer is a distinct concept in the OSI model, its functions are often integrated into the application layer in modern TCP/IP networks. For example, protocols like TCP handle some session-related tasks, such as orderly connection termination. This layer is particularly crucial in applications requiring high data integrity and continuity, such as video conferencing, file transfers, and remote procedure calls.

- ***Layer 6:*** Presentation layer is needed to unify the representation of information between source and destination. It acts as a translator for the network, ensuring that the data exchanged between devices is in a format both systems can understand. Because of its role in ensuring proper data representation, this layer is often referred to as the translation layer or the syntax layer.

When the application layer generates data, the presentation layer converts it into a standard form that can be transmitted across the network. Similarly, when data is received, it translates it into a format the receiving system can process. Some tasks the application layer performs are:

- It maintains proper syntax and semantics of the data.
- It rovides encryption and decryption for secure communication.
- It applies compression techniques to optimize bandwidth usage.

- ✦ It ensures compatibility between different systems and devices.

- ✷ ***Layer 7:*** Application layer represents the high-level service, having direct interactions with the user. The application layer enables the user, whether human or software, to access the network. It provides user interfaces and support for services such as e-Mail, file access and transfer, access to system resources, surfing the world wide web, and network management.

It is important to remark that the protocol specifications for a layer are independent of the specifications of the protocols at the other layers. In other words, it is possible to change a protocol in a layer with another without having to change anything in the protocols of adjacent layers. Of course, the service provided to the adjacent layers must remain unchanged. The protocols from the physical layer to the transport one are related to the network infrastructure and deal with telecommunication aspects from the transmission, to error management, to routing, and, finally, to flow and congestion control, whereas protocols of layers 5–7 are mainly related to software elaboration aspects.

Let us refer to a "system" (i.e., a terminal, a host, etc.) implementing the OSI protocol stack. The generic layer $X \in \{1, 2, \cdots, 7\}$ is composed of functional groups, named entities. A layer may contain more than one entity. For instance, there will be N entities at layer $X = 3$. Each entity provides a service to the upper layer through an interface. Upper layer entities access to this service through a service access point (SAP); there may be different SAPs at the interface between two layers. A unique SAP address identifies each SAP. The exchange of messages between two adjacent layers in a stack is made through primitives. Each entity also receives services from lower layer protocols through the lower level SAP. For example, a transport entity (layer $X = 4$) provides a service to upper layers through a T-SAP and receives a service from lower layers through an N-SAP. As for the interaction between "systems," it occurs through the dialogue of entities of the same layer (i.e., peer entities), according to rules, depending on the protocol of the layer considered.

A protocol is characterized as follows:

1. A set of formats according to which data exchange occurs between peer entities.

2. A set of procedures to exchange data.

Standardization bodies define the different protocols, which a system can use to exchange information. The implementation of interfaces is left free to manufactures, provided that they support the primitives characterizing the service (standard). The protocols of a given layer format their messages in transfer units, generically called PDUs. PDUs are exchanged by end systems through the services provided by lower layers.

The PDUs can be very different at various layers, from the user information at layer 7 to the bits transmitted on the physical link at layer 1. Information is exchanged by means of PDUs through SAPs between adjacent layers. For instance, a PDU of layer $X + 1$ is received by the lower layer X through a SAP and is considered as a service data unit (SDU) of layer X. This SDU can in turn be enriched with a header, containing additional control information of layer X (encapsulation); we have thus obtained a PDU of layer X. If the SDU received from layer $X + 1$ has a length exceeding the maximum value allowed by layer

X, the SDU is fragmented in different segments (the corresponding entity on the receiver side has to reassemble the different segments); conversely, several very short SDUs can be aggregated into a longer one. The process from input PDU to SDU to output PDU repeats at each layer of the OSI protocol stack; see Fig. 4.15.

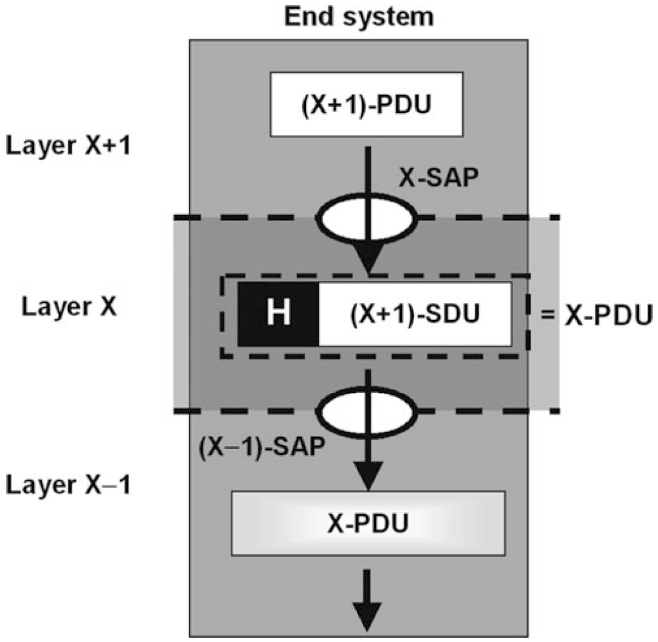

Fig. 4.15: Exchange of data through layer SAPs

Hence, the PDU of a given layer becomes the SDU of the layer below. For instance, an N-entity receives a T-PDU: layer 3 adds a header to this SDU, thus obtaining an N-PDU. Peer entities have a colloquium as if they were directly exchanging PDUs. The protocol of a given layer can perform a multiplexing function: the SDUs received from different SAPs can be addressed to the same SAP of the lower layer; otherwise, parallel transmissions can also be employed by using different SAPs towards the lower layer; see Fig. 4.16.

The header added at the generic layer X is needed to manage the protocol of layer X. The process of exchanging information through different layers is detailed in Fig. 4.17. As already explained, data received from the upper layer (in the form of an SDU) are encapsulated with a header (to form a PDU) and passed to the lower layer. Each protocol layer can provide either a connection-oriented service or a connectionless transfer service with the corresponding peer protocol at the destination. A connection-oriented service is characterized by three phases:

- Connection Establishment,
- Data Transfer,
- Connection Release.

As soon as the connection is obtained, PDUs are exchanged by specifying the identifier of the connection. Connectionless services are characterized by sending independent PDUs,

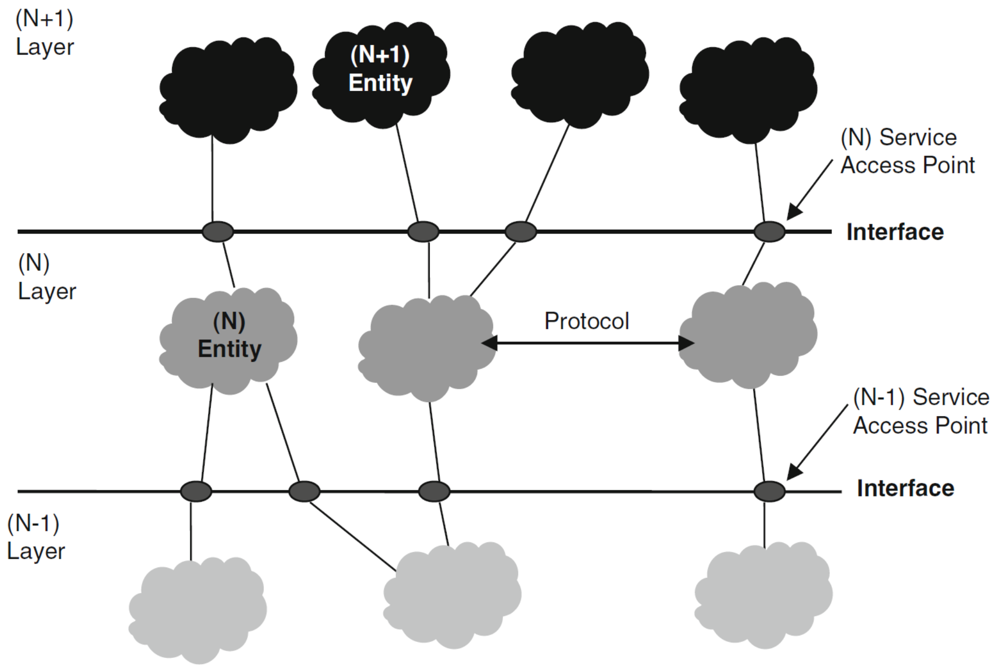

Fig. 4.16: Layers and SAPs

each typically containing the address of both source and destination. Each PDU has an autonomous route in the network: PDUs of the same service may have different paths to reach the same destination; hence, subsequent PDUs could not be received in order due to different delays. The selection between a connection-oriented service and a connectionless one can be performed at the link, network, and transport layers. In particular, on top of layers 2, 3, and 4, there are two SAPs through which the upper layer can access either connection-oriented or connectionless services.

Since the information exchange must occur between two generic terminals connected by the network, an important network functionality is addressing that allows identifying the destination to which information has to be delivered. The network level that receives a PDU with the destination address must decide the SAP towards which to forward the information. This is the routing functionality. In particular, the layer 3 of each intermediate node has to support two essential functions:

- ***Routing:*** Routing is to select the appropriate output SAP for the PDU, depending on the destination address; this is obtained through a routing table.
- ***Forwarding:*** Forwarding is to transfer the PDU from the input SAP to the output one.

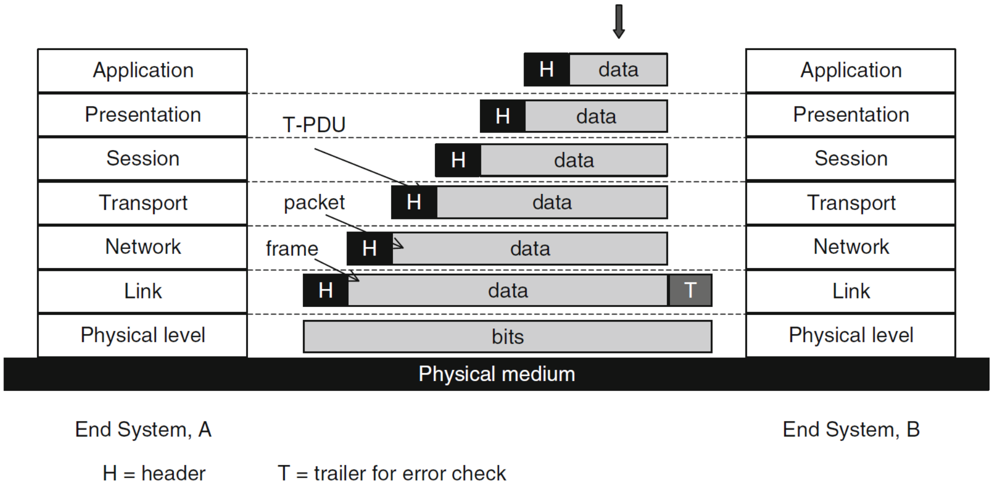

Fig. 4.17: PDUs from source to destination

Chapter 5

Markov Chains and Queuing Theory

5.A Queues and Stochastic Processes

Telecommunication systems are characterized by the transmission of data on wired or wireless links. In these cases, we have that different "messages" share the use of the same transmission resources. Typical examples can be as follows:

- Different phone calls arrive at a switching node and must be routed on a limited set of output links.
- Different packets need to be sent on the same link.

Transmission requests can be different instances of the same process or be generated by concurrent (and uncoordinated) processes, sharing the same transmission resources. All these cases involve the queuing of either different packets or different calls if there are not enough resources for their simultaneous transmissions. In telecommunication networks, the following ones are typical examples of problems that can be tackled by queuing theory:

- Performance analysis for the transmission on links and corresponding buffer dimensioning.
- Network planning (i.e., planning of the capacity needed to interconnect the different nodes of a telecommunication network).
- Performance evaluation of access protocols where different "users" contend for the same resources.

A queue is characterized by the following:

- an ***Arrival Process*** of service requests,
- a ***Waiting List*** of the requests to be processed,
- a ***Discipline*** according to which the requests in the queue are selected to be served,
- a ***Service Process***.

Queues are special cases of stochastic processes that are represented by a state $X(t)$, denoting the number of service requests or "entities" or "customers" queued at time t. As a stochastic process, $X(t)$ is identified by a different distribution of random variable X at different instants t. The function $f_{X(t)}(x)$ will be used to denote the probability density function (PDF) of process X at time t . A stochastic process can be characterized as follows:

- The ***State Space***, which is the set of all the possible values, which can be taken by $X(t)$. Such a space can be continuous or discrete (if the state-space is discrete, the stochastic process is called a chain).
- ***Time*** variable t, which can belong to a continuous set or a discrete one.
- ***Correlation*** characteristics among $X(t)$ random variables at different instants t.

In order to account for the process correlation, we describe $X(t)$ in terms of its joint cumulative distribution function (CDF), sampling the process at different instants defined for any $n \in \overset{+}{\mathbb{I}}$ by the vector

$$\boldsymbol{t} = \begin{bmatrix} t_1 & t_2 & \cdots & t_n \end{bmatrix}^T . \tag{5.1}$$

Let's define the random vector $\boldsymbol{x} \in \mathbb{R}^n$ as

$$\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^T . \tag{5.2}$$

The joint CDF of $X(t)$ will be expressed as

$$F_X(\boldsymbol{x}, \boldsymbol{t}) = \Pr\left\{X(t_1) \le x_1, X(t_2) \le x_2, \cdots, X(t_n) \le x_n\right\} . \tag{5.3}$$

The joint PDF of $X(t)$ can be obtained by differentiating the joint CDF, i.e.,

$$f_{\boldsymbol{X}(\boldsymbol{t})}(\boldsymbol{x}) = \frac{\partial^n F_X(\boldsymbol{x}, \boldsymbol{t})}{\partial x_1 \partial x_2 \cdots \partial x_n}, \tag{5.4}$$

where $\boldsymbol{X}(\boldsymbol{t}) \in \mathbb{R}^n$ is defined as

$$\boldsymbol{X}(\boldsymbol{t}) = \begin{bmatrix} X(t_1) & X(t_2) & \cdots & X(t_n) \end{bmatrix}^T . \tag{5.5}$$

The expected value $\mathrm{E}[X(t)]$ and the autocorrelation function $R(t_1, t_2)$ of process $X(t)$ can be expressed as

$$\mathrm{E}[X(t)] = \int_{-\infty}^{\infty} \tau f_{X(t)}(\tau) d\tau, \tag{5.6}$$

$$R(t_1, t_2) = \mathrm{E}[X(t_1) X(t_2)] . \tag{5.7}$$

A process $X(t)$ is said to be wide-sense stationary if its mean value $\mathrm{E}[X(t)]$ and autocorrelation function $R(t, t+\tau) = \mathrm{E}[X(t) X(t+\tau)]$ are both independent of t, i.e,

$$\mathrm{E}[X(t)] = \mu, \tag{5.8}$$

and

$$R(t, t+\tau) = R(\tau). \tag{5.9}$$

A process is independent if its CDF can be written for all values of n and t as

$$F_X(\boldsymbol{x}, \boldsymbol{t}) = \Pr\{X(t_1) \le x_1\} \Pr\{X(t_2) \le x_2\} \cdots \Pr\{X(t_n) \le x_n\}. \tag{5.10}$$

When (5.10) holds, it can be shown using (5.4) that the joint PDF can be written for all values of n and t as

$$f_{\boldsymbol{X}(\boldsymbol{t})}(\boldsymbol{x}) = f_{X(t_1)}(x_1) f_{X(t_2)}(x_2) \cdots f_{X(t_n)}(x_n). \tag{5.11}$$

The probability mass function (PMF) of a discrete random variable X is defined as

$$P_X(x) = \Pr\{X = x\}. \tag{5.12}$$

The probability generating function (PGF) is a transform adopted for non-negative integer-valued discrete random variables for which the PMF is known. The PGF of random variable X is defined as

$$\begin{aligned} X(z) &= \mathrm{E}\left[X^z\right] \\ &= \sum_{k=0}^{\infty} P_X(k) z^k. \end{aligned} \tag{5.13}$$

5.B Markov Chains

A particular type of stochastic process is a chain that evolves in time by making transitions between states, i.e., discrete values of $X(t)$. These transitions can occur at any instant in continuous-time chains or at specific instants in discrete-time chains. A Markov chain is characterized by the fact that its state value at instant t_{n+1} depends only on its state value at the previous instant t_n, i.e., $X(t_{n+1})$ depends only on $X(t_n)$.

The formal definition of a Markov chain $X(t)$ is

$$\begin{aligned} \Pr\{X(t_{n+1}) = x_{n+1} | X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, \cdots, X(t_1) = x_1\} \\ = \Pr\{X(t_{n+1}) = x_{n+1} | X(t_n) = x_n\}. \end{aligned} \tag{5.14}$$

The evolution of a Markov chain does not depend on how long the chain is in the current state. This memoryless characteristic implies that state sojourn times are exponentially distributed for a continuous-time chain or geometrically distributed for a discrete-time chain. In what follows, we refer mainly to continuous-time Markov chains, where the transitions from one state to another are characterized by mean rates. There are Markovian chains, called birth-death Markov chains, where the transitions from the generic state $X = i$ are only towards state $X = i - 1$ or towards state $X = i + 1$. A birth-death process is a special case of the more general random walk process.

Renewal processes belong to another important type of chains. These are "point" processes (i.e., arrival processes or only-birth processes), like the arrival of points on the time axis. Intervals between adjacent arrivals (points) are independent and identically distributed (IID), according to a general distribution. A generic arrival process can be equivalently characterized by the process $N(t)$ of the number of arrivals in a generic interval t or the distribution of the interarrival times. A special case of renewal processes is the Poisson arrival process, where interarrival times are exponentially distributed with a constant rate. In some cases, each arrival carries multiple "service requests" or "objects": for instance, the arrival of a message that carries multiple packets simultaneously (this could be case of an IP packet fragmented into many layer 2 packets). This group arrival case can have different names in the literature, such as bulk arrival process, batched arrival process, and compound arrival process. These names will be used interchangeably.

Markov chains are characterized by diagrams with states (represented by circles) and transitions (represented by directed arcs). In the case of a continuous-time chain, transitions may occur at any time and are characterized by exponentially distributed intervals with mean rates shown above the arcs of the transitions as shown in Fig. 5.1.

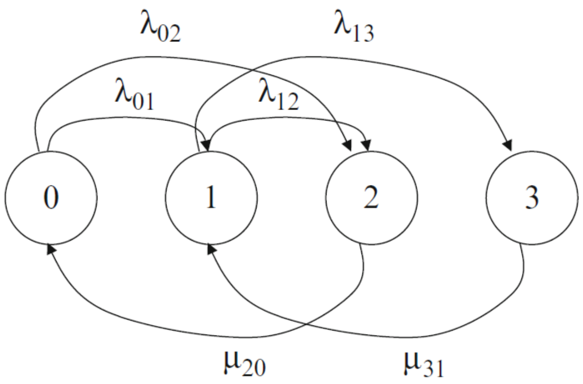

Fig. 5.1: Continuous-time Markov chain with mean transition rates between states

Instead, transitions can only occur at given instants for discrete-time chains; probabilities are used to characterize the transitions that correspond to geometrically distributed intervals. In the discrete-time case, states may have transitions into themselves as shown in Fig. 5.2.

The sum of all the transitional probabilities leaving a state must be equal to 1. A Markov chain is said to be irreducible if it is possible to get to any state from any state. A state i has period k if any return to state i must occur in multiples of k steps. If $k = 1$, then the state is said to be aperiodic. A Markov chain is aperiodic if every state is aperiodic.

5.C Poisson Arrival Process

A Poisson process can be used to describe the number of arrivals $N(t)$ (or equivalently N_t) for any interval of duration t. We have a Poisson arrival process if the following condition

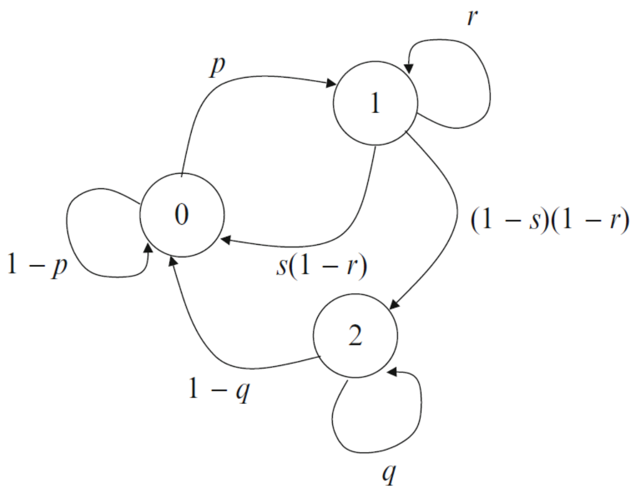

Fig. 5.2: Discrete-time Markov chain with mean transition rates between states

holds:

$$\Pr\left\{N(t)=k\right\}=\frac{(\lambda t)^k}{k!}e^{-\lambda t}, \tag{5.15}$$

where λ is the mean arrival rate. The PGF of the number of arrivals in an interval of duration t is equal to

$$\begin{aligned} N_t(z) &= \sum_{k=0}^{\infty} z^k \frac{(\lambda t)^k}{k!}e^{-\lambda t} \\ &= e^{-\lambda t}\sum_{k=0}^{\infty}\frac{(z\lambda t)^k}{k!} \\ &= e^{-\lambda t}e^{z\lambda t} \\ &= e^{\lambda t(z-1)}. \end{aligned} \tag{5.16}$$

The mean number of arrivals in an interval of duration t is equal to

$$\begin{aligned} \mathrm{E}\left[N_t\right] &= \left.\frac{dN_t(z)}{dz}\right|_{z=1} \\ &= \left.\lambda t e^{\lambda t(z-1)}\right|_{z=1} \\ &= \lambda t. \end{aligned} \tag{5.17}$$

Note that the mean value of the arrival process is not independent of time; meaning that the process cannot be wide sense stationary (WSS). The mean square (MS) value of the number of arrivals in an interval of duration t is equal to

$$\begin{aligned} \mathrm{E}\left[N_t^2\right] &= \left.\frac{d^2N_t(z)}{dz^2}\right|_{z=1} + \left.\frac{dN_t(z)}{dz}\right|_{z=1} \\ &= \left.(\lambda t)^2 e^{\lambda t(z-1)}\right|_{z=1} + \left.\lambda t e^{\lambda t(z-1)}\right|_{z=1} \\ &= \lambda t(\lambda t+1). \end{aligned} \tag{5.18}$$

On the basis of (5.17), it is evident that λ represents the mean arrival rate of the process. Note that the variance of N_t is equal to

$$\begin{aligned}\sigma^2_{N_t} &= \mathrm{E}\left[N_t^2\right] - \left(\mathrm{E}\left[N_t\right]\right)^2 \\ &= \lambda t,\end{aligned} \tag{5.19}$$

which is equal to its mean value. This is a special characteristic of Poisson processes.

The number of Poisson arrivals in disjoint intervals is statistically independent; instead, the number of Poisson arrivals in overlapped intervals is not independent. Hence, N_t and N_s, where t and s are generic instants, are not independent variables. However, even if N_t and N_s are not independent variables, $N_t - N_s$, and N_s are independent variables if $t > s$. The autocorrelation function of a Poisson process for $t > s$ can be obtained as

$$\begin{aligned}R_{NN}(t,s) &= \mathrm{E}\left[N_t N_s\right] \\ &= \mathrm{E}\left[\left(N_t - N_s\right) N_s + N_s^2\right] \\ &= \mathrm{E}\left[\left(N_t - N_s\right)\right] \mathrm{E}\left[N_s\right] + \mathrm{E}\left[N_s^2\right] \\ &= \left(\lambda t - \lambda s\right) \lambda s + \lambda s \left(\lambda s + 1\right) \\ &= \lambda^2 t s + \lambda s.\end{aligned} \tag{5.20}$$

The autocorrelation function of the arrival process $R_{NN}(t,s)$ is not equal to $R_{NN}(t-s)$, meaning that in this case $R_{NN}(t,s)$ depends on the actual measurement times rather than the time difference $t-s$. This result together with the fact that mean value depends on time as clear from (5.17), confirm that the arrival process is not WSS. Nevertheless, increments of Poisson processes are WSS (for instance, $N_t - N_s$).

The autocovariance of the Poisson process can be obtained considering the previous result for $R_{NN}(t,s)$ with $t > s$ according to

$$\begin{aligned}C_{NN}(t,s) &= \mathrm{E}\left[\left(N_t - \overline{N_t}\right)\left(N_s - \overline{N_s}\right)\right] \\ &= \mathrm{E}\left[N_t - N_s\right] - \overline{N_t}\;\overline{N_s} \\ &= \lambda s.\end{aligned} \tag{5.21}$$

Note that $\overline{N_t}$ and $\overline{N_s}$ have been used in the above to denote $\mathrm{E}\left[N_t\right]$ and $\mathrm{E}\left[N_s\right]$, respectively. Let us define the index of dispersion for counts (IDC) for a generic arrival process (or point process) as the ratio between the variance of the number of arrivals in a given interval t and the mean number of arrivals in the same interval, i.e.,

$$I_{N_t} = \frac{\sigma^2_{N_t}}{\overline{N_t}}. \tag{5.22}$$

For a Poisson process, IDC is equal to one for all t. In general, for a renewal process, IDC cannot be equal to one. An arrival process is peaked if IDC is larger than one. An arrival process is smoothed if IDC is smaller than one. The limiting case is when IDC is zero; so that the arrival process is deterministic and arrivals occur at fixed, regular intervals.

Conversely, when IDC is larger than one, arrivals tend to occur in bursts (i.e., bursty arrival process). Bursty arrival processes cause the sudden queuing of requests in queuing systems and consequently high delays. For given resources and mean arrival rate, the mean queuing delay increases with IDC.

Note that a Poisson arrival process is characterized by only one parameter, i.e., the mean rate λ. From measurements on traffic traces, we can consider having a Poisson process when the mean and variance of the number of arrivals in intervals of length t are equal; correspondingly, we derive λ as the ratio of the mean number of arrivals in an interval of length t and time t itself.

Let us study the statistics of interarrival times t_a for the Poisson process. Let $t = 0$ denote the instant of the last arrival. We determine the probability that the next arrival occurs at a generic instant $t > 0$; this is equivalent to considering the probability that there is no Poisson arrival in the interval $(0, t)$, which is equal to $e^{-\lambda t}$. We have thus obtained the complementary cumulative distribution function (CCDF) of t_a as

$$\Pr\left\{t_\mathrm{a} > t\right\} = e^{-\lambda t}. \tag{5.23}$$

Therefore, the CDF and PDF of t_a are, respectively, equal to

$$F_{t_\mathrm{a}}(t) = 1 - e^{-\lambda t}, \tag{5.24}$$

and

$$f_{t_\mathrm{a}}(t) = \lambda e^{-\lambda t}. \tag{5.25}$$

Hence, t_a is exponentially distributed with mean rate λ. Interarrival times are iid. It is possible to prove that we have a Poisson arrival process with mean rate λ if and only if interarrival times are exponentially distributed with mean rate λ (mean value $1/\lambda$).

Poisson processes are quite important in the field of telecommunications since they may model the arrival of several types of events, such as:

- The arrival of new calls at a node of a telephone network (see Fig. 5.3).
- The arrival of email messages in a packet data network.

As a final remark, it is important to point out that the wide adoption of exponential distributions and Poisson arrival processes is not completely linked to the empirical evidence (measurements), but rather to the ease of conditioning with the help of the memoryless property.

5.D SUM OF INDEPENDENT POISSON PROCESSES

Let us consider two independent sources of Poisson arrivals Σ_1 and Σ_2 with related mean rates λ_1 and λ_2. Let $X_1(t)$ and $X_2(t)$ denote the numbers of arrivals from Σ_1 and Σ_2, respectively, in a given interval of duration t. We want to characterize the sum process $X(t)$, where

$$X(t) = X_1(t) + X_2(t). \tag{5.26}$$

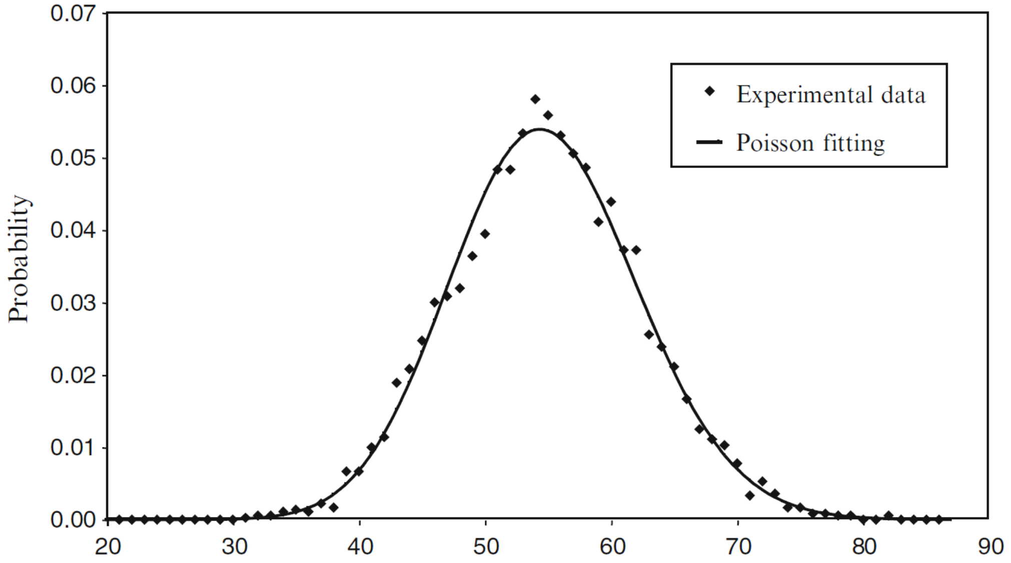

Fig. 5.3: Histogram of arrivals at a switching node in a telephone network

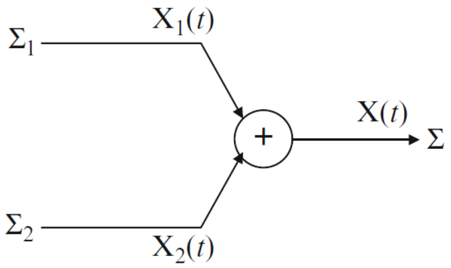

Fig. 5.4: Summing Poisson processes

The summing process is illustrated in Fig. 5.4.

Since $X_1(t)$ and $X_2(t)$ are independent, the PGF of $X(t)$, denoted as $X(z)$, is given by

$$X(z) = X_1(z)X_2(z), \tag{5.27}$$

where $X_1(z)$ and $X_2(z)$ are the PGFs of $X_1(t)$ and $X_2(t)$, respectively. Considering the Poisson characteristic of both $X_1(t)$ and $X_2(t)$ as given in (5.16), we have

$$X_1(z) = e^{\lambda_1 t(z-1)}, \tag{5.28}$$

and

$$X_2(z) = e^{\lambda_2 t(z-1)}. \tag{5.29}$$

Substituting (5.28) and (5.29) into (5.27) we obtain

$$\begin{aligned} X(z) &= e^{\lambda_1 t(z-1)} e^{\lambda_2 t(z-1)} \\ &= e^{(\lambda_1+\lambda_2)t(z-1)}. \end{aligned} \tag{5.30}$$

From (5.30), we note that the PGF $X(z)$ corresponds to that of a Poisson process with mean rate $\lambda_1 + \lambda_2$. In conclusion, the process sum of two independent Poisson processes is still a Poisson process with the mean rate given by the sum of the mean rates of the processes. This is an important property in telecommunication networks since nodes can receive Poisson arrivals of messages from different and independent sources. Another typical example is given by a private branch exchange that collects call arrivals from different phone users.

Bibliography

[1] J. Walrand and S. Parekh, *Communication Networks: A Concise Introduction*, 2nd ed. Morgan Claypool Publishers, 2017.

[2] D. Bertsekas and R. Gallager, *Data Networks: Second Edition*. Athena Scientific, 2021.

[3] M. Newman, *Networks*. OUP Oxford, 2018.

[4] J. Proakis and M. Salehi, *Digital Communications*, ser. McGraw-Hill International Edition. McGraw-Hill, 2008.