EIE 418 LECTURE NOTES

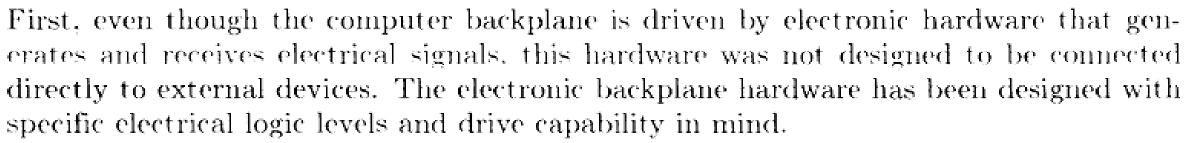
INTERFACING: INTERFACES FOR SIMPLE COMPUTER SYSTEM AND TERMINAL TO TERMINAL.

What is an interface?

An interface is a device and/or set of rules to match the output of one device to send information to the input of another device. An interface usually requires: (i) a physical connection, (ii) the hardware (iii) rules and procedures and last (iv) the software.
Interfacing is the process of connecting devices together so that they can exchange information.

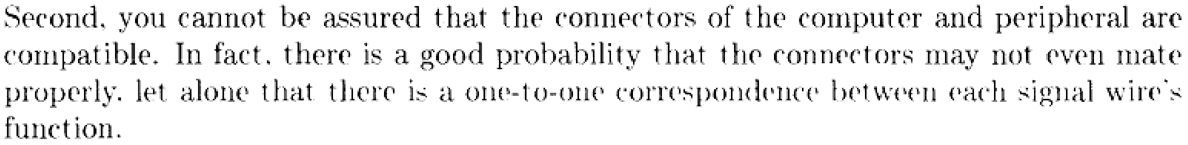
Why do we need an Interface?
The primary function of an interface is obviously to provide a communication path for data and commands between the computer and its resources. Interfaces acts as intermediaries between resources by handling part of the “bookkeeping” work and ensuring that the communication process flows smoothly.
Agreement regarding the signal type, how they should be converted and transmitted is not enough. Agreement is also required regarding the type of connector and the voltage levels they need to support. In other words, the physical and electrical interfaces are important. There is also a logical interface, which defines the significance of the signal. A protocol controls how the signals are built up, how communications are initiated, how they are terminated, the order of transmitting and sending, how to acknowledge a message, etc.
The physical interface defines how equipment is connected as well as the design of the connector. The electrical interface defines the electrical levels and what these denote (ones or zeros). The Logical interface defines what the signals signify.
In summary the following are reasons why we need an interface:

First Reason

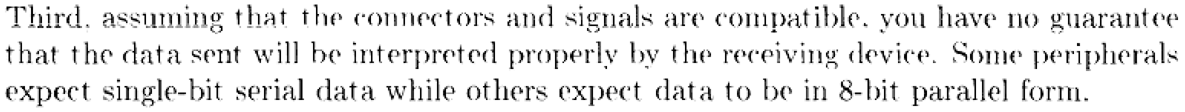

First, even though the computer backplane is driven by electronic hardware that generates and receives electrical signals, this hardware was not designed to be connected directly to external devices. The electronic backplane hardware has been designed with specific electrical logic levels and drive capability in mind.

Exceeding the backplane hardware ratings will damage the electronic hardware.

Second Reason


Second, you cannot be assured that the connectors of the computer and peripheral are compatible. In fact, there is a good probability that the connectors may not even mate properly, let alone that there is a one-to-one correspondence between each signal wire's function.

Third Reason


Third, assuming that the connectors and signals are compatible, you have no guarantee that the data sent will be interpreted properly by the receiving device. Some peripherals expect single-bit serial data while others expect data to be in 8-bit parallel form.

Fourth Reason

There is no reason to believe that the computer and peripheral will be in agreement as to when the data transfer will occur; and when the transfer does begin the transfer rates will probably not match.

From the foregoing it is obvious that interfaces have a great responsibility to oversee the communication between the computer and its resources. Computer intefacing has some advantages as outlined below:

1. Advanced control applications need flexible processing power which is readily provided by the computer. Hence the computer does the complex control processing and sends signals to control the process through appropriate interfaces.
2. We always need to input and output control data. For example we need inputs from sensors (speed, accelaration, temperature, etc.) while we need to give out utput to actuators (motors, switches, valves). The computer can readily receive inputs and provide corresponding outputs once the right interface has been provided.
3. We are able to access the numerous advantages of using the computer for data acquisition and control such as in high speed proceesing, programming flexibility which is usually unavailable in hard wired logic, mass storage of data, data analysis and visualization and relatively low cost.

FUNCTIONS OF AN INTERFACE

The functions of an interface are shown in the block diagram of Figure1. An interface must ensure electrical and mechaninical compatibility, data compatibility, timing compatibility and some other additional functions as is required for data communication to take place between a computer and its peripherals.

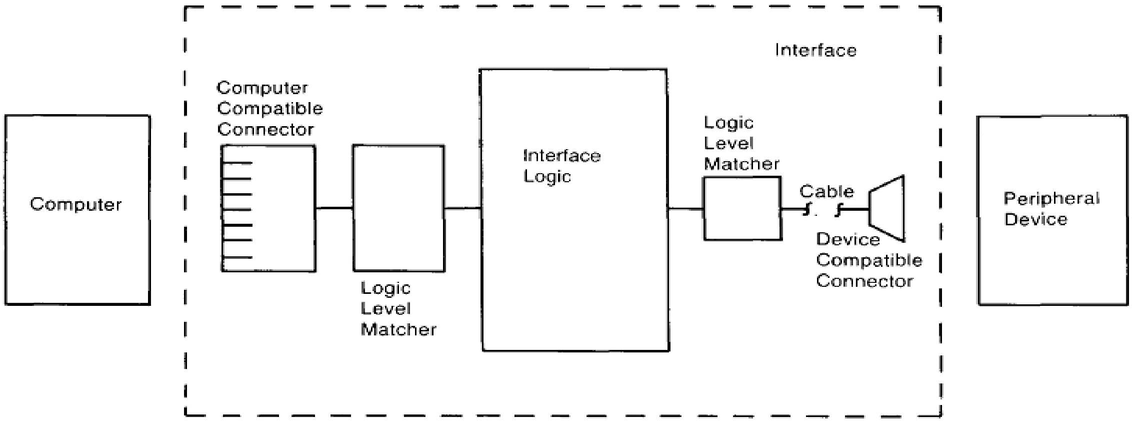


Figure 1: Functional Diagram of an Interface

Electrical and Mechanical Compatibility

Electrical compatibility must be ensured before any thought of connecting two devices occurs. Often the two devices have input and output signals that do not match; if so, the interface serves to match the electrical levels of these signals before the physical connections are made.

Mechanical compatibility simply means that the connector plugs must fit together properly. All of the 9826 interfaces have 100-pin connectors that mate with the computer backplane. The peripheral end of the interfaces may have unique configurations due to the fact that several types of peripherals are available that can be operated with the 9826. Most of the interfaces have cables available that can be connected directly to the device so you don't have to wire the connector yourself.

Data Compatibility

Just as two people must speak a common language, the computer and peripheral must agree upon the form and meaning of data before communicating it. As a programmer, one of the most difficult compatibility requirements to fulfill before exchanging data is that the format and meaning of the data being sent is identical to that anticipated by the receiving device. Even though some interfaces format data, most interfaces have little responsibility for matching data formats: most interfaces merely move agreed-upon quantities of data to or from computer memory. The computer must generally make the necessary changes, if any, so that the receiving device gets meaningful information.

Timing Compatibility

Since all devices do not have standard data-transfer rates, nor do they always agree as to when the transfer will take place, a consensus between sending and receiving device must be made. If the sender and receiver can agree on both the transfer rate and beginning point (in time), the process can be made readily.

If the data transfer is not begun at an agreed-upon point in time and at a known rate, the transfer must proceed one data item at a time with acknowledgement from the receiving device that it has the data and that the sender can transfer the next data item: this process is known as a "handshake". Both types of transfers are utilized with different interfaces and both will be fully described as necessary.

Additional Interface Functions

Another powerful feature of some interface cards is to relieve the computer of low-level tasks, such as performing data-transfer handshakes. This distribution of tasks eases some of the computer's burden and also decreases the otherwise-stringent response-time

requirements of external devices.

MODEM TERMINAL INTERFACES

The word "modem" is a contraction of the words **modulator-demodulator**. A modem is typically used to send digital data over a phone line. The sending modem **modulates** the data into a signal that is compatible with the phone line, and the receiving modem **demodulates** the signal back into digital data. **Wireless modems** convert digital data into radio signals and back. Modems came into existence in the 1960s as a way to allow terminals to connect to computers over the phone lines. A typical arrangement is shown below in Figure 2.

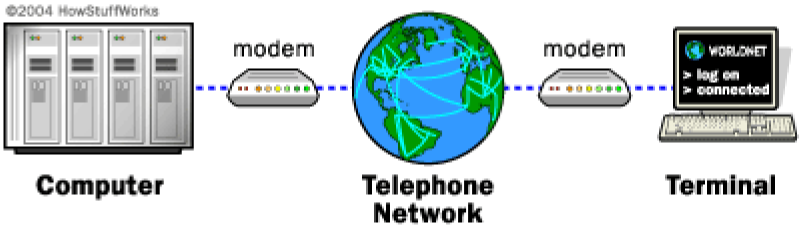


Figure 2: MODEM Interface

In a configuration like this, a **dumb terminal** at an off-site office or store could "dial in" to a large, central computer. The 1960s were the age of **time-shared** computers, so a business would often buy computer time from a time-share facility and connect to it via a 300-bit-per-second (bps) modem. A dumb terminal is simply a keyboard and a screen. A very common dumb terminal at the time was called the **DEC VT-100**, and it became a standard of the day (now memorialized in terminal emulators worldwide). The VT-100 could display 25 lines of 80 characters each. When the user typed a character on the terminal, the modem sent the ASCII code for the character to the computer. The computer then sent the character back to the terminal so it would appear on the screen.

When personal computers started appearing in the late 1970s, **bulletin board systems** (BBS) became the rage. A person would set up a computer with a modem or two and some BBS software, and other people would dial in to connect to the bulletin board. The users would run **terminal emulators** on their computers to emulate a dumb terminal.

People got along at 300 bps for quite a while. The reason this speed was tolerable was because 300 bps represents about 30 characters per second, which is a lot more characters per second than a person can type or read. Once people started transferring large programs and images to and from bulletin board systems, however, 300 bps became intolerable. Modem speeds went through a series of steps at approximately two-year intervals:

- 300 bps - 1960s through 1983 or so
- 1200 bps - Gained popularity in 1984 and 1985
- 2400 bps
- 9600 bps - First appeared in late 1990 and early 1991
- 19.2 kilobits per second (Kbps)
- 28.8 Kbps
- 33.6 Kbps
- 56 Kbps - Became the standard in 1998
- ADSL, with theoretical maximum of up to 8 megabits per second (Mbps) - Gained popularity in 1999

300-bps Modems

We'll use 300-bps modems as a starting point because they are extremely easy to understand. A 300-bps modem is a device that uses **frequency shift keying** (FSK) to transmit digital information over a telephone line. In frequency shift keying, a different tone (frequency) is used for the different bits.

When a terminal's modem dials a computer's modem, the terminal's modem is called the **originate** modem. It transmits a 1,070-hertz tone for a 0 and a 1,270-hertz tone for a 1. The computer's modem is called the **answer** modem, and it transmits a 2,025-hertz tone for a 0 and a 2,225-hertz tone for a 1. Because the originate and answer modems transmit different tones, they can use the line simultaneously. This is known as **full-duplex** operation. Modems that can transmit in only one direction at a time are known as **half-duplex** modems, and they are rare.

Let's say that two 300bps modems are connected, and the user at the terminal types the letter "a." The ASCII code for this letter is 97 decimal or 01100001 binary. A device inside the terminal called a UART (universal asynchronous receiver/transmitter) converts the byte into its bits and sends them out one at a time through the terminal's **RS-232 port** (also known as a **serial port**). The terminal's modem is connected to the RS-232 port, so it receives the bits one at a time and its job is to send them over the phone line.

Faster Modems

In order to create faster modems, modem designers had to use techniques far more sophisticated than frequency-shift keying. First they moved to **phase-shift keying** (PSK), and then **quadrature amplitude modulation** (QAM). These techniques allow an incredible amount of information to be crammed into the 3,000 hertz of bandwidth available on a normal voice-grade phone line. 56K modems, which actually connect at something like 48 Kbps on anything but absolutely perfect lines, are about the limit of these techniques.

All of these high-speed modems incorporate a concept of **gradual degradation**, meaning they can test the phone line and fall back to slower speeds if the line cannot handle the modem's fastest speed.

The next step in the evolution of the modem was **asymmetric digital subscriber line** (ADSL) modems. The word *asymmetric* is used because these modems send data faster in one direction than they do in another. An ADSL modem takes advantage of the fact that any normal home, apartment or office has a **dedicated copper wire** running between it and phone company's nearest mux or central office. This dedicated copper wire can carry far more data than the 3,000-hertz signal needed for your phone's voice channel. If both the phone company's central office and your house are equipped with an ADSL modem on your line, then the section of copper wire between your house and the phone company can act as a purely digital high-speed transmission channel. The capacity is something like 1 million bits per second (Mbps) between the home and the phone company (*upstream*) and 8 Mbps between the phone company and the home (*downstream*) under ideal conditions. The same line can transmit both a phone conversation *and* the digital data.

The approach an ADSL modem takes is very simple in principle. The phone line's bandwidth between 24,000 hertz and 1,100,000 hertz is divided into 4,000-hertz bands, and a **virtual modem** is assigned to each band. Each of these 249 virtual modems tests its band and does the best it can with the slice of bandwidth it is allocated. The aggregate of the 249 virtual modems is the total speed of the pipe.

POINT TO POINT PROTOCOL

Today, no one uses dumb terminals or terminal emulators to connect to an individual computer. Instead, we use our modems to connect to an **Internet service provider** (ISP), and the ISP connects us into the Internet. The Internet lets us connect to any machine in the world. Because of the relationship between your computer, the ISP and the Internet, it is no longer appropriate to send individual characters. Instead, your modem is routing TCP/IP packets between you and your ISP.

The standard technique for routing these packets through your modem is called the **Point-to-Point Protocol** (**PPP**). The basic idea is simple -- your computer's TCP/IP stack forms its TCP/IP datagrams normally, but then the datagrams are handed to the modem for transmission. The ISP receives each datagram and routes it appropriately onto the Internet. The same process occurs to get data from the ISP to your computer.

MODEMS AND ROUTERS

Modems and **routers** are both involved in connecting your home PCs to the Internet. The modem encodes and decodes data so that it can pass between your home network and your Internet Service Provider (ISP). The router, on the other hand, directs the information collected by the modem to devices within that network. The modem brings the information in, and the router distributes (or “routes”) it to different devices like computers and phones.

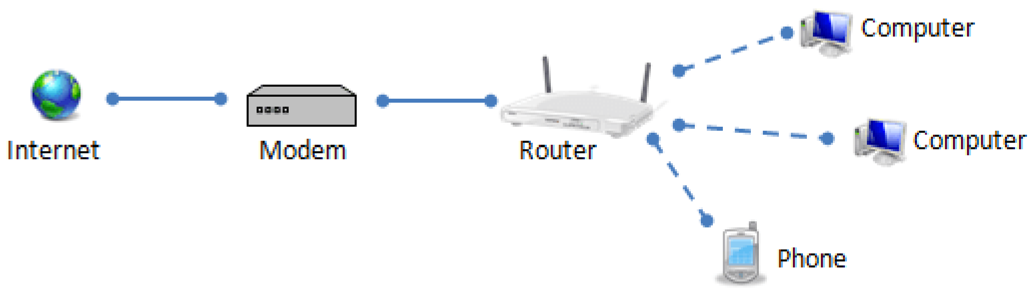
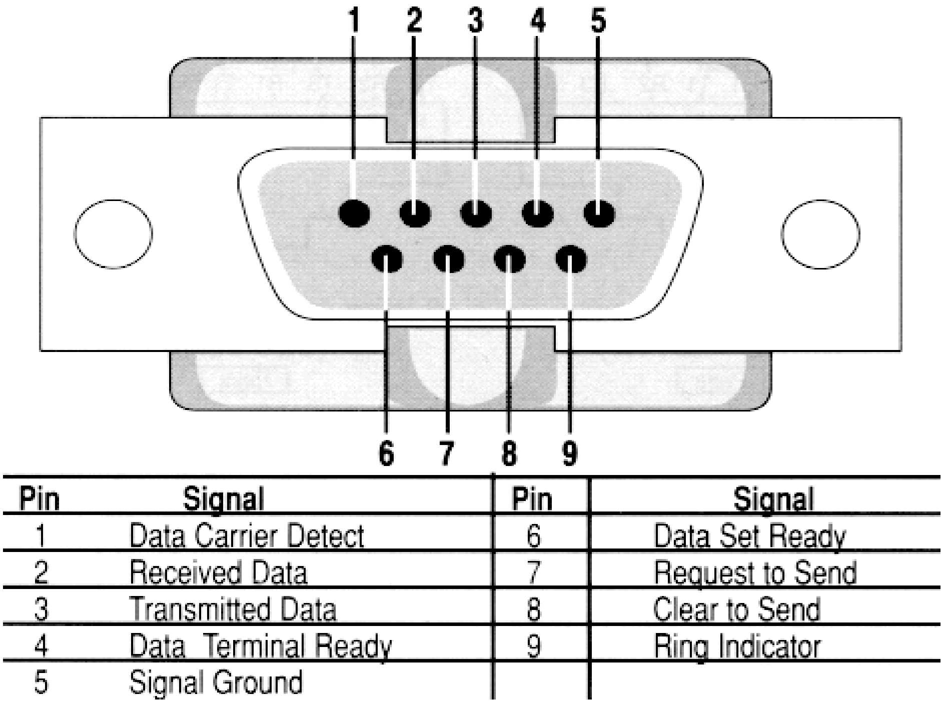


Fig3: Modem and Router

RS 232 Standard

And it’s Variants Including RS232C, RS232D, V24, V28 and V10.

The RS-232 serial interface communications standard has been in use for very many years and is one of the most widely used standards for serial data communications because it is simple and reliable. The RS232 serial interface standard still retains its popularity and remains in widespread use. It is still found on some computers and on many interfaces, often being used for applications ranging from data acquisition to supplying a serial data communications facility in general computer environments. The long term widespread use of the RS232 standard has meant that products are both cheap and freely available, and in these days of new higher speed standards, the reliable, robust RS232 standard still has much to offer. The following figures shows RS232 connections

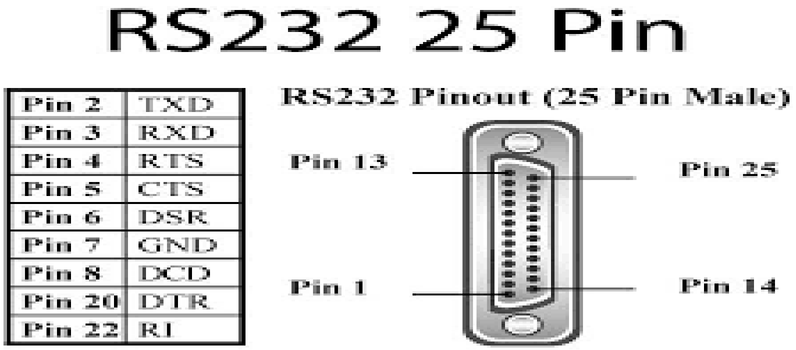


Pin	Signal	Pin	Signal
1	Data Carrier Detect	6	Data Set Ready
2	Received Data	7	Request to Send
3	Transmitted Data	8	Clear to Send
4	Data Terminal Ready	9	Ring Indicator
5	Signal Ground		

Fig 4: RS 232 Interface (9 pins)

Fig 5: RS232C diagram

Fig 6: RS232



RS232 25 Pin

Pin 2	TXD
Pin 3	RXD
Pin 4	RTS
Pin 5	CTS
Pin 6	DSR
Pin 7	GND
Pin 8	DCD
Pin 20	DTR
Pin 22	RI

RS232 Pinout (25 Pin Male)

Fig 7: RS232; (25 Pins)

RS-232 Interface Basics

The interface is intended to operate over distances of up to 15 metres. This is because any modem is likely to be near the terminal. Data rates are also limited with a maximum of 19.2 Kbits per second for RS-232C. However slower rates are often used. In theory it is possible to use any baud rate, but there are a number of standard transmission speeds in bps used as follows: 50, 75, 110, 150, 300, 600, 1200, 2400, 4800, 9600, 19200, 38400, 76800 bps.

Note that speeds up to 19200 bits per second are normally used. Above this speed, noise that is picked up, especially over long cable runs can introduce data errors. Where high speeds and long data runs are required then standards such as RS422 may be used.

RS-232 connections

The RS-232C specification does not include a description of the connector to be used. However, the most common type found is the 25 pin D-type connector.

RS232 signal levels

The voltage levels are one of the main items in the specification. For RS232 data signals a voltage of between -3V and -25V represents a logic 1. The logic 0 is represented by a voltage of between +3V and +25V. Control signals are in the "ON" state if their voltage is between +3V and +25V and "OFF" if they are negative, i.e. between -3V and -25V.

The data is sent serially on RS232, each bit is sent one after the next because there is only one data line in each direction. This mode of data transmission also requires that the receiver knows when the actual data bits are arriving so that it can synchronise itself to the incoming data. To achieve this, a logic 0 is sent as a start bit for the synchronisation. This is followed by the data itself and there are normally seven or eight bits. The receiver obviously has to know how many data bits to expect, and there are often small dual in line switches either on the back of the equipment or inside it to set this information.

Data on RS232 is normally sent using ASCII (American Standard Code for Information Interchange). However other codes including the Murray Code or EBCDIC (Extended Binary Coded Decimal Interchange Code) can be used equally well. After the data itself, a parity bit is sent. Again this requires setting because it is optional and it can be even or odd parity. This is used to check the correctness of the received data and it can indicate whether the data has an odd or even number of logic ones. However there is no facility for error correction unlike what is the case for many systems these days. Finally a stop bit is sent. This is normally one bit long and is used to signify the end of a particular byte. Sometimes two stop bits are required and again this is an option that can often be set on the equipment.

RS232 data transmission is normally asynchronous. However transmit and receive speeds must obviously be the same. A certain degree of tolerance is allowed. Once the start bit is sent the receiver will sample the centre of each bit to see the level. Within each data word the synchronisation must not differ by more than half a bit length otherwise the incorrect data will be seen. Fortunately this is very easy to achieve with today's accurate bit or baud rate generators.

Lines and their usage

There are four types of lines defined in the RS232 specification. They are Data, Control, Timing and Ground. Not all of them are required all the time. It is possible to set up a very simple communication using very few lines. When looking at the lines and their functions it is necessary to remember that they are defined for a connection between a modem (the data set or communications equipment) and a terminal or computer (data terminal equipment) in mind. All the lines have directions, and when used in this way a one to one cable operates correctly.

The most obvious lines are the data lines. There are two of these, one for data travelling in each direction. Transmit data is carried on pin 2 and the receive data is carried on line three (see Figure 7). The most basic of the control circuits is Data Carrier Detected (DCD). This shows when the modem has detected a carrier on the telephone line and a connection appears to have been made. It produces a high, which is maintained until the connection is lost.

Data Terminal Ready (DTR) and Data Set Ready (DSR) are the main control circuits. They convey the main information between the terminal and modem. When the terminal is ready to start handling data it flags this on the DTR line. If the modem is also ready then it returns its signal on the DSR line. These circuits are mainly used for telephone circuits. After a connection has been made the modem will be connected to the line by using DTR. This connection will remain until the terminal is switched off line when the DTR line is dropped. The modem will detect this and release the telephone line.

Sometimes pin 20 is not assigned to DTR. Instead another signal named, Connect Data Set To Line (CDSTL) is used. This is virtually the same as DTR, but differs in that DTR merely enables the modem to be switched onto the telephone line. CDSTL commands the modem to switch, despite anything else it may be doing. A further two circuits, Request To Send (RTS) and Clear To Send (CTS) are also used. These pair of circuits are used together. The terminal equipment will flag that it has data to send. The modem will then return the CTS signal to give the all clear after a short delay.

This signalling is used particularly when switched carriers are used. It means that the carrier is only present on the line when there is data to send. It finds its uses when one central modem is servicing several others at remote locations.

Secondary lines

There are two types of lines that are specified in the RS-232 specification. There are the primary channels that are normally used, and operate at the normal or higher data rates. However, there is also provision for a secondary channel for providing control information. If it is used it will usually send data at a much slower rate than the primary channel. As the secondary lines are rarely used or even implemented on equipment, manufacturers often use these connector pins for other purposes. In view of this it is worth checking that the lines are not being used for other purposes before considering using them. When the secondary system is in use, the handshaking signals operate in the same way as for the primary circuit.

Grounding

The ground connections are also important. There are two. First the protective ground ensures that both equipment are at the same earth potential. This is very useful when there is a possibility that either equipment is not earthed. The signal ground is used as the return for the digital signals travelling along the data link. It is important that large currents that are not part of the signalling do not flow along this line otherwise data errors may occur.

The RS-232 specification is still widely used. Although faster specifications exist, it is likely to remain in use for many years to come. One of the reasons for this is the fact that it is found on most of today's personal computers. Although the parallel "LPT" ports are used almost universally for printers, it still used for many other purposes, including connecting the computer to a modem.

The RS 232 standard has been used in many areas, well beyond its original intended applications. As a result, this has led to uncertainty in the way some applications use the RS232 standard. However the RS 232 standard operates very reliably when correctly set up and for many years it has provided one of the main forms of serial data transmission. Even though many other standards are available for data transmission these days, the RS 232 standard is still widely used, and is likely to remain so for many years to come.

Development of the RS 232 standard

The RS 232 standard for data communications was devised in 1962 when the need to be able to transmit data along a variety of types of lines started to grow. The idea for a standard had grown out of the realisation in the USA that a common approach was required to allow interoperability. As a result the Electrical Industries Association in the USA created a standard for serial data transfer or communication known as RS232. It defined the electrical characteristics for transmission of data between a Data Terminal Equipment (DTE) and the Data Communications Equipment (DCE). Normally the data communications equipment is the modem (modulator/demodulator) which encodes the data into a form that can be transferred along the telephone line. A Data Terminal Equipment could be a computer.

The RS 232 standard underwent several revisions, the C issue known as RS232C was issued in 1969 to accommodate the electrical characteristics of the terminals and devices that were being used at the time. The RS 232 standard underwent further revisions and in 1986 Revision D was released (often referred to as RS232D). This revision of the RS 232 standard was required to incorporate various timing elements and to ensure that the RS 232 standard harmonised with the CCITT standard V.24, while still ensuring interoperability with older versions of RS 232 standard. Further updates and revisions have occurred since then and a newer version is TIA-232-F issued in 1997 under the title: "Interface Between Data Terminal Equipment and Data Circuit-Terminating Equipment Employing Serial Binary Data Interchange." The name of the RS 232 standard has changed during its history, several times as a result of the sponsoring organisation. As a result it has variously been known as EIA RS-232, EIA 232, and most recently as TIA 232.

Variations of the RS 232 standard

There are number of different specifications and standards that relate to RS 232. The RS 232 standard is often referred to by the other related standards and in particular V.24 which is the ITU / CCITT designation for the serial data communications standard. A description of some of the RS 232 standards and the various names and references used is given below:

- ***EIA/TIA-232:*** This reference to the RS 232 standard includes the names of the first and current sponsoring organisations, the Electronic Industries Alliance (EIA) the Telecommunications Industry Alliance (TIA).

- ***RS-232C:*** This was the designation given to the release of RS 232 standard updated in 1969 to incorporate many of the device characteristics.
- ***RS-232D:*** This was the release of the RS 232 standard that occurred in 1986. It was revised to incorporate various timing elements and to ensure that the RS 232 standard harmonised with the CCITT standard V.24.
- ***RS-232F:*** This version of the RS 232 standard was released in 1997 to accommodate further revisions to the standard. It is also known as TIA-232-F.
- ***V24:*** The International Telecommunications Union (ITU) / CCITT (International Telegraph and Telephone Consultative Committee) of the ITU developed a standard known as ITU v.24, often just written as V24. This standard is compatible with RS232, and its aim was to enable manufacturers to conform to global standards and thereby allow products that would work in all countries around the world. It is entitled "List of definitions for interchange circuits between data terminal equipment (DTE) and data circuit-terminating equipment (DCE)."
- ***V28:*** V.28 is an ITU standard defining the electrical characteristics for unbalanced double current interchange circuits, i.e. a list of definitions for interchange circuits between data terminal equipment (DTE) and data circuit-terminating equipment (DCE).
- ***V10:*** V.10 is an ITU standard or recommendation for unbalanced data communications circuits for data rates up to 100 kbps that was first released in 1976. It can inter-work with V.28 provided that the signals do not exceed 12 volts. Using a 37 pin ISO 4902 connector it is actually compatible with RS423.

RS-232 Applications

The RS-232 standard has come a long way since its initial release in 1962. Since then the standard has seen a number of revisions, but more importantly, RS232 has been used in an ever increasing number of applications. Originally it was devised as a method of connecting telephone modems to teleprinters or teletypes. This enabled messages to be sent along telephone lines - the use of computers was still some way off.

As computers started to be used, links to printers were required. The RS-232 standard provided an ideal method of connection and therefore it started to be used in a rather different way. However its use really started to take off when personal computers were first introduced. Here the RS-232 standard provided an ideal method of linking the PC to the printer.

The RS-232 standard provided an ideal method of linking many other remote items to computers and data recorders. As a result, RS-232 became an industry standard, used in a host of applications that were never conceived when it was first launched in 1962.

The RS 232 standard is very widely used and is probably the most widely used standard for serial data communications over distances. The RS 232 standard has stood the test of time, and being introduced in 1962 it has been in use for well over 45 years.

Serial Port for PC

Standard PC serial ports come in to versions: 9 pin and 25 pin one The functions of those both version are exactly the same, only different kind of connectors and different pinout PC serial port is nowadays usually used for interfacing PC to modem or mouse. Original PC serial port was designed to operate up to 19.2 kbit/s (maximum speed defined in RS-232C standard) but nowadays they can typically go up to 115.2 kbit/s (some special cards can do even faster than that). PC serial port sends and receives data in serial format. In serial, asynchronous data transfer the individual bits which comprise each data byte are sent one after the other over a single line. In this context, asynchronous means that the clock information is not included with the transmission, so that frequent re-synchronization using start/stop bits is required.

The maximum length specified by RS-232 is only 50 feet (around 15 meters), however much longer lengths are possible with proper shielding on the cable. Generally you can run 9600 bps communication up to 250 feet (80 meters) over shielded data cable or unshielded twisted pair cable in good environment. When using shielded cable and slower data rate longer lengths are possible (up to hundreds of meters in good conditions)

RS449 Basics, Interface and Pinout

The basics of RS449 data communications standard, what it is, the RS449 pinout and the RS 449 interface.

The RS449 or RS-449 interface is a further enhancement of RS232 and RS423. It is aimed at catering for very fast serial data communications at speeds up to 2 Mbps. In order to achieve this RS449 makes some changes when compared to RS232 to the way in which the signals are referenced, while still being able to retain some compatibility with RS232.

The RS499 standard which has now been discontinued is also known by the references EAI-449, TIA-449 and ISO 4902

RS449 interface

One of the ways in which the RS449 data communications standard is able to send at high speeds without stray noise causing interference is to use a differential form of signalling. Earlier data communications standards such as RS232 used signalling that was referenced to earth and while this was easier to implement and cheaper to cable, it introduced limitations into the system.

By using twisted wire pairs for the data lines, any unwanted noise will be picked up by both wires together. As the RS449 receivers use a differential input, and they are not referenced to ground, any noise that is picked up does not affect the input. This means that higher levels of noise can be tolerated without any degradation to the performance to the data communications system.

For the RS449 interface, ten additional circuits functions have been provided when compared to RS232. Additionally three of the original interchange circuits have been abandoned.

In order to minimise any confusion that could easily occur, the circuit abbreviations have been changed. In addition to this the RS449 interface requires the use of 37 way D-type connectors and 9 way D-type connectors, the latter being necessary when use is made of the secondary channel interchange circuits.

RS449 Primary connector pinout and interface connections

The RS449 primary connector, which is used the one that is used as standard uses a 37 way D-type connector. The pinout and connections are given in the table below:

Pin	Signal Name	Description
1		Shield
2	SI	Signal Rate Indicator
3	n/a	Unused
4	SD-	Send Data (A)
5	ST-	Send Timing (A)
6	RD-	Receive Data (A)
7	RS-	Request To Send (A)
8	RT-	Receive Timing (A)
9	CS-	Clear To Send (A)
10	LL	Local Loopback
11	DM-	Data Mode (A)
12	TR-	Terminal Ready (A)
13	RR-	Receiver Ready (A)
14	RL	Remote Loopback
15	IC	Incoming Call
16	SF/SR+	Signal Freq./Sig. Rate Select.
17	TT-	Terminal Timing (A)
18	TM-	Test Mode (A)
19	SG	Signal Ground
20	RC	Receive Common
21	n/a	Unused
22	SD+	Send Data (B)
23	ST+	Send Timing (B)
24	RD+	Receive Data (B)
25	RS+	Request To Send (B)
26	RT+	Receive Timing (B)
27	CS+	Clear To Send (B)
28	IS	Terminal In Service
29	DM+	Data Mode (B)
30	TR+	Terminal Ready (B)
31	RR+	Receiver Ready (B)
32	SS	Select Standby
33	SQ	Signal Quality
34	NS	New Signal
35	TT+	Terminal Timing (B)
36	SB	Standby Indicator
37	SC	Send Common

RS449 primary connector pinout and connections

Within the RS449 interface a number of differential connections are defined. In the pinout table above they are labelled as either "A and B" or "+" and "-". When setting up a connection, it is necessary to ensure that the correct polarities are used. As twisted pairs are used for the A and B connections, it is often possible to mix them. If this happens the interface will not work.

RS449 Auxilliary connector

A second connector is defined for use when the secondary channel interchange circuits are needed.. This connector uses a 9 way D-type connector.

Pin	Signal Name	Description
1		Shield
2	SRR	Secondary Receive Ready
3	SSD	Secondary Send Data
4	SRD	Secondary Receive Data
5	SG	Signal Ground
6	RC	Receive Common
7	SRS	Secondary Request to Send
8	SCS	Secondary Clear to Send
9	SC	Send Common

RS449 secondary connector

The RS449 data communications interface is an interface standard that is able to provide data communications with speeds of up to 2 Mbps. Retaining some similarities to RS232, it is a more comprehensive interface capable of greater speeds and operation with greater levels of data integrity.

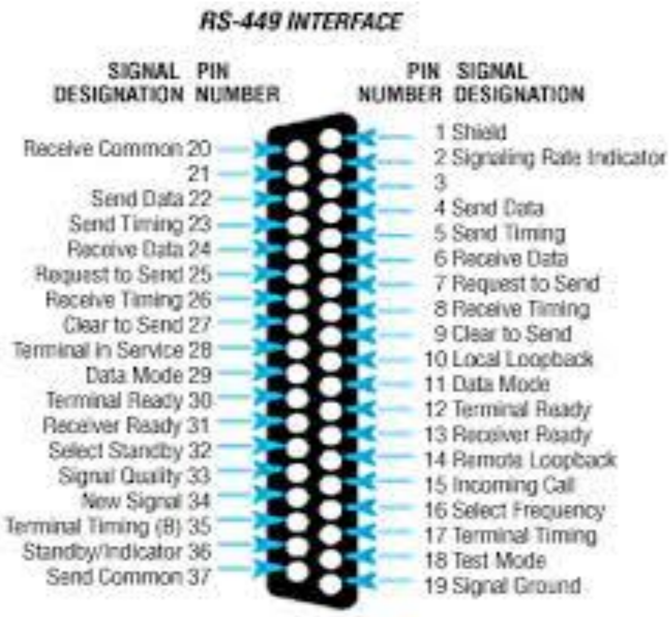


Fig: RS 449

LECTURE TWO: Channel Coding and Error Control:

Forward Error Control; Error Detection Methods; Parity Checking; Linear Block Codes, Cyclic Redundancy Checking; Feedback Error Control.

INTRODUCTION

The main aim of any communication schemes is to provide error-free data transmission. In a communication system, information can be transmitted by analog or digital signals. For analog means, the amplitude of the signal reflects the information of the source, whereas for digital case, the information will first be translated into a stream of ‘0’ and ‘1’. Then two different signals will be used to represent ‘0’ and ‘1’ respectively. The main advantage of using digital signal communication is that errors introduced by noise during the transmission can be detected and possibly corrected. For communication using cables, the random motion of charges in conducting (e.g. resistors), known as thermal noise, is the major source of noise. For wireless communication channels, noise can be introduced in various ways. In the case of mobile phones, noise also includes the signals sent by other mobile phone users in the system. Figure 1and Figure 2 show the flow of a simple digital communication system.

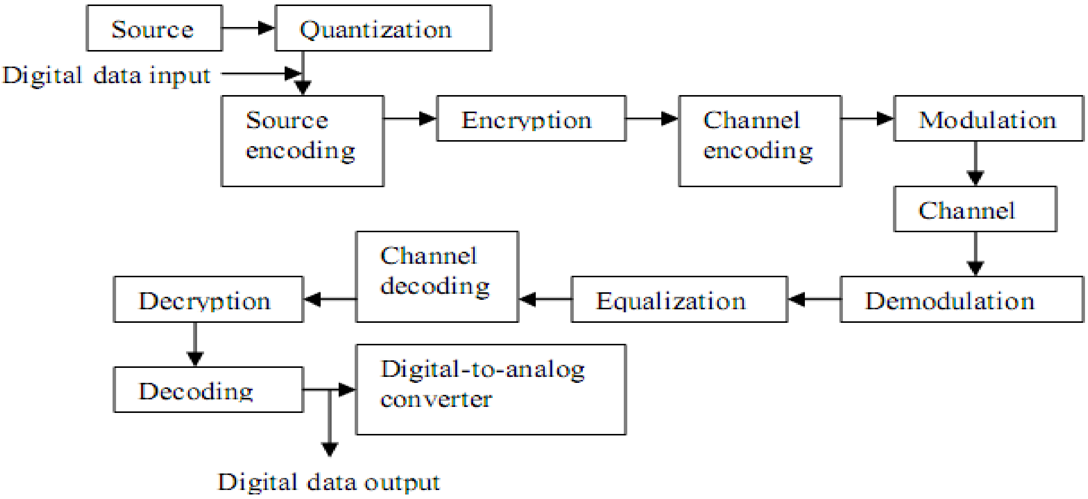


Fig 1: The flow diagram of a simple digital communication system.

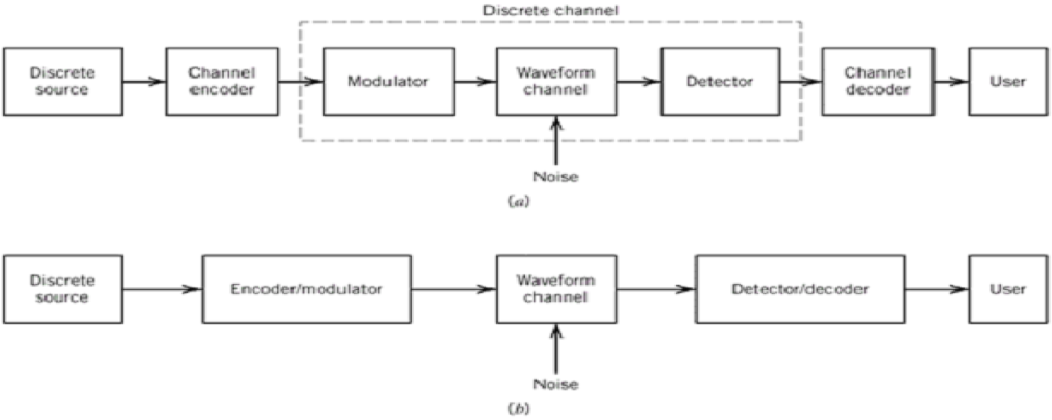
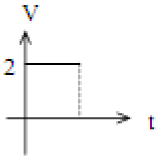
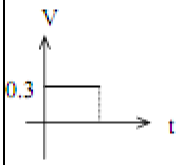
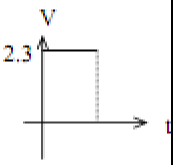
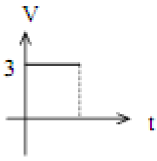
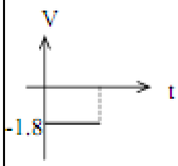
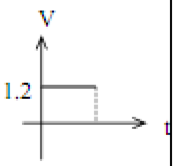


Fig 2: The flow diagram of a simple digital communication system showing noise addition.

Table 1 shows the difference between signal detection for analogue and digital systems. From the Table, it can be seen that error detection and correction as well as better accuracy of transmitted information are possible for digital communication due to encoding and decoding.

Table 1: Difference between signal detection for analogue and digital systems.

Type	Signal transmitted	Channel	Signal received	Information Detected
Analog: All amplitudes are possible. The value represents the information (e.g. loudness of the voice)	Information: 2V  	Noise: 0.3V (The mean of noise is 0V)  	2.3V  	2.3V
Digital: Only two amplitudes (±3V) will be transmitted to represent '1' and '0' respectively	Information: '2' which has been encoded as '1' and will be using 3V pass through the channel.  	Noise: -1.8V (The mean of noise is 0V)  	1.2V  	Since 1.2V > 0V, so it is still detected as 3V. That means the code '1' is transmitted and therefore the information is '2'.

The fundamental resources at the disposal of a communications engineer are signal power, time and bandwidth. For a given communications environment, these three resources can be traded against each other. A general objective, however, is often to achieve maximum data transfer, in a minimum bandwidth *while maintaining an acceptable quality of transmission.* The quality of transmission, in the context of digital communications, is essentially concerned with the probability of bit error, *Pe,* at the receiver.

The Shannon-Hartley law shown in equation 1 for the capacity of a communications channel demonstrates two things.

(i) Firstly it shows (quantitatively) how bandwidth (B) and signal power (S/N) may be traded in an ideal system,

(ii) Secondly it gives a theoretical limit for the transmission rate of (reliable, i.e. error free) data (R) from a transmitter of given power, over a channel with a given bandwidth, operating in a given noise environment.

$$R_{\max} = B \log_2\left(1 + \frac{S}{N}\right) \quad \text{bit/s}$$

……………………………………………….1

In order to realize this theoretical limit, however, an appropriate coding scheme (which the Shannon-Hartley law assures us exists) must be found.

Coding in Engineering

There are basically two types of coding used in Engineering: (i) Source Coding and Channel coding. Irrespective of the type of coding used, it is generally aimed at achieving the following:

(i) To encrypt information for security purposes (Encryption)

(ii) To reduce space for the data stream (Data Compression)

(iii) To change the form of representation of the information so that it can be transmitted over a communication channel.

(iv) To encode a signal so that any error that occurs during transmission can be detected and possibly corrected.

In practice, the objective of the design engineer is to realize the required data rate (often determined by the service being provided) within the bandwidth constraint of the available channel and the power constraint of the particular application. For a fixed S*/N*, the only practical option available for changing data quality from problematic to acceptable is to use *error-control coding.* Another practical motivation for the use of coding is to reduce the required S*/N* for a fixed bit error rate. This reduction in S*/N* may, in turn, be exploited to reduce the required transmitted power or reduce the hardware costs by requiring a smaller antenna size in the case of radio communications.

Moreover, the use of error-control coding adds *complexity* to the system, especially for the implementation of decoding operations in the receiver. Thus, the design trade-offs in the use of error-control coding to achieve acceptable error performance include considerations of bandwidth and system complexity.

Bit Error rates (BER) can be made smaller by the following methods:

1. By increasing transmitter power but this may not always be desirable, for example in man-portable systems where the required extra battery weight may be unacceptable.
2. Use of diversity which is effective against burst errors caused by signal fading. There are three main types of diversity: space diversity, frequency diversity, and time diversity. All these schemes incorporate redundancy in that data is, effectively, transmitted twice: i.e. via two paths, at two frequencies, or at two different times. In space diversity two or more antennas are used which are sited sufficiently far apart for fading at their outputs to be de-correlated. Frequency diversity employs two different frequencies to transmit the same information. (Frequency diversity can be in-band or out-band depending upon the frequency spacing between the carriers.) In time diversity systems, the same message is transmitted more than once at different times.
3. By introducing full deplex transmission, implying simultaneous 2-way transmission. Here when a transmitter sends information to a receiver, the information is 'echoed' back to the transmitter on a separate feedback channel. Information echoed back which contains errors can then be retransmitted. This technique requires twice the bandwidth of single direction (simplex) transmission, which may be unacceptable in terms of spectrum utilization.
4. A fourth method for coping with poor BER is automatic repeat request (ARQ). Here a simple error *detecting* code is used and, if an error is detected in a given data block, and then a request is sent via a feedback channel to retransmit that block. ARQ is very effective, for example in facsimile transmission. On long links with fast transmission rates, however, such as is typical in satellite communications, ARQ can be very difficult to implement.
5. The fifth technique for coping with high BER is to employ forward error correction coding (FECC). In common with three of the other four techniques FECC introduces redundancy, this time with data check bits interleaved with the information traffic bits. It relies on the number of errors in a long block of data being close to the statistical average and, being a forward technique, requires no return channel. The widespread adoption of FECC was delayed, historically, because of its complexity and high cost of implementation relative to the other possible solutions. Complexity is

now less of a problem following the proliferation of VLSI custom coder/decoder chips.

Source Coding

Suppose a word ‘Zebra’ is going to be sent out. Before this information can be transmitted to the channel, it is first translated into a stream of bits (‘0’ and ‘1’). The process is called source coding. There are many commonly used ways to translate that. For example, if ASCII code is used, each alphabet will be represented by 7-bit so called the code word. The alphabets ‘Z’, ‘e’, ‘b’, ‘r’, ‘a’, will be encoded as:
‘1010101’, ‘0110110’, ‘0010110’, ‘0010111’, ‘0001110’
The **ASCII code** is an example of fixed-length code, because each of the code word is of the same length (7 bits). However, in the view of efficient communication, the occurrence of ‘Z’ is not as often as that of ‘e’ and ‘a’. If there is a way of encoding information such that the alphabets with higher probability of occurrence are assigned with shorter code words, and longer for the other letters which seldom come out, then on the whole it may be able to conserve the number of bits to be sent to the channel while sending the same information. This is what the variable length code can do. Example of this type os codes is the **Huffman Codes**.

Channel Coding

As already mentioned, error control coding is a method to detect and possibly correct errors by introducing redundancy to the stream of bits to be sent to the channel. The design goal of channel coding (also referred as Error Control Coding) is basically to increase the resistance of a digital communication system to a channel noise. Error control coding is used to detect and often correct symbols which are received in error. The two main methods of error control are: Automatic repeat request and Forward error control techniques.

Figure 3 shows different types of error control codes (or channel coding methods).

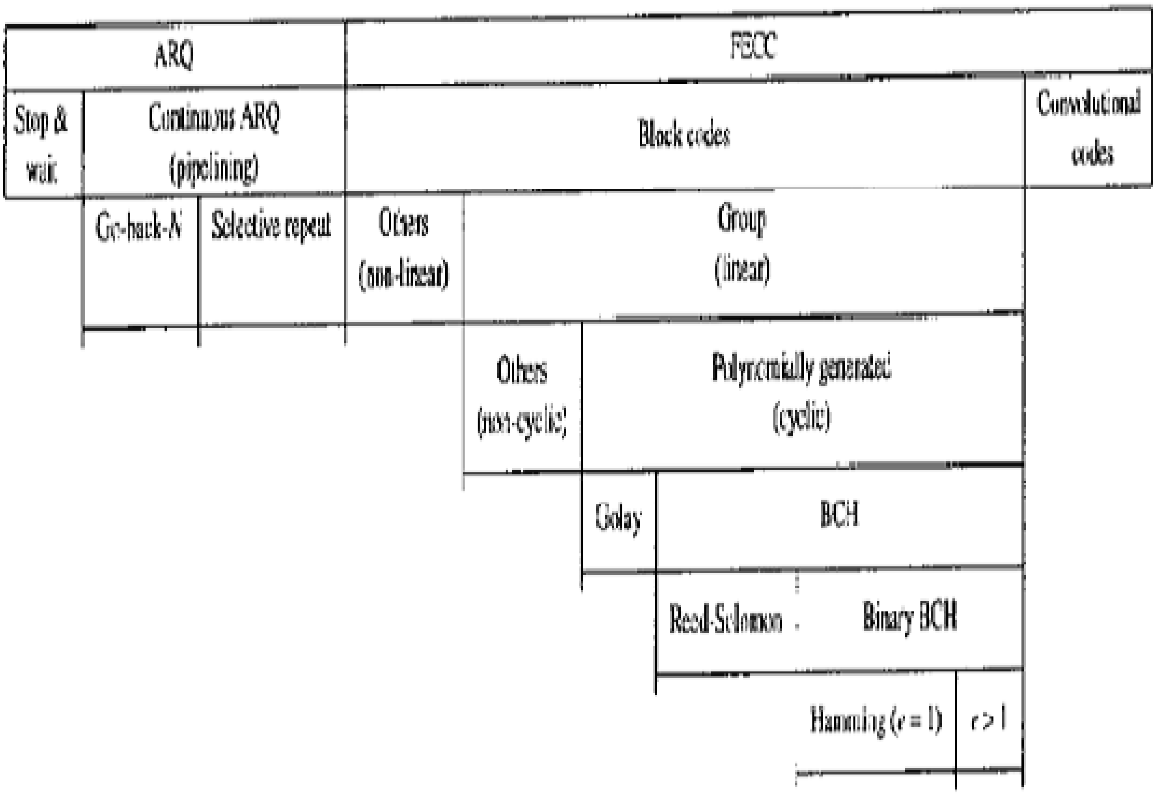
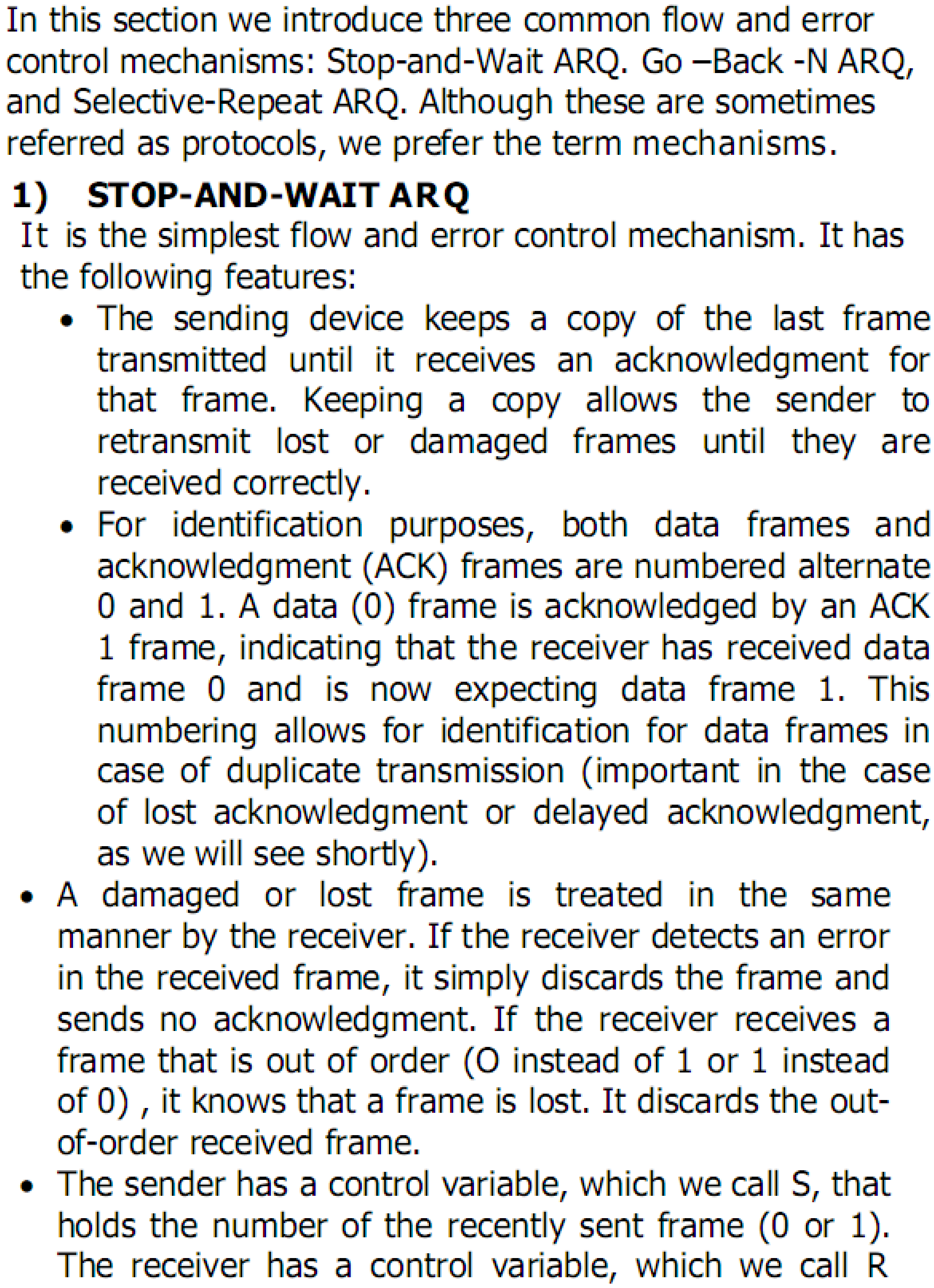


Figure 3: Different types of error control codes (or channel coding methods).

Automatic Repeat Request (ARQ).

In this method when a receiver circuit detects errors in a block of data, it request that the data is retransmitted.

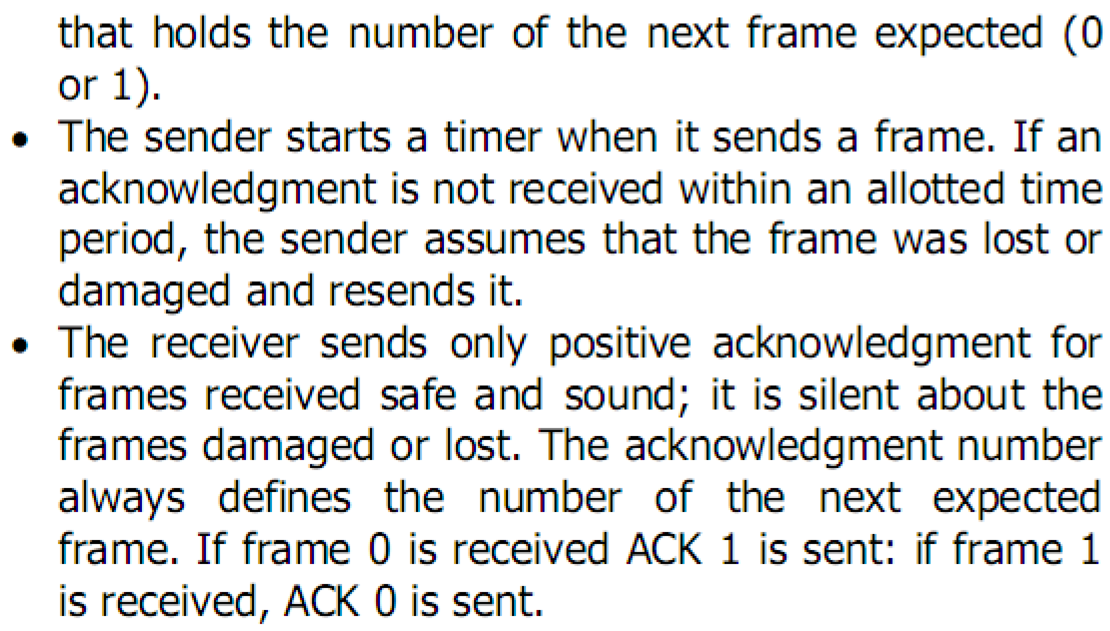


In this section we introduce three common flow and error control mechanisms: Stop-and-Wait ARQ. Go –Back -N ARQ, and Selective-Repeat ARQ. Although these are sometimes referred as protocols, we prefer the term mechanisms.

1) STOP-AND-WAIT ARQ

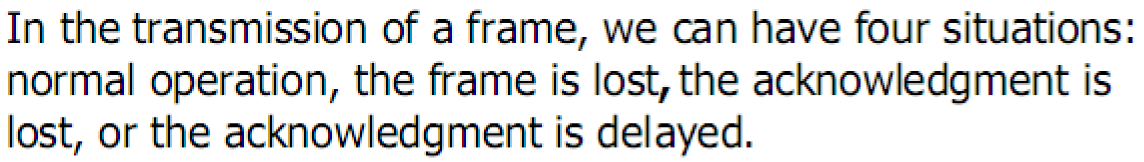
It is the simplest flow and error control mechanism. It has the following features:

- The sending device keeps a copy of the last frame transmitted until it receives an acknowledgment for that frame. Keeping a copy allows the sender to retransmit lost or damaged frames until they are received correctly.
- For identification purposes, both data frames and acknowledgment (ACK) frames are numbered alternate 0 and 1. A data (0) frame is acknowledged by an ACK 1 frame, indicating that the receiver has received data frame 0 and is now expecting data frame 1. This numbering allows for identification for data frames in case of duplicate transmission (important in the case of lost acknowledgment or delayed acknowledgment, as we will see shortly).
- A damaged or lost frame is treated in the same manner by the receiver. If the receiver detects an error in the received frame, it simply discards the frame and sends no acknowledgment. If the receiver receives a frame that is out of order (O instead of 1 or 1 instead of 0) , it knows that a frame is lost. It discards the out-of-order received frame.
- The sender has a control variable, which we call S, that holds the number of the recently sent frame (0 or 1). The receiver has a control variable, which we call R



that holds the number of the next frame expected (0 or 1).

- The sender starts a timer when it sends a frame. If an acknowledgment is not received within an allotted time period, the sender assumes that the frame was lost or damaged and resends it.
- The receiver sends only positive acknowledgment for frames received safe and sound; it is silent about the frames damaged or lost. The acknowledgment number always defines the number of the next expected frame. If frame 0 is received ACK 1 is sent: if frame 1 is received, ACK 0 is sent.



In the transmission of a frame, we can have four situations: normal operation, the frame is lost, the acknowledgment is lost, or the acknowledgment is delayed.

CASE 1: Normal Operation In a normal transmission, the sender sends frame 0 and waits to receive ACK 1. When ACK 1 is received, it sends frame 1 and then waits to receive ACK 0, and so on. The ACK must be received before the timer set for each frame expires. Figure 4 shows successful frame transmissions.

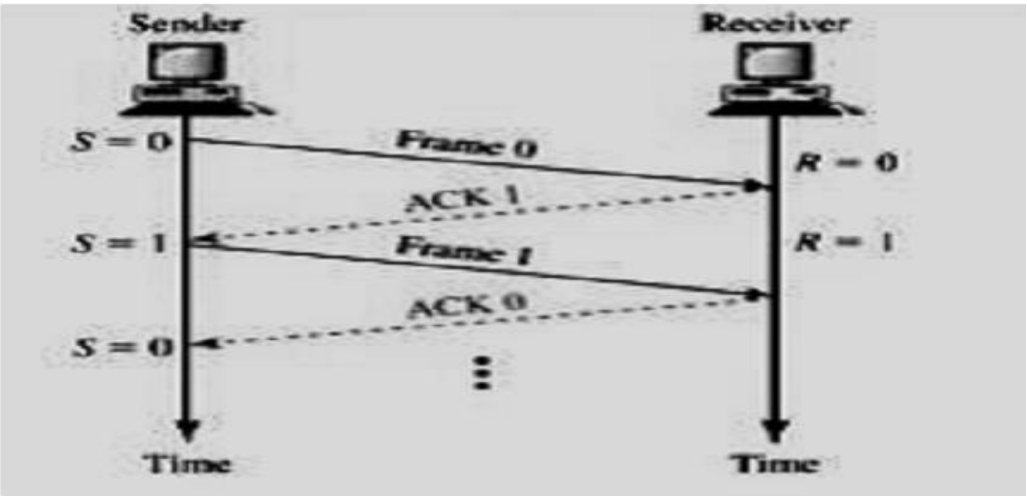


Fig 4

CASE 2: Lost or Damaged Frame A lost or damaged frame is handled in the same way by the receiver; when the receiver receives a damaged frame, it discards it, which essentially means the frame is lost. The receiver remains silent about a lost frame and keeps its value of *R*. For example, in Fig 5, the sender transmits frame 1, but it is lost. The receiver does nothing, retaining the value of R (1) (. After the timer at the sender site expires, another copy of frame 1 is sent.

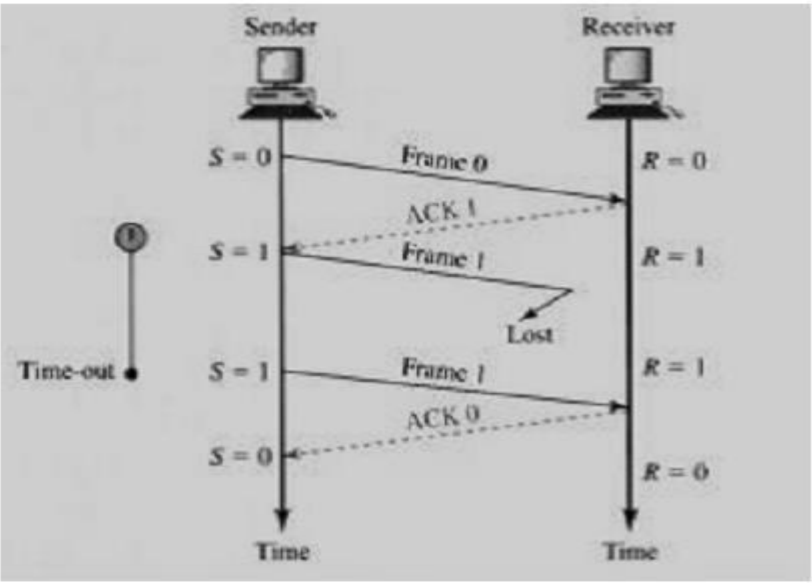


Figure 5

CASE 3* : *Lost Acknowledgment A lost or damaged acknowledgment is handled in the same way by the sender; if the sender receives a damaged acknowledgment, it discards it Figure 6 shows a lost ACK 0. The waiting sender does not know if frame 1 has been received. When the timer for frame 1 expires the sender retransmits frame 1. Note that the receiver has already received frame 1 and it was expecting to receive frame 0 (*R*= 0). Therefore, it silently discards the second copy of frame 1.

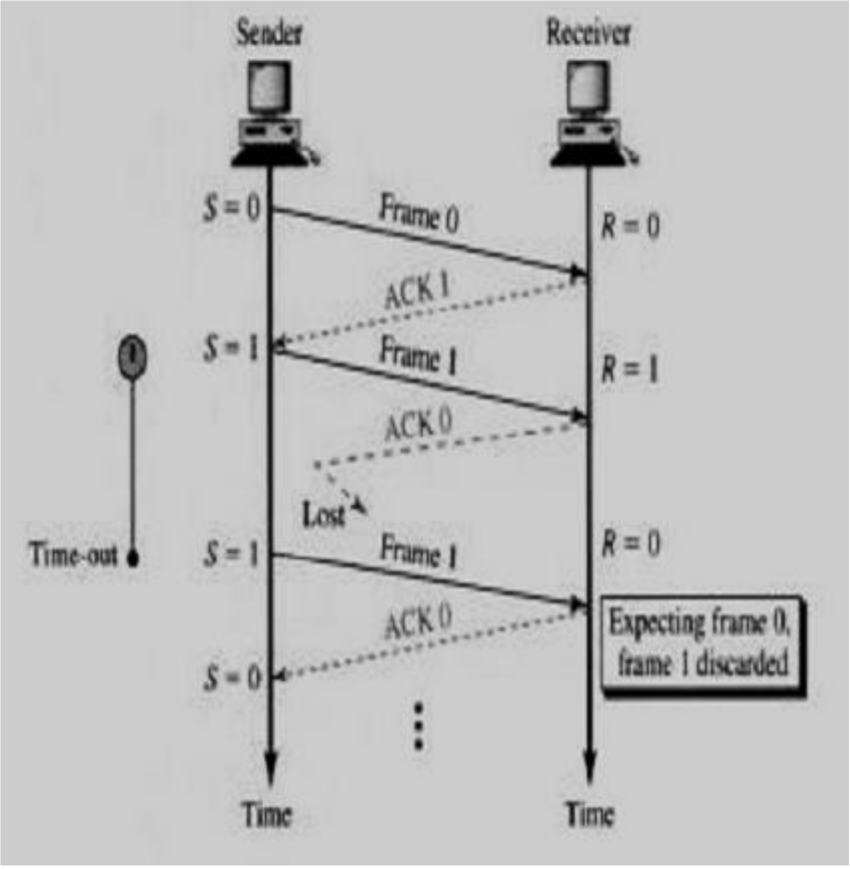


Figure 6

CASE 4: Delayed Acknowledgment Another problem that may occur is delayed acknowledgment. An acknowledgment can be delayed at the receiver or by some problem with the link. Figure 7 shows the delay of ACK 1; it is received after the timer for frame 0 has already expired. The sender has already retransmitted a copy of frame 0. However, the value of *R* at the receiver site is still 1, which means that the receiver expects to see frame 1, the receiver, therefore, discards the duplicate frame 0. The sender has now received two ACKs, one that was delayed and one that was sent after the duplicate frame 0 arrived. The second ACK 1 is discarded.

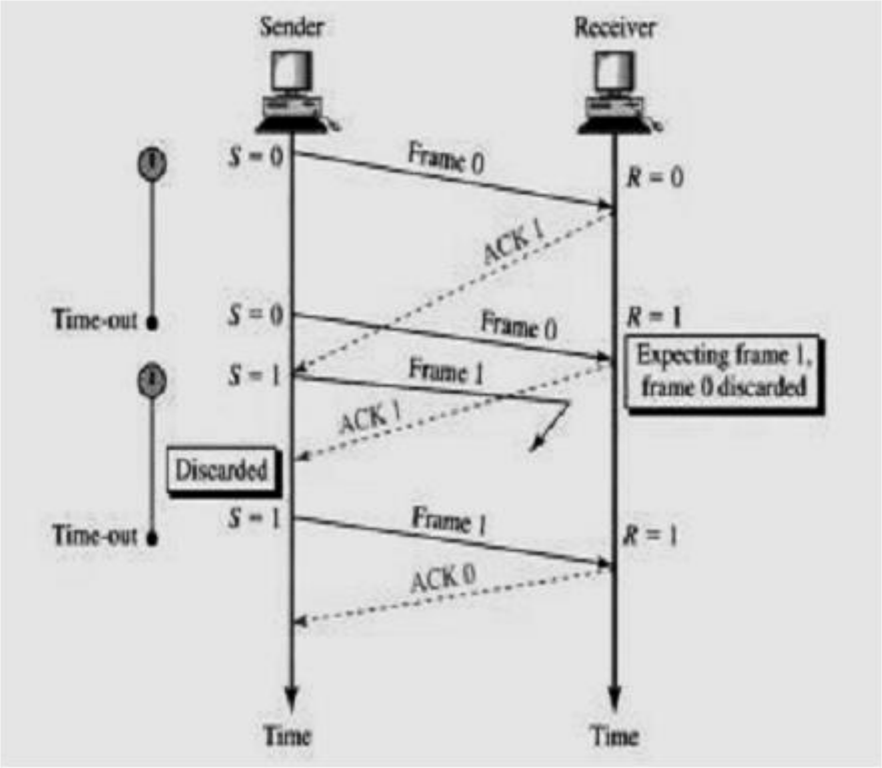


Figure: 7

2. Go-Back-N ARQ (Sliding Window)

CASE 1: Normal Operation Figure 8 shows a normal operation of this mechanism. The sender keeps track of the outstanding frames and updates the variables and windows as the acknowledgments arrive

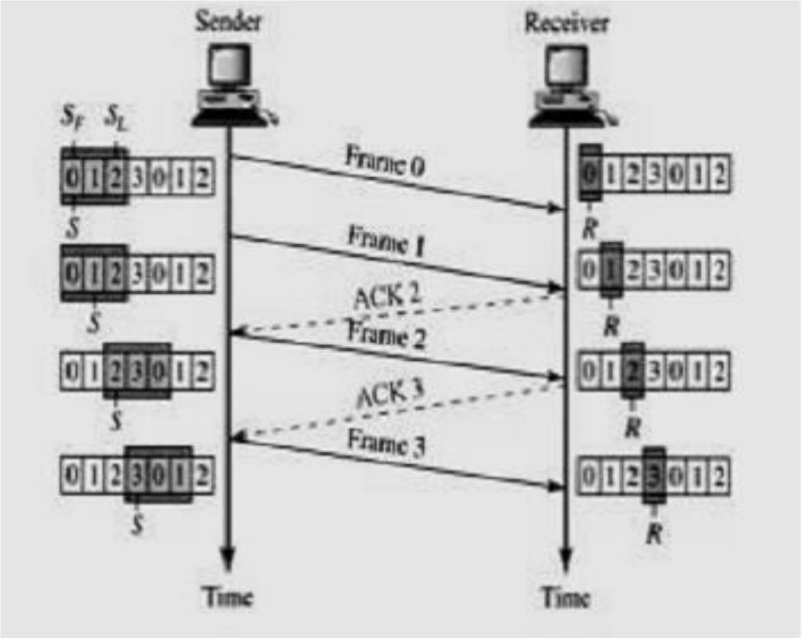


Figure 8

CASE 2: Damaged or Lost Frame Now let us see what happens if a frame is lost. Figure 9 shows that frame 2 is lost. Note that when the receiver receives frame 3 it is discarded because the receiver is expecting frame 2 not frame 3 (according to its window). After the timer for frame 2 expires at the sender site, the sender sends frames 2 and 3 (it goes back to 2).

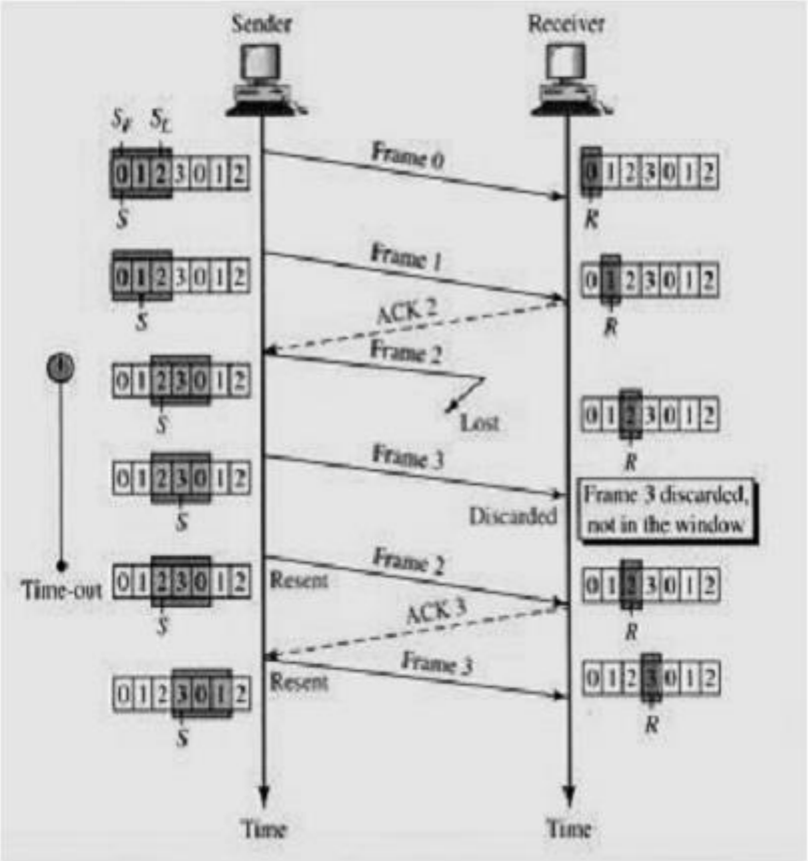


Figure 9

CASE 3: Damaged or Lost Acknowledgment

If an acknowledgment is damaged or lost, we can have two situations. If the next acknowledgment arrives before the expiration of any timer, there is no need for retransmission of frames because acknowledgments are cumulative in this protocol. ACK 4 means ACK 1 to ACK 4. So if ACK 1, ACK 2, and ACK 3 are lost. ACK 4 covers them. However, if the next ACK arrives after the time-out, the frame and all the frames after that are resent. Note that the receiver never resends an ACK. The figure and details are left out as an exercise.

3. SELECTIVE REPEAT ARQ

Go-Back-N ARQ simplifies the process at the receiver site. The receiver keeps track of only one variable, and there is no need to buffer out-of-order frames; they are simply discarded. However, this protocol is very inefficient for a noisy link. In a noisy link a frame has a higher probability of damage, which means the resending of multiple frames. This resending uses up the bandwidth and slows down the transmission. For noisy links, there is another mechanism that does not resend N frames when just one frame is damaged: only the damaged frame is resent. This mechanism is called Selective Repeat ARQ. It is more efficient for noisy links, but the processing at the receiver is more complex. Let us show the operation of the mechanism with an example of a lost frame, as shown in Figure 10. Frames 0 and 1 are accepted and received because they are in the range specified by the receiver window. When frame 3 is received, it is also accepted for the same reason. However, the receiver sends a NAK 2 to show that frame 2 has not been received. When the sender receives the NAK 2, it resends only frame 2, which is then accepted because it is in the range of the window.

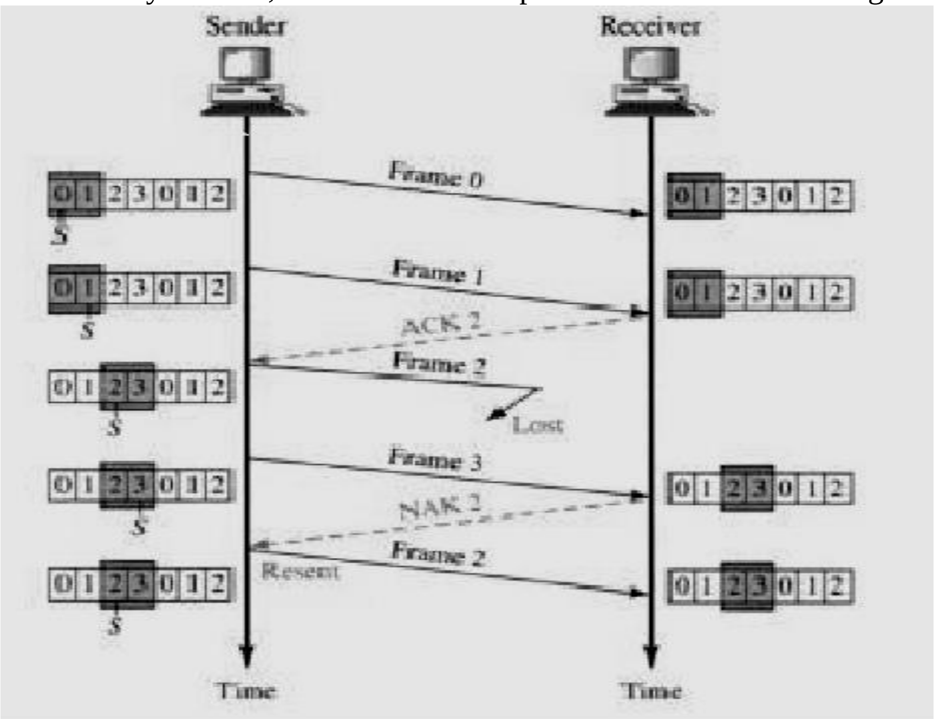


Figure 10

Forward Error Correction (FEC)

In this method, the transmitted data is encoded so that the data can correct as well as detect errors caused by channel noise. The choice of ARQ or FEC depends on the particular application. ARQ is often used when there is a full duplex (2-way) channel because it is relatively inexpensive to implement. FEC is used when the channel is not full duplex or where ARQ is not desirable because real time is required.

When we talk about digital systems, be it a digital computer or a digital communication set-up, the issue of error detection and correction is of great practical significance. Errors creep into the bit stream owing to noise or other impairments during the course of its transmission from the transmitter to the receiver. Any such error, if not detected and subsequently corrected, can be disastrous, as digital systems are sensitive to errors and tend to malfunction if the bit error rate is more than a certain threshold level. Error detection and correction, involves the addition of extra bits, called check bits, to the information-carrying bit stream to give the resulting bit sequence a unique characteristic that helps in detection and localization of errors. These additional bits are also called redundant bits as they do not carry any information. While the addition of redundant bits helps in achieving the goal of making transmission of information from one place to another error free or reliable, it also makes it inefficient. The Channel Encoder will add bits to the message bits to be transmitted systematically. After passing through the channel, the Channel decoder will detect and correct the errors. A simple example is to send '000' ('111' correspondingly) instead of sending only one '0' ('1' correspondingly) to the channel. Due to noise in the channel, the received bits may become '001'. But since either '000' or '111' could have been sent. By majority logic decoding scheme, it will be decoded as '000' and therefore the message has been a '0'. In general the channel encoder will divide the input message bits into blocks of k messages bits and replaces each k message bits block with a n-bit code word by introducing (n-k) check bits to each message block. In this section, we will examine some common error detection and correction codes.

PARITY CODE

A parity bit is an extra bit added to a string of data bits in order to detect any error that might have crept into it while it was being stored or processed and moved from one place to another in a digital system. We have an even parity, where the added bit is such that the total number of "l"s in the data bit string becomes even, and an odd parity, where the added bit makes the total number of "l"s in the data bit string odd. This added bit could be a '0' or a '1'. As an example, if we have to add an even parity bit to 01000001 (the eight-bit ASCII code for 'A'), it will be a '0' and the number will become 001000001. If we have to add an odd parity bit to the same number, it will be a 'l' and the number will become 101000001. The odd parity bit is a complement of the even parity bit. The most common convention is to use even parity, that is, the total number of 1s in the bit stream, including the parity bit, is even.

The parity check can be made at different points to look for any possible single-bit error, as it would disturb the parity. This simple parity code suffers from two limitations. Firstly, it cannot detect the error if the number of bits having undergone a change is even. Although the number of bits in error being equal to or greater than 4 is a very rare occurrence, the addition of a single parity cannot be used to detect two-bit errors, which is a distinct possibility in data storage media such as magnetic tapes. Secondly, the single-bit parity code cannot be used to localize or identify the error bit even if one bit is in error. There are several codes that provide self-single-bit error detection and correction mechanisms.

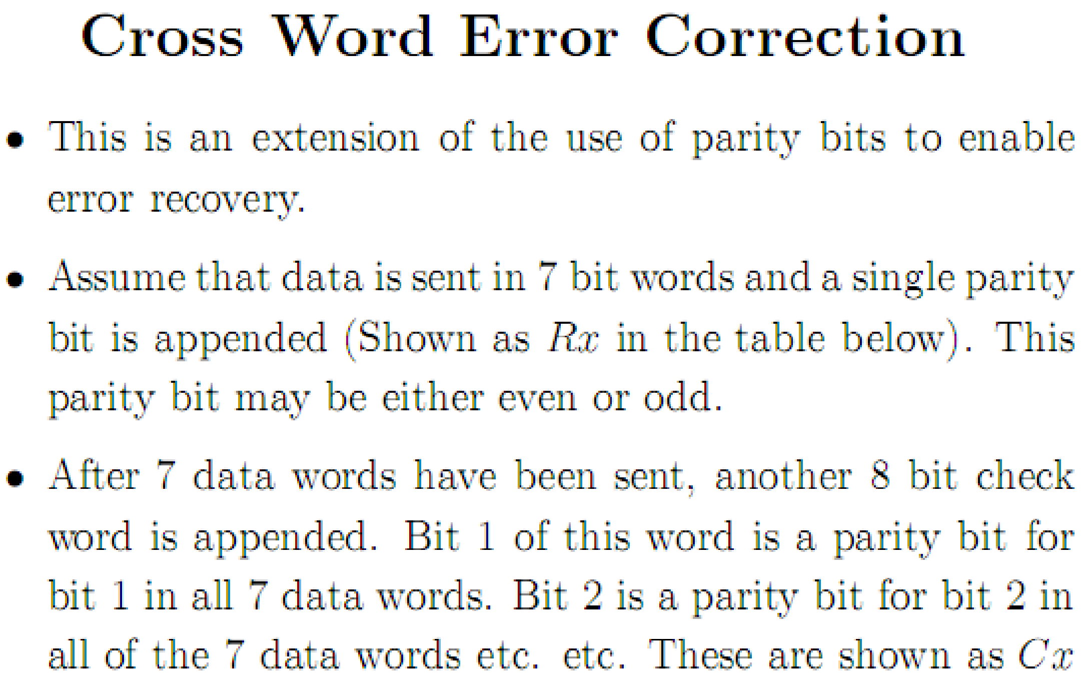
REPETITION CODE

The repetition code makes use of repetitive transmission of each data bit in the bit stream. In the case of threefold repetition, '1' and '0' would be transmitted as '111' and '000' respectively. If, in the received data bit stream, bits are examined in groups of three bits, the occurrence of an error can be detected. In the case of single-bit errors, '1' would be received as 011 or 101 or 110 instead of 111, and a '0' would be received as 100 or 010 or 001 instead of 000. In both cases, the code becomes self-correcting if the bit in the majority is taken as the correct bit. There are various forms in which the data are sent using the repetition code. Usually, the data bit stream is broken into blocks of bits, and then each block of data is sent some predetermined number of times. For example, if we want to send eight-bit data given by 11011001, it may be broken into two blocks of four bits each. In the case of threefold repetition, the transmitted data bit stream would be 110111011101100110011001. However, such a repetition code where the bit or block of bits is repeated 3 times is not capable of correcting two-bit errors, although it can detect the occurrence of error. For this, we have to increase the number of times each bit in the bit stream needs to be repeated. For example, by repeating each data bit 5 times, we can detect and correct all two-bit errors. The repetition code is highly inefficient and the information throughput drops rapidly as we increase the number of times each data bit needs to be repeated to build error detection and correction capability.

Cyclic Redundancy Code

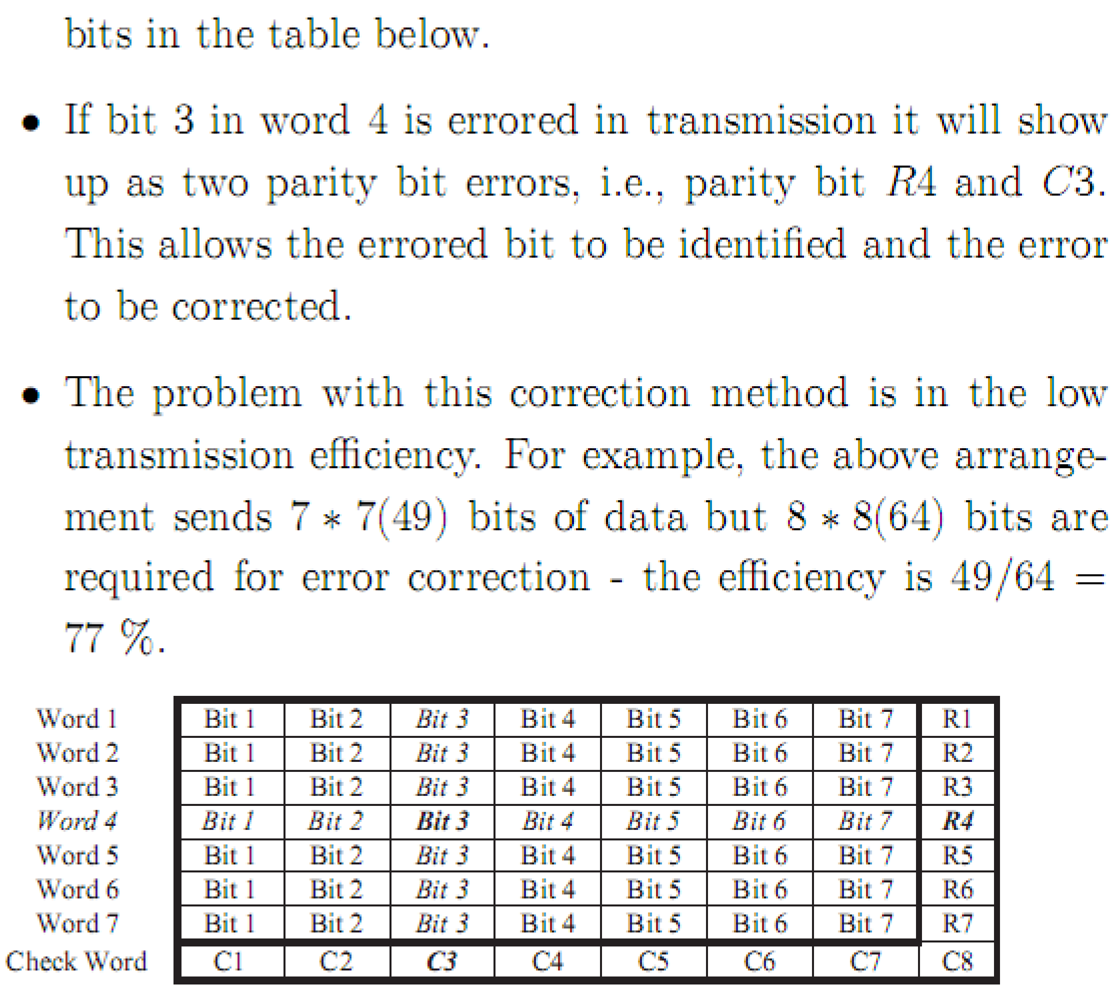
Cyclic redundancy check (CRC) codes provide a reasonably high level of protection at low redundancy level. The cycle code for a given data word is generated as follows. The data word is first appended by a number of 0s equal to the number of check bits to be added. This new data bit sequence is then divided by a special binary word whose length equals n+1, n being the number of check bits to be added. The remainder obtained as a result of modulo-2 division is then added to the dividend bit sequence to get the cyclic code. The code word so generated is completely divisible by the divisor used in the generation of the code. Thus, when the received code word is again divided by the same divisor, an error-free reception should lead to an all ‘0’ remainder. A nonzero remainder is indicative of the presence of errors.

The probability of error detection depends upon the number of check bits, n, used to construct the cyclic code. It is 100 % for single-bit and two-bit errors. It is also 100 % when an odd number of bits are in error and the error bursts have a length less than n+1. The probability of detection reduces to $1 – (1/2)^{n-1}$ for an error burst length equal to n+1, and to $1 – (1/2)^{n}$ for an error burst length greater than n+1.



Cross Word Error Correction

- This is an extension of the use of parity bits to enable error recovery.
- Assume that data is sent in 7 bit words and a single parity bit is appended (Shown as Rx in the table below). This parity bit may be either even or odd.
- After 7 data words have been sent, another 8 bit check word is appended. Bit 1 of this word is a parity bit for bit 1 in all 7 data words. Bit 2 is a parity bit for bit 2 in all of the 7 data words etc. etc. These are shown as Cx


bits in the table below.

- If bit 3 in word 4 is errored in transmission it will show up as two parity bit errors, i.e., parity bit $R4$ and $C3$. This allows the errored bit to be identified and the error to be corrected.
- The problem with this correction method is in the low transmission efficiency. For example, the above arrangement sends $7 * 7(49)$ bits of data but $8 * 8(64)$ bits are required for error correction - the efficiency is $49/64 =$ 77 %.

Word 1	Bit 1	Bit 2	*Bit 3*	Bit 4	Bit 5	Bit 6	Bit 7	R1
Word 2	Bit 1	Bit 2	*Bit 3*	Bit 4	Bit 5	Bit 6	Bit 7	R2
Word 3	Bit 1	Bit 2	*Bit 3*	Bit 4	Bit 5	Bit 6	Bit 7	R3
Word 4	*Bit 1*	*Bit 2*	**Bit 3**	*Bit 4*	*Bit 5*	*Bit 6*	*Bit 7*	***R4***
Word 5	Bit 1	Bit 2	*Bit 3*	Bit 4	Bit 5	Bit 6	Bit 7	R5
Word 6	Bit 1	Bit 2	*Bit 3*	Bit 4	Bit 5	Bit 6	Bit 7	R6
Word 7	Bit 1	Bit 2	*Bit 3*	Bit 4	Bit 5	Bit 6	Bit 7	R7
Check Word	C1	C2	***C3***	C4	C5	C6	C7	C8

Hamming Code

We have seen, in the case of the error detection and correction codes described above, how an increase in the number of redundant bits added to message bits can enhance the capability of the code to detect and correct errors. If we have a sufficient number of redundant bits, and if these bits can be arranged such that different error bits produce different error results, then it should be possible not only to detect the error bit but also to identify its location. In fact, the addition of redundant bits alters the ‘distance’ code parameter, which has come to be known as the Hamming distance. The Hamming distance is nothing but the number of bit disagreements between two code words. For example, the addition of single-bit parity results in a code with a Hamming distance of at least 2. The smallest Hamming distance in the case of a threefold repetition code would be 3. Hamming noticed that an increase in distance enhanced the code’s ability to detect and correct errors. Hamming’s code was therefore an attempt at increasing the Hamming distance and at the same time having as high an information throughput rate as possible.

The algorithm for writing the generalized Hamming code is as follows:

1. The generalized form of code is $P_1P_2D_1P_3D_2D_3D_4P_4D_5D_6D_7D_8D_9D_{10}D_{11}P_5\ldots$, where P and D respectively represent parity and data bits.
2. We can see from the generalized form of the code that all bit positions that are powers of 2 (positions 1, 2, 4, 8, 16, . . .) are used as parity bits.
3. All other bit positions (positions 3, 5, 6, 7, 9, 10, 11, . . .) are used to encode data.
4. Each parity bit is allotted a group of bits from the data bits in the code word, and the value of the parity bit (0 or 1) is used to give it certain parity.
5. Groups are formed by first checking $N-1$ bits and then alternately skipping and checking N bits following the parity bit. Here, N is the position of the parity bit; 1 for P_1, 2 for P_2, 4 for P_3, 8 for P_4 and so on. For example, for the generalized form of code given above, various groups of bits formed with different parity bits would be $P_1D_1D_2D_4D_5\ldots$, $P_2D_1D_3D_4D_6D_7\ldots$, $P_3D_2D_3D_4D_8D_9\ldots$, $P_4D_5D_6D_7D_8D_9D_{10}D_{11}\ldots$ and so on. To illustrate the formation of groups further, let us examine the group corresponding to parity bit P_3. Now, the position of P_3 is at number 4. In order to form the group, we check the first three bits ($N-1=3$) and then follow it up by alternately skipping and checking four bits ($N=4$).

The Hamming code is capable of correcting single-bit errors on messages of any length. Although the Hamming code can detect two-bit errors, it cannot give the error locations. The number of parity bits required to be transmitted along with the message, however, depends upon the message length, as shown above. The number of parity bits n required to encode m message bits is the smallest integer that satisfies the condition $(2^n - n) > m$.

Table 2.9 Generation of Hamming code.

	P_1	P_2	D_1	P_3	D_2	D_3	D_4
Data bits (without parity)			0		1	1	0
Data bits with parity bit P_1	1		0		1		0
Data bits with parity bit P_2		1	0			1	0
Data bits with parity bit P_3				0	1	1	0
Data bits with parity	1	1	0	0	1	1	0

The most commonly used Hamming code is the one that has a code word length of seven bits with four message bits and three parity bits. It is also referred to as the Hamming (7, 4) code. The code word sequence for this code is written as $P_1P_2D_1P_3D_2D_3D_4$, with P_1, P_2 and P_3 being the parity bits and D_1, D_2, D_3 and D_4 being the data bits. We will illustrate step by step the process of writing the Hamming code for a certain group of message bits and then the process of detection and identification of error bits with the help of an example. We will write the Hamming code for the four-bit message 0110 representing numeral '6'. The process of writing the code is illustrated in Table 2.9, with even parity.

Thus, the Hamming code for 0110 is 1100110. Let us assume that the data bit D_1 gets corrupted in the transmission channel. The received code in that case is 1110110. In order to detect the error, the parity is checked for the three parity relations mentioned above. During the parity check operation at the receiving end, three additional bits X, Y and Z are generated by checking the parity status of $P_1D_1D_2D_4$, $P_2D_1D_3D_4$ and $P_3D_2D_3D_4$ respectively. These bits are a '0' if the parity status is okay, and a '1' if it is disturbed. In that case, ZYX gives the position of the bit that needs correction. The process can be best explained with the help of an example.

Examination of the first parity relation gives $X=1$ as the even parity is disturbed. The second parity relation yields $Y=1$ as the even parity is disturbed here too. Examination of the third relation gives $Z=0$ as the even parity is maintained. Thus, the bit that is in error is positioned at 011 which is the binary equivalent of '3'. This implies that the third bit from the MSB needs to be corrected. After correcting the third bit, the received message becomes 1100110 which is the correct code.

Example 2.6

By writing the parity code (even) and threefold repetition code for all possible four-bit straight binary numbers, prove that the Hamming distance in the two cases is at least 2 in the case of the parity code and 3 in the case of the repetition code.

Solution

The generation of codes is shown in Table 2.10. An examination of the parity code numbers reveals that the number of bit disagreements between any pair of code words is not less than 2. It is either 2 or 4. It is 4, for example, between 00000 and 10111, 00000 and 11011, 00000 and 11101, 00000 and 11110 and 00000 and 01111. In the case of the threefold repetition code, it is either 3, 6, 9 or 12 and therefore not less than 3 under any circumstances.

Example 2.7

It is required to transmit letter 'A' expressed in the seven-bit ASCII code with the help of the Hamming (11, 7) code. Given that the seven-bit ASCII notation for 'A' is 1000001 and that the data word gets corrupted to 1010001 in the transmission channel, show how the Hamming code can be used to identify the error. Use even parity.

Table 2.10 Example 2.6.

Binary number	Parity code	Three-time repetition Code	Binary number	Parity code	Three-time repetition code
0000	00000	000000000000	1000	11000	100010001000
0001	10001	000100010001	1001	01001	100110011001
0010	10010	001000100010	1010	01010	101010101010
0011	00011	001100110011	1011	11011	101110111011
0100	10100	010001000100	1100	01100	110011001100
0101	00101	010101010101	1101	11101	110111011101
0110	00110	011001100110	1110	11110	111011101110
0111	10111	011101110111	1111	01111	111111111111

Solution

- The generalized form of the Hamming code in this case is $P_1P_2D_1P_3D_2D_3D_4P_4D_5D_6D_7 = P_1P_21P_3000P_4001$.
- The four groups of bits using different parity bits are $P_1D_1D_2D_4D_5D_7$, $P_2D_1D_3D_4D_6D_7$, $P_3D_2D_3D_4$ and $P_4D_5D_6D_7$.
- This gives $P_1 = 0$, $P_2 = 0$, $P_3 = 0$ and $P_4 = 1$.
- Therefore, the transmitted Hamming code for 'A' is 00100001001.
- The received Hamming code is 00100101001.
- Checking the parity for the P_1 group gives '0' as it passes the test.
- Checking the parity for the P_2 group gives '1' as it fails the test.
- Checking the parity for the P_3 group gives '1' as it fails the test.
- Checking the parity for the P_4 group gives '0' as it passes the test.
- The bits resulting from the parity check, written in reverse order, constitute 0110, which is the binary equivalent of '6'. This shows that the bit in error is the sixth from the MSB.
- Therefore, the corrected Hamming code is 00100001001, which is the same as the transmitted code.
- The received data word is 1000001.

Block Codes

Denoted by (n, k) a block code is a collection of code words each with length n, k information bits and r = n – k check bits. It is linear if it is closed under addition mod 2. A **Generator Matrix** G (of order k × n) is used to generate the code.

$G = [\, I_k \quad P \,]_{k \times n}$ ……………………………………………………………………..

where I_k is the k × k identity matrix and P is a k × (n – k) matrix selected to give desirable properties to the code produced.

For example, denote D to be the message, G to be the generator matrix, C to be code word. For

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

We get the collection of code words in the Table below

Table: Collection of Code words

Messages (D)	Code words (C)
0 0 0	0 0 0 0 0 0
0 0 1	0 0 1 1 1 0
0 1 0	0 1 0 1 0 1
0 1 1	0 1 1 0 1 1
1 0 0	1 0 0 0 1 1
1 0 1	1 0 1 1 0 1
1 1 0	1 1 0 1 1 0
1 1 1	1 1 1 0 0 0

In particular when D = [0 1 1],

$$C = DG = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 & 1 \end{bmatrix}.$$

Now define the **Parity Check Matrix** to be

$H = \lfloor\, P^T \quad I_{n-k} \,\rfloor_{(n-k) \times n}.$ …………………………………………..

As long as the code word C is generated by G, the product $CH^T = 0$.

Denote the **received code word** after passing through the channel be R. It is made up of the original code word C and error bits E from the channel. Further define the **Error Syndrome** to be

$S = RH^T$ Then,

$$R = C + E$$
$$S = RH^T = (C + E)\, H^T = CH^T + EH^T = EH^T$$

If $S = 0$, $R \equiv C$ and D is the first k bits of R. If $S \neq 0$ and S is the jth row of H^T, then it imply an error occurs in the jth position of R.

To illustrate, suppose after passing through the channel, the received code word is R = [0 1 1 1 1 0].

Thus RHT = [0 1 1 1 1 0] H^T = [1 0 1] which is the second row of H^T. Thus it implies there exists an error in the second bit of the received code word. So we can correct it and detect that [0 0 1 1 1 0] should have been sent.

3. Applications for error control codes:

1) Compact disc players provide a growing application area for FECC.

2) In CD applications the powerful Reed-Solomon code is used since it works at a symbol level, rather than at a bit level, and is very effective against burst errors.

3) The Reed-Solomon code is also used in computers for data storage and retrieval.

4) Digital audio and video systems are also areas in which FEC is applied.

5) Error control coding, generally, is applied widely in control and communications systems for aerospace applications, in mobile (GSM).

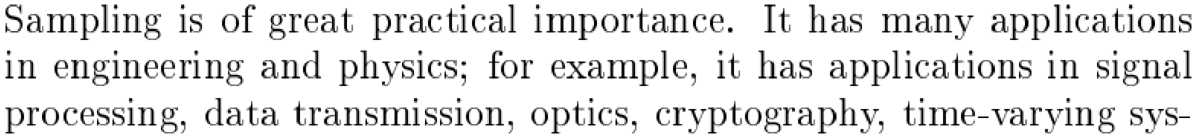
6) Cellular telephony and for enhancing security in banking and barcode readers.

Digitalization (sampling theorem, Shannon theorem, PCM and Quantization Error, Multiplexing, FDM, TDM, Higher Order Multiplexing; Frame Formatting Time slot

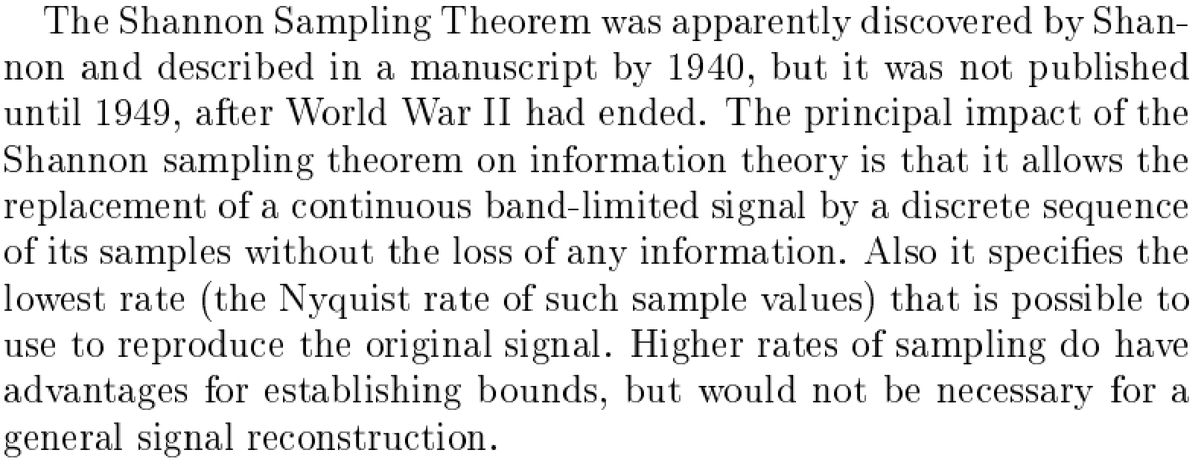
Introduction

With rapid advancement in data acquistion technology (i.e. analog-to-digital and digital-to-analog converters) and the explosive introduction of micro-computers, selected complex linear and nonlinear functions currently implemented with analog circuitry are being alternately implemented with sample data systems.

Though more costly than their analog counterpart, these sampled data systems feature programmability. Additionally, many of the algorithms employed are a result of developments made in the area of signal processing and are in some cases capable of functions unrealizable by current analog techniques. With increased usage a proportional demand has evolved to understand the theoretical basis required in interfacing these sampled data-systems to the analog world.



Sampling is of great practical importance. It has many applications in engineering and physics; for example, it has applications in signal processing, data transmission, optics, cryptography, time-varying systems and boundary value problems.



The Shannon Sampling Theorem was apparently discovered by Shannon and described in a manuscript by 1940, but it was not published until 1949, after World War II had ended. The principal impact of the Shannon sampling theorem on information theory is that it allows the replacement of a continuous band-limited signal by a discrete sequence of its samples without the loss of any information. Also it specifies the lowest rate (the Nyquist rate of such sample values) that is possible to use to reproduce the original signal. Higher rates of sampling do have advantages for establishing bounds, but would not be necessary for a general signal reconstruction.

Shannon's original statement of the Shannon's sampling Theorem states that:

"If a function contains no frequencies greater than ω cycles per second, then it is completely determined by giving its ordinates at a series of points spaced $^1/_{2\omega}$ *apart".*

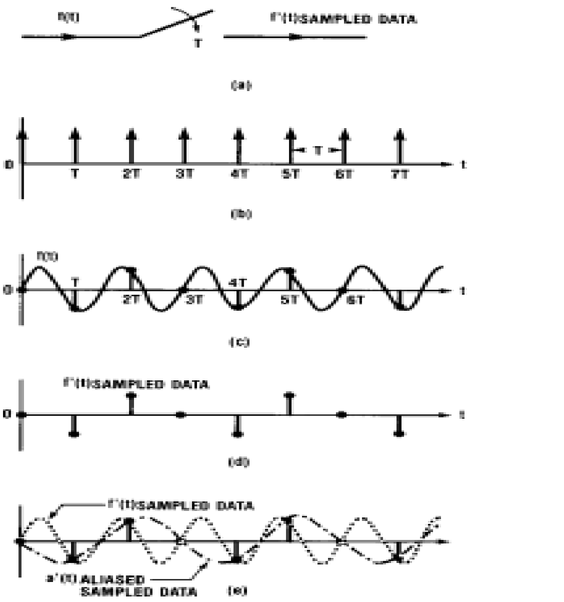
The limitation on frequencies in this statement translates with our choice of notation of saying that the Inverse Fourier Transform (INVFT) of the function has support in $[-2\pi\omega, 2\pi\omega]$.

Shannon's sampling theorem can be rephrased as follows:

"In order to recover the signal function f(t) exactly, it is necessary to sample f(t) at a rate greater than twice its highest frequency component".

The sampling theorem by C.E. Shannon in 1949 places restrictions on the frequency content of the time function signal, f(t). Practically speaking for example, to sample an analog signal having a maximum frequency of 2Kc Hz requires sampling at a frequency greater than 4Kc Hz to preserve and recover the waveform exactly.

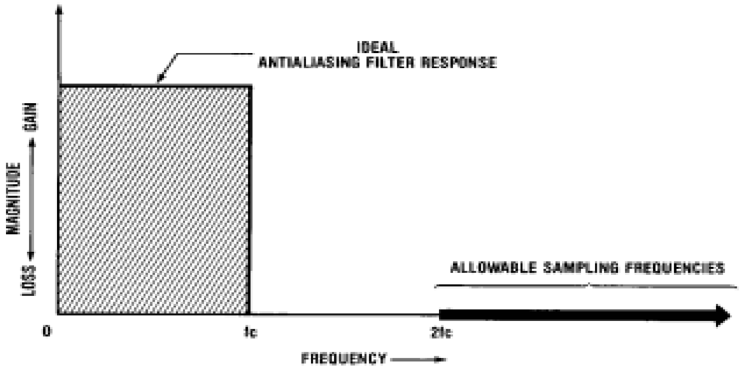
The consequences of sampling a signal at a rate below its highest frequency component result in a phenomenon known as aliasing. This concept results in a frequency mistakenly taking on the identity of an entirely different frequency when recovered. In an attempt to clarify this, envision the ideal sampler of Figure 1(a), with a sample period of T shown in (b), sampling the waveform f(t) as pictured in (c). The sampled data points of f'(t) are shown in (d) and can be defined as the sample set of the continuous function f(t). Note in Figure 1(e) that another frequency component, a'(t), can be found that has the same sample set of data points as f'(t) in (d). Because of this it is difficult to determine which frequency a'(t), is truly being observed. This effect is similar to that observed in western movies when watching the spoked wheels of a rapidly moving stagecoach rotate backwards at a slow rate. The effect is a result of each individual frame of film resembling a discrete strobed sampling operation flashing at a rate slightly faster than that of the rotating wheel. Each observed sample point or frame catches the spoked wheel slightly displaced from its previous position giving the effective appearance of a wheel rotating backwards. Again, aliasing is evidenced and in this example it becomes difficult to determine which is the true rotational frequency being observed. On the surface it is easily said that anti-aliasing designs can be achieved by sampling at a rate greater than twice the maximum frequency found within the signal to be sampled. In the real world, however, most signals contain the entire spectrum of frequency components; from the desired to those present in white noise. To recover such information accurately the system would require an unrealizably high sample rate.



TL/H/5620-1

FIGURE 1. When sampling, many signals may be found to have the same set of data points. These are called aliases of each other.

This difficulty can be easily overcome by preconditioning the input signal, the means of which would be a band-limiting or frequency filtering function performed prior to the sample data input. The pre-filter, typically called anti-aliasing filter guarantees, for example in the low pass filter case, that the sampled data system receives analog signals having a spectral content no greater than those frequencies allowed by the filter. As illustrated in Figure 2, it thus becomes a simple matter to sample at greater than twice the maximum frequency content of a given signal.



TL/H/5620–2

FIGURE 2. Shown in the shaded area is an ideal, low pass, anti-aliasing filter response. Signals passed through the filter are bandlimited to frequencies no greater than the cutoff frequency, fc. In accordance with the sampling theorem, to recover the bandlimited signal exactly the sampling rate must be chosen to be greater than 2fc.

A parallel analog of band-limiting can be made to the world of perception when considering the spectrum of white light. It can be realized that the study of violet light wavelengths generated from a white light source would be vastly simplified if initial band-limiting were performed through the use of a prism or white light filter.

Mathematical Presentation of the Sampling Theorem

To solidify some of the intuitive thoughts presented in the previous section, the sampling theorem will be presented applying the rigor of mathematics supported by an illustrative proof. This should hopefully leave the student with a comfortable understanding of the sampling theorem.

Theorem: If the Fourier transform $F(\omega)$ of a signal function $f(t)$ is zero for all frequencies above $|\omega| \geq \omega_c$, then $f(t)$ can be uniquely determined from its sampled values

$$f_n = f(nT) \qquad (1)$$

These values are a sequence of equidistant sample points spaced $\frac{1}{2fc} = \frac{Tc}{2} = T$ apart. $f(t)$ is thus given by

$$f(t) = \sum_{n=-\infty}^{\infty} f(nT) \frac{\sin \omega_c (t - nT)}{\omega_c (t - nT)} \qquad (2)$$

Proof: Using the inverse Fourier transform formula:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) \varepsilon^{j\omega t} \, d\omega \qquad (3)$$

the band limited function, $f(t)$, takes the form, *Figure 3a*,

$$f(t) = \frac{1}{2\pi} \int_{-\omega_c}^{\omega_c} F(\omega) \varepsilon^{j\omega t} \, d\omega \qquad (4)$$

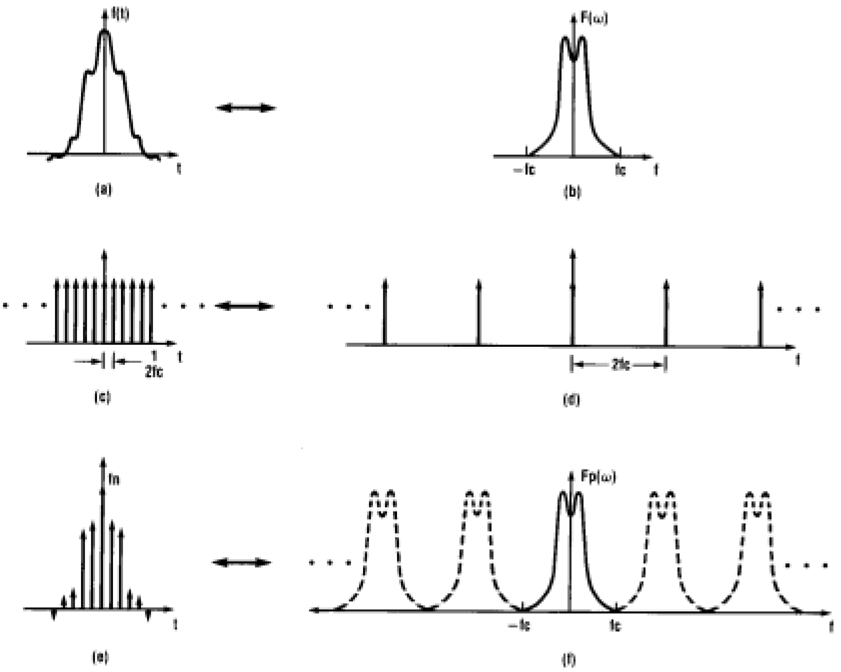
$f_n = f\left(n \frac{\pi}{\omega_c}\right)$ is then given as

$$f_n = \frac{1}{2\pi} \int_{-\omega_c}^{\omega_c} F(\omega) \varepsilon^{j\omega \frac{n\pi}{\omega_c}} \, d\omega \qquad (5)$$

See *Figure 3c* and *e*.

Expressing $F(\omega)$ as a Fourier series in the interval $-\omega_c \leq \omega \leq \omega_c$ we have

$$F(\omega) = \sum_{n=-\infty}^{\infty} C_n \varepsilon^{-j\omega \frac{n\pi}{\omega_c}} \qquad (6)$$



TL/H/5620–3

FIGURE 3. Fourier transform of a sampled signal.

Where,

$$C_n = \frac{1}{2\omega_c}\int_{-\omega_c}^{\omega_c} F(\omega)\epsilon^{j\omega\frac{n\pi}{\omega_c}}d\omega \tag{7}$$

Further manipulating eq. (7)

$$C_n = \frac{2\pi}{2\omega_c}\frac{1}{2\pi}\int_{-\omega_c}^{\omega_c} F(\omega)\epsilon^{j\omega\frac{n\pi}{\omega_c}}d\omega \tag{8}$$

C_n can be written as

$$C_n = \frac{\pi}{\omega_c}f_n \tag{9}$$

Substituting eq. (9) into eq. (6) gives the periodic Fourier Transform

$$Fp(\omega) = \sum_{n=-\infty}^{\infty} \frac{\pi}{\omega_c}f_n\epsilon^{-j\omega\frac{n\pi}{\omega c}} \tag{10}$$

of *Figure 3f*. Using Poisson's sum formula[1] $F(\omega)$ can be stated more clearly as

$$F(\omega) = \sum_{n=-\infty}^{\infty} F(\omega - 2_n\omega_c) \tag{11}$$

Interestingly for the interval $-\omega_c \leq \omega \leq \omega_c$ the periodic function $F_p(\omega)$ and *Figure 3f*. equals $F(\omega)$ and *Figure 3b*. respectively. Analogously if $F_p(\omega)$ were multiplied by a rectangular pulse defined,

$$H(\omega) = 1 \qquad -\omega_c \leq \omega \leq \omega \tag{12}$$

and

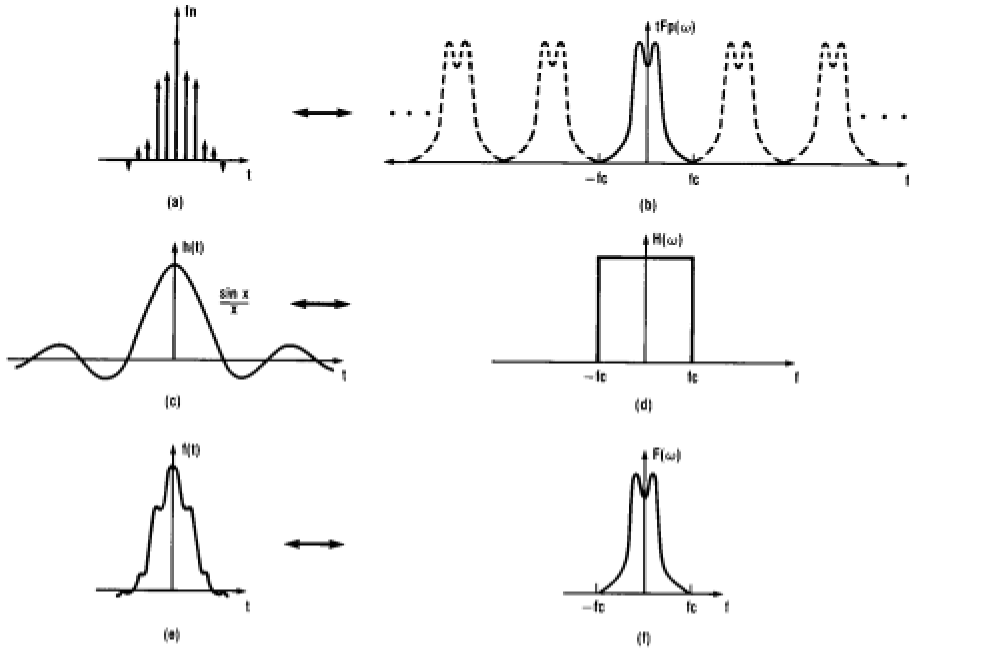
$$H(\omega) = 0 \qquad |\omega| \geq \omega_C \tag{13}$$

then as pictured in *Figures 4b*, *d*, and *f*,

[1] Poisson's sum formula

$$\frac{1}{T}\sum_{n=-\infty}^{\infty} F(\omega - n\omega_s) = \sum_{n=-\infty}^{\infty} f(nT)\epsilon^{-j\omega nT}$$

where $T = \frac{1}{f_s}$ and f_s is the sampling frequency



TL/H/5620–4

FIGURE 4. Recovery of a signal f(t) from sampled data information.

$$F(\omega) = H(\omega) \bullet F_p(\omega) = H(\omega) \sum_{n=-\infty}^{\infty} \frac{\pi}{\omega_c} f_n \epsilon^{-j\omega \frac{n\pi}{\omega_c}} \qquad (14)$$

Solving for f(t) the inverse Fourier transform eq (3) is applied to eq (14)

$$f(t) = \frac{1}{2\pi} \int_{-\omega_c}^{\omega_c} F(\omega) \epsilon^{j\omega t}\, d\omega \qquad (3)$$

$$= \frac{1}{2\pi} \int_{-\omega_c}^{\omega_c} \left[H(\omega) \sum_{n=-\infty}^{\infty} \frac{\pi}{\omega_c} f_n \epsilon^{-j\omega \frac{n\pi}{\omega c}} \right] \epsilon^{j\omega t}\, d\omega$$

$$= \sum_{n=-\infty}^{\infty} fn \frac{1}{2\omega_c} \int_{-\omega_c}^{\omega_c} \epsilon^{j\omega\left(t - \frac{n\pi}{\omega}\right)}\, d\omega$$

giving

$$f(t) = \sum_{n=-\infty}^{\infty} f_n \frac{\sin \omega_c \left(t - \frac{n\pi}{\omega_c}\right)}{\omega_c \left(t - \frac{n\pi}{\omega_c}\right)} \qquad (15)$$

As observed in *Figures 3* and *4*, each step of the sampling theorem proof was also illustrated with its Fourier transform pair. This was done to present alternate illustrative proofs.

Recalling the convolution[2] theorem, the convolution of F(ω), *Figure 3b*, with a set of equidistant impulses, *Figure 3d*, yields the same periodic frequency function $F_p(\omega)$, *Figure 3f*, as did the Fourier transform of f_n, *Figure 3e*, the product of f(t), *Figure 3a*, and its equidistant sample impulses, *Figure 3c*.

In the same light the original time function f(t), *Figure 4e*, could have been recovered from its sampled waveform by convolving f_n, *Figure 4a*, with h(t), *Figure 4c*, rather than multiplying $F_p(\omega)$, *Figure 4b*, by the rectangular function H(ω), *Figure 4d*, to get F(ω), *Figure 4f*, and finally inverse transforming to achieve f(t), *Figure 4e*, as done in the mathematic proof.

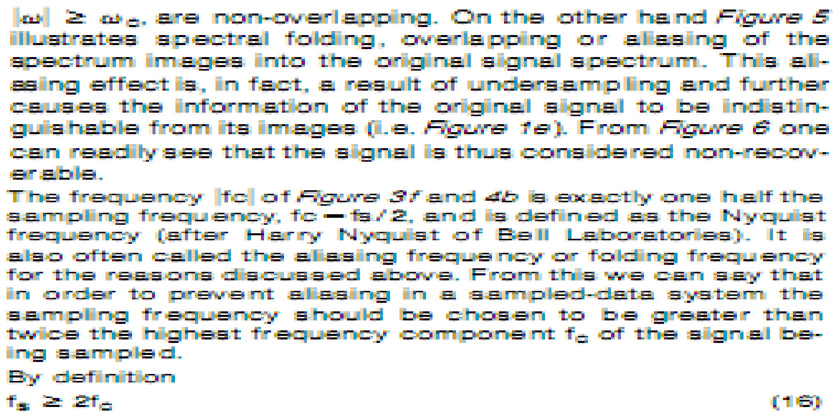
Some Observations and Definitions

If *Figures 3f* or *4b* are re-examined it can be noted that the original spectrum $F_p(\omega)$, $|\omega| \leq \omega_c$, and its images $F_p(\omega)$,

[2] The convolution theorem allows one to mathematically convolve in the time domain by simply multiplying in the frequency domain. That is, if f(t) has the Fourier transform F(ω), and x(t) has the Fourier transform X(ω), then the convolution f(t)*x(t) has the Fourier transform F(ω)•X(ω).

$$f(t) * x(t) \longleftrightarrow F(\omega) \bullet X(\omega)$$

$$f(t) \bullet x(t) \longleftrightarrow F(\omega) * X(\omega)$$



$|\omega| \geq \omega_c$, are non-overlapping. On the other hand *Figure 5* illustrates spectral folding, overlapping or aliasing of the spectrum images into the original signal spectrum. This aliasing effect is, in fact, a result of undersampling and further causes the information of the original signal to be indistinguishable from its images (i.e. *Figure 1e*). From *Figure 6* one can readily see that the signal is thus considered non-recoverable.

The frequency |fc| of *Figure 3f* and *4b* is exactly one half the sampling frequency, fc — fs/2, and is defined as the Nyquist frequency (after Harry Nyquist of Bell Laboratories). It is also often called the aliasing frequency or folding frequency for the reasons discussed above. From this we can say that in order to prevent aliasing in a sampled-data system the sampling frequency should be chosen to be greater than twice the highest frequency component f_c of the signal being sampled.

By definition

$$f_s \geq 2f_c \quad (16)$$

Note, however, that no mention has been made to sample at precisely the Nyquist rate since in actual practice it is impossible to sample at fs
=2f$_c$ unless one can guarantee there are absolutely no signal components above fc. This can only be achieved by filtering the signal prior to sampling with a filter having infinite rolloff . . . a physical impossibility, see Figure 2.

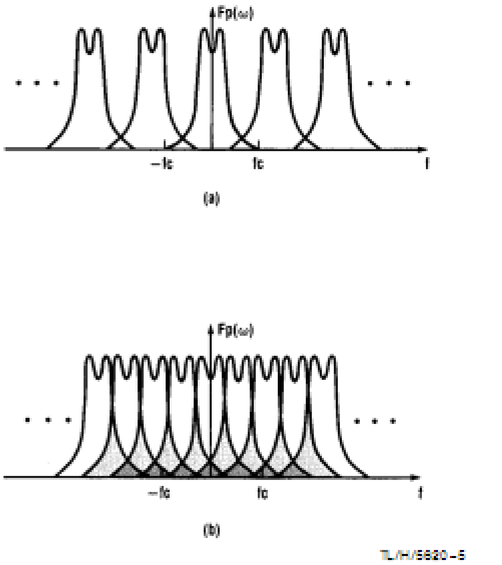


FIGURE 5. Spectral folding or aliasing caused by:
(a) under sampling
(b) exaggerated under sampling.

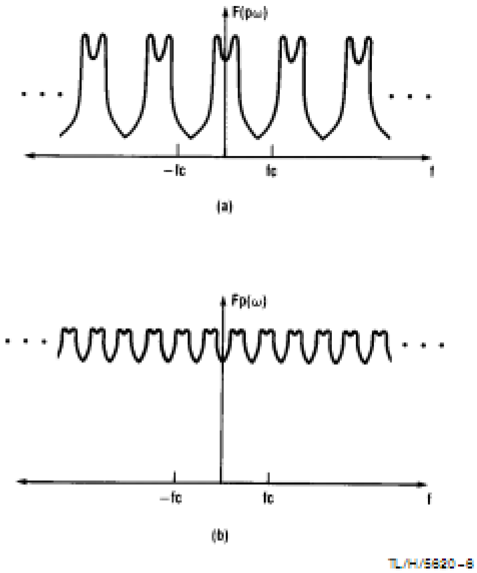


FIGURE 6. Aliased spectral envelope (a) and (b) of Figures 5a and 5b respectively.

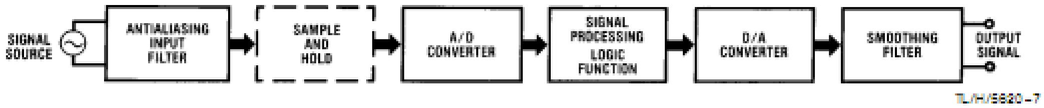


FIGURE 7. Generalized single channel sample data system.

1.1 Pulse Code Modulation:

Pulse code Modulation: The pulse code modulator technique samples the input signal x(t) at a sampling frequency. This sampled variable amplitude pulse is then digitalized by the analog to digital converter. Figure.(1) shows the PCM generator.

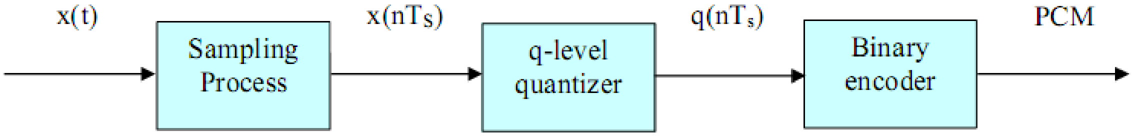


Figure.(1): PCM modulator

In the PCM generator, the signal is first passed through sampler which is sampled at a rate of (f_s) where:

$$\boxed{f_s \geq 2f_m} \qquad (1)$$

The output of the sampler x(nT_s) which is discrete in time is fed to a q-level quantizer. The quantizer compares the input x(nT_s) with it's fixed levels. It assigns any one of the digital level to x(nT_s) that results in minimum distortion or error. The error is called quantization error, thus the output of the quantizer is a digital level called q(nT_s). The quantized signal level q(nT_s) is binary encode. The encoder converts the input signal to $\boldsymbol{v}$ digits binary word.

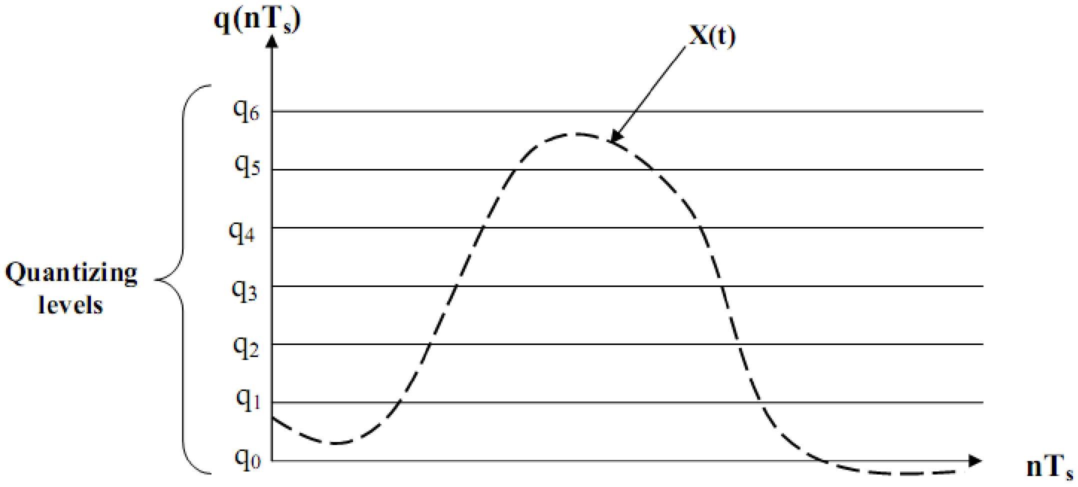


Figure.(2) A sampled signal and the quantized levels

Figure.(3) shows the block diagram of the PCM receiver. The receiver starts by reshaping the received pulses, removes the noise and then converts the binary bits to analog. The received samples are then filtered by a low pass filter; the cut off frequency is at f_c.

$$\boxed{f_c = f_m} \qquad (2)$$

where f_m: is the highest frequency component in the original signal.

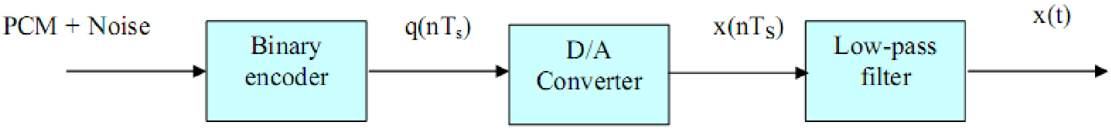


Figure.(3): PCM demodulator

It is impossible to reconstruct the original signal x(t) because of the permanent quantization error introduced during quantization at the transmitter. The quantization error can be reduced by the increasing quantization levels. This corresponds to the increase of bits per sample(more information). But increasing bits ***(v)*** increases the signaling rate and requires a large transmission bandwidth. The choice of the parameter for the number of quantization levels must be acceptable with the quantization noise (quantization error). Figure.(4) shows the reconstructed signal.

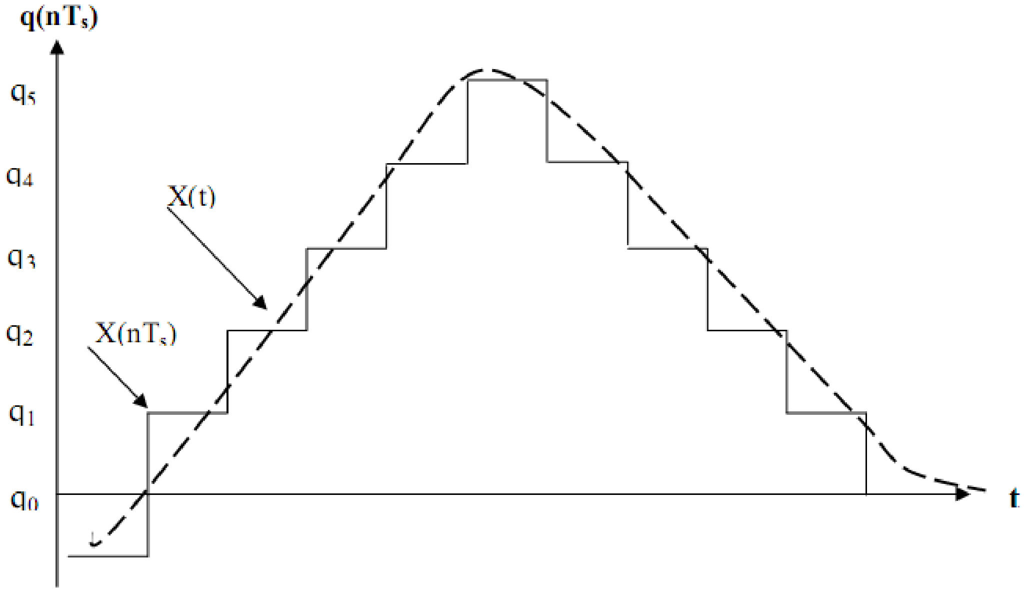


Figure.(4):The reconstructed signal

1.1.1 Signaling Rate in PCM

Let the quantizer use 'v' number of binary digits to represent each level. Then the number of levels that can be represented by ν digits will be :

$$\boxed{q=2^{\nu}} \qquad (3)$$

The number of bits per second is given by :

(Number of bits per second)=(Number of bits per samples)x(number of samples per second)

= ν (bits per sample) x f_s (samples per second)

The number of bits per second is also called signaling rate of PCM and is denoted by 'r':

$$\boxed{\text{Signaling rate}= \nu\, f_s} \qquad (4)$$

Where:

$f_s \geq f_m$

Example

If the number of binary bits = 3 and the sampling rate is 2 sample/sec find the signaling rate, number of quantization levels?

Solution:

$f_s=2, \quad \nu=3$

signaling rate(r) = $\nu\, f_s$

=3*2

=6 bits/sec

Number of quantization(q)=2^{ν}

=2^3

=8 levels

1.1.2 Quantization Noise in PCM System

Errors are introduced in the signal because of the quantization process. This error is called "quantization error". We define the quantization error as:

$$\boxed{\varepsilon = x_q\,(nT_s) - x(nT_s)} \qquad (5)$$

Let an input signal $x(nT_s)$ have an amplitude in the range of x_{max} to - x_{max} The total amplitude range is :

$$\text{Total amplitude} = x_{max} - (-\,x_{max})$$
$$= 2\,x_{max}$$

If the amplitude range is divided into 'q' levels of quantizer, then the step size 'Δ'.

$$\boxed{\Delta = \frac{2\,X_{max}}{q}} \qquad (6)$$

If the minimum and maximum values are equal to 1, x_{max},=1, - x_{max}=-1, then the equation (6)will be:

$$\boxed{\Delta = \frac{2}{q}} \qquad (7)$$

If Δ is small it can be assumed that the quantization error is uniformly distributed. The quantization noise is uniformly distributed in the interval [-Δ/2, Δ/2]. The figure.(5) shows the uniform distribution of quantization noise:

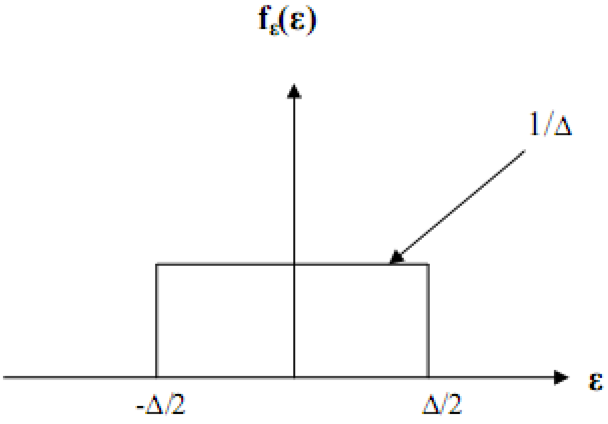


Fig.(5) The uniform distribution of quantization error

The noise power is given by:

$$\boxed{\text{Noise power} = V_{noise}^{2} / R} \qquad (8)$$

V_{noise}^{2} : is the mean square value of noise voltage, since noise is defined by random variable "ε" and PDF $\mathbf{f_{\varepsilon}(\varepsilon)}$, it's mean square value is given by :

$$\boxed{V_{noise}^{2} = \int_{-\Delta/2}^{\Delta/2} \varepsilon^{2} f_{\varepsilon}(\varepsilon).d\varepsilon} \qquad (9)$$

Substitute the value of $f_{\varepsilon}(\varepsilon)$ =1/Δ in eq(9):

$$V_{noise}^{2} = \int_{-\Delta/2}^{\Delta/2} \varepsilon^{2} 1/\Delta.d\varepsilon$$

$$= \frac{1}{3\Delta}\left[\frac{\Delta^{3}}{8} + \frac{\Delta^{3}}{8}\right]$$

$$= \frac{\Delta^{2}}{12} \qquad \text{If R=1}$$

$$\boxed{\text{Quantization noise power} = \frac{\Delta^{2}}{12}} \qquad (10)$$

1.1.3 Signal to Quantization Noise ratio in PCM

The signal to quantization noise ratio is given as:

$$\boxed{\frac{S}{N_q} = \frac{\text{Normalized signal power}}{\text{Normalized noise power}}} \qquad \textbf{(11)}$$

$$= \frac{\text{Normalized signal power}}{\frac{\Delta^2}{12}} \qquad \textbf{(12)}$$

The number of quantization value is equal to:

$$q=2^v$$

Putting this value in eq(6), we get:

$$\Delta = \frac{2X_{\max}}{2^V}$$

Substitute this value in eq(12), we get

$$\frac{S}{N_q} = \frac{\text{Normalized signal power}}{\left[\frac{2X_{\max}}{2^v}\right]^2 * \frac{1}{12}}$$

Let the normalized signal power is equal to P then the signal to quantization noise will be given by:

$$\boxed{\frac{S}{N_q} = \frac{3P * 2^{2v}}{X_{\max}^2}} \qquad (13)$$

Examples(2)

A signal that has the highest frequency component of 4.2MHz and a peak to peak value of 4 volts is transmitted using a binary PCM. The number of quantization levels is 512 and P=0.04W calculate:

1. Code word length.
2. Bite rate.
3. output signal to quantization noise ratio.

Solution:

$$f_m=4.2\text{ MHz},\qquad q=512$$

$$512=2^{v}$$

$$\log 512=v\log 2$$

1. length of code word= v=9 bits

2. Bit rate $r= v\, f_s$

$$= v*(2\ f_m)$$

$$=9*2*4.2\text{MHz}$$

$$=75.6*10^{6}\text{ bits/sec}$$

3. $$\frac{S}{N_q}=\frac{3P*2^{2v}}{X^{2}_{\max}}$$

$$=\frac{3*0.04*2^{16}}{4}$$

$$=1966.08\approx 33\text{dB}$$

1.1.4 Advantages of PCM

1. Effect of noise is reduced.
2. PCM permits the use of pulse regeneration.
3. Multiplexing of various PCM signals is possible.

MULTIPLEXING

Multiplexing is the process of simultaneously transmitting two or more individual signals over a single communication channel. Multiplexing has the effect of increasing the number of communications channels in which more information can be transmitted. There are many instances in communications where it is necessary or desirable to transmit more than one voice or data signal. The application itself may require multiple signals, and money can be saved by using a single communication channel to send multiple information signals. Telephone and satellite systems use multiplexing to make the system practical and less expensive.

There are two basic types of multiplexing frequency division multiplexing (FDM) and Time division Multiplexing (TDM). Generally speaking, FDM systems are used to deal with analog systems, while TDM systems are used for digital systems. Of course, TDM are also found in many analog systems. The primary difference between these techniques is that in FDM individual signals to be transmitted are assigned a different frequency within a common bandwidth. In TDM, the multiple signals are transmitted in different time slots.

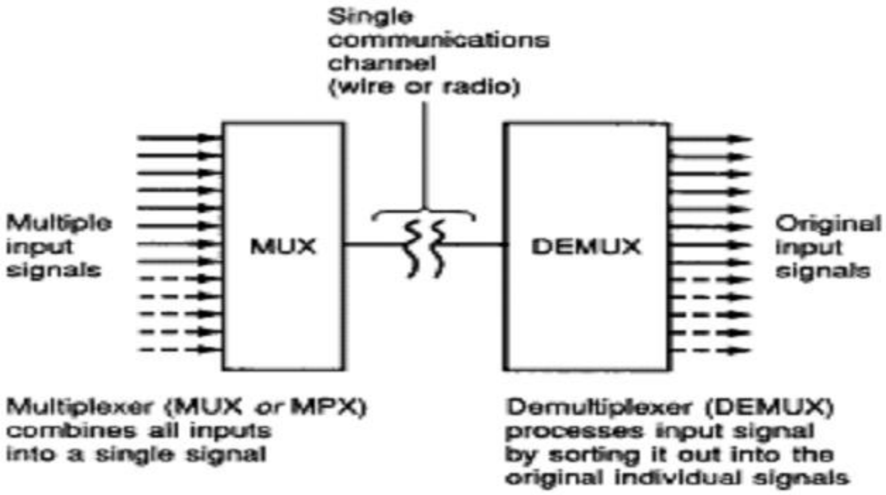

Single communications channel (wire or radio)
Multiple input signals
MUX
DEMUX
Original input signals
Multiplexer (MUX or MPX) combines all inputs into a single signal
Demultiplexer (DEMUX) processes input signal by sorting it out into the original individual signals

Figure: Concept of Multiplexing

Frequency Division Multiplexing (FDM)

Frequency Division Multiplexing (FDM) is a technique for transmitting multiple messages simultaneously over a wideband channel by first modulating the message signals onto several sub-carriers and forming a composite baseband signal that consists of the sum of these modulated sub-carriers. This composite signal may then be modulated on to the main carrier. Any type of modulation, such as AM, PM, FM can be used. The composite baseband signal then modulates a main transmitter to produce the FDM signal that is transmitted over the wide band channel.

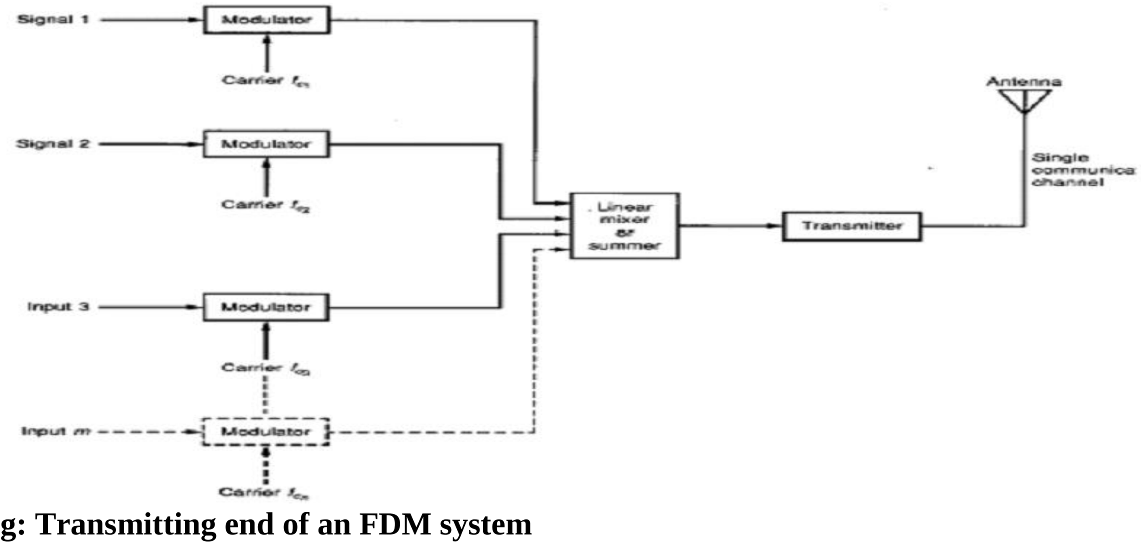


Fig: Transmitting end of an FDM system

The receiver portion of the system is shown below. It picks up the signal and demodulates it into the FDM signal. This is sent to a group of bandpass filters (BPF), each centered on one of the carrier frequencies. Each filter passes only its channel and rejects all others. A channel demodulator then recovers each original input signal.

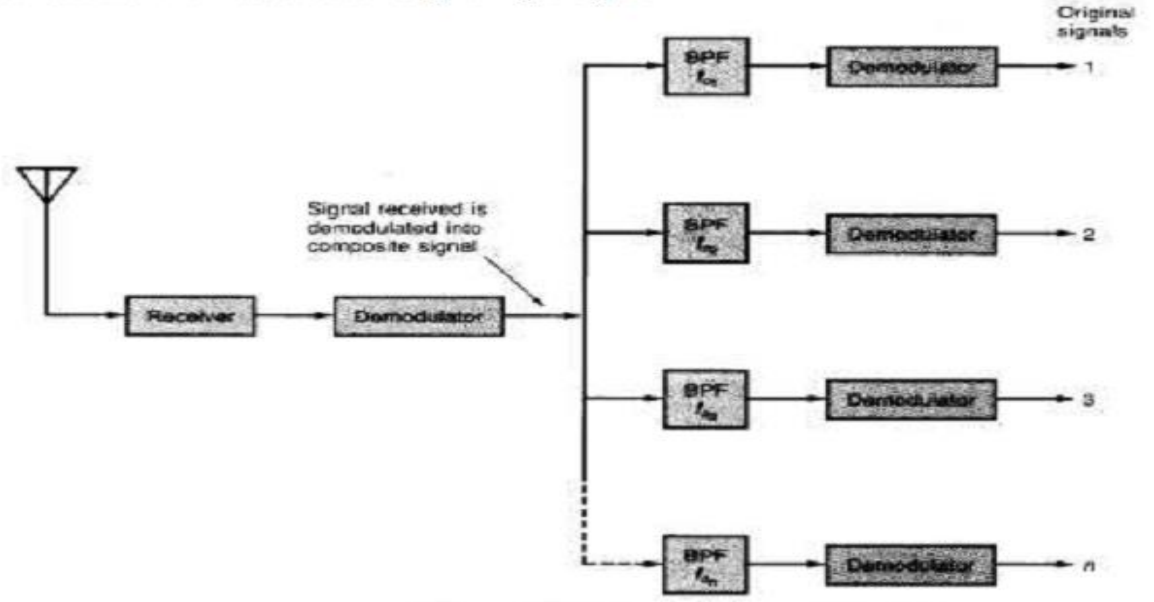


Fig: The receiving end of an FDM system.

An example of a commonly used FDM system is the telephone system. Telephone companies have been using FDM to send multiple telephone conversations over a minimum number of cables. The original signal is voice in the 300-3000Hz range. The voice is used to modulate a sub-carrier. Each sub-carrier is on a different frequency. These sub-carrier are then added together to form a single group. This multiplexing process is repeated at several levels so that very large number of users can communicate over a single communication channel.

FDM Telephone Hierarchy

In toll telephone service, voice signals are transmitted over high capacity channels using either TDM or FDM. TDM has become dominant and FDM is being phased out. The telephone FDM technique illustrated below is adapted by the American telephones telegraph company (AT&T) FDM hierarchy.

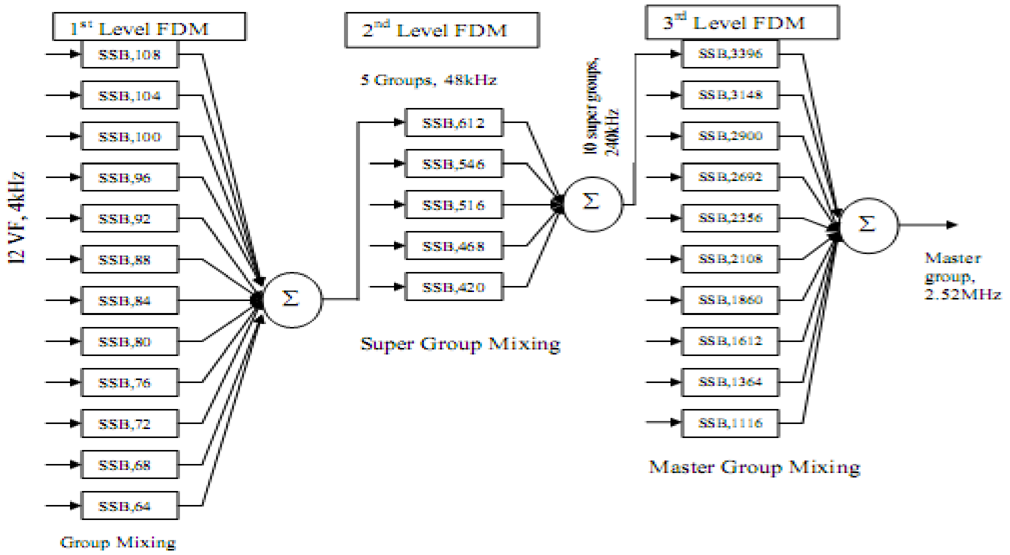
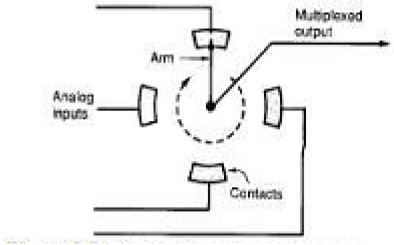


Fig: FDM Telephone Hierarchy

Time Division Multiplexing (TDM)

Time Division Multiplexing (TDM) is the time interleaving of samples from several sources so that the information from these sources can be transmitted serially over a single communication channel. Unlike FDM, in TDM each signal can occupy the entire bandwidth of the channel. However, each signal is transmitted for only a brief period of time. Each signal is allowed to use the channel for fixed period of time, one after another. Once all the signals have been transmitted, the cycle repeats again and again. One transmission of each channel completes one cycle of operation called a frame. The cycle repeats itself at a high rate of speed. In this way, the data bytes of the individual channels are simply interleaved. Frame synchronization is needed at the TDM receiver so that the received multiplexed data can be sorted and directed to the appropriate output channel. The frame sync can be provided to the receiver demultiplexer circuit either by sending a frame sync signal from the transmitter over a separate channel or by deriving the frame sync from the TDM signals itself. Any analog signal, be it voice or video can readily be transmitted by TDM techniques. This is accomplished by sampling the analog signal repeatedly at a high rate. By combining the concepts of TDM and PAM, you can see how multiple analog signals can be transmitted over a single channel. This type multiplexer is known as PAM-TDM system and it is shown below.



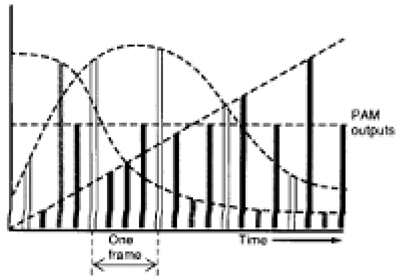


Fig: Simple rotary switch multiplier FigFour Channel PAM time division Multiplexer

In practical TDM/PAM systems, electronic circuits are used instead of mechanical switches or commutators. The multiplexer itself is usually implemented with FETs which are nearly ideal on/off switch that can switch at very high speed.

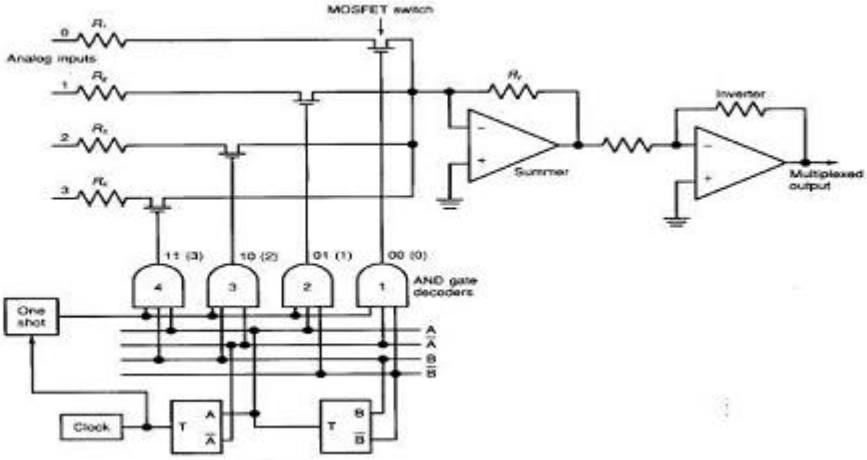
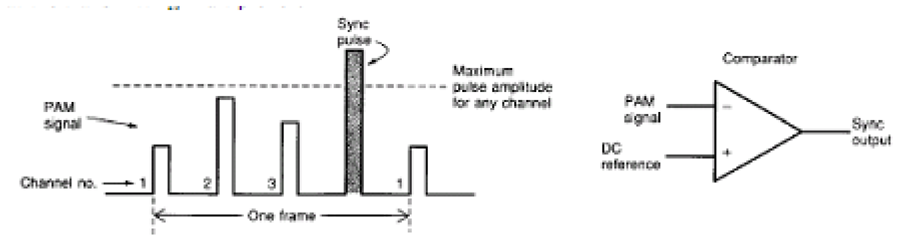


Figure: A PAM Multiplier

The main problem encountered in demultiplexing is synchronization. That is, in order for the PAM signal to be accurately demultiplexed into the original sampled signal, some method must be used to ensure that the clock frequency used on the DEMUX is identical to that used at the transmitting multiplexer.



Frame Synchronous Pulse and Comparator Detector

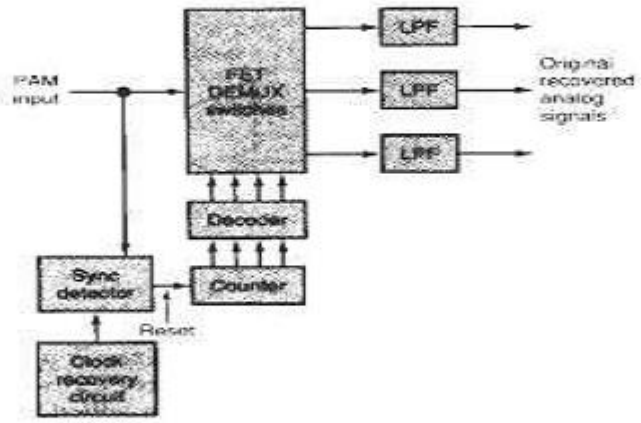


Figure: Complete PAM demultiplexer

Exercise:- Describe how the Complete PAM demultiplexer works. The most popular form of pulse modulation used in TDM systems is pulse code modulation. This system is known as PCM/TDM system. In PCM/TDM system multiple channels of serial digital data are transmitted with TDM by allowing each channel a timeslot in which to transmit one binary word of data. The system is depicted below. The multiplexer is done with a simple digital multiplexer. Since the entire signals to be transmitted are binary in nature, a multiplexer constructed of standard AND or NAND gates can be used. A binary counter drives a decoder that selects the desired input channel.

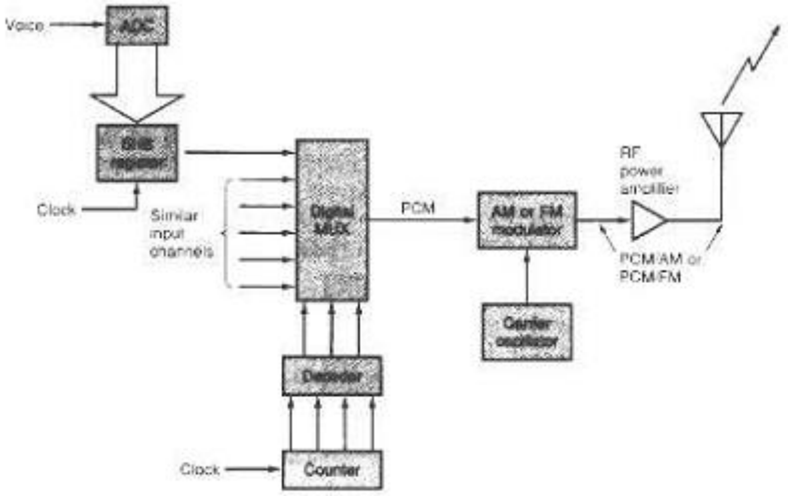


Figure: A PCM system

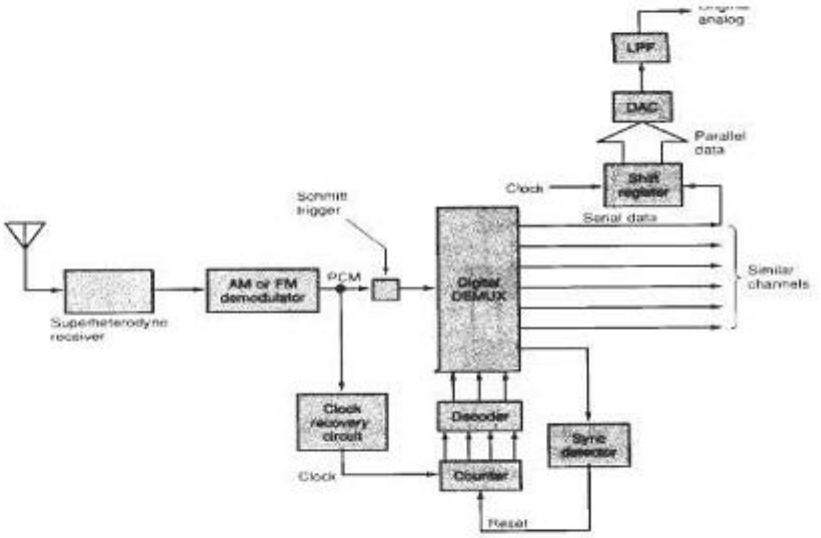


Figure: A PCM receiver demultiplexer

Synchronous and Asynchronous lines

In a synchronous system, each device is designed so that its internal clock is relatively stable for a long period of time, and it is synchronized to a system master clock. Each bit of data is clocked in synchronous with the master clock. In Asynchronous systems, the timing is precise only for the bits within each character or word. This is also called start stop signaling, because each character consists of start bit that start the receiver clock and concludes with one or two stop bits that terminates the clocking. The synchronous transmission is more efficient because start and stop bits are not required. However the synchronous mode of transmission requires that the clocking signal be passed along with the data and that the receiver synchronize to the clocking signal. Intelligent TDM may be used to concentrate data arriving from many different terminals or sources. They are capable of providing speed, code and protocol conversion.

TDM Telephone Hierarchy

In practice, TDM may be grouped in to two categories. The first category consists of multiplexer used in conjunction with digital computer systems to merge digital signals from several sources for TDM transmission over high-speed line to a digital computer. The second category of TDM is used by the common carriers such as the ATT, to combine different sources into a high-speed digital TDM signal for transmission over toll networks. Unfortunately, the standards adopted by North America and Japan are different from those that have been adopted in other parts of the world. The telephone industry has standardized the bit rates to 1.544Mbit/sec, 6.312Mbits/sec etc. and designates them as Ds-1 for digital signal, type 1, Ds-2 for digital signal type 2 etc.

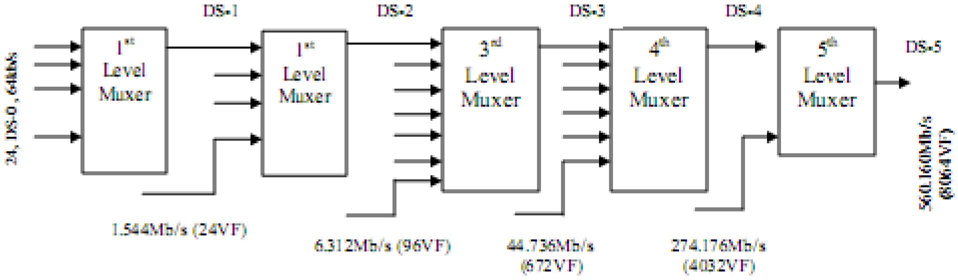


Figure: The N-American digital TDM Hierarchy

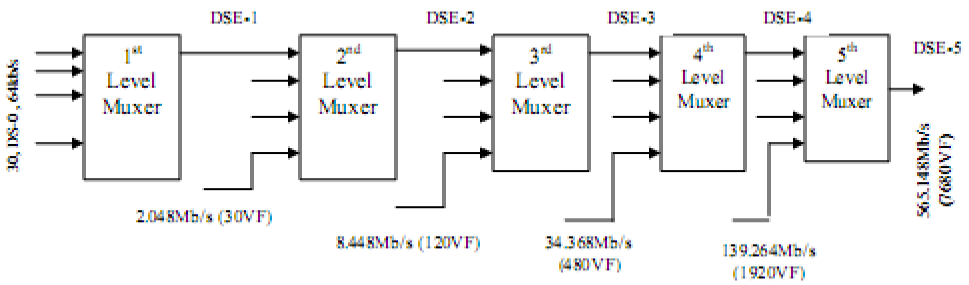


Figure: CCTTI digital TDM Hierarchy

The higher multiplexing level inputs are not always derived from lower level multiplexer. For example one analog television signal can be converted directly to Ds-3 data stream (47.73Mb/sec). Similarly, the Ds streams can carry a mixture of information from a variety of sources such as video, VF and computers. Lower Ds levels may be transmitted using twisted pair and higher once over coaxial, fiber optic cable, microwave radio or via satellite.

Other Multiple Access Methods

Code Division Multiple Access (CDMA)

In CDMA the information signals of different users are modulated by orthogonal or nonorthogonal spreading codes. The resulting spread signals simultaneously occupy the same time and bandwidth, as shown in figure 5.14. The receiver uses the spreading code structure to separate out the different users. The most common form of CDMA is multiuser spread spectrum with either DS or FH.

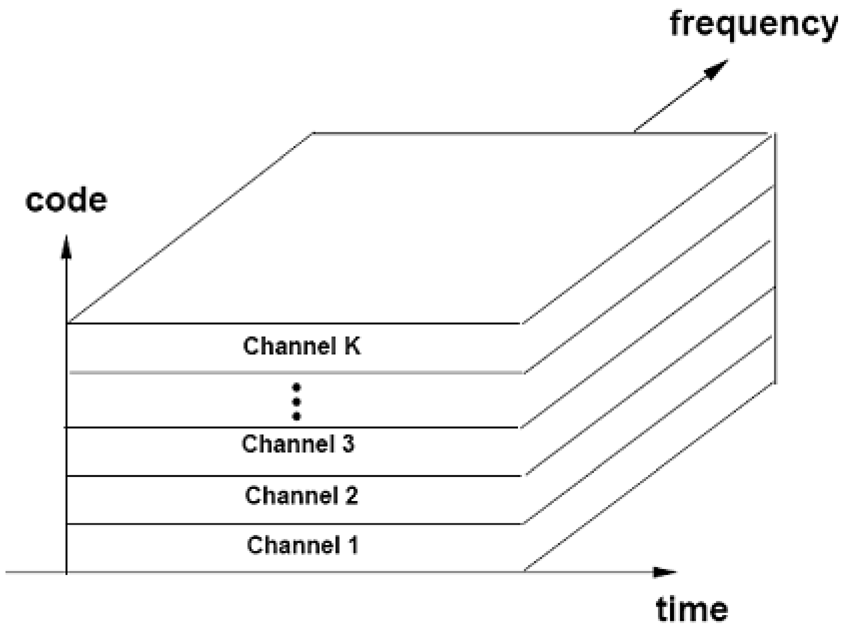


Figure: Code division multiplexing

Space Division Multiplexing (SDM)

Space-division multiplexing (SDM) uses direction (angle) as another dimension in signal space, which can be channelized and assigned to different users. This is generally done with directional antennas, as shown in figure 14.5. In practice SDM is often implemented using sectorized antenna arrays. In these arrays the 360^0 angular range is divided into N sectors. There is high directional gain in each sector and little interference between sectors. TDM or FDM is used to channelize users within a sector. For mobile users SDM must adapt as user angles change or, if directionality is achieved via sectorized antennas, then a user must be handed off to a new sector when it moves out of its original sector.

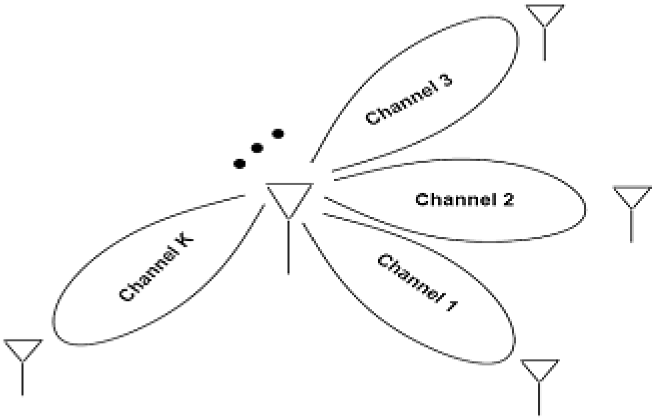


Figure: Space Division Multiplexing

Spread Spectrum Technology

Spread spectrum technology has blossomed from a military technology into one of the fundamental building blocks in current and next-generation wireless systems. From cellular to cordless to wireless LAN (WLAN) systems, spectrum is a vital component in the system design process.

Since spread-spectrum is such an integral ingredient, it's vital for designers to have an understanding of how this technology functions. In this lecture, we'll take on that task, addressing the basic operating characteristics of a spread-spectrum system. We'll also examine the key differentiators between frequency-hop spread spectrum (FHSS) and direct-sequence spread spectrum (DSSS) implementations.

How It Works

Spread spectrum uses wideband, noise-like signals that are hard to detect, intercept, or demodulate. Additionally, spread-spectrum signals are harder to jam (interfere with) than narrow band signals. These low probability of intercept (LPI) and anti-jam (AJ) features are why the military has used the spread spectrum scheme for so many years. Spread-spectrum signals are intentionally made to be a much wider band than the information they are carrying to make them more noise-like.

Spread-spectrum transmitters use similar transmit power levels to narrowband transmitters. Because spread-spectrum signals are so wide, they transmit at a much lower spectral power density, measured in watts per hertz, than narrow band transmitters. This lower transmitted power density characteristic gives spread-spectrum signals a big plus. Spread-spectrum and narrowband signals can occupy the same band, with little or no interference. This capability is the main reason for all the interest in spread spectrum today.

The use of special pseudo noise (PN) codes in spread-spectrum communications makes signals appear wide band and noise-like. It is this very characteristic that makes spread-spectrum signals possess a low LPI. Spread-spectrum signals are hard to detect on narrow band equipment because the signal's energy is spread over a bandwidth of maybe 100 times the information bandwidth **(Figure 1)**.

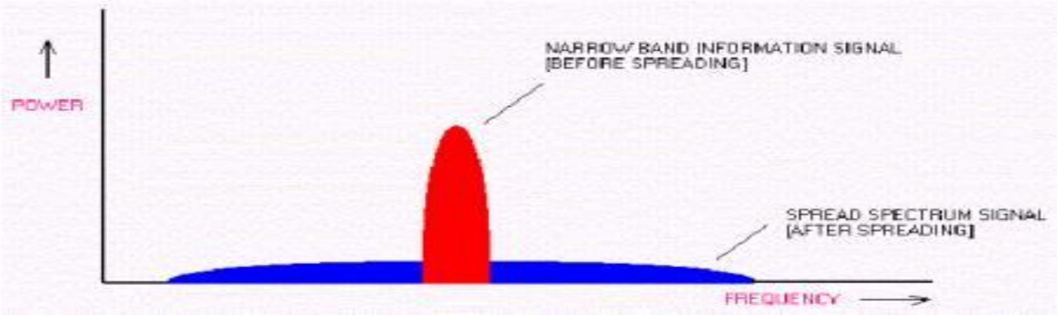


Figure 1: In a spread-spectrum system, signals are spread across a wide bandwidth, making them difficult to intercept, demodulate, and intercept.

The spread of energy over a wide band, or lower spectral power density, also makes spread-spectrum signals less likely to interfere with narrowband communications. Narrowband communications, conversely, cause little or no interference to spread spectrum systems because the correlation receiver effectively integrates over a very wide bandwidth to recover a spread spectrum signal. The correlator then "spreads" out a narrowband interferer over the receiver's total detection bandwidth.

Since the total integrated signal density or signal-to-noise ratio (SNR) at the correlator's input determines whether there will be interference or not. All spread spectrum systems have a threshold or tolerance level of interference beyond which useful communication ceases. This tolerance or threshold is related to the spread-spectrum processing gain, which is essentially the ratio of the RF bandwidth to the information bandwidth.

Direct or Hopping
Direct sequence and frequency hopping are the most commonly used methods for the spread spectrum technology. Although the basic idea is the same, these two methods have many distinctive characteristics that result in complete different radio performances.

The carrier of the direct-sequence radio stays at a fixed frequency. Narrowband information is spread out into a much larger (at least 10 times) bandwidth by using a pseudo-random chip sequence. The generation of the direct sequence spread spectrum signal (spreading) is shown in **Figure 2**.

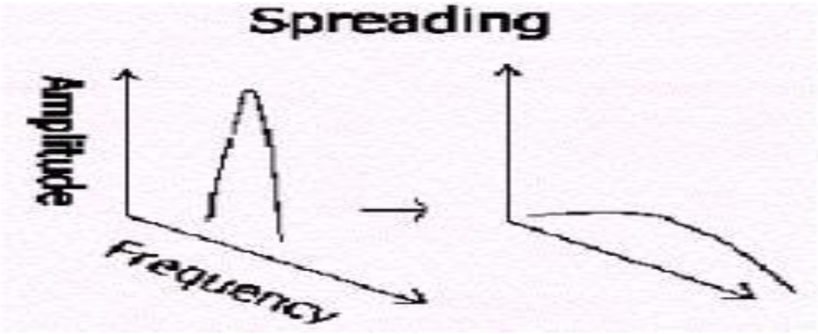


Figure 2: Comparison of the generation of a narrowband and direct-sequence spread spectrum signals.

In Figure 2, the narrowband signal and the spread-spectrum signal both use the same amount of transmit power and carry the same information. However, the power density of the spread-spectrum signal is much lower than the narrowband signal. As a result, it is more difficult to detect the presence of the spread spectrum signal. The power density is the amount of power over a certain frequency. In the case of Figure 2, the narrowband signal's power density is 10 times higher than the spread spectrum signal, assuming the spread ratio is 10.

At the receiving end of a direct-sequence system, the spread spectrum signal is de-spread to generate the original narrowband signal. **Figure 3** shows the de-spreading process.

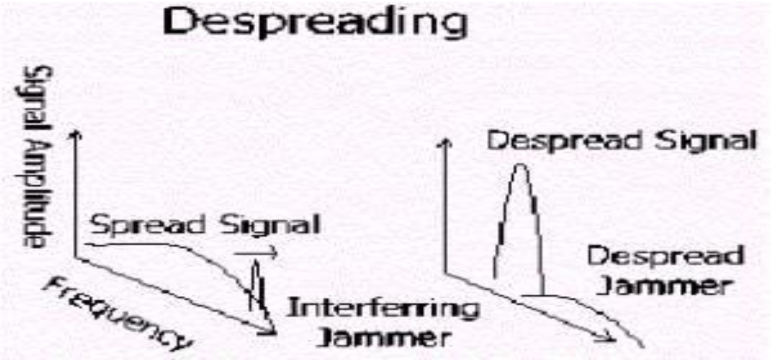


Figure 3: Diagram illustrating the despreading process in a direct-sequence system.

If there is an interference jammer in the same band, it will be spread out during the de-spreading. As a result, the jammer's impact is greatly reduced. This is the way that the direct-sequence spread-spectrum (DSSS) radio fights the interference. It spreads out the offending jammer by the spreading factor **(Figure 4)**. Since the spreading factor is at least a factor of 10, the offending jammer's amplitude is greatly reduced by at least 90%.

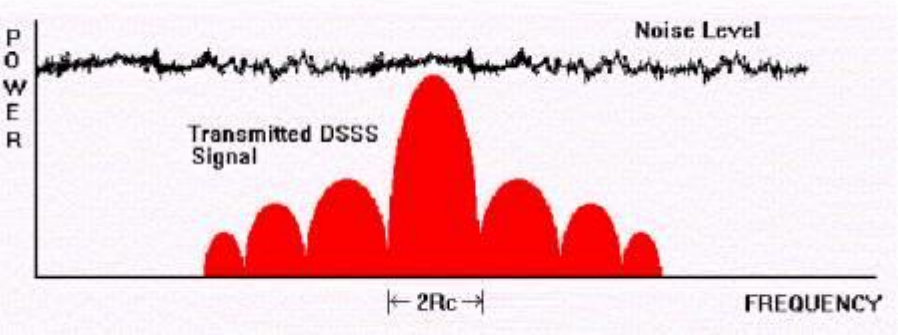


Figure 4: Direct-sequence systems combat noise problems by spreading jammers across a wideband as shown in this figure.

The Hopping Approach

Frequency-hopping systems achieve the same results provided by direct-sequence systems by using different carrier frequency at different time. The frequency-hop system's carrier will hop around within the band so that hopefully it will avoid the jammer at some frequencies. A frequency-hopping signal is shown in **Figure 5a and 5b**.

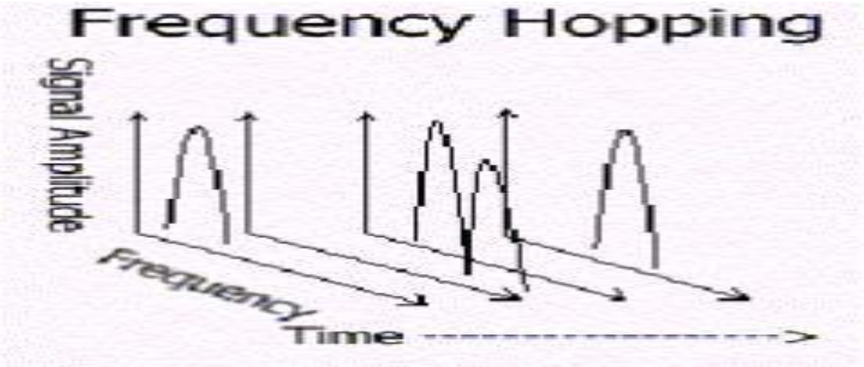


Figure 5: Diagram showing how a frequency-hop system works.

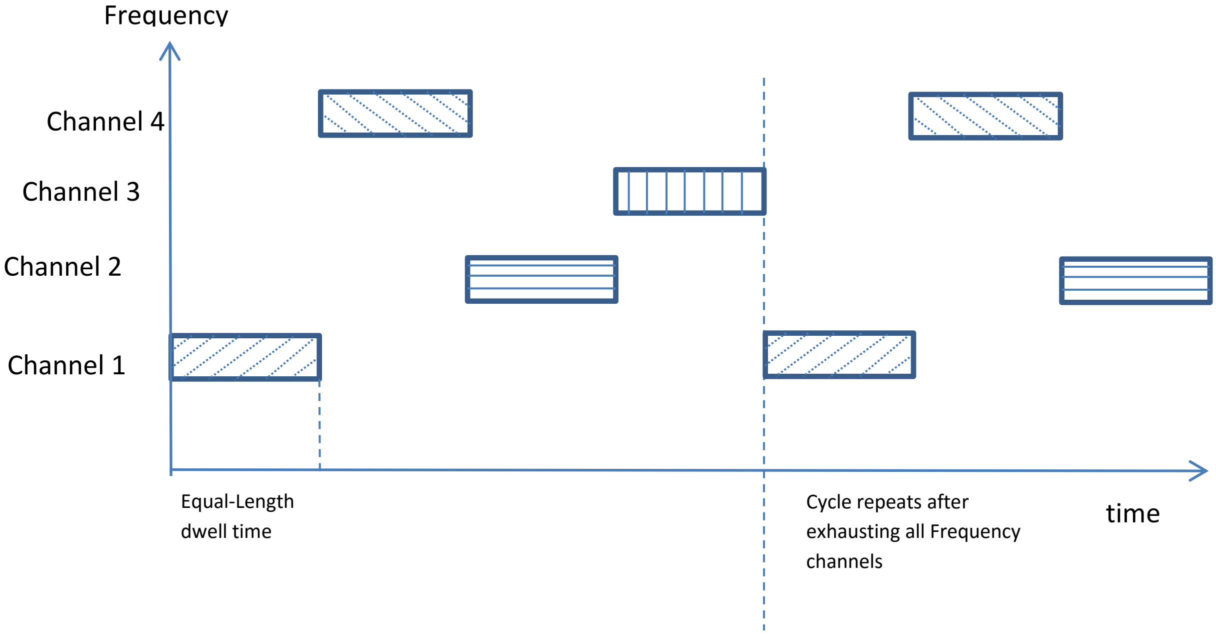


Figure 5b: A four channel FHSS system (Obaidat, et al, 2011).

The frequency-hopping technique does not spread the signal, as a result, there is no processing gain. The processing gain is the increase in power density when the signal is de-spread and it will improve the received signal's Signal-to-noise ratio (SNR). In other words, the frequency hopper needs to put out more power in order to have the same SNR as a direct-sequence radio.

The frequency hopper is also more difficult to synchronize. In these architectures, the receiver and the transmitter must be synchronized in time and frequency in order to ensure proper transmission and reception of signals. In a direct-sequence radio, on the other hand, only the timing of the chips needs to be synchronized.

The frequency hopper also needs more time to search the signal and lock to it. As a result, the latency time is usually longer. While a direct-sequence radio can lock in the chip sequence in just a few bits.

To make the initial synchronization possible, the frequency hopper will typically park at a fixed frequency before hopping or communication begins. If the jammer happens to locate at the same frequency as the parking frequency, the hopper will not be able to hop at all. And once it hops, it will be very difficult, if not impossible to re-synchronize if the receiver ever lost the sync.

The frequency hopper, however, is better than the direct-sequence radio when dealing with multipath. Since the hopper does not stay at the same frequency and a null at one frequency is usually not a null at another frequency if it is not too close to the original frequency. So a hopper can usually deal with multipath fading issues better than direct-sequence radio.

The hopper itself, however, could suffer performance problems if it interferes with another radio. In these scenarios, the system that survives depends upon which can suffer more data loss. In general, a voice system can survive an error rate as high as 10^{-2} while a data system must have an error rate better than 10^{-4}. Voice system can tolerate more data loss because human brain can "guess" between the words while a dumb microprocessor can't.

Modulation and Demodulation

For direct-sequence systems the encoding signal is used to modulate a carrier, usually by phase-shift keying (PSK; for example, bi-phase or quad-phase) at the code rate. Frequency-hopping systems generate their wide band by transmitting at different frequencies, hopping from one frequency to another according to the code sequence. Typically such a system may have a few thousand frequencies to choose from, and unlike direct sequence signal, it has only one output rather than symmetrically distributed outputs.

It's important to note that for both direct-sequencing and frequency-hopping systems generate wideband signals controlled by the code sequence generator. For one the code is the direct carrier modulation (direct sequence) and the other commands the carrier frequency (frequency hopping).

There are several different modulation techniques that designers can employ when developing frequency-hop or direct-sequence systems. Information modulation can be accomplished using amplitude (AM) or frequency modulation (FM) techniques. AM is normally used because it tends to be detectable when examining the spectrum. FM is more useful because it is a constant-envelope signal, but information is still readily observed. In both AM and FM, no knowledge of the code is needed to receive the transmitted information.

Clock modulation, which is actually frequency modulation of the code clock, is another option in spread-spectrum designs. In most cases (including frequency hopping), clock modulation is not used because the loss in correlation due to phase slippage between received and local clocks, could cause degraded performance.

Code modification is another modulation technique that designers can use when building a spread-spectrum system. Under this approach, the code is changed in such a way that the information is embedded in it, then modulated by phase transitions on a RF carrier.

In direct-sequence designs, balance modulation can be used in any suppressed carrier system used to generate the transmitted signal. Balanced modulation helps to hide the signal, as well as there are no power wasted in transmitting a carrier that would contribute to interference rejection or information transfer. When a signal has poor balance in either code or carrier, spikes are seen in its spectrum. With these spikes, or spurs, the signal is easily detectable, since these spikes are noticed above the noise and thus provide a path for detecting the hidden signal.

Once the signal is coded, modulated and then sent, the receiver must demodulate the signal. This is usually done in two steps. The first step entails removing the spectrum-spreading modulation. Then, the remaining information-bearing signal is demodulated by multiplying with a local reference identical in structure and synchronized with the received signal.

Coding Techniques

In order to transmit anything, codes used for data transmission have to be considered. However, this section will not discuss the coding of information (like error correction coding) but those that act as noise-like carriers for the information being transferred. These codes are of much greater length than those for the usual areas of data transfer, since it is intended for bandwidth spreading.

Codes in a spread-spectrum system are used for:

1. Protection against interference: Coding enables a bandwidth trade for processing gain against interfering signals.
2. Provision for privacy: Coding enables protection of signals from eaves dropping, so that even the code is secure.
3. Noise-effect reduction: error-detection and correction codes can reduce the effects of noise and interference.

Maximal sequencing is one of the more popular coding methods in a spread-spectrum system. Maximal codes can be generated by a given shift register or a delay element of given length. In binary shift register sequence generators, the maximum length sequence is (2^n-1) chips, where n is the number of stages in the shift register.

A shift register generator consists of a shift register in conjunction with the appropriate logic, which feeds back a logical combination of the state of two or more of its stages to its input. The output, and its contents of its n stages at any clock time, is its function of the outputs of the stages fed back at the proceeding sample time. Some maximal codes can be of length 7 to [(2^{36}-1] chips.

Error detection and correction codes (EDAC) must be used in frequency-hopping systems in order to overcome the high rates of error induced by partial band jamming. These codes usefulness has a threshold that must be exceeded before satisfactory performance is achieved.

In direct-sequence systems, EDACs may not be advisable because of the effect it has on the code, increasing the apparent data transmission rate, and may increase jamming threshold. Some demodulators can operate detecting errors at approximately the same accuracy as an EDAC, so it may not be worthwhile to include a complex coding/decoding scheme in the system.

Advantages of Spread Spectrum
Spread-spectrum systems provide some clear advantages to designers. As a recap, here are nine benefits that designers can expect when using a spread-spectrum-based wireless system.

1. Reduced crosstalk interference: In spread-spectrum systems, crosstalk interference is greatly attenuated due to the processing gain of the spread spectrum system as described earlier. The effect of the suppressed crosstalk interference can be essentially removed with digital processing where noise below certain threshold results in negligible bit errors. These negligible bit errors will have little effect on voice transmissions.

2. Better voice quality/data integrity and less static noise: Due to the processing gain and digital processing nature of spread spectrum technology, a spread-spectrum-based system is more immune to interference and noise. This greatly reduces consumer electronic device-induced static noise that is commonly experienced by conventional analog wireless system users.

3. Lowered susceptibility to multipath fading: Because of its inherent frequency diversity properties (thanks to wide spectrum spread), a spread spectrum system is much less susceptible to multipath fading.

4. Inherent security: In a spread spectrum system, a PN sequence is used to either modulate the signal in the time domain (direct sequence systems) or select the carrier frequency (frequency hopping systems). Due to the pseudo-random nature of the PN sequence, the signal in the air has been "randomized". Only the receiver having the exact same pseudo-random sequence and synchronous timing can de-spread and retrieve the original signal. Consequently, a spread spectrum system provides signal security that is not available to conventional analog wireless systems.

5. Co-existence: A spread spectrum system is less susceptible to interference than other non-spread spectrum systems. In addition, with the proper designing of pseudo-random sequences, multiple spread spectrum systems can co-exist without creating severe interference to other systems. This further increases the system capacity for spread spectrum systems or devices.

6. Longer operating distances: A spread spectrum device operated in the ISM band is allowed to have higher transmit power due to its non-interfering nature. Because of the higher transmit power, the operating distance of such a device can be significantly longer than that of a traditional analog wireless communication device.

7. Hard to detect: Spread-spectrum signals are much wider than conventional narrowband transmission (of the order of 20 to 254 times the bandwidth of narrowband transmissions). Since the communication band is spread, it can be transmitted at a low power without being detrimentally affected by background noise. This is because when de-spreading takes place, the noise at one frequency is rejected, leaving the desired signal.

8. Hard to intercept or demodulate: The very foundation of the spreading technique is the code use to spread the signal. Without knowing the code it is impossible to decipher the transmission. Also, because the codes are so long (and quick) simply viewing the code would still be next to impossible to solve the code, hence interception is very hard.

9. Harder to jam: The most important feature of spread spectrum is its ability to reject interference. At first glance, it may be considered that spread spectrum would be most affected by interference. However, any signal is spread in the bandwidth, and after it passes through the correlator, the bandwidth signal is equal to its original bandwidth, plus the bandwidth of the local interference. An interference signal with 2 MHz bandwidth being input into a direct-sequence receiver whose signal is 10 MHz wide gives an output from the correlator of 12 MHz. The wider the interference bandwidth, the wider the output signal. Thus the wider the input signal, the less its effect on the system because the power density of the signal after processing is lower, and less power falls in the band pass filter.

Code-Division Multiple Access (CDMA)

CDMA (Code-Division Multiple Access) refers to any of several protocols used in so-called second-generation (2G) and third-generation (3G) wireless communications. As the term implies, CDMA is a form of multiplexing, which allows numerous signals to occupy a single transmission channel, optimizing the use of available bandwidth. The technology is used in ultra-high-frequency (UHF) cellular telephone systems in the 800-MHz and 1.9-GHz bands. CDMA employs analog-to-digital conversion (ADC) in combination with spread spectrum technology. Audio input is first digitized into binary elements. CDMA can either use DSSS of FHSS.

In FHSS based CDMA, after the ADC conversion, the frequency of the transmitted signal is made to vary according to a defined pattern (code), so it can be intercepted only by a receiver whose frequency response is programmed with the same code so that it follows exactly along with the transmitter frequency. There are trillions of possible frequency-sequencing codes, which enhance privacy and makes cloning difficult.

CDMA can also be implemented using *Direct-Sequence Spread-Spectrum* (DSSS) and radio technology that originally achieved prominence in military communications systems, and then in early wireless LANs. The idea in DSSS is simple -- instead of sending 0s and 1s over the air directly, we convert each 0 and 1 to a longer string of bits, which is the "code." This may appear to waste bandwidth, but the technique in fact improves reliability because damage to one or two bits during transmission need not require that the entire packet of data be resent. Rather, when the signal is de-spread by the receiver, we can use statistical techniques to guess what the original bit was. Using the right codes, we can often guess correctly (and we still use error-checking codes at the end of each packet, regardless). Now, suppose we pick the codes that are *orthogonal* to one another, meaning that two properly designed orthogonal codes can actually exist in the same spectrum at the same time and not -- really! -- interfere with each other. We'd give one code to one user and another to a second user and so on, and then, assuming everyone transmits at the same power level relative to one another so that no one station drowns out the others. This is how the DSSS based CDMA is implemented.

The CDMA channel is nominally 1.23 MHz wide. CDMA networks use a scheme called soft handoff, which minimizes signal breakup as a handset passes from one cell to another. The combination of digital and spread-spectrum modes supports several times as many signals per unit bandwidth as analog modes. CDMA is compatible with other cellular technologies; this allows for nationwide roaming.

The original CDMA standard, also known as CDMA One and still common in cellular telephones in the U.S., offers a transmission speed of only up to 14.4 Kbps in its single channel form and up to 115 Kbps in an eight-channel form. CDMA2000 and Wideband CDMA deliver data many times faster.

References

1. ISSCE—Circuits Systems. "Principles of Spread Spectrum Communication", target="_new">http://olt.et.tudelf.nl/~glas/ssc/techn/techniques.html
2. Dixon, Robert C. "Spread Spectrum Techniques". John Wiley & Sons.
3. Internet Magazine. "Spread Spectrum Scene: ABC's of Spread Spectrum", http://sss-mag.com/ss.html
4. http://searchtelecom.techtarget.com/definition/CDMA
5. http://searchtelecom.techtarget.com/tip/3G-The-CDMA-alternative

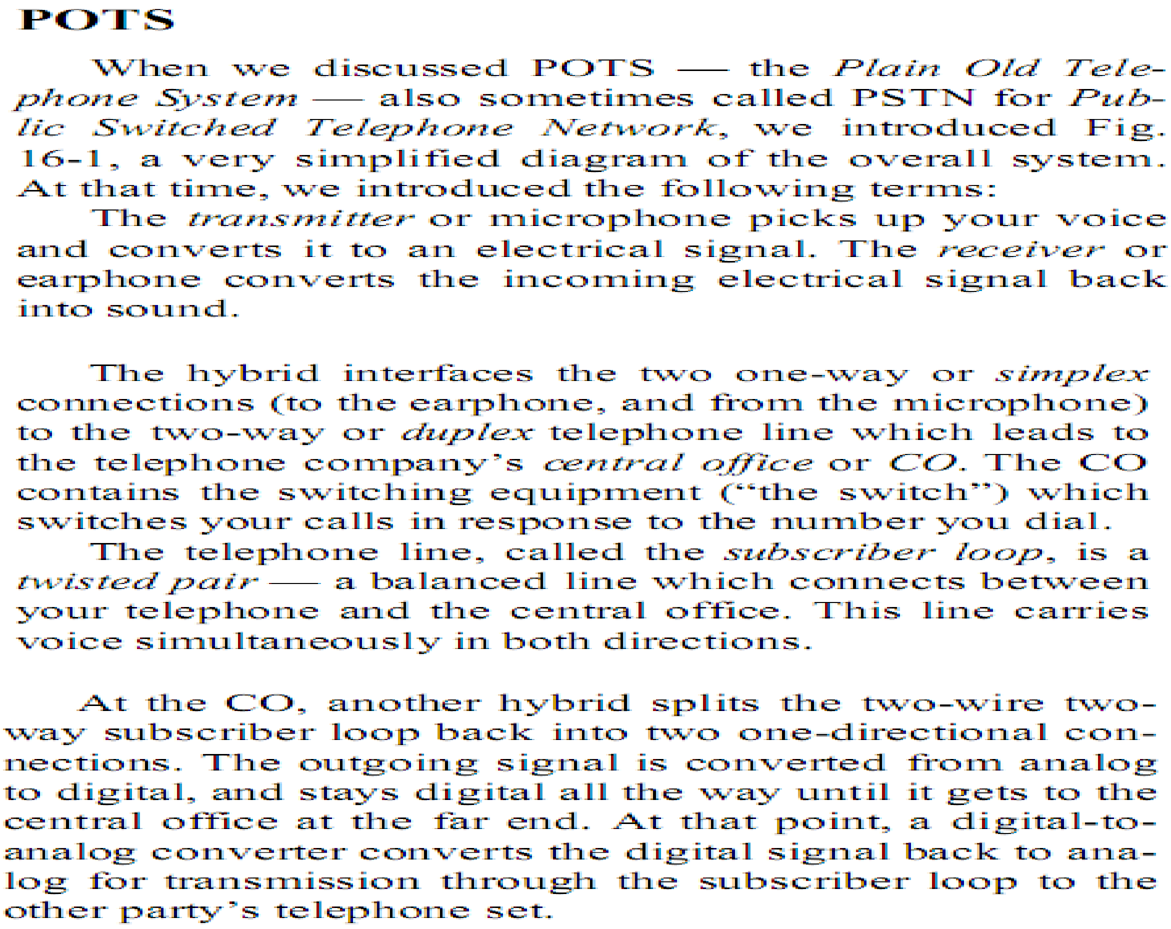
TELEPHONY

Telephony is the technology associated with the electronic transmission of voice, fax, or other information between distant parties using systems historically associated with the telephone, a handheld device containing both a speaker or transmitter and a receiver. With the arrival of computers and the transmittal of digital information over telephone systems and the use of radio to transmit telephone signals, the distinction between *telephony* and *telecommunication* has become difficult to make.

Internet telephony is the use of the Internet rather than the traditional telephone company infrastructure and rate structure to exchange spoken or other telephone information. Since access to the Internet is available at local phone connection rates, an international or other long-distance call will be much less expensive than through the traditional call arrangement.

On the Internet, three new services are now available:

- The ability to make a normal voice phone call (whether or not the person called is immediately available; that is, the phone will ring at the location of the person called) through the Internet at the price of a local call
- The ability to send fax transmissions at very low cost (at local call prices) through a gateway point on the Internet in major cities
- The ability to send voice messages along with text e-mail



POTS

When we discussed POTS — the *Plain Old Telephone System* — also sometimes called PSTN for *Public Switched Telephone Network*, we introduced Fig. 16-1, a very simplified diagram of the overall system. At that time, we introduced the following terms:

The *transmitter* or microphone picks up your voice and converts it to an electrical signal. The *receiver* or earphone converts the incoming electrical signal back into sound.

The hybrid interfaces the two one-way or *simplex* connections (to the earphone, and from the microphone) to the two-way or *duplex* telephone line which leads to the telephone company's *central office* or *CO*. The CO contains the switching equipment ("the switch") which switches your calls in response to the number you dial.

The telephone line, called the *subscriber loop*, is a *twisted pair* — a balanced line which connects between your telephone and the central office. This line carries voice simultaneously in both directions.

At the CO, another hybrid splits the two-wire two-way subscriber loop back into two one-directional connections. The outgoing signal is converted from analog to digital, and stays digital all the way until it gets to the central office at the far end. At that point, a digital-to-analog converter converts the digital signal back to analog for transmission through the subscriber loop to the other party's telephone set.

The conversion is done at a sampling rate of 8000 times per second and therefore the anti-aliasing filters in the system cut off all audio above roughly 3500.

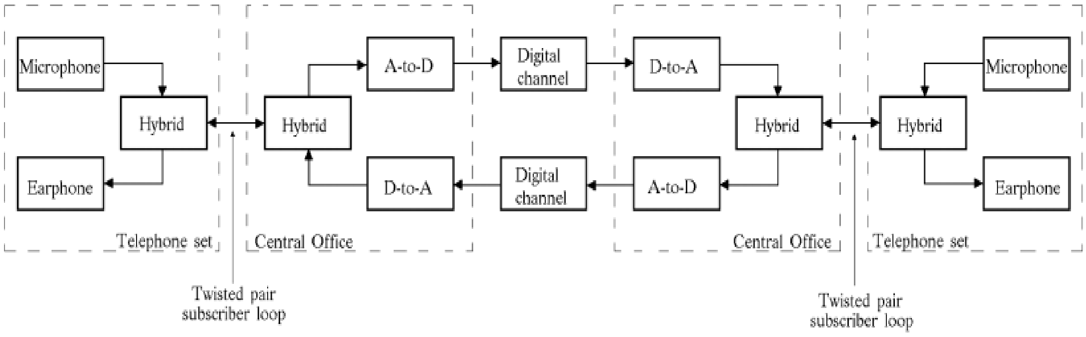


Fig. 16-1. Overall block diagram of the telephone system

The subscriber loop

The subscriber loop is an unshielded pair of (generally) 24- or 26-gauge wires. It is a balanced and twisted pair; the balanced connection helps to reduce the pickup of outside noise and hum, as well as crosstalk from other, adjacent wire pairs.

Because the wires are thin and close together, there is a sizable capacitance between them. This makes the circuit into a low-pass filter, which reduces the high-frequency response. Even with short cables, at just 3000 or 3500 Hz these high frequencies are attenuated and result in a noticeable lack of treble. The curve labelled "plain cable" in Fig. 16-2 shows the typical frequency response of a subscriber loop several thousand feet long.

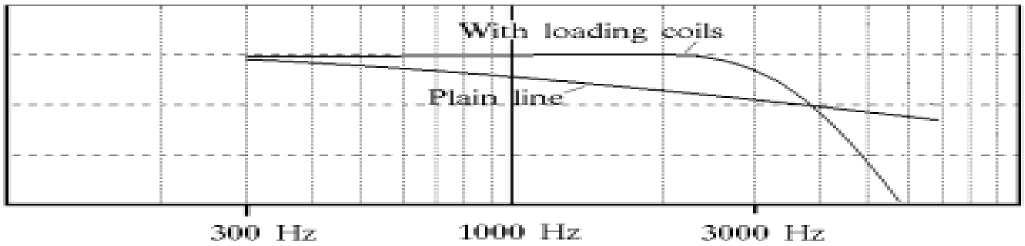


Fig. 16-2. Subscriber loop frequency response

In order to improve the high frequency response, the telephone company therefore often inserts *loading coils* into the subscriber loop. These are toroidal inductors, most often 44, 66, or 88 millihenries (see Fig. 16-3), which are connected in series with the line. These are not as common in large cities, where the distance from your phone to the nearest CO may be fairly small, but appear quite often in the suburbs or out in the country.

The loading coil is also sometimes called a *peaking coil*; it resonates with the line capacitance and produces a peak in the frequency response somewhere between 3000 and 4000 Hz; this increases the high-frequency re-

Fig. 16-3. 88-mhy loading coil

sponse of the line, as shown in the "with loading coils" curve in Fig. 16-2. But you can see that, although the loading coil improves the frequency response in the high audio range up to about 3500 Hz or so, it actually makes things worse above that. (It also affects the phase of signals.) Signals above 4000 or 5000 Hz (as well as harmonics of any pulse or digital signals) are now almost totally attenuated. This explains why, when the telephone company decides to use an existing pair of wires for a digital circuit, it must remove any loading coils that were previously used for voice circuits.

Your telephone set

Also called a subscriber set (part of the CPE or *Customer Premises Equipment*), today your telephone is an analog instrument which converts between sound and electrical signals. In addition to the microphone and earphone, the telephone set also contains the dial (either a pulse dial or a tone dial which emits *DTMF* — *Dual Tone Multiple Frequencies*) and a bell (called a *ringer* in telephone parlance.) Fig. 16-4 shows a very simplified circuit of how the instrument connects to the rest of the network.

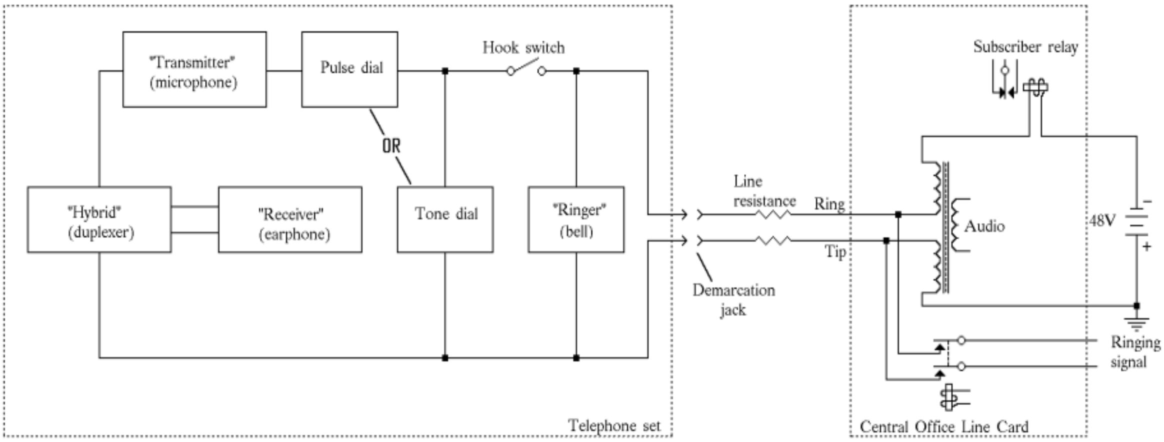


Fig. 16-4. Block diagram of a typical telephone set

In the central office (shown at the right of Fig. 16-4) is a 48-volt battery. This is generally a series of wet cells (i.e., batteries with a wet electrolyte, much like the lead-acid battery in a car) connected in series to provide 48 volts. The advantage of using batteries is that they provide power even if the local electric utility fails. These batteries are constantly being charged, however, so that the actual voltage is a bit higher — typically 50 volts or so.

When you are not using the telephone, the hook switch (*switch hook* in telephone talk) is open, and only the ringer is connected to the line. A capacitor in series with the ringer prevents dc current flow through the ringer; hence the telephone set looks like an open dc circuit and there is no current through the loop. Even though there is some resistance in the subscriber loop (as well as in the circuitry in the CO), the absence of current means there is no voltage drop, and so the full 50 volts appears across your telephone set. The lack of current tells the CO that your phone is *on hook*.

When there is an incoming call, the telephone company rings your phone by sending an ac ringing voltage of 100 volts at 20 Hz. Since there is a capacitor in series with your ringer, the ac ringing signal is sent to the ringer. The ringer requires very little current, and thus rings.

When you pick up the handset to answer the phone, you close the hook switch under the handset. The name dates back to when an earphone or handset would hang on a hook at the side of the phone; the switch would close when you picked up the handset, and this was called *going off hook*). When the hook switch closes, dc current can now pass through the telephone. We show in Fig. 16-4 that the dc current passes through a subscriber relay; its closing then tells the telephone company that you have picked up the phone, at which point it stops ringing your bell. (In modern central offices, an integrated circuit is used instead of a relay to sense that dc current, as we will see shortly.)

Fig. 16-5 shows the schematic diagram of an actual (older!) telephone set. Keep in mind that there are many ways to wire a phone; this diagram shows an older 500-style rotary (pulse) dial set of the type in Fig. 16-6, and your phone may vary.

The ringer circuit consists of two coils separated by a 0.5 µF capacitor. The capacitor prevents dc current from flowing through the coils. This makes sure that the CO doesn't mistake the bell for an off-hook telephone.

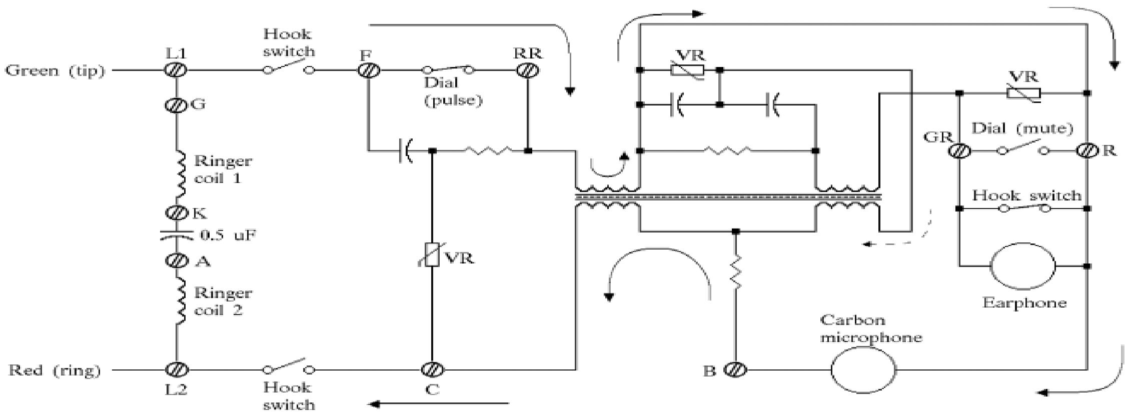
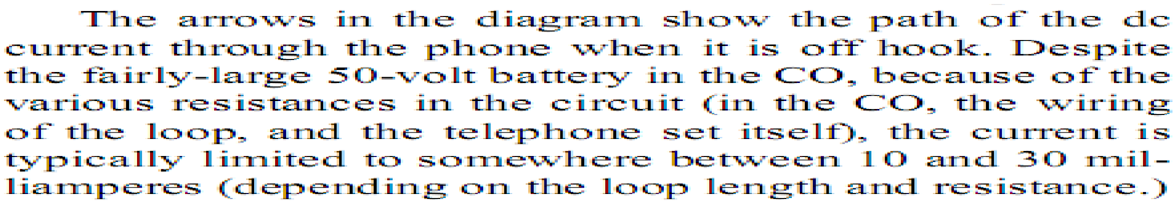


Fig. 16-5. Actual telephone schematic diagram



The arrows in the diagram show the path of the dc current through the phone when it is off hook. Despite the fairly-large 50-volt battery in the CO, because of the various resistances in the circuit (in the CO, the wiring of the loop, and the telephone set itself), the current is typically limited to somewhere between 10 and 30 milliamperes (depending on the loop length and resistance.)

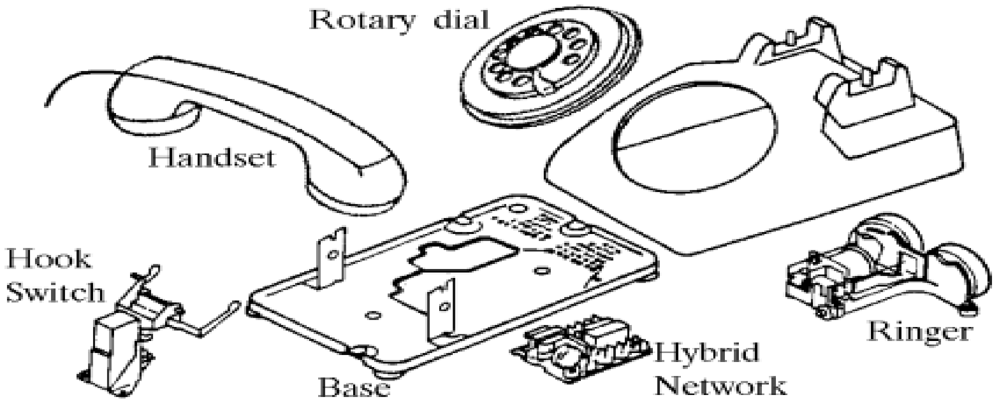
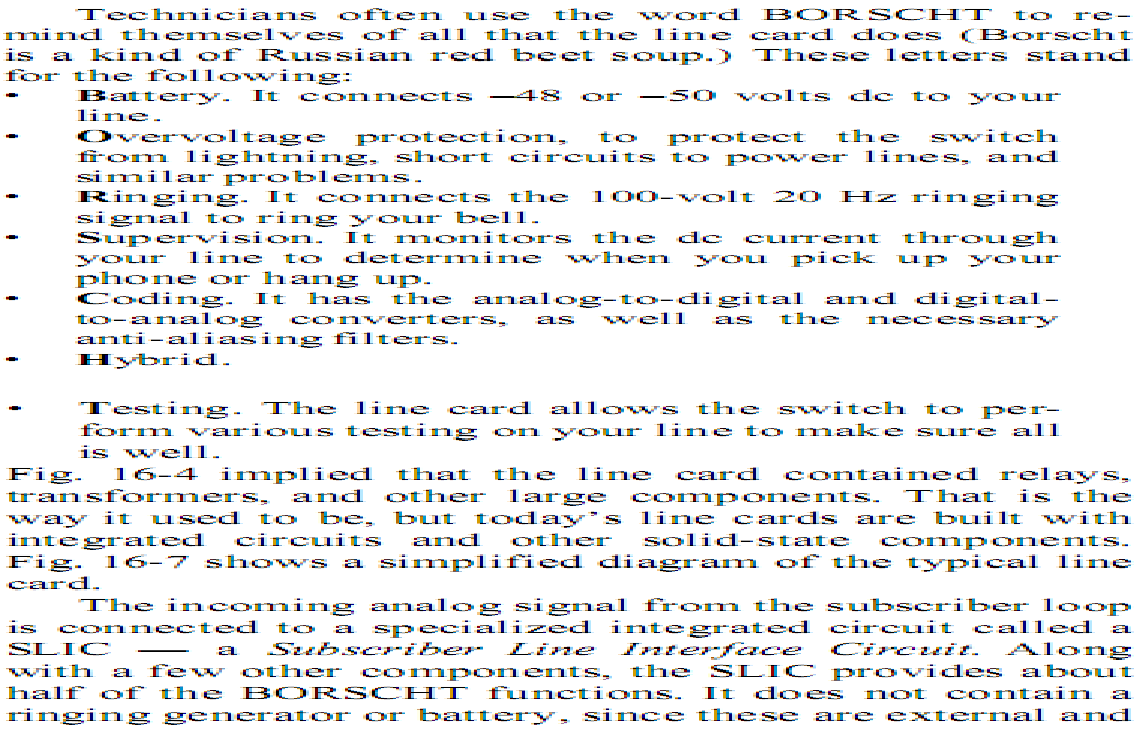


Fig. 16-6. A disassembled older telephone instrument

The Line Card

Located at the Telephone Company's Central Office, the switch does the switching of calls. Much of the circuitry in the switch is common to all users, but some circuitry must be duplicated for each line. As a result all the in-coming subscriber loops terminate in line cards, which handle just a small number of subscribers each.



Technicians often use the word BORSCHT to remind themselves of all that the line card does (Borscht is a kind of Russian red beet soup.) These letters stand for the following:

- Battery. It connects —48 or —50 volts dc to your line.
- Overvoltage protection, to protect the switch from lightning, short circuits to power lines, and similar problems.
- Ringing. It connects the 100-volt 20 Hz ringing signal to ring your bell.
- Supervision. It monitors the dc current through your line to determine when you pick up your phone or hang up.
- Coding. It has the analog-to-digital and digital-to-analog converters, as well as the necessary anti-aliasing filters.
- Hybrid.

- Testing. The line card allows the switch to perform various testing on your line to make sure all is well.

Fig. 16-4 implied that the line card contained relays, transformers, and other large components. That is the way it used to be, but today's line cards are built with integrated circuits and other solid-state components. Fig. 16-7 shows a simplified diagram of the typical line card.

The incoming analog signal from the subscriber loop is connected to a specialized integrated circuit called a SLIC — a *Subscriber Line Interface Circuit*. Along with a few other components, the SLIC provides about half of the BORSCHT functions. It does not contain a ringing generator or battery, since these are external and

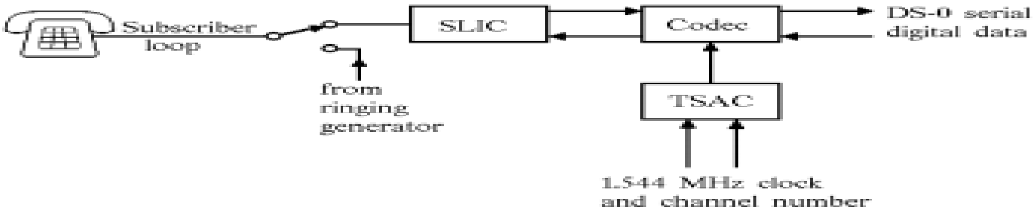
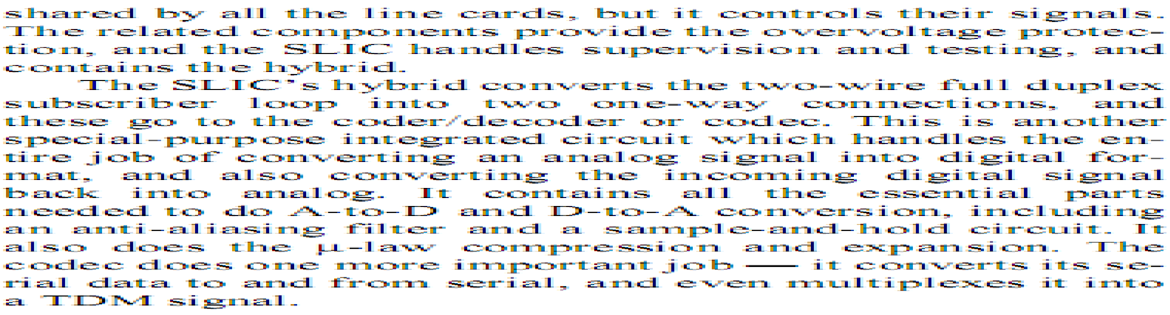


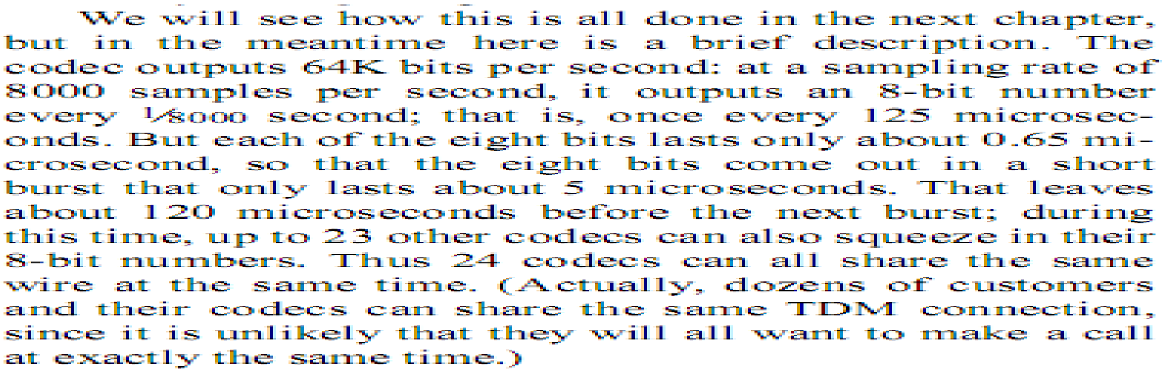
Fig. 16-7. Simplified block diagram of a line card



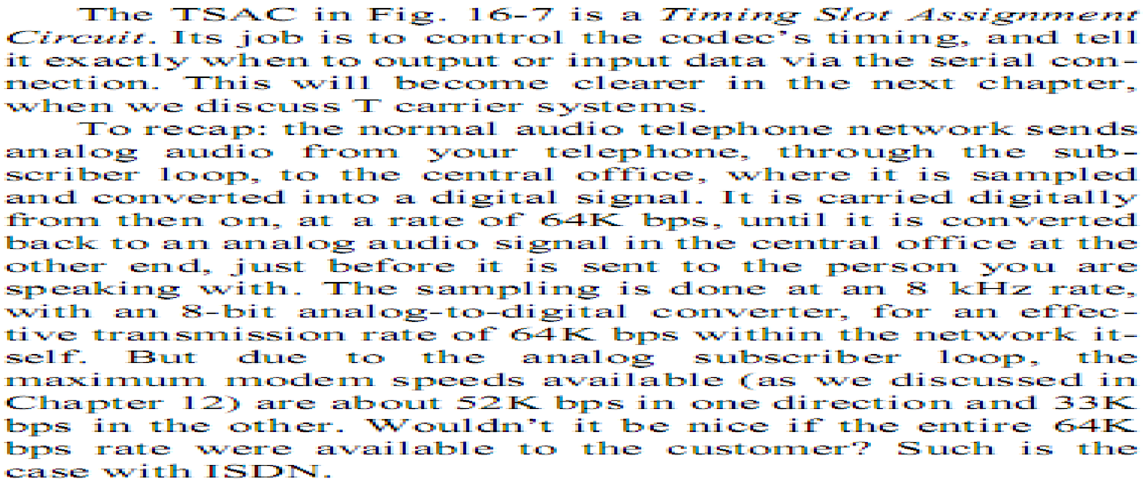
shared by all the line cards, but it controls their signals. The related components provide the overvoltage protection, and the SLIC handles supervision and testing, and contains the hybrid.

The SLIC's hybrid converts the two-wire full duplex subscriber loop into two one-way connections, and these go to the coder/decoder or codec. This is another special-purpose integrated circuit which handles the entire job of converting an analog signal into digital format, and also converting the incoming digital signal back into analog. It contains all the essential parts needed to do A-to-D and D-to-A conversion, including an anti-aliasing filter and a sample-and-hold circuit. It also does the μ-law compression and expansion. The codec does one more important job — it converts its serial data to and from serial, and even multiplexes it into a TDM signal.

The Codec converts the telephone signal into 8 bit μ-coded binary numbers. It therefore outputs 8-bit numbers for the outgoing signals and inputs 8 bit numbers for the returning signal for a total of 16 bits. There is one codec for each telephone customer; if every codec had 16 wires carrying parallel data back and forth, the switch will be terribly complex. Thus a multiplexed serial connection is absolutely necessary to keep things within reason.



We will see how this is all done in the next chapter, but in the meantime here is a brief description. The codec outputs 64K bits per second: at a sampling rate of 8000 samples per second, it outputs an 8-bit number every 1/8000 second; that is, once every 125 microseconds. But each of the eight bits lasts only about 0.65 microsecond, so that the eight bits come out in a short burst that only lasts about 5 microseconds. That leaves about 120 microseconds before the next burst; during this time, up to 23 other codecs can also squeeze in their 8-bit numbers. Thus 24 codecs can all share the same wire at the same time. (Actually, dozens of customers and their codecs can share the same TDM connection, since it is unlikely that they will all want to make a call at exactly the same time.)



The TSAC in Fig. 16-7 is a *Timing Slot Assignment Circuit*. Its job is to control the codec's timing, and tell it exactly when to output or input data via the serial connection. This will become clearer in the next chapter, when we discuss T carrier systems.

To recap: the normal audio telephone network sends analog audio from your telephone, through the subscriber loop, to the central office, where it is sampled and converted into a digital signal. It is carried digitally from then on, at a rate of 64K bps, until it is converted back to an analog audio signal in the central office at the other end, just before it is sent to the person you are speaking with. The sampling is done at an 8 kHz rate, with an 8-bit analog-to-digital converter, for an effective transmission rate of 64K bps within the network itself. But due to the analog subscriber loop, the maximum modem speeds available (as we discussed in Chapter 12) are about 52K bps in one direction and 33K bps in the other. Wouldn't it be nice if the entire 64K bps rate were available to the customer? Such is the case with ISDN.

Digital Switching: Digital Switching Systems, Space Switching, Time Switching Module; Time-Space-Time Switch Structure, Circuit switching networks; Packet switching networks; X.25 packet switched networks

DIGITAL SWITCHING SYSTEMS

3.1 INTRODUCTION

3.1.0 Digital Switch

A digital switch is a device that handles digital signals generated at or passed through a telephone company central office and forwards them across the company's backbone network. It receives the digital signals from the office's channel banks that have been converted from users' analog signals and switches them with other incoming signals out to the wide area network.

Digital switches are described in terms of classes based on the number of lines and features that are provided. A private branch exchange (PBX) is a digital switch owned by a private company. A centrex is a digital switch at the central office that manages switching for the private company from the central office.

3.1.1 Concepts

The fundamental task of telecommunications is to transfer messages. The communication system must ensure that the messages arrive at the correct receiver. The message transfer consists of the conversion of a message into signal units, the transport of these signal units, and the reconstruction of the message from these signal units.

Strictly speaking, the message transfer consists of switching as well as transmission. The transmission technology makes channels available for information transmission for long periods of time. But even this availability though, is flexible and can be varied. In the early days of transmission technology, flexibility was guaranteed by the distribution frame. Nowadays management commands are used to establish and direct transmission pathways. Following the further development of the control systems, transmission systems have begun to develop characteristics that have become more and more similar to those of switching technology. The major remaining difference is the control system, which uses measures of the network management (transmission technology) or signalling during connection set-up (switching technology). Both technologies are rapidly converging.

Switching network

The connection of terminal equipments, between which messages are to be exchanged, is performed by a switching network.

The switching network must be able to perform the following basic tasks:

- At any time, a connection from every piece of terminal equipment or from every point to all terminal equipment on the network or the transfer to other networks must be possible in principle.

- Every connection must be controllable by the user.

On one hand, the network must be in the position to fulfil the expected connection requests with sufficiently high probability, and to satisfy guaranteed quality parameters.
The technical effort to satisfy connection requests must, on the other hand, be reasonably limited.

The switching network is structured according to different points of view:

- requirements of the switching principle employed,
- amount of traffic,
- technical and economic parameters of the technology utilised,
- regulatory requirements.

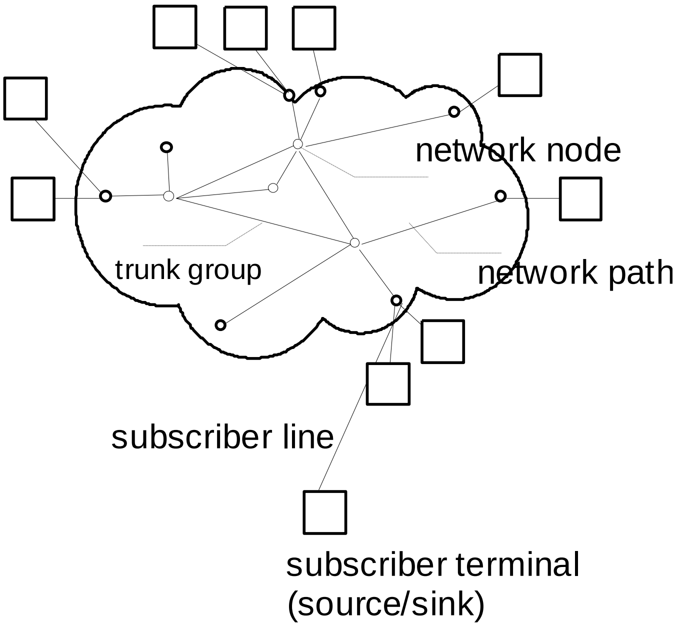


Figure 3.1 - Switching network

The most important elements of the network are the nodes and paths. The payload between the network nodes is transported in the paths. Network edges are connection lines which link the terminal equipment on the network and are connection trunks between the network nodes and users. Groups of connections or channels between these same network nodes are brought together in trunk groups. The payload is determined in the network nodes.

Connections

A connection is a coupling of at least two pieces of terminal equipment using network access interfaces, network paths and network nodes of a network for the purpose of exchanging information.

For all forms of information exchange the rule is: at first, a connection through the network must be created. This connection can exist continuously or it can be created for a certain time period. If the connection has been created for a limited period of time, then there must be switching. A connection then exists for the duration of the complete information transmission (for example, in a telephone network) or the time for the transmission of a part of the information (for example, in ATM networks). The switching is carried out in the network nodes.

A switching process is always carried out in connection with a definite communication relationship.

Switching

Switching is the creation of connections for a limited period of time in a network by means of connecting channels, which make up the partial segments of the connection. Switching is the creation of the connection by means of control signalling.

Switching technology

All technical equipment which is used for the switching in a network can be designated switching technology.

The switching technology ensures that the information in a network, according to the switching principles current in this network, reach exactly those network nodes or subscribers for which they were designated.

From the point of view of the user of a network, switching is a service that can be employed in order to exchange information with one or many other users on the network.

A switching node is that part of a network where partial segments of the network are put together for a connection by evaluating technical switching information. Simultaneously, depending on the traffic volume, the traffic of many terminals on the network is concentrated on a few paths of the network by switching.

The place where a switching node is located is called an exchange.

Switching nodes are distinguished according their location in the network hierarchy as well as well as by their technical configuration.

3.1.2 Switching Principles

The switching principle is the way the switching of connections or messages is carried out.

Connectionless transmission

The connectionless mode is appropriate for networks in which sporadic, short information segments must be exchanged between the terminals, such that the time required for setting up and terminating a connection can be reduced. For this reason, these networks have mainly developed for communication between computers. The disadvantage of this kind of network is that all nodes are loaded with traffic, even if the information is not intended for them.

Connection-oriented transmission

If the time required for the set-up of a connection is short compared with the time period that the connection exists, then connection-oriented service modes are more advantageous. Information is transported only to nodes that are necessarily involved with the communication. Telephone networks have evolved on this model. Connection-oriented networks can work with switched channels (channel switching) or the message switching (packet switching or virtual connections).

Connection-oriented channel switching includes switching in the spatial domain (spatial separation of the channels - spatial switching) and in the time domain (time multiplexing of the channels).

Message Switching

Message switching consists of packet switching (a number of packets per message) and consignment switching (one packet per message).

A special position must be given to ATM switching, which is gaining in importance and will be described in a section 3.3.3.

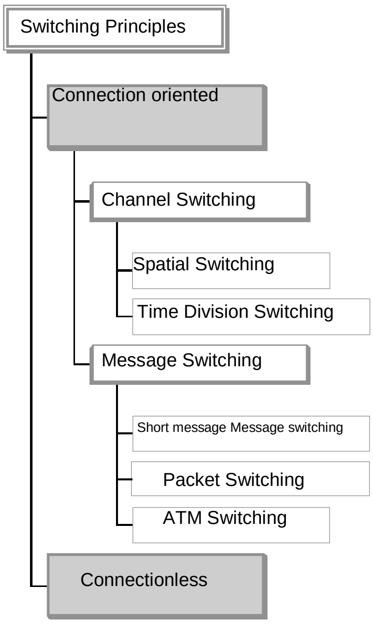


Figure 3.2 - Overview of switching principles

3.1.3 References

Walrand, J.: Communication Networks.- Boston: Irwin, 1991

Schwartz, M.: Telecommunication Networks.- Reading: Addison-Wesley, 1988

3.2 CHANNEL SWITCHING

For channel switching, the relationship between the communication partners is implemented by connecting channels. After the relationship is created, the subscribers are directly connected with each other for the complete duration of the communication.

The spatial switched channel is the "classical" form of the connection. In the simplest cases, they are made with electrical connections, which are switched together with contacts. Switched channels can be either switched or fixed connections. For switched connections, the participating terminals are automatically connected together for a certain period of time, based on the destination information of the source (using switching technology and signalling). Dedicated connections are created by network management measures for a certain period of time. The oldest network working on the connection-oriented principle is the telephone network.

Spatial switching is the switching of physically separated electrical channels.

Time switching is the switching (rearrangement) of time slots in systems, in which the information from individual channels is transported in time slots.

Channel switching is also designated as circuit switching. For circuit switching, the creation of a connection is necessary before the actual communication is made; after the communication, the connection must be terminated again. Therefore the connection is divided into phases.

3.2.1 Connection phases

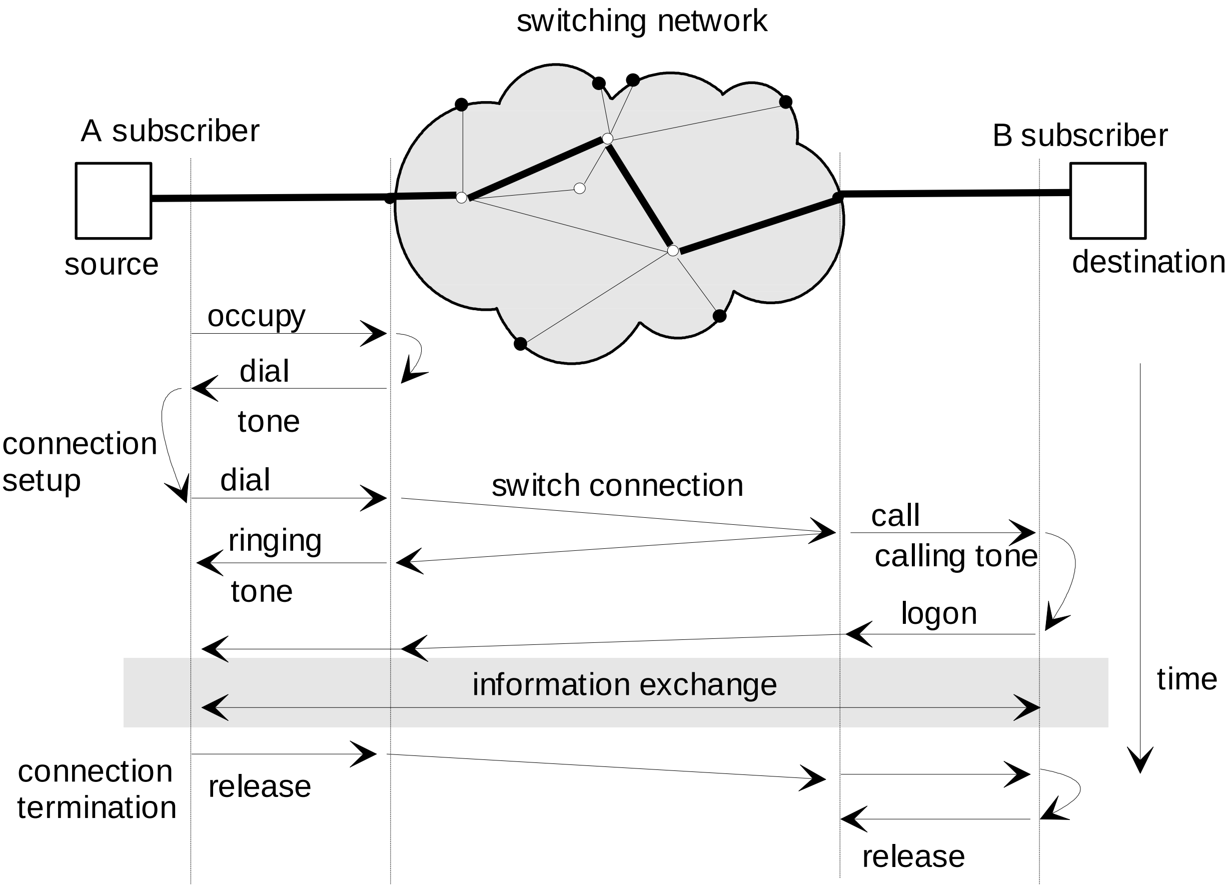


Figure 3.3 - Schematic representation of the phases of a circuit switched connection

Connection set-up. The connection set-up is carried out by an exchange of signalling information between the active terminal equipment and the exchange, and between the exchanges. The initiative is taken by the terminal equipment which wants to set up the communication relationship (in telecommunication technology and in the above example in Figure 3.3: 'A'-subscriber). Thereafter follows the reservation of the switching device equipment to which the A-subscriber is connected. If this reservation is accepted, that is, if a facility is free to process the connection request, then the terminal equipment is informed (in the telephone network: using dial tone). Next, the terminal equipment notifies, by dialling, which other terminal it desires to connect to (dial information, address information). Then an attempt is made to establish a path to the destination terminal (B-subscriber). If this is successful, then the B-subscriber is called, and the A-subscriber is informed of the connection set-up (call display, in telephone network: ringing tone). After the B-subscriber has acknowledged the call (logon), the connection enters into the second phase. The created occupancy is, from the point of view of the A- subscriber, an outgoing call and, from the point of view of the B-subscriber, an incoming call.

In general, the requested connection extends over a number of switching configurations, and signalling is also necessary between them.

Information exchange. In the second phase of the connection the actual information exchange occurs which also can be accompanied by signalling. Thus, during the course of a connection, service components can be switched on and off and teleservices can be managed.

Connection release. The third phase of the connection is the connection release, which one of the terminals initiates by means of signalling. The switching equipment engaged and the occupied channels are released again. Data is collected for the recording of connection-dependent fees.

3.2.2 Structure of a switching system

Functional blocks. A switching configuration has a variety of functional blocks, which are either involved in or support the actual switching process:

- Switching: Connection of subscribers by means of subscriber lines and link lines, in order to create individual communication relationships.
- Administration: Administration of the subscriber lines associated with the exchange, trunk lines, the equipment of the exchange and the processes which run on this equipment. The collection and processing of fee and traffic data is also included.
- Maintenance: The ensuring of equipment availability of the central unit.
- Operation: communication between the central units and their operation personnel.

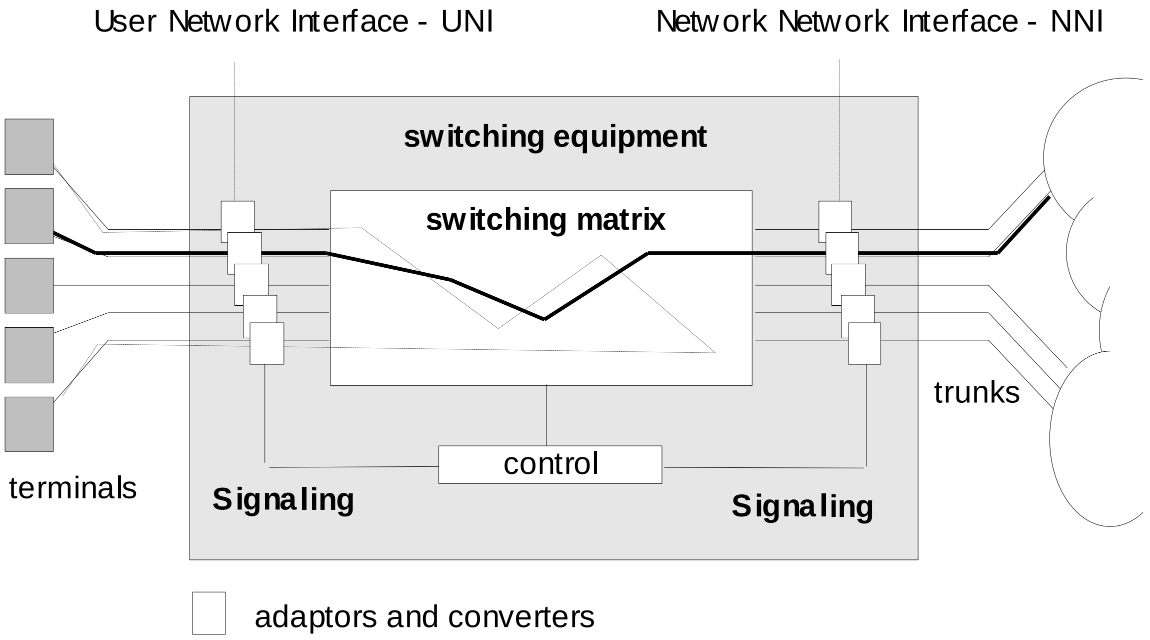


Figure 3.4 - Principle elements of a switching configuration from the point of view of the switching process

Figure 3.4 represents a local exchange. This is the most general case of a switching system, because here connections to subscribers, as well as connections to other exchanges, are represented. On the left side, subscriber lines connecting terminal equipment are represented, using the user network interface (User Network Interface - UNI). On the right side are trunk lines between the switching stations. Exchanges are connected by means of network interfaces (Network Network Interface - NNI).

A connection between two terminals attached to the same switching station is called an internal connection, and is represented with dotted lines in Fig.3.4. A connection from or to a subscriber, which is attached to another exchange is called an external connection. This kind of a connection is drawn in bold lines in the Figure.

Control. An important element of the switching system is the control, which processes the signalling information from and to the terminal equipment and between the exchanges. The control system obtains the necessary information for adaptation from adapters and converters and from subscriber lines and trunk lines.

Switching matrix. The actual creation of connections takes place in the switching matrix, also called switching network. It is the basic element of a switching system and is set up by the control system.

Periphery. The periphery of the switching system must provide additional functionality so that the switching node can successfully integrate into the rest of the environment. The most important task requirements of this periphery are:

- the supply of power to the subscribers line, i.e. supplying the electrical energy,
- the protection of the switching system from electrical influences on the connections (for example, due to cable error, voltage overload, lightning etc.),
- the separation of payload and control signals for inband signalling (for example, from and to subscribers in a telephone network),
- the interference suppression of payload and control signals,
- the conversion of message forms (e.g. 2 wire, 4 wire conversion),
- recognition of incoming signalling,
- creation of signalling,
- recognition of errors for maintenance purposes.

The above functions are implemented in so-called trunk circuits and subscriber circuits. The subscriber circuit carries out the so-called BORSCHT function. BORSCHT is an English acronym for the functions

- Battery (loading),
- Over voltage protection,
- Ringing,
- Signalling,
- Coding (e.g. analogue- digital- conversion),
- Hybrid (2- wire, 4- wire conversion),
- Test (error detection).

3.2.3 Task requirements of the function unit 'switching' of a switching system

For the task requirements of the most important functional units of a switching central unit, the elements of the service "switching" available to the user are described. The most important task requirements are:

- Search for a free unit for carrying out a function. Such a unit can be a free link in a certain direction (path seek), but also can be a software procedure instance for realising a service characteristic.
- Testing of identifications and access privileges.
- The occupation of a long-distance unit upon request. This unit is assigned to a connection to be created and locked for any other attempts at occupation.
- Switching on of dial tones.
- Receiving and evaluation of dialling information.
  Reception of dialling information and evaluation in terms of the selected direction, of the subscriber or of service characteristics.
- Signalling transmission, i.e. transmission of a telephone number from the switching system to another switching system or to terminal equipment.
- Connection, i.e. creation of a connection in the switching network.
- Connection termination, i.e. determination of fees, the signalling of the connection completion, release of the equipment.
- The disabling of a facility from use in case of malfunction, during maintenance or for other reasons (for example, to prevent traffic overload of other elements of the central unit or of the network).
- Release of allocated or disabled equipment within the exchange.

3.2.4 Switching matrix

The switching matrix is an arrangement of switching elements which are used to connect payload channels in a switching system.

The switching network is the central element of a switching facility. With switching networks, the required connections of transmission channels between the switching exchanges are created.

Based on the signalling information and available channels, the switching arrangement connects input ports and output ports. The task of the switching matrix is the set-up and release of connections, as well as handling the administration of the simultaneously existing connections.

In general, a switching network consists of a number of connecting stages. They are individual layers with a multiplicity of switching elements which are functionally parallel.

Function groups

The complete switching network is divided into three important functional groups, in which the traffic to be switched is concentrated, distributed, and finally expanded. The most important function is the distribution of the traffic. The required technical equipment in general is very complex and can be better utilised with concentration. The concentration / distribution / expansion structure is functional. This basic structure of switching systems is the same for all principles that can be applied to switching, independent of whether it is switching between a variety of spatial connections, time slots or packets.

Concentration. Concentrating switching networks are used when more inputs than outputs are involved. Concentration is the switching of a number of input lines onto a few output lines. The traffic of the lightly utilised input lines is concentrated on more heavily utilised output lines. The expensive equipment assigned to the output lines is also better utilised.

Distribution. Linear switching networks are used when an equal number of inputs and outputs are involved. In distribution, the traffic is distributed according to its direction.

Expansion. Expanding switching networks are used when more outputs than inputs are involved. After distribution, the traffic must be reconstituted to the separate individual subscriber lines at the destination local exchange. The traffic is expanded.

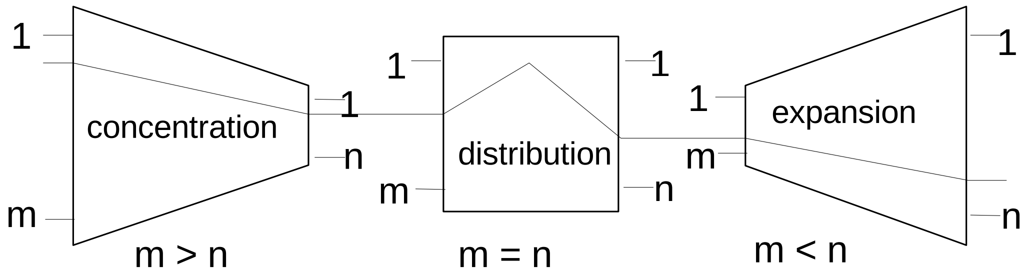


Figure 3.5 - Concentrating, distributing and expanding in a switching network

A connection in a switching system is processed at first with a concentrating, then a distributing, and finally with an expanding, switching arrangement. This arrangement of the individual components of the coupling network is purely functional. For the practical realisation of switching network, a concentrating and expanding switching arrangement can comprise the same physical elements.

Spatially-separated switching matrixes

Spatially-separated switching is the oldest form of switching. A channel is made up of a certain number of lines (wires), which are connected with electrical contacts to one another. These contacts can be implemented by means of

- relays,
- selectors (lift-rotate selector, motor selector),
- co-ordinate switches or
- electronic building blocks (transistors).

A switching matrix for three wires per channel and with 4 x 4 channels on the basis of a Strowger selector appears in Figure 3.6. An arrangement of three coupled mechanical switches represents one crosspoint.

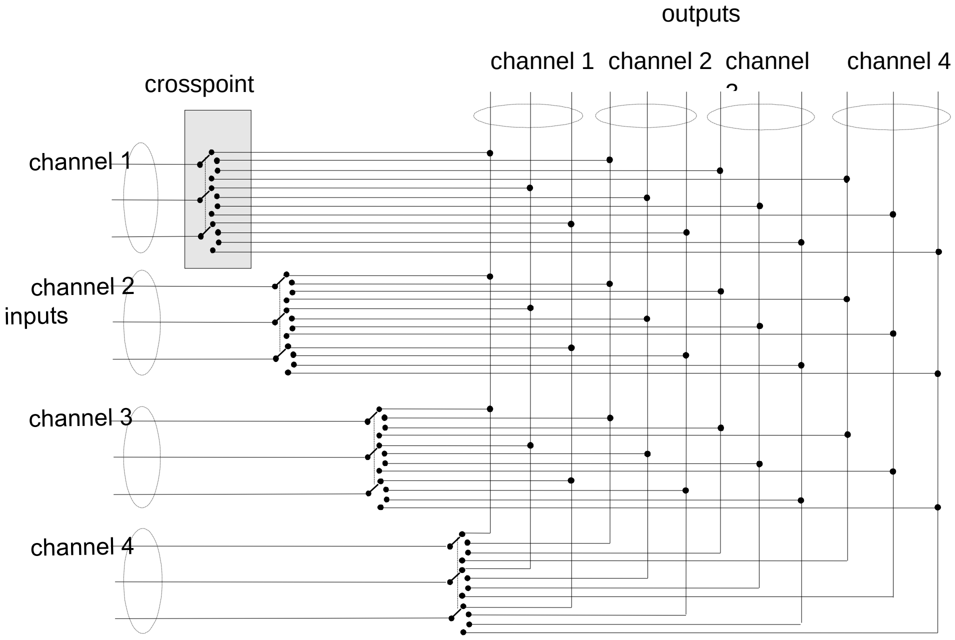


Figure 3.6: Representation of the operating principles of a mechanical switching matrix

Switching arrangement. The switching arrangement itself is a matrix, and connections can be created at the crosspoints. Figure 3.7 shows this kind of a coupling matrix in a so-called stretched representation. One crosspoint is required for a connection of an input to an output. Therefore, for m inputs and n outputs, m*n crosspoints are required. The switching network is free of blockage, which means that already existing connections cannot block new connections. Part a) of the diagram shows all coupling points, while the simplified representation in part b) of the diagram symbolises only the number of the inputs and outputs.

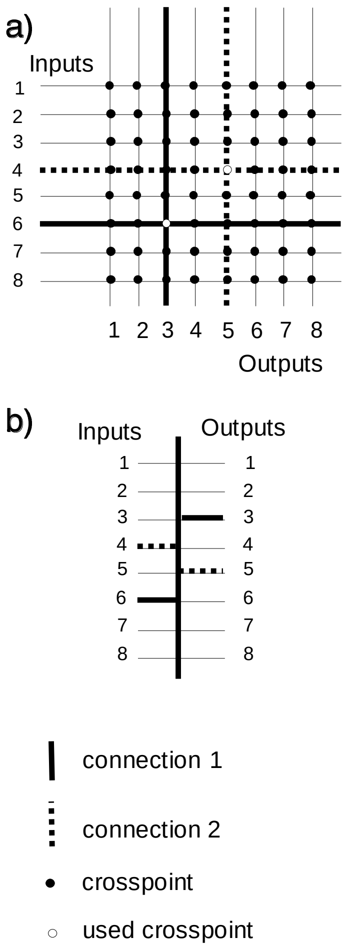


Figure 3.7 - Single-level switching matrix in a stretched arrangement; a) complete representation; b) simplified representation

Example: The coupling arrangement displayed in Figure 3.7 has m = 8 inputs and n = 8 outputs. Therefore m * n = 64 crosspoints are necessary. Every input can be connected with every output. Existing connections do not prevent other connections from being switched when other inputs and outputs are involved. In the example, connections exist between input 4 and output 5 as well as between input 6 and output 3.

Apart from the stretched arrangement, switching matrices can also be operated in the so-called reversal arrangement. In this case, inputs as well as outputs are connected on the same side (rows) of the matrix. The columns of the matrix serve to connect rows. For p columns of the matrix (m+n) * p coupling points are required. Two crosspoints are required for a connection. A maximum of p connections can exist at the same time. The disadvantage of this coupling matrix is that the connection between certain inputs and outputs cannot be created under certain conditions, because other connections already exist (internal blockage).

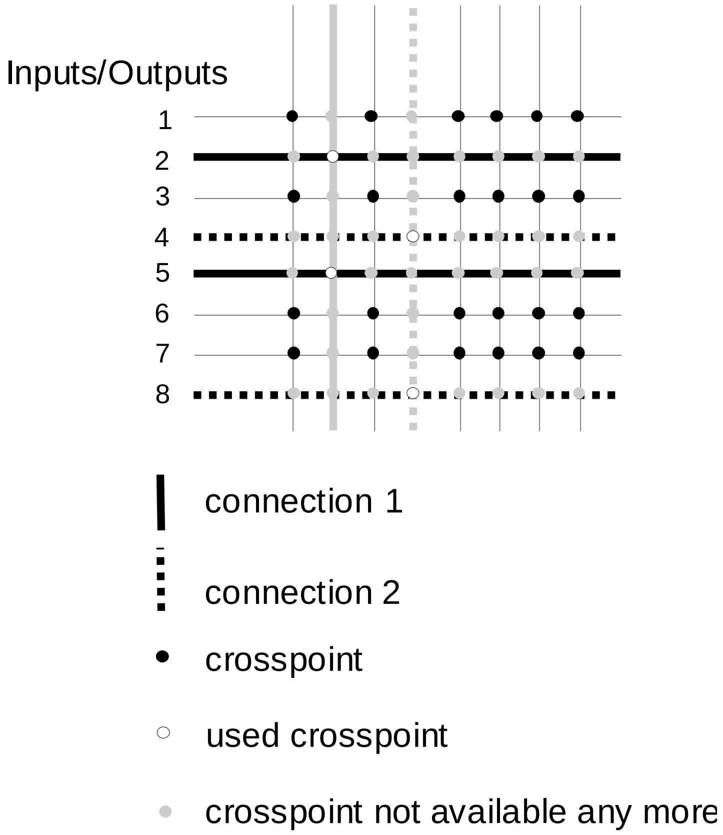


Figure 3.8 - Switching matrix in reverse arrangement

Example: The coupling arrangement shown in Figure 3.8 has m + n = 8 connections which could be inputs or outputs. Determined by p = 8 columns of the coupling matrix, p * (m + n) = 64 crosspoints are necessary. Every connection uses a column of the matrix (in this case, drawn in grey) to complete the circuit. Therefore a maximum of p connections can be switched. Every switched connection effects that the coupling points of the rows and columns required for the completion of the circuit cannot be used for other connections. The coupling points no longer in use are also drawn in grey.

This configuration of the coupling matrix meets an important requirement for the configuration of switching matrixes: the number of the employed technical elements should be approximately proportional to connection capacity; this not the case for a coupling matrix in a stretched arrangement, in this case it is a quadratic dependency.

Because of the necessary requirement for extensibility, switching networks should be modularly designed. This can be achieved by dividing up large switching matrixes into smaller matrixes and then switching these matrixes together over a number of levels. With multi-level switching networks and the switching together of smaller matrixes, fewer crosspoints are required than for single-level switching networks. But in the case of multi-level switching arrangements, internal blockages are possible. The probability of an internal blockage goes up with the concentration factor of the switching matrix and declines with the size of the individual switching matrix.

Example: The switching network, which is represented in Figure 3.9, allows for the connection of up to 100 subscribers. A maximum of five internal connections can be simultaneously set up, as well as up to three external trunk groups with up to five connections each.

For the case m = 10 inputs and n = 5 outputs, per switching matrix in a stretched arrangement in layer 1 m * n * 10 = 500 crosspoints are required.
In layer 2, the switching matrices also have m = 10 inputs and n = 5 outputs. The 5 switching matrices of this level thus have a total of 10 * 5 * 5 = 250 crosspoints.
In layer 3, the number of the crosspoints can be calculated from m = 5, n = 5 and the number of matrices which is five. This yields 5 * 5 * 5 = 125 crosspoints.
In total, 875 crosspoints will be required.
Hence because of internal blockage, it is not possible to create more than five connections for a subscriber group out of 10 subscribers which belong to one and the same switching matrix of the first layer. Furthermore, not more than five internal connections, and not more than five connections to the external trunk group, can be created simultaneously.
Two connections are displayed:

- line 10 of the first matrix of layer 1 connects to line 1 of the second matrix of layer 1 (internal connection) and;
- connection 3 of the 10th matrix of layer 1 connects to line 1 of the external trunk group (external connection).

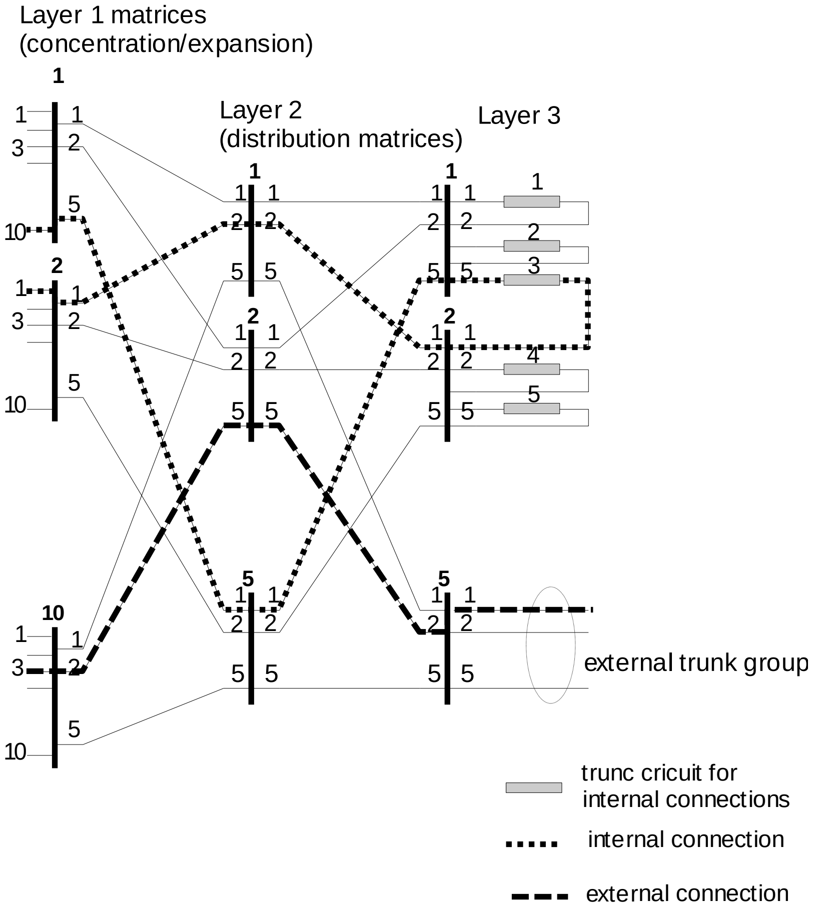


Figure 3.9 - Multi-layer switching network

By carefully designing the switching network layers and the connections between the layers, a compromise can be found between crosspoint number and blockage probability. This information on switching networks mainly refers to the switching of spatially separated channels, which can be implemented with Strowger selectors or co-ordinate switches.

But channels can also be in different forms. It is possible to assign a channel a fixed carrier frequency and switch this carrier in the switching system. Another possibility is the assignment of a time slot to a channel. In digital switching technology, the spatial and the temporal domains are utilised. In switching devices, spatial and time switching arrangements are often used in combination.

Time-division switching networks

Time-division synchronous channel splitting. In the case of time-division channel splitting, individual time slots are assigned the information to be transmitted in the channels. This technology, for example, is applied for Pulse-Code-Modulation (PCM). The assignment of individual channels to time slots is shown in Figure 3.10. The assignment is rigidly defined in a frame structure. The position of individual bits in the frame determine to which information relationship they belong. The synchronisation which is carried out for a frame must last for the time it takes for a complete pass through the frames. A time frame is represented in Figure 3.10 a) as a complete cycle of the rotating switch. 32 channels are nested in it and a cycle requires 125 ms.

Time-switching arrangement. The principle of a simple switching arrangement for switching the time position of an individual channel is shown in Figure 3.10 b). In this case, the information which arrives at the input in individual time slots is written to specific fields of the switching memory by a controller and temporarily stored. This writing process is controlled by a control table. The reading of information from the memory occurs in a fixed sequence. The control table contains the assignment of the time slots of the output lines to those of the input lines. It is also conceivable that the data is written in a fixed sequence and read out with a control table.

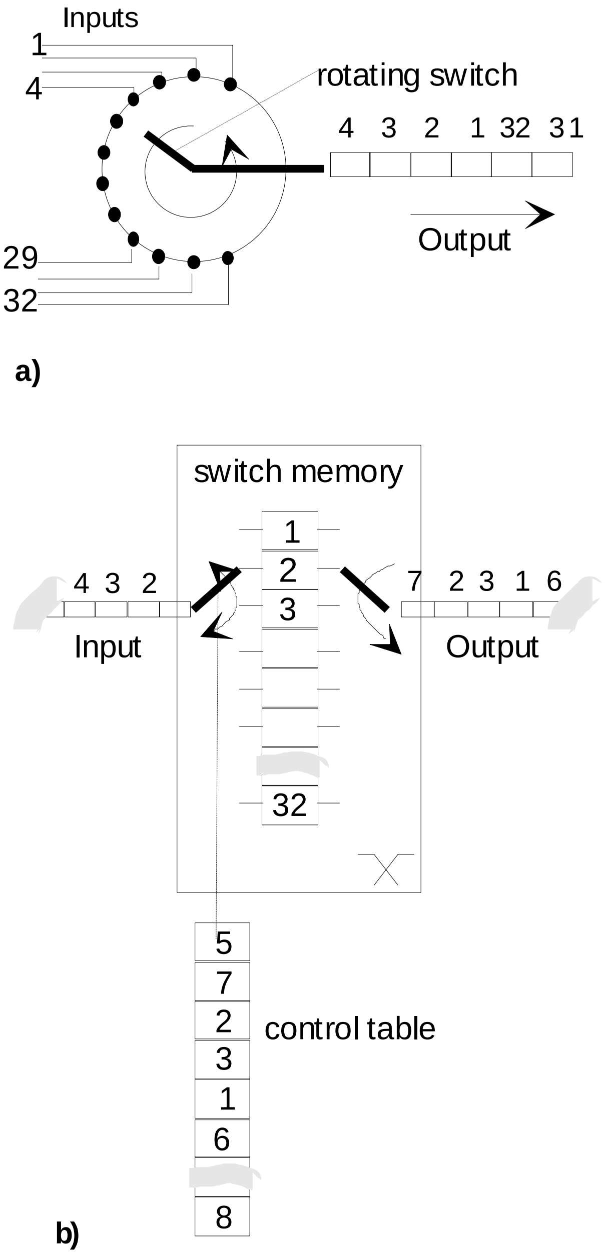


Figure 3.10: a) Assignment from channels to time slots;
b) Rearrangement of time slots because of intermediate storage

In every case, storage of the time slot information is required for the rearrangement of time slots. This can occur at the input of the coupling field, at the output of the coupling field, centrally for the complete coupling matrix, or distributed for every coupling position. For a detailed representation of the storage types, refer to the section on ATM switching, because the same principles will be applied there.

Time/Space Switching. In general, a number of PCM input lines reach a switching network. The job of the switching network consists of executing the rearrangement of the time slots as well as co-ordinating between PCM connections. For this, a space and time switching network (Figure 3.11) is required.

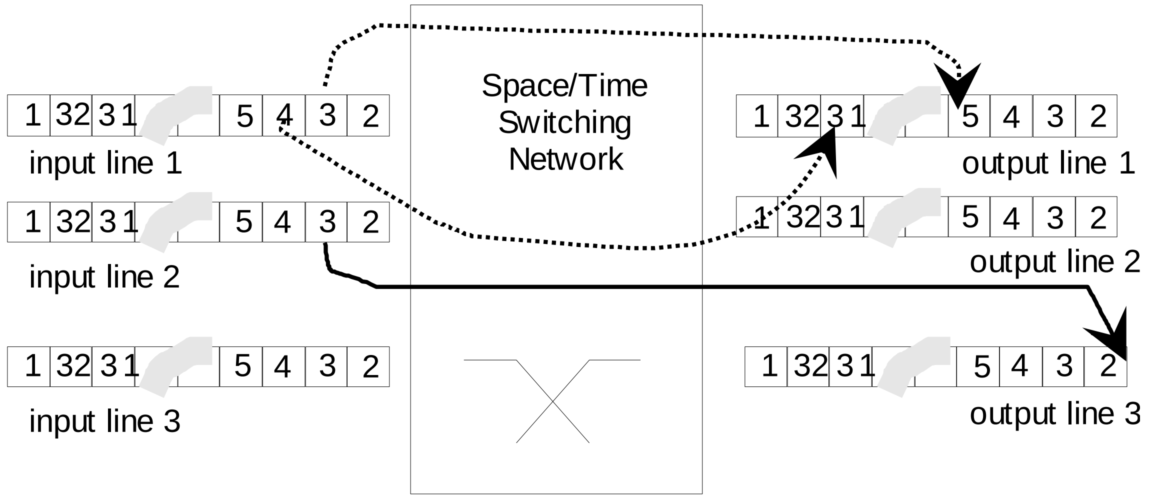


Figure 3.11: Spatial temporal switching

Example for Figure 3.11: Time slot 3 of the input line 2 is to be assigned to time slot 2 of the output line 3 (solid line). For this task, a temporal and a spatial switching process are necessary. The dotted-line rearrangements, in contrast, require only time switching. Technically, the spatial and the temporal switching can be carried out at the same time. For this purpose, all spatially-separated input lines on a line are multiplexed (note: for inputs, this line must have more than an n-fold processing speed) and stored; the individual spatially-separated output lines of the coupling arrangement, parallel to each other, are read out of the correct time slots from the common memory.

3.2.5 Control of switching devices

The special feature of the switching device control system is that a connection almost always pass through a number of network nodes and therefore a number of switching stations, and all of these switching stations are incorporated into the control system of the connection.

The transmission of control information between switching stations and from/to the terminal equipment is carried out by signalling.

Every connection is built up piece by piece by selecting channels. This selection subdivides into

- a forced selection, which determines the direction in which the connection will continue to be built, and
- a free selection, which automatically dials up a free channel in this direction.
- The forced selection is always controlled by the dial information.

The dial information required for the control system of the participating switching device is created in the calling terminal. If this dial information is used directly to control the switching system, this is called direct control. If the dial information first goes to temporary storage and then is evaluated, this is called indirect control.

The direct control system was introduced with the introduction of the lift-rotate Strowger selector. The impulses of a dialler directly control the lift steps. In the pause between two dialled digits, the free selection of a channel in the selected direction can be carried out. The next dialled digit now directly controls a selector in the next selection level or in another switching station.

The indirect control system has applications mainly in SPC switching and computer-controlled switching systems.

The direct control system is no longer used today. The indirect control has the following advantages:

- Before individual segments of a connection become occupied, it can be determined if a path can be found through the network up to the destination terminal equipment, thus avoiding the stepwise occupancy of channels before the actual effective connections can be completely made;
- Considerably more complex methods of path searching (routing) for a connection through the network can be applied than with the stepwise connection set-up.

3.2.6 References

ITU-T References

exchanges - Introduction and field of application

[Q.511] (11/88) - Exchange interfaces towards other exchanges

[Q.512] (02/95) - Digital exchange interfaces for subscriber access

[Q.513] (03/93) - Digital exchange interfaces for operations,

administration and maintenance

[Q.521] (03/93) - Digital exchange functions

[Q.522] (11/88) - Digital exchange connections, signalling and

ancillary functions

[Q.541] (03/93) - Digital exchange design objectives - General

[Q.542] (03/93) - Digital exchange design objectives - Operations

and maintenance

[Q.543] (03/93) - Digital exchange performance design objectives

[Q.544] (11/88) - Digital exchange measurements

[Q.551] (11/96) - Transmission characteristics of digital exchanges

[Q.552] (11/96) - Transmission characteristics at 2-wire analogue

interfaces of digital exchanges

[Q.553] (11/96) - Transmission characteristics at 4-wire analogue

interfaces of digital exchanges

[Q.554] (11/96) - Transmission characteristics at digital

interfaces of digital exchanges

[Q.700] (03/93) - Introduction to CCITT Signalling System No. 7

(Series, Q.700 - Q.788)

[Q.920] (03/93) - Digital Subscriber Signalling System No. 1 (DSS1) -

ISDN user-network interface data link layer - General aspects (Series Q.920 - Q.957)

[Q.1200] (09/97) - General series Intelligent Network Recommendation structure

[Q.2010] (02/95) - Broadband integrated services digital network

overview - Signalling capability set 1, release 1

3.3 MESSAGE SWITCHING

In the case of message switching, no channels are established on which the information is exchanged, but rather individual messages units, most often packets, which contain all or a part of the information to be transmitted, are switched.

This occurs exactly like one would imagine the "switching" of postal letters in a network of post offices: The packets are supplied with addresses which give information about the receiver. In each switching station, the address is evaluated and the message is forwarded in a direction which brings it closer to its destination. The switching is carried out separately for each individual message unit. Therefore no connection set-up is required. Packets, which belong to the same information relationship, can take different paths through the network.

Store and forward switching. Message switching is often called store and forward switching. Typical for this configuration is that the packets are lead step for step (from switching system to switching system) through the network. The packets are stored temporarily in each of the network nodes.

3.3.1 Packet switching

Packet switching switches information that is divided into a number of packets.
A packet in this sense has the following basic set-up:

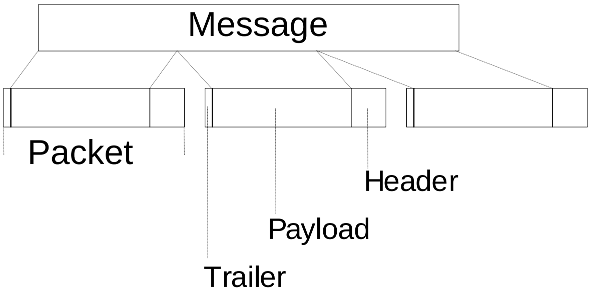


Figure 3.12 - Set-up of packets in packet switching

Packet. A message is divided into a number of units. These units are supplied with a header and a trailer. The header, payload and trailer form a packet. Packets can be of fixed or variable length. The packet trailer is not necessary for certain switching procedures.

The packets are created at the transmitting terminal equipment. At the network nodes, the addresses of the packets are analysed and are forwarded in a direction which will bring them closer to their destination. For this purpose, packets need not necessarily take the same path. The forwarding process is dependent on the traffic load which is currently on the network.

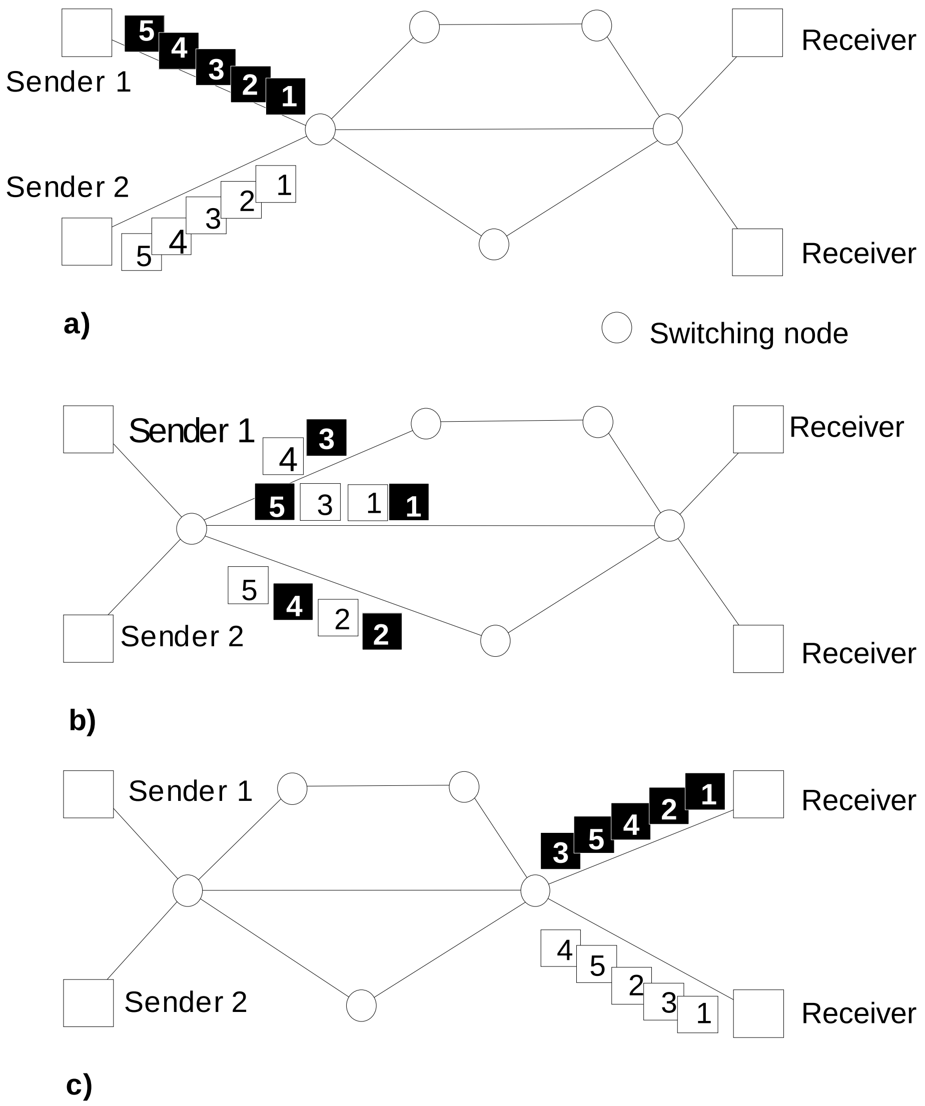


Figure 3.13 - Switching of packets in a packet switching network;

a) phase 1: transmission of the packets;
b) phase 2: switching of packets to a network node;

c) reception of the packets by the receiver

Comment to Figure 3.13: The simultaneous but independent transport of two message units is described. At first, both transmitters allocate the transmission information into the packets 1 to 5. These are passed on to the network in the order of their numbering (Figure 3.13 a)). The first switching node test attempts to direct the packets on the shortest path in the direction of the receiver. Both receivers are connected to the same switching node. The expedition of the packets is first of all successful for both of the first packets. Now the transmission capacity on the direct connection to the receiver is exhausted for the moment and so the respective second packets are sent over the alternative lower part of the network. With this transmission, this path is also fully utilised. The third packet of the information relationship 1 must now be sent on a longer alternative along the upper part of the network, because now the first alternative also has no further transmission capacity available. Now a packet along the direct path can be accepted (packet 3 of information relationship 2), the next packet (packet 4 of information relationship 1) is once again sent on the shortest alternative. Packet 4 of connection 2 takes the long alternative. Once more a packet can be sent along the direct path and the last packet (packet 5 of the relationship 2) can take the short alternative (Figure 3.13 b). Because of the different transmission times for each route, the packets arrive at their receivers in the order shown (Figure 3.13 c).

The advantages of packet switching are:

- rapid transmission without connection set-up times, especially appropriate for short, sporadic information transmission and a low number of packets,
- good time and space capacity utilisation of the network resources, especially for sporadic, burst-mode traffic.

The disadvantages of packet switching are:

- transmission time varies and cannot be guaranteed,
- resource cannot be guaranteed (bandwidth),
- packets can overtake each other (see Figure 3.13 c)),
- higher computing power requirements for the routing of the packets.

3.3.2 Message switching

Message switching conveys packets which contain the complete contents of an information relationship.
A message packet which is conveyed in transmission switching, has the following design.

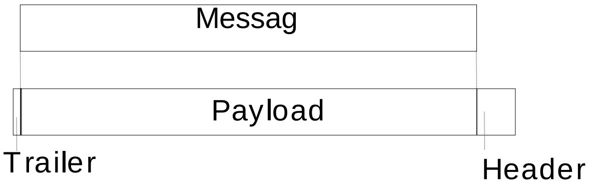


Figure 3.14 - Set-up of a packet for message switching

The packets have a variable length. The complete contents of a message are contained in a packet. Therefore, in contrast to packet switching, there is no need for the division of the message into data blocks and the protocol overhead that results.
The process does not differ from packet switching from a technical point of view. It is used, for example, for the short message service (SMS) in GSM networks.

3.3.3 ATM switching

In the case of ATM switching, the composition of information packets is similar to that for packet switching. They all have the same length of 53 bytes. All packets of an ATM connection take the same path through the network, for which the transmission capacity has been reserved in advance.

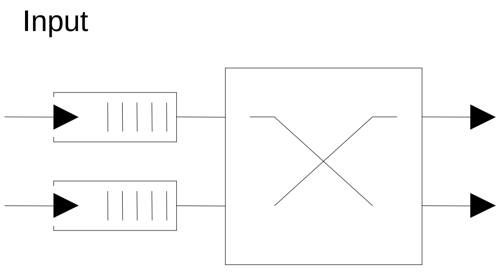
ATM switching differs from classical packet switching by the constant packet length and the determination of a connection path. This allows the switching of ATM cells to be simpler and computationally easier to control.

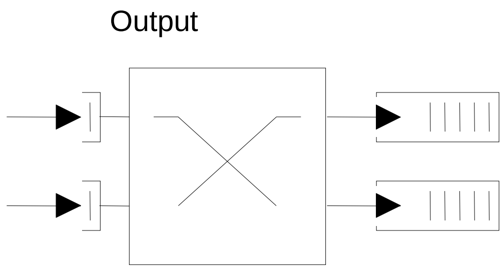
Storage principles

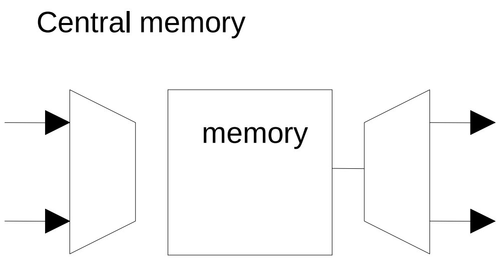
A requirement for the switching of ATM cells is that the cells in every switching system are temporarily stored. For this purpose, the following basic principles can be applied:

- Input memory: Per input, the incoming cells are stored in memory on the principle first-in-first-out (FIFO). For the switching process, an internal blocking-free matrix is employed. The disadvantage of this storage method is the possible blockage of waiting cells in the FIFO, so that even though the respective output is free, it is possible that a cell must wait for switching because previous cells to other outputs must be handled first.
- Output memory: Immediately after arriving, the cells are switched to a FIFO per output, and read out from there with the output line cycles. On the input, only the storage of one cell per lead is necessary. The disadvantage of this storage method is that the internal speed of the switching matrix must be greater than the speed of all incoming cells.
- Central memory: All incoming cells are stored in a common memory. This can be smaller than the sum of all separate memory requirements, but the control system for memory access is complex and very high-speed memory access is required.

Distributed memory: In a matrix made up of input and output lines, memory is allocated at every crosspoint to allow the multiplexing of the cells on the output lines. The disadvantage of this method is the large memory requirement.







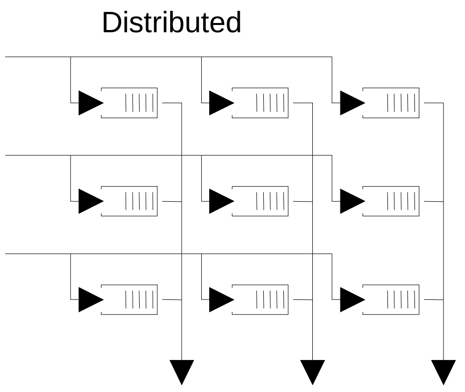


Figure 3.15 - Storage principles in ATM- switching

3.3.4 Virtual connections

In the case of virtual connections, individual packets are switched, but all packets of an information relationship are transmitted along only one path which is established at connection set-up.

Connection orientation. Before the information exchange begins, there is a connection set-up which determines if a path with adequate transmission capacity is available between source and sink. This channel is not occupied during the total connection time, but only when the transmission capacity is required. If no packets are available for some duration, the transmission channel can be used for other virtual connections. The capacity of transmission sections can even, within certain limits, be overbooked (statistical multiplex gain), nevertheless, all virtual connections have access to guaranteed resources and at times even have the use of more bandwidth than they were guaranteed.

Virtual connections combine the advantages of packet switching and channel switching. They:

- do a good job of utilising the resource of the network (an advantage of packet switching);
- can quickly make available large transmission capacities (an advantage of packet switching);
- guarantee resources (an advantage of channel switching), and;
- have a control system which is inherently less complex to realise than with a strict packet switching system.

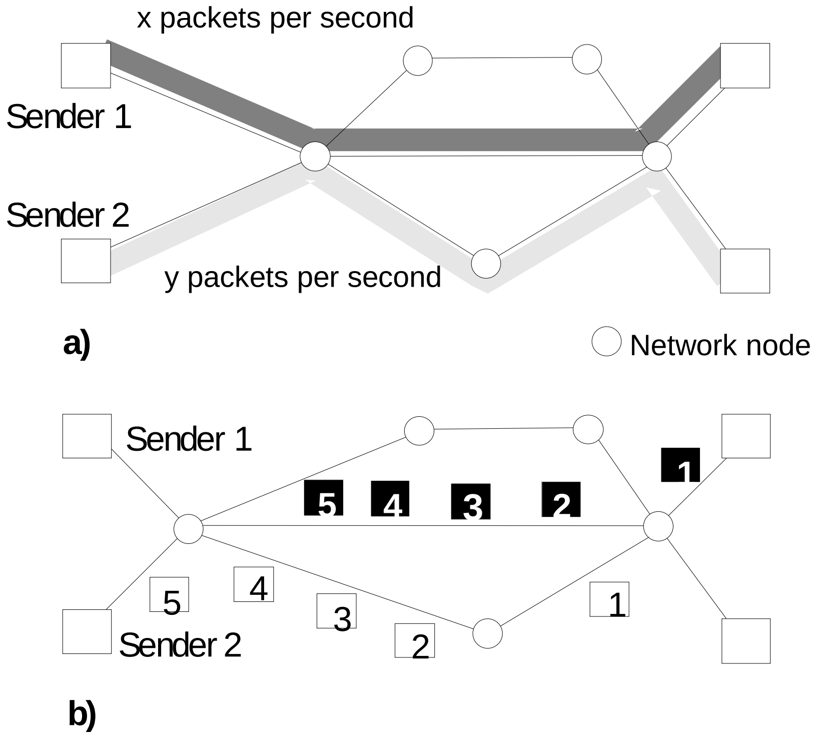


Figure 3.16 - The switching of packets in a switching network with virtual connections;

a) phase 1: connections set-up;
b) phase 2: switching of the packets along the set paths.

In phase 1 (connection set-up), the transmission capacity along both designated paths is reserved. In order to guarantee the desired bandwidth, as in the example, both connections must be led along different paths. For the transport of the packets in phase 2 (information exchange), the reservation of path and bandwidth of service quality (Quality of Service - QoS) ensures that packets cannot overtake each other and are delivered within the timing requirements.

3.3.5 Switching and routing

Switching

Switching is the creation of connections in a classical telecommunications network for a limited period of time by the interconnection of channels (line or circuit switching). During connection set-up, which is carried out before the actual information transmission occurs, the creation of the connection is controlled by signalling. Connections can also be virtual as is the case with ATM.

Switching is carried out at layer 2 of the OSI Reference Model.

Routing

Routing is the directing of data packets, based on the complete address of the destination of the sender contained in the data header, to the receiver over a varying number of nodes (routers) through the network. The job of the routing function is, for example, to transport datagrams in a packet network from a transmitter to one (unicast) or numerous (multicast, broadcast) destinations. For this, two sub-tasks must be performed:

- the construction of routing tables, and;
- the forwarding of the datagrams using the routing tables.

The routing process described here is the forwarding of data packets. It has nothing to do with path searching for switched circuits under certain network conditions, such as in the case of overload, errors, or for optimising the costs of a connection (least-cost routing).

The datagrams are transferred from one router (next-hop) to the next (hop-by-hop). A given router knows the next router which lies in the direction of the destination. The decision on the next router (next-hop) depends on the destination address of the datagram (destination based routing). An entry in the routing tables contains the destination and the next-hops that belong with it, as well as supplementary data.

The routing table determines the next node that a data packet must reach in order to get to the desired destination. Routing tables can be:

- static, or;
- dynamic.

In the case of static routing, the next-hop of a route is entered as a fixed location in the tables. Static routing is appropriate for smaller networks and networks with a simple topology. In the case of dynamic routing, the next hop is determined from network state information. Employment makes sense for larger networks with a complex topology and for the automatic path adaptation in case of error (backup), and in case of the overloading of the network parts.

3.3.6 References

ITU-T References

[I.232.1] (11/88) - Packet-mode bearer service categories: Virtual call and permanent virtual circuit bearer service category

[I.232.2] (11/88) - Packet-mode bearer service categories: Connectionless bearer service category

[I.232.3] (03/93) - Packet-mode bearer service categories: User

signalling bearer service category (USBS)

[I.233] (10/91) - Frame mode bearer services, ISDN frame relaying bearer service and ISDN frame switching bearer service

[I.233.1 Annex] (07/96) - Frame mode bearer services: ISDN frame relaying bearer service - Annex F: Frame relay multicast

General References

Schwartz, M.: Telecommunication Networks.- Reading: Addison-Wesley, 1988

3.4 TELEPHONE SWITCHING TECHNOLOGY

Telephone switching technology is the technical basis of what is applied for the switching of connections in analogue and digital networks for the telephony service and in ISDN. It is characterised by the switching of narrow band channels.

The telephone network is the oldest telecommunication network in the world. The first switching functions were also introduced into this network.

1877	First telephone switching (manual switching in USA)
1892	First automatic switching (USA)
1965	First fully electronic local switching system (USA)

Table 3.1A - Development of the telephone switching technology

1881	First telephone exchange in Germany (Berlin, 8 subscribers)
1908	First automatic switching in Europe (Hildesheim, 900 subscribers)
1923	First fully automatic switching beyond the local region (Weilheim)
1970	Total-area coverage self-dialling service in Germany
1975	Computer-controlled local switching technology in Germany
1984	First digital remote switching station in Germany
1985	First digital local switching station in Germany
1998	Completion of the total digitalisation of the telephone network in Germany

Table 3.1B - Development of the telephone switching technology in Germany

The worldwide telephone network today has a structure as shown in Figure 3.17.

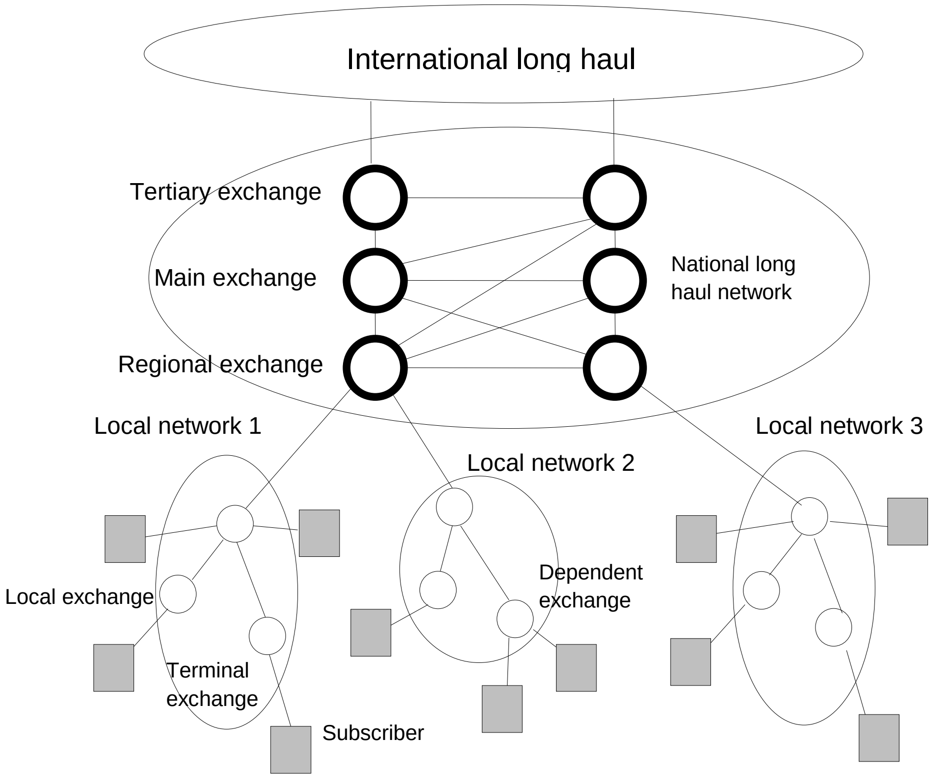


Figure 3.17 - Structure of the worldwide telephone network

3.4.1 Local network

In the lowest level of the telephone network is the local network to which the subscriber is connected. It is made up of local exchanges, terminal exchanges and dependent exchanges which are controlled remotely from local exchanges.

Local networks can be of different sizes. While on the one hand, digital concentrators can be employed for very small local networks with up to a few hundred subscribers, if the number of subscribers is a few thousand then remote-controlled switching stations are employed. Very large local networks can have up to 100,000 subscribers. They are implemented with independent local exchanges.

The subscriber is connected to the local network by means of subscriber lines. The local exchanges are tied together by local trunk lines.

3.4.2 Long haul network

Local networks are connected through national long haul networks. These are mapped by the regional exchanges, main exchanges and tertiary exchanges.

This structure can also be expressed in the subscriber numbering; i.e. within a local network, only the telephone number of the subscriber is selected in order to connect to a another subscriber in the same local network. From outside the local network, the user must dial the local network code and furthermore, for a subscriber in another country, the country code.

Local networks and long haul networks internally can contain a number of hierarchical levels; in some countries, though, no difference is made between the local and the long-distance level.

It is possible, that the actual path that a connection takes in the network does not follow the hierarchy set by the numbering. By means of so-called traffic routing, shorter and therefore more efficient paths are possible. Digital, computer-controlled telecommunications systems contain numbering schemes that are independent of the hierarchical structure of the network.

The national long haul networks of the individual countries are again networked through the international long haul network. This is subdivided once more into two network levels: the intercontinental long-distance network has exchanges in New York, London, Sydney, Moscow and Tokyo. The sub-level is constructed by the continental long-distance networks. The continental long-distance networks have the following codes:

1:	North America,
2:	Africa
3 & 4:	Europe
5:	North America
6:	Australia, Oceania
7:	Russian Federation
8:	Asia without Russia, India and the Arabic countries
9:	India and the Arabic countries

3.5 CONNECTIONLESS MESSAGES TRANSFER

3.5.1 Principles

In connectionless message transfer, the transfer is carried out in packets that include both the source and the destination addresses. All packets reach all network nodes and terminals of the respective network. Every receiver looks for and retrieves "his own" messages based on the address information given.

This form of message transfer is especially used in networks for data transmission, for example, LAN or WAN applications. The advantage of this method lies in the ability to send information without previously setting up a connection. Additionally, no routing mechanism is required. This is especially advantageous for sporadically occurring, short information relationships.

The transmission is possible only in frames or packets. Since the packets contain source and sink addresses, no connection set-up and termination is required. The packets are transmitted spontaneously. But the availability of sufficient resources in the entire network cannot be guaranteed, nor whether the sink has the ability to accept the transmitted information. Therefore measures are required to ensure that a message has really reached the sink.

This is implemented with protocols, at higher levels of the OSI reference model.

A shared medium is a transmission medium that is used by a number of communication relationships. The transmission capacity for a specific connection is dependant upon the traffic of all other communication relationships.

Media access. Since no connections for individual information relationships have been made, all existing information relationships must share the transmission medium (shared medium). For this reason, there is always a time frame and regulation for the media access. This can either be based on chance and uncoordinated, i.e. access is not previously agreed upon with other stations (random access), or the stations are given transmission rights at predetermined time slots (token access).

Network topologies. Figure 3.18 shows the possible network configurations for connectionless message transfer. One can see that no hierarchical composition of the network is possible as would be the case with tree or meshed networks.

The interconnection of connectionless networks, which would imply the creation of hierarchies, makes it necessary to selectively make a distinction between internal traffic (source and sink are contained in the same network) and external traffic (source and sink are located in different networks). For this purpose, bridges and routers have become typical network elements of LANs and WANs. They analyse the address information of the data packets and filter the external traffic for the transfer to the next higher network level.
If connectionless service in networks with connection set-up is offered, special network nodes (servers) are required which accept connectionless traffic and after analysing the address information, pass it on. This causes a logical sub-network of fixed address connections to be created for the connectionless traffic.

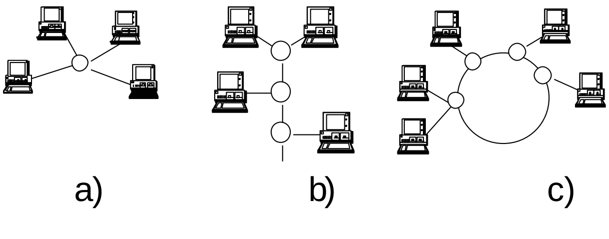


Figure 3.18 - Network topologies for connectionless messages transfer

a) Star network, b) Bus network, c) Ring network

3.5.2 Individual techniques

CSMA/CD (carrier sense multiple access / with collision detection). This method applies the probabilistic access on the transmission medium which is not synchronised with other stations. The medium is queried for a short period of time before transmission. If it is free, the station transmits, otherwise a waiting period must pass and then the medium is queried again. A collision can occur if a number of stations have 'queried' at the same time and then begun to transmit as soon as the medium is free. CSMA/CD is standardised in IEEE 802.3. The network topology is a bus (Figure 3.18 b)). A typical example of CSMA/CD networks is Ethernet. Ethernet can reach a transmission capacity in excess of 10 Mbit/s. Currently work is being conducted on the standardisation of a gigabit Ethernet which should reach a transmission capacity of 1Gbit/s and will also be applicable for wide-area networks.

Token Ring. The token model is a deterministic media access process with a decentralised control system. A transmission permission (token) is passed on from station to station. A station ready to transmit occupies a free token and sends a message. In this way, a new token is created. In the token ring process, the token circulates on a physical ring. The network topology is represented by Figure 3.18 c). The token transfer is carried out along the physical ring. A typical token ring process is the IBM token ring as described in IEEE 802.5.

Token Bus. With this method, all stations connected on a bus (see Figure 3.18 b)) form a logical ring. The token transfer forward is carried out with the addresses of the connected stations. The addresses of the previous and subsequent stations must be known. An example of a token bus process is described in IEEE 802.4.

FDDI (Fibre Distributed Data Interface). FDDI uses the token bus process in a double ring structure with counter-directional rings constructed of fibre optic connections. The data is transported in packets of variable length. FDDI systems are designed to be error-tolerant and are conceived for a high transmission capacity in a High Speed LAN –(HSLAN) of up to 100 Mbit/s. The access procedure permits synchronous service as well as asynchronous data transmission. In this case, every station is assigned a fixed part of the bandwidth.

DQDB (Distributed Queue Dual Bus). While FDDI, token model, and CSMA/CD were developed for the transmission in local area networks, DQDB is the transmission procedure in MAN (Metropolitan Area Networks). It is described in the standard IEEE 802.6.

For DQDB, the transmission is carried out with a frame structure on a double bus running in opposite directions. Depending on which direction the sink is located which is to receive messages from a station, a transmission is requested on the bus of the opposite direction. If a free slot in the desired transmission direction arrives, then it is occupied. With this process, a distributed wait queue develops at each of the stations. The stations can transmit their information with equal rights and without conflicts depending on the general state of the network.

3.6 ABBREVIATIONS

ATM	Asynchronous Transfer Mode
BORSCHT	Battery, Over voltage protection, Ringing, Signalling, Coding, Hybrid, Test
CSMA/CD	Carrier Sense Multiple Access / with Collision Detection)
DQDB	Distributed Queue Dual Bus
DSS1	Digital Subscriber Signaling system No.1
ETSI	European Telecommunications Standards Institute
FDDI	Fibre Distributed Data Interface).
FIFO	First In First Out (normally relating to buffers)
GSM	Group Special Mobile (ETSI committee on second generation cellular systems)
HSLAN	High Speed Local Area Network
IEEE	Institute of Electrical and Electronic Engineers
ISDN	Intergrated Services Digital Network
LAN	Local Area Network
MAN	Metropolitan Area Networks
NNI	Network Network Interface
OSI	Open Systems Interconnection
PCM	Pulse Code Modulation
QoS	Quality of Service
SMS	Short Message Service
UNI	User Network Interface
USBS	User Signalling Bearer Service
WAN	Wide Area Network