COMMUNICATION ENGINEERING

About the Author

S Vijayachitra obtained a BE degree in Electronics and Communication Engineering from Mookambigai College of Engineering, Pudukkottai. She then obtained her ME degree on Process Control and Instrumentation Engineering from Annamalai University, Chidambaram; and PhD in Electrical Engineering from Anna University, Chennai, in the years 1994, 2001 and 2009 respectively. During her postgraduate degree programme, she received the first rank for academic excellence.

Currently, she is serving as Professor at Kongu Engineering College, Perundurai, Tamil Nadu. During her teaching service, she received the Best Staff award in 2005 from Kongu Engineering College, Perundurai.

She has published three textbooks and more than 50 papers in various international journals, and international and national conference proceedings.

Dr Vijayachitra has organised many seminars, workshops, STTP programmes funded by AICTE, ICMR, CSIR, BRNS, etc. Currently, she is guiding 12 research scholars under Anna University, Chennai. Her areas of interest include Neural Networks, Fuzzy Logic, Genetic Algorithms, Process Modeling, etc.

COMMUNICATION ENGINEERING

S Vijayachitra
Professor
Kongu Engineering College
Perundurai, Tamil Nadu

McGraw Hill Education (India) Private Limited
NEW DELHI

McGraw Hill Education Offices
New Delhi New York St Louis San Francisco Auckland Bogotá Caracas
Kuala Lumpur Lisbon London Madrid Mexico City Milan Montreal
San Juan Santiago Singapore Sydney Tokyo Toronto

McGraw Hill Education (India) Private Limited

Published by McGraw Hill Education (India) Private Limited
P-24, Green Park Extension, New Delhi 110016

Communication Engineering

ISBN (13) : 978-1-25900686-9
ISBN (10) : 1-25-900686-7

Vice President and Managing Director: *Ajay Shukla*

Head—Higher Education (Publishing and Marketing): *Vibha Mahajan*

Publishing Manager (SEM & Tech. Ed.): *Shalini Jha*
Editorial Executive: *Koyel Ghosh*

Manager—Production Systems: *Satinder S Baveja*
Assistant Manager—Editorial Services: *Sohini Mukherjee*
Production Executive: *Anuj K Shriwastava*

Assistant General Manager (Marketing)—Higher Education: *Vijay Sarathi*
Senior Product Specialist: *Tina Jajoriya*
Senior Graphic Designer—Cover: *Meenu Raghav*

General Manager—Production: *Rajender P Ghansela*
Manager—Production: *Reji Kumar*

Information contained in this work has been obtained by McGraw Hill Education (India), from sources believed to be reliable. However, neither McGraw Hill Education (India) nor its authors guarantee the accuracy or completeness of any information published herein, and neither McGraw Hill Education (India) nor its authors shall be responsible for any errors, omissions, or damages arising out of use of this information. This work is published with the understanding that McGraw Hill Education (India) and its authors are supplying information but are not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought.

Typeset at Print-O-World, 2579, Mandir Lane, Shadipur, New Delhi 110 008, and printed at

Cover Printer :

Brief Content

1. **Introduction to Communication Systems** 1

2. **Amplitude Modulation** 49

3. **Amplitude Demodulation** 104

4. **Angle Modulation** 135

5. **Frequency Demodulation** 165

6. **Pulse Modulation** 186

7. **Digital Modulation** 224

8. **Information Theory and Coding** 278

9. **Data Communication and Networks** 329

10. **Optical-Fibre Communication** 387

11. **Satellite Communication** 461

12. **Radar Principles** 516

13. **Wireless Communication** 564

14. **Transmission Lines** 626

Appendix 643

Index 679

Content

Preface *xiii*

1. Introduction to Communication Systems **1**

1.1 Introduction *1*
1.2 Importance of Electronic Communication *2*
1.3 History of Electronic Communication *3*
1.4 Components of Communication *3*
1.5 Types of Communication *5*
1.6 Types of Electrical Signals *5*
1.7 Classification of Signals *7*
1.8 Representation of Signals *19*
1.9 Fourier Series *21*
1.10 Fourier Transform *25*
1.11 Noise in Communication Systems *33*
1.12 Mathematical Models for Communication Channels *38*
1.13 Electromagnetic Spectrum *41*
Summary 45
Review Questions 47

2. Amplitude Modulation **49**

2.1 Introduction *49*
2.2 Need for Modulation *50*
2.3 Types of Modulation *51*
2.4 Amplitude Modulation *53*
2.5 Modulation by Several Sine Waves *70*
2.6 Double-Side-Band Suppressed-Carrier Amplitude Modulation (DSB-SC-AM) *71*
2.7 Single Side-Band Suppressed-Carrier Amplitude Modulation (SSB-SC AM) *75*
2.8 Vestigial Side-Band AM (VSB AM) *80*
2.9 Applications of AM *82*
2.10 Generation of AM Waves *82*
2.11 AM Transmitters *93*
Summary 101
Review Questions 101
Problems 103

3. Amplitude Demodulation **104**
3.1 Introduction *104*
3.2 Types of AM Detectors *104*
3.3 AM Receivers *113*
Summary 132
Review Questions 133

4. Angle Modulation **135**
4.1 Introduction *135*
4.2 Frequency Modulation *136*
4.3 Phase Modulation *149*
4.4 Generation of FM *153*
4.5 FM Transmitters *158*
4.6 Advantages and Disadvantages of FM *162*
Summary 163
Review Questions 163

5. Frequency Demodulation **165**
5.1 Introduction *165*
5.2 Types of FM Detectors *165*
5.3 Demodulation of Phase-Modulated Waves *176*
5.4 FM Receivers *177*
Summary 183
Review Questions 184

6. Pulse Modulation **186**
6.1 Introduction *186*
6.2 Definition *186*
6.3 Types of Pulse Modulation *187*
6.4 Analog Pulse Modulation *187*
6.5 Pulse Amplitude Modulation (PAM) *190*
6.6 Pulse Duration Modulation (PDM) *197*
6.7 Pulse Position Modulation (PPM) *203*
6.8 Digital Pulse Modulation *205*
Summary 221
Review Questions 222

7. Digital Modulation **224**
7.1 Introduction *224*
7.2 Need for Multiplexing *224*
7.3 Concept of Multiplexing *225*
7.4 Time-Division Multiplexing (TDM) *225*
7.5 Frequency-Division Multiplexing (FDM) *232*
7.6 Comparison Between TDM and FDM Methods *237*
7.7 Digital Modulation *237*
7.8 Amplitude Shift Keying (ASK) *240*
7.9 Frequency-Shift Keying (FSK) *246*
7.10 Phase-Shift Keying (PSK) *250*
7.11 Modems *263*
7.12 Quanisation Process *270*
Summary 275
Review Questions 276

8. Information Theory and Coding **278**
8.1 Introduction *278*
8.2 Information Theory *278*
8.3 Entropy *282*
8.4 Conditional Entropy *284*
8.5 Information Rate *284*
8.6 Channel Capacity *286*
8.7 Examples of Channel Capacity *288*
8.8 Source Coding *290*
8.9 Shannon–Fano Coding *293*
8.10 Huffman Coding *298*
8.11 Intersymbol Interference (ISI) *301*
8.12 Eye Pattern *304*
8.13 Error Control *306*
8.14 Methods of Error Detection *309*
8.15 Methods of Error Correction *317*
Summary 325
Review Questions 326

9. Data Communication and Networks **329**
9.1 Introduction *329*
9.2 Data-Communication Codes *329*
9.3 Data Transmission *334*

9.4 Bit Rate *336*
9.5 Reception of Data Bits *336*
9.6 Modes of Data Transmission *337*
9.7 Transmission Channel *339*
9.8 Directional Capabilities of Data Exchange *340*
9.9 Data Communication Hardware *341*
9.10 Serial Interfaces *341*
9.11 Parallel Interface *346*
9.12 Data Communication Networks *347*
9.13 Applications of Networking *348*
9.14 Components of a Computer Network *349*
9.15 Network Functions *350*
9.16 Network Models *353*
9.17 Network Topologies *356*
9.18 Network Classifications *362*
9.19 Layered Architecture of a Computer Network *364*
9.20 Open System Interconnection (OSI) *367*
9.21 Layered Architecture of the OSI Reference *367*
9.22 Integrated Services Digital Network (ISDN) *371*
9.23 Local Area Networks (LAN) *375*
Summary 382
Review Questions 384

10. Optical-Fibre Communication 387
10.1 Introduction *387*
10.2 Optical-Fibre Communication *388*
10.3 Generations of Optical-Fibre Communication *389*
10.4 Wavelength and Spectra *389*
10.5 Advantages of Optical-Fibre Communication *391*
10.6 Disadvantages of Optical-Fibre Communication *392*
10.7 Propagation of Light Waves in an Optical Fibre *393*
10.8 Optical Fibres *406*
10.9 Structure of an Optical Fibre *407*
10.10 Types of Optical Fibres *408*
10.11 Optical-Fibre Configurations *421*
10.12 Losses in Optical Fibres *425*
10.13 Optical Connectors and Splices *429*
10.14 Optical Sources *432*
10.15 Optical Detectors *443*

10.16 Limitations of Optical-Cable Communication *447*
10.17 Applications of Optical Fibres *447*
10.18 Wavelength Division Multiplexing (WDM) *453*
Summary 458
Review Questions 459

11. Satellite Communication 461
11.1 Introduction *461*
11.2 Features of Satellite Communication *462*
11.3 Satellite-Frequency Bands *462*
11.4 Advantages of Using Satellite Communication *463*
11.5 Basic Elements of Satellite Communication *464*
11.6 Applications of Satellite Communications *464*
11.7 Block Diagram of a Satellite-Communication System *466*
11.8 Types of Satellites *470*
11.9 Theory of Satellite Communication *471*
11.10 Satellite Orbits *473*
11.11 Satellite Earth-Station Subsystem *493*
11.12 Satellite Transponder Subsystem *496*
11.13 Satellite Launching *497*
11.14 Radio-Wave Transmission *499*
11.15 Tropospheric Scattering *501*
11.16 Satellite Antennas *502*
11.17 Radiation Pattern *506*
Summary 511
Review Questions 513

12. Radar Principles 516
12.1 Introduction *516*
12.2 What is Radar? *517*
12.3 Principles of Radar Measurement *517*
12.4 Basic Radar System *518*
12.5 Classification of Radar Systems *526*
12.6 Basic Pulse Radar System *529*
12.7 Antennas Used in Radar System *532*
12.8 Radar Displays *538*
12.9 Search Radar *543*
12.10 Tracking Radar *543*

12.11 Moving Target Indicator (MTI) *543*
12.12 Continuous-Wave (CW) Radar *552*
12.13 Frequency-Modulated CW Radar *553*
12.14 Tracking Radar *555*
Summary 560
Review Questions 561

13. Wireless Communication 564
13.1 Introduction *564*
13.2 Needs of Wireless Communication *564*
13.3 Examples of Wireless Equipment *565*
13.4 Transmission Media for Wireless Communication *565*
13.5 Applications of Wireless Technology *565*
13.6 Range of Wireless Services *566*
13.7 History of Mobile Telephones *566*
13.8 Principles of Mobile Communication *567*
13.9 Advanced Mobile Phone Service (AMPS) 575
13.10 Digital Cellular System *581*
13.11 Multiple-Access Techniques *581*
13.12 Wireless LANs *586*
13.13 Wireless Personal Area Network (PAN) *591*
13.14 Bluetooth *591*
13.15 Zigbee *597*
13.16 Caller ID *605*
13.17 Cordless Telephones *606*
13.18 Pager *610*
13.19 Fascimile (Fax) *613*
Summary 622
Review Questions 624

14. Transmission Lines 626
14.1 Introduction *626*
14.2 Types of Transmission Lines *626*
14.3 A Model of Electrical Transmission Line *627*
14.4 Transmission-Line Theory *628*
14.5 Characteristic Impedance *630*
14.6 Line Losses *632*

14.7 Transmission of Energy *633*
14.8 Standing-Wave Ratio *633*
14.9 Transmission Media *634*
Summary 640
Review Questions 641

Appendix 643

Index 679

Preface

In various ways, computerised concepts, broadband communication and fibre-optic cables have helped us establish ways of effective communication over long distances. To meet the challenges and rapid advancements in our day-to-day lives, we await new revolutions in communication technology. This book has been designed to initiate the students and incite in them an eagerness to learn more about this subject area.

About the Book

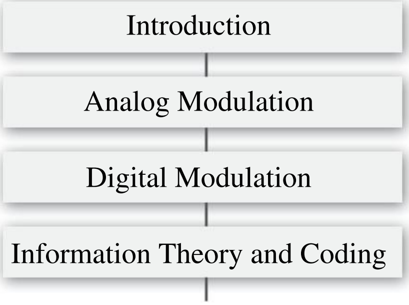
This textbook is specially prepared to give complete knowledge to students of Communication Engineering. It will definitely be useful for undergraduate and postgraduate students (both communication and non-communication majors) who want to acquire in-depth knowledge in both analog and digital communication systems. Academicians and professionals who want to gain thorough details of various communication systems will also be benefitted by this text. To provide maximum topical coverage, the contents of the book have been decided carefully by considering the curriculum of all reputed universities.

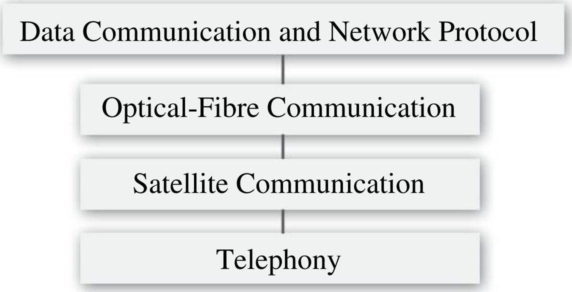
For simplifying the subject matter and solving complicated problems, a large number of examples have been discussed and presented in this book. Another important feature of this book is that it contains numerous diagrams. Every chapter begins with an introduction and ends with a detailed summary, added for an elaborative explanation of the subject contents.

I believe that the features mentioned above will help students learn and understand the subject in a clear manner and hope that the task has been carried out successfully.

Highlights

Multiple years of teaching experience and careful analysis of the syllabus, as offered in various technical universities across India, reveals the following major components of a course on Communication Engineering.




Data Communication and Network Protocol
Optical-Fibre Communication
Satellite Communication
Telephony

In the table below, I have provided a roadmap that will help students and teachers alike in covering the course using the contents of this book.

Modules	Chapters
Introduction	Chapter 1: Introduction to Communication Systems
Analog Modulation	Chapter 2: Amplitude Modulation Chapter 3: Amplitude Demodulation Chapter 4: Angle Modulation Chapter 5: Frequency Demodulation
Digital Modulation	Chapter 6: Pulse Modulation Chapter 7: Digital Modulation
Information Theory and Coding	Chapter 8: Information Theory and Coding
Data Communication and Network Protocol	Chapter 9: Data Communication and Networks
Optical Fibre Communication	Chapter 10: Optical-Fibre Communication
Satellite Communication	Chapter 11: Satellite Communication
Telephony	Chapter 12: Radar Principles Chapter 13: Wireless Communication Chapter 14: Transmission Lines

Salient Features

- ➢ All-round topical coverage of analog modulation, satellite communication, network protocol, optical fibre communication
- ➢ Dedicated coverage of *data communication systems* and *data communication networks*
- ➢ Concepts explained in a lucid style
- ➢ Rich pedagogy including
  - 464 diagrams
  - 175 Solved Examples
  - Over 545 Review Questions
  - Over 70 Problems

Chapter Organisation

The oraganisation of the book is as follows.

Chapter 1 discusses the purpose and evolution of communication. It also deals with the components involved in communication systems and various signals involved with their representation. Furthermore, the mathematical models of various communication channels and electromagnetic spectrum are also described.

Chapter 2 describes the need of modulation and its various types. It specially discusses Amplitude Modulation (AM), Double Side Band Suppressed Carrier AM (DSB-SC-AM), Single Side Band Suppressed Carrier AM (SSB-SC-AM) and Vestigial Side Band AM (VSB-AM) in detail. Also, the details of AM generation and AM transmitters are discussed.

Chapter 3 introduces various types of AM detectors and AM receivers. Furthermore, Automatic Gain Control (AGC) and its different types are described in detail.

Chapter 4 discusses details of Angle Modulation such as Frequency Modulation (FM) and Phase Modulation (PM). For each modulation type, a comprehensive mathematical and theoretical description is provided. It also deals with FM generation and FM transmitters.

Chapter 5 describes FM detection by different methods and FM transmitters.

Chapter 6 introduces Pulse Modulation and its types such as Pulse Amplitude Modulation (PAM), Pulse Duration Modulation (PDM) and Pulse Position Modulation (PPM) with their mathematical representation, graphical form, different generation methods, transmission and reception in detail. In addition, digital modulation, comprising of Pulse Code Modulation (PCM) and Delta Modulation (DM), is also dealt with in detail.

Chapter 7 provides detailed description about Amplitude Shift Keying (ASK), Frequency Shift Keying (FSK) and Phase Shift Keying (PSK). Furthermore, the process of quantisation is also discussed.

Chapter 8 gives comprehensive theoretical description about information theory in a communication system and also provides details about various error-control and error-detection methods.

Chapter 9 discusses data-communication networks with their functions, components and topologies and also provides details about Integrated Services Digital Network (ISDN) and Local Area Network (LAN).

Chapter 10 discusses fundamental concepts behind fibre-optic communication. It deals with the propagation characteristics, structure and various types of an optical fibre in a clear manner. Furthermore, details about different optical sources and optical detectors used in optical-fibre communication and various applications are also described.

Chapter 11 describes necessary fundamentals about satellite communication. The types, various orbits, earth-station subsystem and transponder subsystem, satellite-launching procedures, satellite antennas and radiation patterns are also discussed in an elaborative manner.

Chapter 12 introduces principles and functions of a radar system. Pulse radar system, radar displays and various antennas used in a radar system are also discussed in detail.

Furthermore, the details about moving-target indicator, continuous-wave radar and frequency-modulated CW radar and tracking radar are also covered.

Chapter 13 discusses wireless communication, detailed view of mobile communication, wireless LAN, PAN, Bluetooth, Zigbee and Caller ID. It also covers cordless telephones, pager and facsimile systems.

Finally, **Chapter 14** describes the fundamental concepts behind the transmission lines. Specifically, this chapter deals with different types, model of a transmission line, transmission line losses, Standing Wave Ratio (SWR) and its types, and different transmission media.

Online Learning Centre

The OLC for this book can be accessed at http://www.mhhe.com/vijayachitra/ce1 and contains the Solution Manual.

Acknowledgements

I gratefully acknowledge the contributions made towards the preparation of this textbook by colleagues at Kongu Engineering College, Perundurai. My sincere thanks to Prof. S Kuppuswamy, Principal, Kongu Engineering College, and the Management, for their continuous encouragement and valuable support given to me to carry out this work.

I wish to express my gratitude to Prof. K Krishnaswamy, Dean (CDSA), Kongu Engineering College, for his valuable suggestions and motivation to begin and successfully complete the book. My special thanks to Dr U S Ragupathy, Head of the Department, and beloved staff members of Electronics and Instrumentation Engineering for their kind help and support in all ways.

My sincere appreciation goes out to the professors who reviewed the contents of the book and offered their feedback, criticism and suggestions to shape the book. Their names are given below:

Vikas Gupta	*Technocrats Institute of Technology (TIT)* *Bhopal, Madhya Pradesh*
Rita Jain	*Lakshmi Narain College of Technology (LNCT)* *Bhopal, Madhya Pradesh*
R N Yadav	*Maulana Azad National Institute of Technology (MANIT)* *Bhopal, Madhya Pradesh*
Sameer Srivastava	*Bansal Institute of Engineering and Technology* *Lucknow, Uttar Pradesh*
Manish Kumar Jha	*Mody Institute of Technology and Science* *Lakshmangarh, Rajasthan*
Subhankar Bhattacharya	*Techno India College of Technology* *Kolkata, West Bengal*

Goutam Nandi	*Siliguri Government Polytechnic* *Siliguri, West Bengal*
A H Ansari	*Pravara Rural College of Engineering* *Loni, Maharashtra*
Joshi Maulin Mahesh	*Sarvajanik College of Engineering and Technology* *Surat, Gujarat*
V Dinesh	*Kongu Engineering College* *Perundurai, Tamil Nadu*
K Vasudevan	*Cochin University of Science and Technology* *Kochi, Kerala*
M Arun	*Bannari Amman Institute of Technology* *Sathyamangalam, Tamil Nadu*
R Saranya	*Dr N G P Institute of Technology* *Coimbatore, Tamil Nadu*
M Palanivelan	*Rajalakshmi Engineering College* *Chennai, Tamil Nadu*
S Narayanan	*R V College of Engineering, Bangalore, Karnataka*

I am indebted to McGraw Hill Education (India) Private Limited, New Delhi, for their intense interest in this project from the very beginning, starting at the receipt of chapter contents, review with experts, to-and-fro feedback mechanism and finally for publishing the textbook in time. I wish to thank Koyel Ghosh and Anuj Shriwastava for their commitment and involvement that has resulted in the successful publication of the book.

My thanks with love goes out to my husband, Mr S Palaniappan; sweet children, S P Madhumitha and S P Sharmistha, and lovable parents for their invaluable moral support and dedication during this work. Finally, I thank all those who have helped me directly or indirectly in the successful completion of my research work.

Feedback

Much care has been taken to avoid errors as far as possible. Some errors may still exist in the text and I will be grateful to the readers who give their feedback and suggestions for improving the next edition.

S Vijayachitra

Publisher's Note

Do you have any further request or suggestion? We are always open to new ideas (the best ones come from you!). You may send your comments to *tmh.ecefeedback@gmail.com*

Piracy-related issues may also be reported!

GUIDED TOUR

All-round topical coverage of **Analog Modulation, Satellite Communication, Data Communication, Network Protocol, Optical-Fibre Communication**

2

AMPLITUDE MODULATION

Objectives

- To know the need for modulation in a communication system
- To discuss the different types of modulation in detail
- To provide the process of Amplitude Modulation (AM) and its repr in detail.
- To provide the process of Double Side-Band Suppressed Carrier— Modulation (DSB-SC-AM) and its representation in detail
- To provide the process of Single Side-Band Suppressed Carrier— Modulation (SSB-SC-AM) and its representation in detail
- To know about the purpose of Vestigial Side Band (VSB) and its rep
- To discuss various methods of generation of AM and also functioning of AM transmitters in detail

10

OPTICAL-FIBRE COMMUNICATION

Objectives

- To know the purpose and different generations and wavelength spectra of optical fibre communication
- To discuss details about the propagation characteristics of optical fibres
- To discuss the structure and types of optical fibres and optical-fibre connectors
- To provide details about various losses to be considered in optical fibres
- provide the details about different optical sources and optical detectors n optical-fibre communication and also about various applications

11

SATELLITE COMMUNICATION

Objectives

- To know about the features and different satellite-frequency band
- To discuss the details of the satellite communication systems
- To discuss the types, various orbits, earth-station subsystem and tr subsystem
- To provide details about satellite-launching procedures, satellit and radiation patterns
- To provide details about radio-wave transmission

9

DATA COMMUNICATION AND NETWORKS

Objectives

- To discuss different data communication codes in detail
- To provide details about data communication, its hardware and interfaces
- To discuss about data communication networks with their functions, mponents and topologies
- ovide details about the ISDN and LAN

13

WIRELESS COMMUNICATION

Objectives

- To know the needs, examples, media and applications of wireless communication
- To discuss details about mobile communication and advanced mobile communication
- To discuss wireless LAN, PAN, Bluetooth and Zigbee
- To provide the details, the needs and principle of caller ID
- To provide the details of cordless telephones, and pager and facsimile systems.

9.3 DATA TRANSMISSION

There is always need to exchange data, commands and other control information between a computer and its terminals or between two computers. This information is in the form of bits.

Data transmission refers to movement of the bits over some physical medium connecting two or more digital devices. There are two options of transmitting the bits, namely, parallel transmission or serial transmission.

9.3.1 Parallel Transmission

In parallel transmission, all the bits of a byte are transmitted simultaneously on separate wires as shown in Figure 9.2 and hence multiple circuits interconnecting the two devices are required. It is practical only if the two devices are close to each other like a computer and its associated printer.

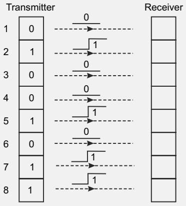


Fig. 9.2 Parallel transmission

Dedicated coverage of
Data Communication Systems and
Data Communication Networks

9.12 DATA COMMUNICATION NETWORKS

Data communication network is defined as any group of computer terminals connected together and the process of sharing resources between computer terminals over a data communication network is called **networking**. In other words, networking is two or more computer terminals linked together by means of a common transmission medium for the purpose of sharing the information or data.

The number of links L required between N nodes is calculated as follows.

$$L = \frac{N(N-1)}{2} \tag{9.3}$$

More than 550
illustrations and diagrams
to enhance the concepts

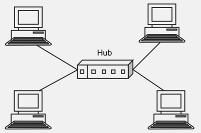


Fig. 9.21 Star configuration

install and reconfigure. Far less cabling needs to be housed and additions, moves and deletions involve only one connection between that device and the hub.

Other advantages include robustness. If one link fails, only that link is affected. Other links remain active. This factor also lends itself to easy fault identification and fault isolation. As long as the hub is working, it can be used to monitor link problems and bypass defective links.

However, although a star requires far less cable than a mesh, each node must be linked to a central hub. For this reason, more cabling is required in a star configuration.

3. Tree Topology

A tree topol
that controls
and the turn

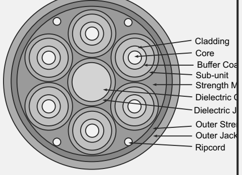


Fig. 10.23 A typical six-fibre inside-plant ca

provides protection to the entire enclosed fibre system. Kevl
constructing the outer strength member for premise cable sy
rated, plenum-rated or low-smoke, zero-halogen-rated.

Figure 10.24 shows a typical armoured outside-plant cable

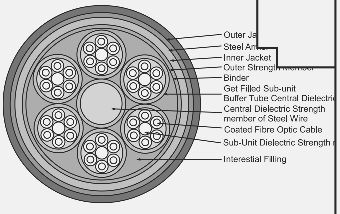


Fig. 10.24 A typical armoured outside-plant cable system

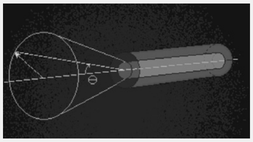


Fig. 10.13 Visualisation of acceptance angle in an optical fibre

10.7.9 Numerical Aperture (NA)

Numerical Aperture (NA) of the fibre is the light-collecting efficiency of the fibre and it is the measure of the amount of light rays that can be accepted by the fibre. This factor gives the relationship between the acceptance angle and the refractive indices of the three media involved, namely the core, cladding and the air. The NA is related to the acceptance angle θ_a, which indicates the size of a cone of light that can be accepted by the fibre.

Figure 10.14 shows a light ray incident on the fibre core at an angle θ_1 to the fibre axis which is less than the acceptance angle for the fibre θ_a. The ray enters the fibre from a medium (air) of refractive index n_0 and the fibre core has refractive index n_1 which is slightly greater than the cladding refractive index n_2.

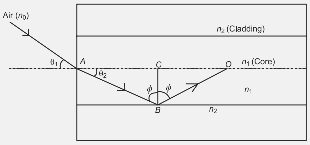
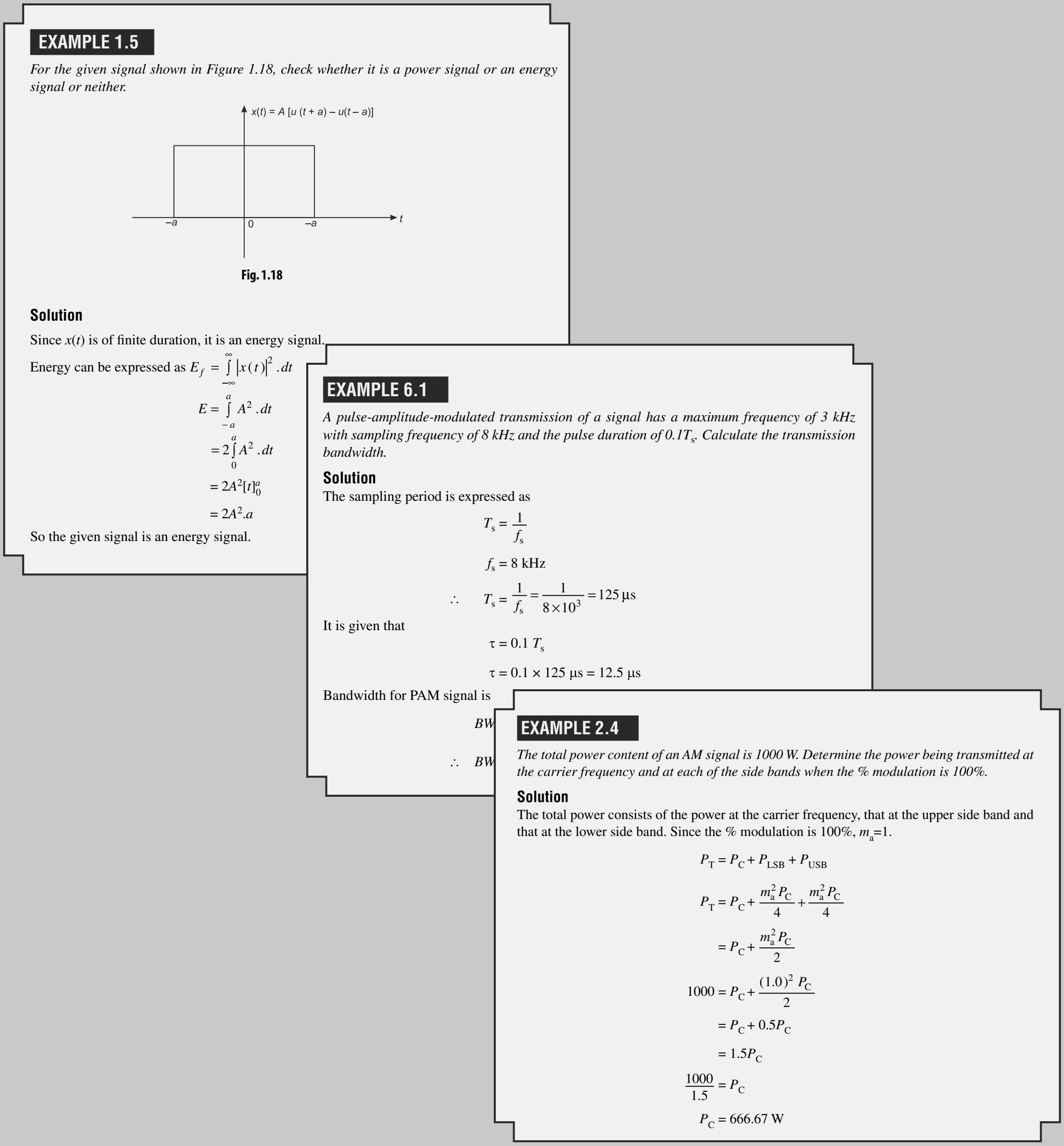


Fig. 10.14 Light ray path for a meridional ray

In each chapter, the theoretical derivations and technical descriptions are followed by a set of carefully chosen **Solved Examples**



EXAMPLE 1.5

For the given signal shown in Figure 1.18, check whether it is a power signal or an energy signal or neither.

Fig. 1.18

Solution

Since $x(t)$ is of finite duration, it is an energy signal.

Energy can be expressed as $E_f = \int_{-\infty}^{\infty} |x(t)|^2 .dt$

$$E = \int_{-a}^{a} A^2 .dt$$

$$= 2\int_{0}^{a} A^2 .dt$$

$$= 2A^2[t]_0^a$$

$$= 2A^2.a$$

So the given signal is an energy signal.

EXAMPLE 6.1

A pulse-amplitude-modulated transmission of a signal has a maximum frequency of 3 kHz with sampling frequency of 8 kHz and the pulse duration of 0.1T_s. Calculate the transmission bandwidth.

Solution

The sampling period is expressed as

$$T_s = \frac{1}{f_s}$$

$$f_s = 8 \text{ kHz}$$

$$\therefore \quad T_s = \frac{1}{f_s} = \frac{1}{8\times 10^3} = 125\,\mu\text{s}$$

It is given that

$$\tau = 0.1\ T_s$$

$$\tau = 0.1 \times 125\ \mu\text{s} = 12.5\ \mu\text{s}$$

Bandwidth for PAM signal is

BW

$\therefore \quad BW$

EXAMPLE 2.4

The total power content of an AM signal is 1000 W. Determine the power being transmitted at the carrier frequency and at each of the side bands when the % modulation is 100%.

Solution

The total power consists of the power at the carrier frequency, that at the upper side band and that at the lower side band. Since the % modulation is 100%, m_a=1.

$$P_T = P_C + P_{LSB} + P_{USB}$$

$$P_T = P_C + \frac{m_a^2 P_C}{4} + \frac{m_a^2 P_C}{4}$$

$$= P_C + \frac{m_a^2 P_C}{2}$$

$$1000 = P_C + \frac{(1.0)^2\ P_C}{2}$$

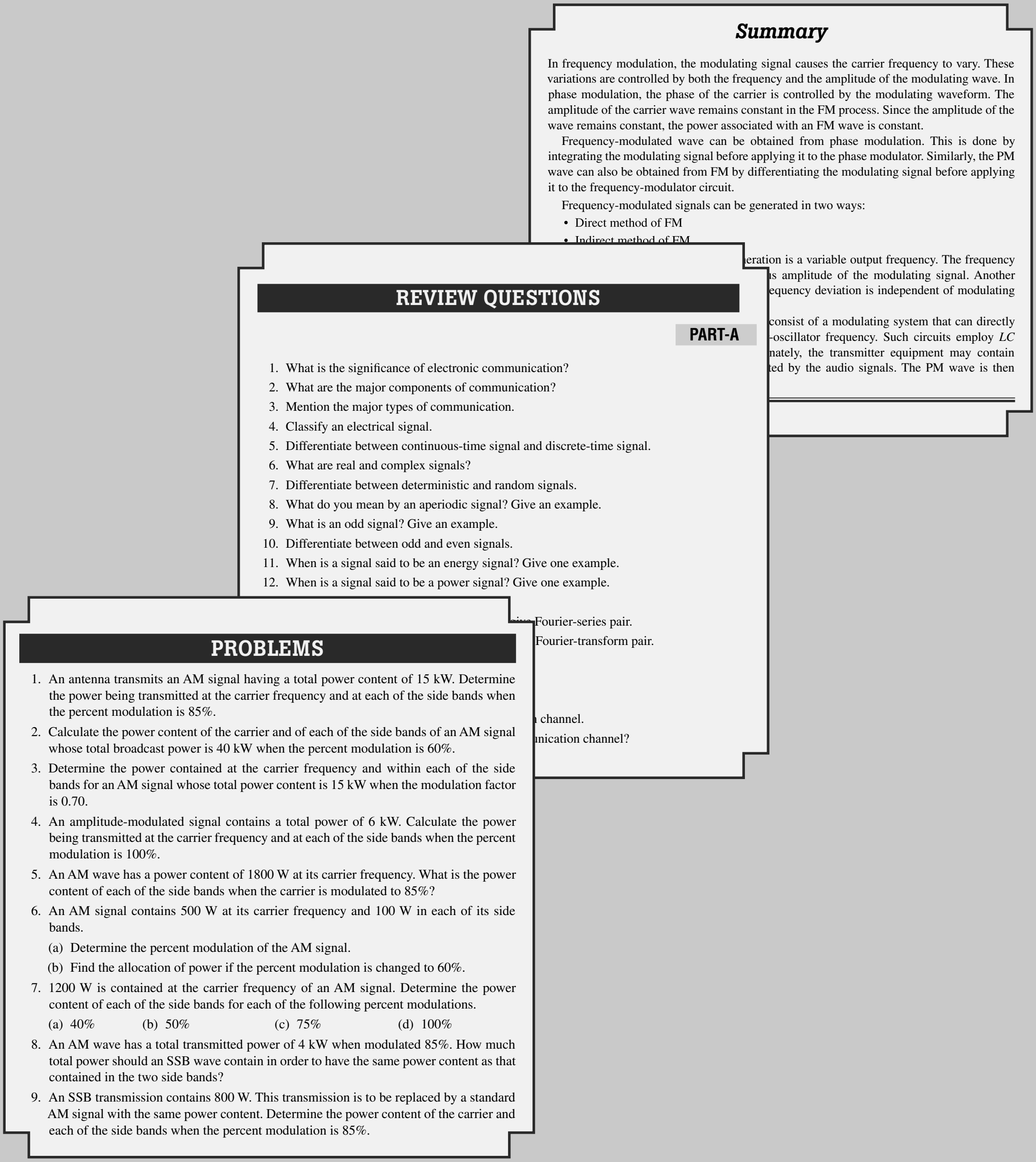
$$= P_C + 0.5P_C$$

$$= 1.5P_C$$

$$\frac{1000}{1.5} = P_C$$

$$P_C = 666.67 \text{ W}$$

- **Summary** in each chapter for quick recap of the concepts
- **Review Questions** to test the student's subjective grasp on the topics
- **Problems** for practice



Summary

In frequency modulation, the modulating signal causes the carrier frequency to vary. These variations are controlled by both the frequency and the amplitude of the modulating wave. In phase modulation, the phase of the carrier is controlled by the modulating waveform. The amplitude of the carrier wave remains constant in the FM process. Since the amplitude of the wave remains constant, the power associated with an FM wave is constant.

Frequency-modulated wave can be obtained from phase modulation. This is done by integrating the modulating signal before applying it to the phase modulator. Similarly, the PM wave can also be obtained from FM by differentiating the modulating signal before applying it to the frequency-modulator circuit.

Frequency-modulated signals can be generated in two ways:

- Direct method of FM
- Indirect method of FM

eration is a variable output frequency. The frequency
s amplitude of the modulating signal. Another
equency deviation is independent of modulating

consist of a modulating system that can directly
-oscillator frequency. Such circuits employ *LC*
nately, the transmitter equipment may contain
ted by the audio signals. The PM wave is then

REVIEW QUESTIONS

PART-A

1. What is the significance of electronic communication?
2. What are the major components of communication?
3. Mention the major types of communication.
4. Classify an electrical signal.
5. Differentiate between continuous-time signal and discrete-time signal.
6. What are real and complex signals?
7. Differentiate between deterministic and random signals.
8. What do you mean by an aperiodic signal? Give an example.
9. What is an odd signal? Give an example.
10. Differentiate between odd and even signals.
11. When is a signal said to be an energy signal? Give one example.
12. When is a signal said to be a power signal? Give one example.

Fourier-series pair.

Fourier-transform pair.

channel.

nication channel?

PROBLEMS

1. An antenna transmits an AM signal having a total power content of 15 kW. Determine the power being transmitted at the carrier frequency and at each of the side bands when the percent modulation is 85%.
2. Calculate the power content of the carrier and of each of the side bands of an AM signal whose total broadcast power is 40 kW when the percent modulation is 60%.
3. Determine the power contained at the carrier frequency and within each of the side bands for an AM signal whose total power content is 15 kW when the modulation factor is 0.70.
4. An amplitude-modulated signal contains a total power of 6 kW. Calculate the power being transmitted at the carrier frequency and at each of the side bands when the percent modulation is 100%.
5. An AM wave has a power content of 1800 W at its carrier frequency. What is the power content of each of the side bands when the carrier is modulated to 85%?
6. An AM signal contains 500 W at its carrier frequency and 100 W in each of its side bands.
   (a) Determine the percent modulation of the AM signal.
   (b) Find the allocation of power if the percent modulation is changed to 60%.
7. 1200 W is contained at the carrier frequency of an AM signal. Determine the power content of each of the side bands for each of the following percent modulations.
   (a) 40% (b) 50% (c) 75% (d) 100%
8. An AM wave has a total transmitted power of 4 kW when modulated 85%. How much total power should an SSB wave contain in order to have the same power content as that contained in the two side bands?
9. An SSB transmission contains 800 W. This transmission is to be replaced by a standard AM signal with the same power content. Determine the power content of the carrier and each of the side bands when the percent modulation is 85%.

1

Introduction to Communication Systems

Objectives

- To know the purpose of communication and evolution of electronic communication
- To provide the components of communication and the types of communication
- To provide the different types of signals and their representations
- To discuss various methods of frequency-domain representation of signals
- To discuss the importance and development of mathematical models for communication channels
- To understand the electromagnetic spectrum with its different regions of different wavelengths

1.1 INTRODUCTION

Generally speaking, the meaning of communication is sharing one's thoughts with others. Today, communication in our daily lives involves so many different forms.

Communication is a bidirectional process. One must have the feedback from the opposite end in order to know what to say next.

The process of establishing a connection or link between two points for information exchange is called **communication**. Simply, it is the process of conveying messages at a distance.

A **communication system** is an assembly of different electronic equipment which are mainly used for communication purpose.

Examples of Communication Systems

- Line telephony and line telegraphy
- Radio broadcasting
- Mobile communication
- Computer communication
- Satellite communication
- Point-to-point communication

1.2 IMPORTANCE OF ELECTRONIC COMMUNICATION

Electronic communication is the transmission, reception and processing of information between two or more locations using electronic circuits. The information can be in analog (continuous) form or in digital (discrete) form. All forms of information, however, must be converted to electromagnetic energy before being propagated. Figure 1.1 shows the basic block diagram of a communication system, and Figure 1.2 shows radio communication, a type of communication method.

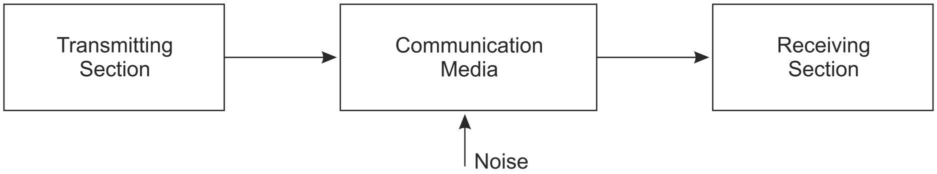


Fig. 1.1 Block diagram of a general communication system

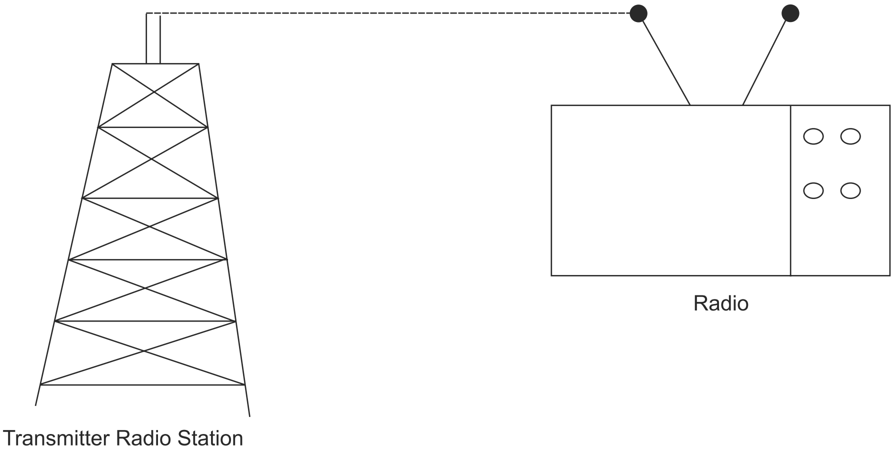


Fig. 1.2 Radio communication system

1.3 HISTORY OF ELECTRONIC COMMUNICATION

The first electronic communication was developed in 1837 by Samuel Morse. He used electromagnetic induction to transmit the information in the form of dots, dashes and spaces across a length of metallic wire and he named this invention the **Telegraph**.

In 1876, Alexander Graham Bell and Thomas A Watson were the first to successfully transmit human conversation over a crude telephone system.

Radio communication began in 1894 when Marchese Guglielmo Marconi transmitted the first wireless signals through the earth's atmosphere. AM radio broadcasting started in 1920 and commercial FM broadcasting started in 1936, after the invention of FM in 1933 by Major Edwin Howard Armstrong.

It became more widely used through the invention of the transistor, integrated, circuits and other semiconductor devices in the subsequent years. Today, linear integrated circuits have simplified circuit design, allowed for miniaturisation, improved performance and reliability and reduced overall cost.

The tremendous growth in electronic communication systems is primarily due to digital, microwave and satellite as well as optical fibre systems. There has been an increasing emphasis on the use of computers in communication.

1.4 COMPONENTS OF COMMUNICATION

A communication system is shown in Figure 1.3, which involves the following components.

1. Process of information
2. Transmission
3. Reception

The **process of information** includes

- Collection of data
- Processing of data
- Storage of the information

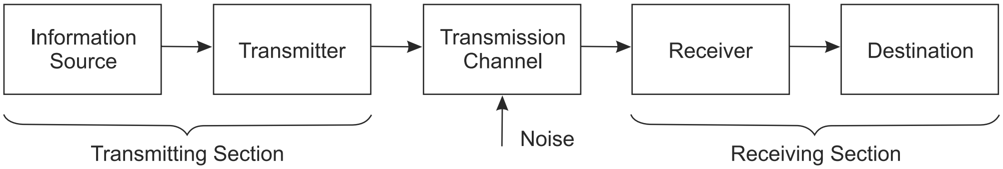


Fig. 1.3 A communication system

Next, **transmission** includes the further processing of data such as encoding. Finally, **reception** includes

- Decoding
- Storage of the data
- Interception

1.4.1 Information Source

The message produced by the information source is not electrical in nature, but it may be a voice signal, a picture signal, etc. So an input transducer is required to convert the original message into a time-varying electrical signal. These signals are called baseband signals (or) message signals (or) modulating signals.

At the destination, another transducer is used to convert the electrical signal into the appropriate message.

1.4.2 Transmitter

The transmitter comprising electrical and electronic components converts the message signal into a suitable form for propagating over the communication medium. This is often achieved by modulating the carrier signal which may be an electromagnetic wave. The wave is often referred as a **modulated signal**.

1.4.3 Channel

The transmitter and the receiver are usually separated in space. The channel provides a connection between the source and destination.

The channel can be of many forms like coaxial cable, microwave link, radio wave link (or) an optical fibre. Regardless of its type, the channel degrades the fixed signal in a number of ways which produces the signal distortion. This occurs in a channel

1. due to imperfect response of the channel and system,
2. undesirable electrical interference in the channel,
3. insufficient channel bandwidth, and
4. contamination of signals due to no use.

1.4.4 Receiver

The main function of the receiver is to extract the message signal from the degraded version of the transmitted signal. The transmitter and receiver are carefully designed to avoid distortion and minimise the effect of the noise from the receiver. The receiver has the task of operating on the received signal so as to reconstruct the original form of the original message signal and deliver it to the user destination.

1.5 TYPES OF COMMUNICATION

Various types of communication systems are listed in Figure 1.4.

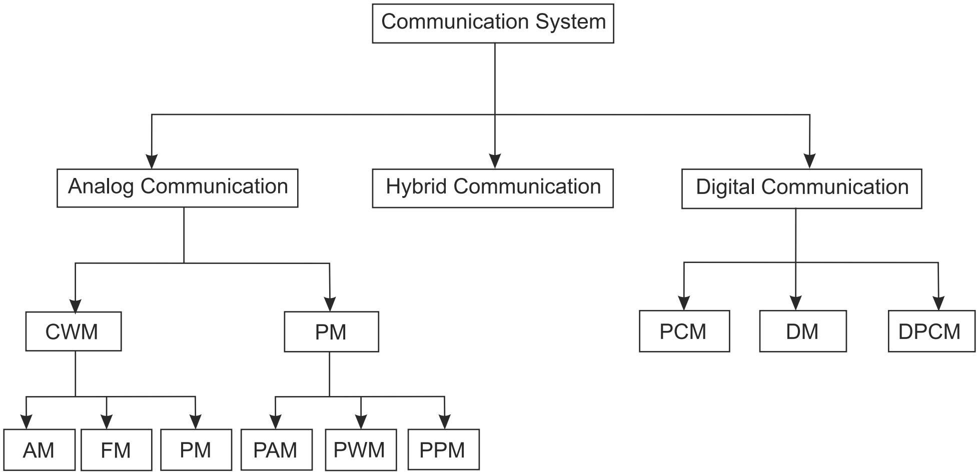


Fig. 1.4 Various communication systems

1.6 TYPES OF ELECTRICAL SIGNALS

The electrical signals produced by transducers at the transmitting section are of two types.

1. Analog signals that continuously vary with time, and
2. Digital signals which are not continuous.

The analog signals are analog in nature with or without harmonics and represent the variations of a physical quantity like a sound wave. The digital signals comprise a series of pulses that occur at discrete intervals of time.

1.6.1 Analog Signals

An analog signal is a continuous signal for which the time-varying feature of the signal is a representation of some other time-varying quantity. Telephone, radio broadcast or television signals are the common types of analog signals. The representations of those signals are voltage or current waveforms that have different amplitudes at different instants of time.

1. Telephone Signals

A telephone signal comprises speech sounds which produce audio waves making the diaphragm of the microphone vibrate. The diaphragm is attached to a coil surrounded by a static magnetic field on all the sides. The motion of the coil in this field causes an emf to be induced in the coil which is the electrical equivalent to the sound waves. Figure 1.5 shows a sample electrical signal equivalent to a speech signal.

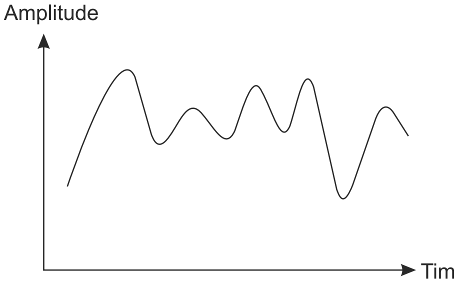


Fig. 1.5 Speech signal

2. *TV Picture Signals*

A TV picture signal comprises bright and dark spots called picture elements arranged in a particular sequence. These elements are systematically scanned by an electron beam and converted into an electric signal in a particular order. Figure 1.6 shows the electric signal equivalent to a TV picture signal produced by a TV camera.

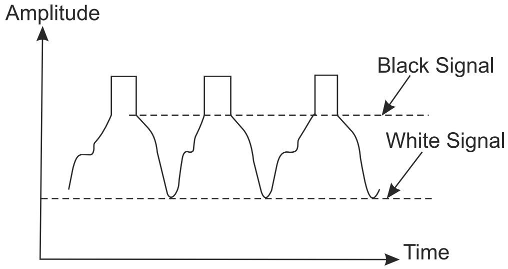


Fig. 1.6 TV picture signal

3. *Radio-Broadcast Signals*

The most popular analog signal used for entertainment and education is the radio-broadcast signal. These signals may be in the form of speech or music. If it is of the music type, the music occupies a larger bandwidth than for a speech-type signal. A frequency bandwidth of 5 kHz is usually employed for an ordinary radio-broadcasting programme.

1.6.2 Digital Signals

A digital signal is a physical signal that is a representation of a sequence of discrete values. They comprise pulses occurring at discrete intervals of time. The pulse may be notified at definite periods of time. These digital signals are very important in the transmission and reception of coded messages. Examples for digital signals include telegraph and teleprinter signals.

Sometimes analog signals are also converted into digital signals with the help of A/D converters before being transmitted.

1. *Telegraph Signal*

Signals from telegraphs and teleprinters are used to transmit written texts in the form of coded signals where codes are allotted to different characters.

A simple telegraph circuit consists of a telegraph key which switches on the current into the line when it is pressed and stops the current when the key is enclosed. The interval for which the current flows is termed the **MARK interval**, and the state of zero current is termed **SPACE interval**. Such a telegraph signal is shown in Figure 1.7.

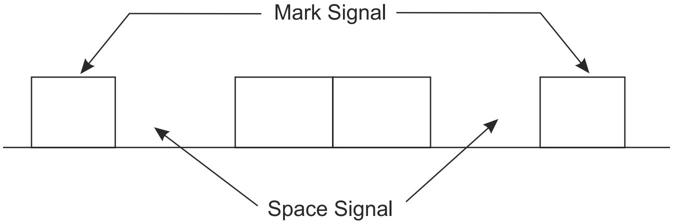


Fig. 1.7 A telegraph signal

2. Radar Signal

To find out the location of a distant object, a very popular device called radar is used. This is done by transmitting a short-period signal and focusing it to the target. The reflected signal is picked up by the radar receiver and is used to determine the location of the object.

Figure 1.8 shows a radar signal and it is basically a train of rectangular, pulses transmitted at a low pulses of repetitive frequency of around 1 kHz.

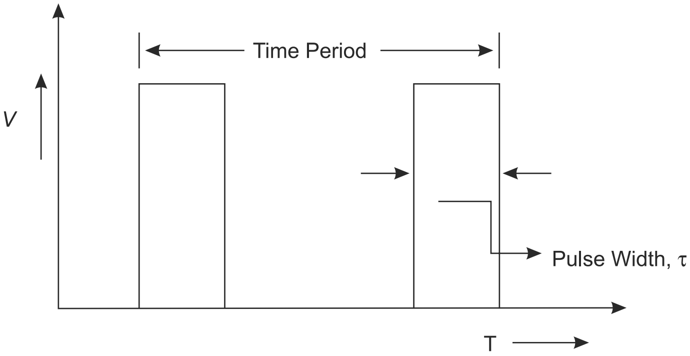


Fig. 1.8 A radar signal

3. Data Signals

Data signals are required to transmit very important data from one place to another. They include transmission of business, industrial, medical and statistical data for analysis via computers. It also includes transmission of data by satellites to earth stations where they are analysed. The data to be transmitted are converted into pulses and, therefore, data signals are in the form of digital signals. The bandwidth required varies from one system to another.

1.7 CLASSIFICATION OF SIGNALS

Signals can be classified under the following categories.

1. Continuous-time and discrete-time signals

2. Real and complex signals
3. Deterministic and random signals
4. Periodic and aperiodic signals
5. Even and odd signals
6. Energy and power signals

1.7.1 Continuous-Time and Discrete-Time Signals

A signal $x(t)$ is a continuous-time signal if t is a continuous variable. This continuous-time signal is defined continuously in the time domain.

If time t is a discrete variable, $x(t)$ is defined at discrete times and then $x(t)$ is a discrete-time signal. The discrete-time signal is identified as a sequence of numbers and is denoted by $x(n)$ where n is an integer. Figure 1.9(a) and (b) show a continuous-time signal and a discrete-time signal.

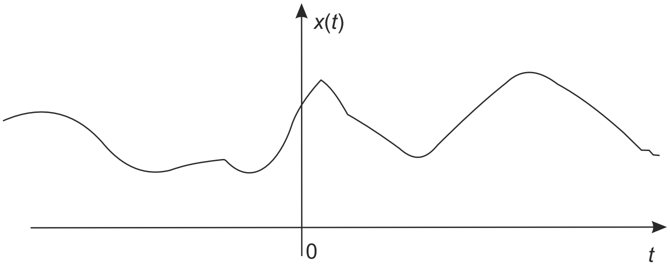


Fig. 1.9(a) A continuous-time signal

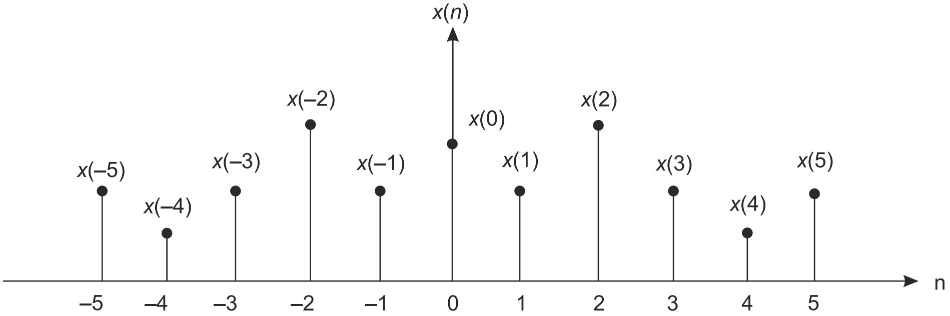


Fig. 1.9(b) A discrete-time signal

1.7.2 Real and Complex Signals

A signal $x(t)$ is a **real signal** if its value is a real number. Similarly, a signal $x(t)$ is a **complex signal** if its value is a complex number.

A continuous-time complex signal $x(t)$ and a discrete-time complex signal $x(k)$ are represented as follows.

$$x(t) = x_I(t) + jx_Q(t), \text{ and} \tag{1.1}$$

$$x(k) = x_I(k) + jx_Q(k) \tag{1.2}$$

where I and Q denote in-phase and quadrature components of the signal.

1.7.3 Deterministic and Random Signals

Deterministic signals are those signals which can be completely specified in time. The pattern of this kind of signal is regular and it can be characterised mathematically. The nature and amplitude of such a signal at any time can be predicted. For example, a deterministic signal $x(t) = kt$ is a ramp signal whose amplitude varies linearly with respect to time and the slope of the signal is represented as k.

Also, a deterministic signal has no uncertainty with respect to its value at any value of the independent variable. Figure 1.10 shows a rectangular pulse which is a deterministic signal.

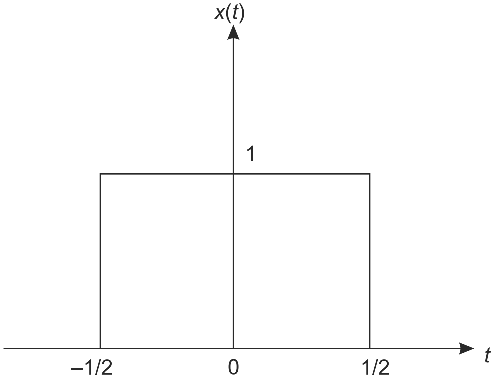


Fig. 1.10 A rectangular pulse

A **random signal** is one whose occurrence is always random in nature. The pattern of such a signal is quite irregular. They are also called nondeterministic signals. A random signal is a signal which has some degree of uncertainty with respect to its value at any value of the independent variable, namely time. For example, thermal agitation noise in conductors is a random signal. Figure 1.11 shows a random signal.

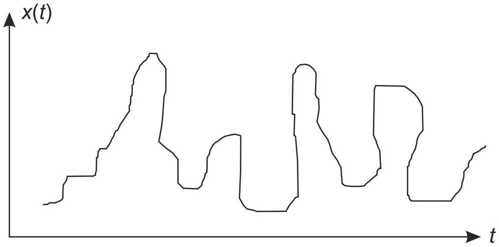


Fig. 1.11 A random signal

1.7.4 Periodic and Aperiodic Signals

A **periodic signal** has a definite pattern and repeats over and over with a repetition period of T. Simply, a signal is said to be periodic if it satisfies periodicity such as,

$$x(t + T) = x(t), -\infty \ < t < \infty \tag{1.3}$$

The smallest value of the period T which satisfies the above equation is called the **fundamental period**. Figure 1.12 shows a periodic signal.

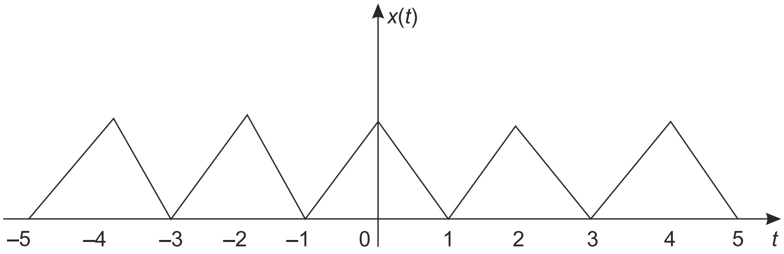


Fig. 1.12 A periodic signal

A signal $x(t)$ is called **aperiodic** if it does not repeat. It is said to have a period equal to infinity. Figure 1.13 shows an aperiodic signal.

$$x(t) = e^{-at} \tag{1.4}$$

$$x(t + T_0) = e^{-a(t+T_0)}$$

$$x(t + T_0) = e^{-at}.0$$

$$x(t + T_0) = 0 \neq x(t) \tag{1.5}$$

The signal with period $T_0 = \infty$ is an aperiodic signal.

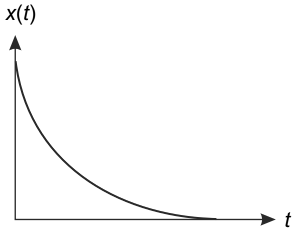


Fig. 1.13 Aperiodic signal

EXAMPLE 1.1

Determine whether the following signal is periodic or not: x(t) = sin 15 πt

Solution

$$x(t) = \sin 15\ \pi \text{t}$$

$$\omega = 15\ \pi \text{t}$$

The fundamental period is $T = \dfrac{2\pi}{\omega}$

$$\frac{2\pi}{15\pi} = 0.133 \text{ seconds}$$

Hence, the given signal is a periodic signal.

EXAMPLE 1.2

Check whether the following signal is periodic or not: x(t) = sin $\sqrt{2}\,\pi t$

Solution

The fundamental period is $T = \dfrac{2\pi}{\omega}$

where $\omega = \sqrt{2}\ \pi t$

$$= \frac{2\pi}{\sqrt{2\pi}} = 1.41 \text{ seconds}$$

Hence, the given signal is a periodic signal.

1.7.5 Even and Odd Signals

An **even signal** satisfies symmetry in the time domain. This type of signal is identical about the origin. Mathematically, an even signal must satisfy the following condition.

$$x(t) = x(-t) \tag{1.6}$$

Figure 1.14 shows an even signal.

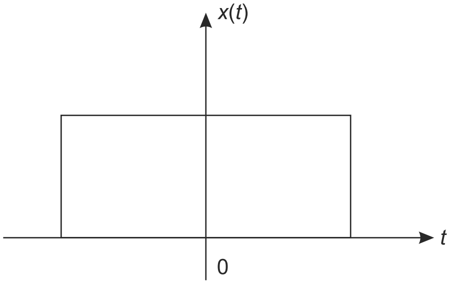


Fig. 1.14 Even signal

On the other hand, an **odd signal** does not satisfy the symmetry. This type of signal is not identical about the origin. Actually, this signal is identical to its negative. Mathematically, an odd signal must satisfy the following condition.

$$x(t) = -x\ (-t) \tag{1.7}$$

Figure 1.15 shows an odd signal.

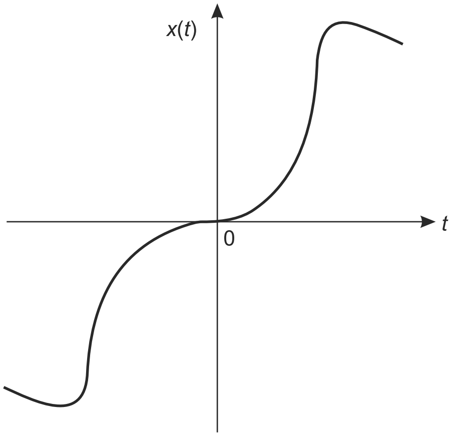


Fig. 1.15 Odd signal

Odd signals must necessarily be zero at $t = 0$ since $x(t) = -x(-t) \Rightarrow x(0) = -x(0) = 0$. An important fact is that any signal can be broken into a sum of two signals, one even and one odd.

$$\text{Even }\{x(t)\} = \frac{1}{2}[x(t) + x(-t)] \tag{1.8}$$

$$\text{Odd }\{x(t)\} = \frac{1}{2}[x(t) - x(-t)] \tag{1.9}$$

1.7.6 Energy and Power Signals

A signal can be classified to be an energy signal or a power signal. The **energy signal** is one which has finite energy and zero average power. Hence, $x(t)$ is an energy signal if

$$0 < E < \infty \text{ and } P = 0 \tag{1.10}$$

where E is the energy and P is the power of the signal $x(t)$.

Given a continuous-time signal $x(t)$, the energy contained over a finite time interval is defined as follows.

$$E_{(T_1, T_2)} = \int_{T_1}^{T_2} |x(t)|^2 .dt, T_2 > T_1 \tag{1.11}$$

$$E_f = \int_{-\infty}^{\infty} |x(t)|^2 .dt \tag{1.12}$$

Equation (1.11) defines the energy contained in the signal over the time interval from T_1 till T_2. On the other hand, Equation (1.12) defines the total energy contained in the signal. If the total energy of a signal is a finite nonzero value then that signal is classified as an energy signal. Typically, the signals which are not periodic turn out to be energy signals. For example, a single rectangular pulse and a decaying exponential signal are energy signals. Figure 1.16 shows a single rectangular pulse.

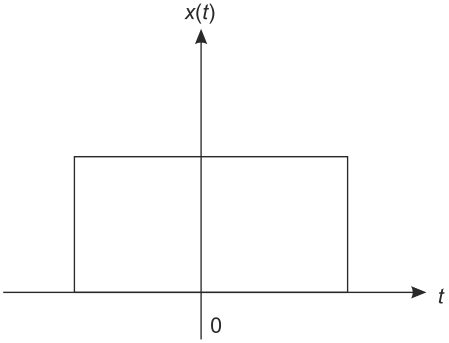


Fig. 1.16 A single rectangular pulse

The power signal is one which has finite average power and infinite energy. Hence, $x(t)$ is a power signal if

$$0 < P < \infty \quad \text{and} \quad E = \infty \tag{1.13}$$

For a continuous-time signal, the power of the signal can be expressed as follows from Equation (1.9).

$$P_f = \underset{T\to\infty}{\text{Lt}} \frac{1}{2T} \int_{-T}^{T} |x(t)|^2 \,.dt \tag{1.14}$$

$$P_f = \frac{1}{T} \int_{-T/2}^{T/2} |x(t)|^2 \,.dt \tag{1.15}$$

Most periodic signals tend to be power signals. Given the period of a cycle, the power of a periodic signal can be defined by Equation (1.11). Equation (1.11) can be used to find the power of a dc signal also. The dc signal is also a power signal. If the power of a signal is a finite nonzero value and its energy is infinite then that signal is classified as a power signal. Figure 1.17 shows a power signal which is a periodic pulse train.

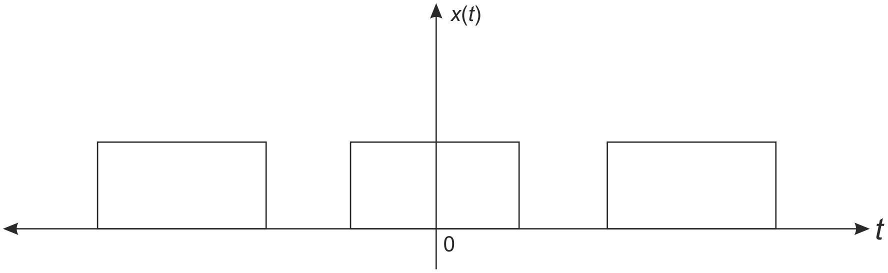


Fig. 1.17 A power signal

There are some signals which can be classified neither as power signals nor as energy signals. For example, a ramp signal defined from zero till infinity is neither a power signal nor an energy signal, since both power and energy of a ramp signal are not bounded.

EXAMPLE 1.3

Consider a signal $x(t) = e^{-3t}\,u(t)$ passed through a lowpass filter with a cut-off frequency of 1 rad/s. Find whether the above signal is an energy signal or not.

Solution

$$x(t) = e^{-3t}\,u(t)$$

$$\text{Energy } E_f = \int_{-\infty}^{\infty} |x(t)|^2 \,.dt$$

$$= \int_0^{\infty} (e^{-3t})^2 \, .dt \qquad = \int_0^{\infty} e^{-6t} \, .dt$$

$$= 0.167 \text{ joules}$$

Since the energy is finite, the given input signal is an energy signal.

EXAMPLE 1.4

Find whether the following signal is a power signal or not: x(t) = A sin ωt

Solution

Power of a signal can be expressed as

$$P_{\text{f}} = \frac{1}{T}\int_0^{T} |x(t)|^2 \, .dt$$

$$= \frac{1}{2\pi}\int_0^{2\pi} A^2 \sin^2 t \, .dt$$

$$= \frac{A^2}{2\pi}\int_0^{2\pi} \frac{1}{2}(1 - \cos 2t).dt$$

$$= \frac{A^2}{2\pi}\int_0^{2\pi} \left(\frac{1}{2} - \frac{1}{2}\cos 2t\right).dt$$

$$= \frac{A^2}{2\pi}\left[\frac{t}{2} - \frac{\sin 2t}{4}\right]_0^{2\pi}$$

$$= \frac{A^2}{8\pi}[2t - \sin 2t]_0^{2\pi}$$

$$= \frac{A^2}{8\pi}[(4\pi - \sin 4\pi] - (0 - 0)]$$

$$= \frac{A^2}{8\pi}[(4\pi - 0 - 0)]$$

$$= \frac{A^2}{2}$$

So the given signal is a power signal.

EXAMPLE 1.5

For the given signal shown in Figure 1.18, check whether it is a power signal or an energy signal or neither.

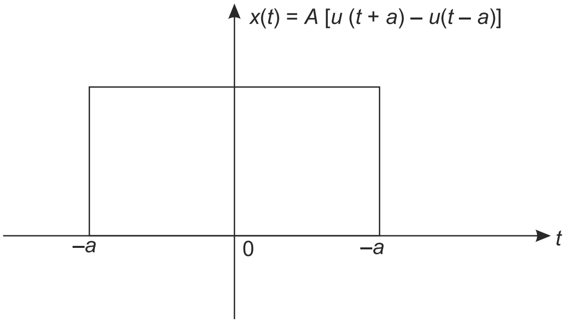


Fig. 1.18

Solution

Since $x(t)$ is of finite duration, it is an energy signal.

Energy can be expressed as $E_f = \int\limits_{-\infty}^{\infty} |x(t)|^2 .dt$

$$E = \int\limits_{-a}^{a} A^2 .dt$$

$$= 2\int\limits_{0}^{a} A^2 .dt$$

$$= 2A^2[t]_0^a$$

$$= 2A^2.a$$

So the given signal is an energy signal.

EXAMPLE 1.6

For the signal $x(t) = e^{-a|t|}$, check whether it is an energy signal or not.

Solution

$$E = \int\limits_{-a}^{a} x^2 (t).dt$$

$$E = \int\limits_{-\infty}^{\infty} e^{-2a|t|} .dt$$

$$= \int_{-\infty}^{0} e^{-2a(-t)}.dt + \int_{0}^{\infty} e^{-2a(t)}.dt$$

$$= \int_{-\infty}^{0} e^{2at}.dt + \int_{0}^{\infty} e^{-2at}.dt$$

$$= \int_{-\infty}^{0} e^{2at}.dt + \int_{0}^{\infty} e^{-2at}.dt$$

$$= 2\int_{0}^{\infty} e^{-2at}.dt$$

$$= 2\left[\frac{e^{-2at}}{-2a}\right]_0^{\infty}$$

$$= -\frac{1}{a}[e^{-\infty} - e^{0}]$$

$$= -\frac{1}{a}[0-1]$$

$$= \frac{1}{a}$$

Since the energy is finite, the given signal is an energy signal.

EXAMPLE 1.7

For the given signal shown in Figure 1.19, check whether it is a power signal or an energy signal or neither.

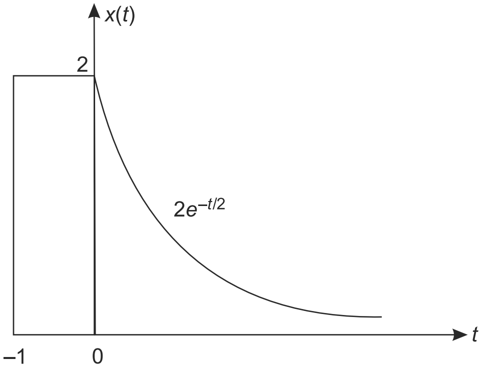


Fig. 1.19

Solution

From the given signal $x(t)$, it is clear that the signal amplitude $\rightarrow 0$ as $|t| \rightarrow \infty$. Hence, the given signal is an energy signal.

Energy $E = \int_{-\infty}^{\infty} x^2(t).dt$

$$E = \int_{-1}^{0} (2)^2\, dt + \int_{0}^{\infty} (2e^{-t/2})^2\, dt$$

$$= \int_{-1}^{0} 4\,dt + \int_{0}^{\infty} 4e^{-t}\, dt$$

$$= 4 + 4 = 8$$

Since the energy is finite, the given signal is an energy signal.

EXAMPLE 1.8

Check whether the given signal x(t) = A cos (ωt + θ) is a power signal or not. If yes, find the power.

Solution

The given signal *x(t)* = *A* cos *(*ω*t* + θ*)* is a periodic signal having period

$$T = \frac{2\pi}{\omega}$$

Hence, the given signal is a power signal.

Power can be expressed as

$$P_{\mathrm{f}} = \underset{T\to\infty}{\mathrm{Lt}} \frac{1}{T} \int_{-T/2}^{T/2} |x(t)|^2 .dt$$

Given that $x(t) = A\cos(\omega t + \theta)$

$$\therefore P_{\mathrm{f}} = \underset{T\to\infty}{\mathrm{Lt}} \frac{1}{T} \int_{-T/2}^{T/2} A^2 \cos^2 |\omega t + \theta|.dt$$

$$= \underset{T\to\infty}{\mathrm{Lt}} \frac{1}{T} \int_{-T/2}^{T/2} \frac{A^2}{2}[2\cos^2(\omega t + \theta)].dt$$

$$= \underset{T\to\infty}{\mathrm{Lt}} \frac{1}{T} \int_{-T/2}^{T/2} \frac{A^2}{2}[1 + \cos(2\omega t + 2\theta)].dt$$

$$= \underset{T\to\infty}{\mathrm{Lt}} \frac{1}{T} \int_{-T/2}^{T/2} \frac{A^2}{2} dt + \underset{T\to\infty}{\mathrm{Lt}} \frac{1}{T} \int_{-T/2}^{T/2} \frac{A^2}{2} \cos(2\omega t + 2\theta).dt$$

$$= \underset{T\to\infty}{\text{Lt}} \frac{A^2}{2T} \int_{-T/2}^{T/2} dt + \underset{T\to\infty}{\text{Lt}} \frac{A^2}{2T} \int_{-T/2}^{T/2} \cos(2\omega t + 2\theta).dt$$

$$= \underset{T\to\infty}{\text{Lt}} \frac{A^2}{2T} T + \underset{T\to\infty}{\text{Lt}} \frac{A^2}{2T} \int_{-T/2}^{T/2} \cos(2\omega t + 2\theta).dt$$

$$= \frac{A^2}{2} + 0$$

$$\therefore \qquad P = \frac{A^2}{2}$$

EXAMPLE 1.9

Given an exponential signal as $x(t) = Ae^{-kt}.u(t)$. Find its energy.

Solution

$$x(t) = Ae^{-kt}.u(t)$$

Energy $$E_f = \int_{-\infty}^{\infty} |x(t)|^2 .dt$$

$$E_f = \int_0^{\infty} (Ae^{-kt})^2 .dt$$

$$\therefore \qquad E_f = \frac{A^2}{2k}$$

EXAMPLE 1.10

Given an exponential signal as $x(t) = A \sin(\omega t)$. Find its power.

Solution

Power can be expressed as

$$P_f = \underset{T\to\infty}{\text{Lt}} \frac{1}{T} \int_{-T/2}^{T/2} |x(t)|^2 .dt$$

$$P_f = \frac{1}{T} \int_0^T A\sin(\omega t).dt$$

$$= \frac{A^2}{2}$$

Since the value of the power of the given signal over a cycle is a finite, nonzero value, it is a power signal.

EXAMPLE 1.11

Given a square-wave signal defined by $x(t)=\begin{cases}-A & \text{for} \quad -T/2<t<0 \\ A & \text{for} \quad 0<t<T/2\end{cases}$ *find its power.*

Solution

$$x(t)=\begin{cases}-A & \text{for} \quad -T/2<t<0 \\ A & \text{for} \quad 0<t<T/2\end{cases}$$

Power can be expressed as

$$P_{\text{f}} = \underset{T\to\infty}{\text{Lt}} \frac{1}{T}\int_{-T/2}^{T/2} |x(t)|^2 \,.dt$$

$$P_{\text{f}} = \frac{1}{T}\int_{-T/2}^{T/2} A^2 \,.dt = A^2$$

1.8 REPRESENTATION OF SIGNALS

Signals contain information about a variety of things and activities in the physical world. Broadly, there are two types of signal representation. They are

1. Time-domain representation
2. Frequency-domain representation

1.8.1 Time-Domain Representation

The signal is a time-varying quantity in time-domain representation. This means that, in time-domain representation, the amplitude of the signal varies with respect to time. Figure 1.20 shows a time-domain signal.

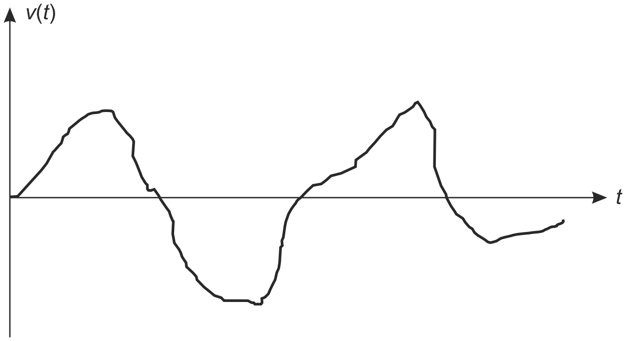


Fig. 1.20 A time-domain signal

Time-domain representation can be used to describe the analysis of mathematical functions, physical signals or any real-world data with respect to time. In the time domain, the value of the signal is known for all real numbers for the case of continuous time or at various separate instants in the case of discrete time.

Time-domain representations of signals are relatively comfortable to use. Signals measured on an oscilloscope are displayed in the time domain, and digital information is often represented by a voltage as a function of time.

1.8.2 Frequency-Domain Representation

In the frequency domain, a signal is represented by its frequency spectrum. They can be represented by a magnitude and phase as a function of frequency.

A time-domain signal shows how a signal changes over time, whereas a frequency-domain signal shows how much of the signal lies within each given frequency band over a range of frequencies. A frequency-domain representation can also include information on the phase shift that must be applied to every sinusoidal signal to recover the original time signal. Figure 1.21 shows a frequency-domain signal.

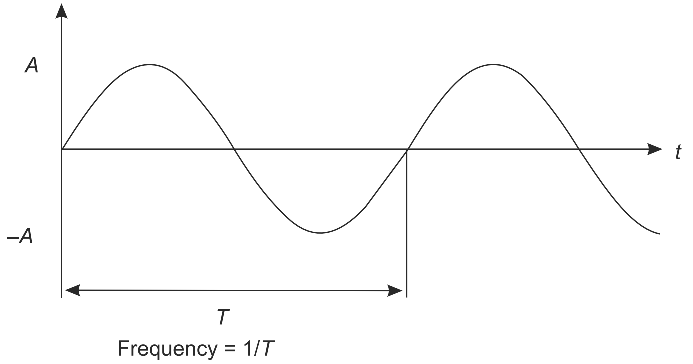


Fig. 1.21 A frequency-domain signal

1.8.3 Methods of Frequency Domain

There are a number of different mathematical transforms which are used to analyse time functions and are called frequency-domain methods. The following methods are the most common and they are in use with different fields.

1. Fourier series is specifically useful for repetitive signals
2. Fourier transform is specifically useful for nonrepetitive signals
3. Laplace transform is useful for control systems, etc.
4. Z-transform is specifically useful for discrete signals

5. Wavelet transform is specifically useful for signal compression, etc.
6. Fourier series and Fourier transform are briefly discussed as follows.

1.9 FOURIER SERIES

Fourier series is used to get frequency spectrum of a time-domain signal when the signal is a periodic function of time. With the help of Fourier series, a given periodic function of time may be expressed as the sum of an infinite number of sinusoids whose frequencies are harmonically related.

Let $x_p(t)$ be a periodic signal with period T_0. Then $f_0 = \frac{1}{T_0}$ is called the **fundamental frequency** and nf_0 is called the ***n*** ***n*th harmonic**, where n is an integer. Generally, Fourier transform is a mathematical operation that decomposes a signal into its constituent frequencies. Here, it decomposes $x_p(t)$ into DC, fundamental and its various higher harmonics, namely,

$$x_p(t) = \sum_{n=-\infty}^{\infty} x_n e^{j2\pi n f_0 t} \tag{1.16}$$

This equation is referred to as the **exponential form of the Fourier series**.

The coefficients $\{x_n\}$ constitute the Fourier series and are related to $x_p(t)$ as

$$x_n = \frac{1}{T_0} \int_{T_0} x_p(t)_e^{-j2\pi n f_0 t} .dt \tag{1.17}$$

The coefficients $\{x_n\}$ are in general complex. Hence,

$$x_n = |x_n|.e^{j\varphi_n} \tag{1.18}$$

where x_n denotes the magnitude of the complex number and φ_n denotes the angle.

$$x_p(t) = \sum_{n=-\infty}^{\infty} |x_n| e^{j(2\pi n f_0 t + \varphi_n)} \tag{1.19}$$

In general, $x_p(t)$ is composed of the frequency components at DC, fundamental and its higher harmonics. x_n is the magnitude of the component in $x_p(t)$ at frequency nf_0 and φ_n, its phase. The plot of x_n versus nf_0 is called the **magnitude spectrum** and φ_n versus nf_0 is called the **phase spectrum**.

The spectrum of a periodic signal exists only at discrete frequencies at nf_0, $n = 0, \pm 1, \pm 2$, ...etc.

Let $x_p(t)$ be real. Then

$$x_{-n} = \frac{1}{T_0} \int x_p(t) e^{j2\pi f_0 t} dt \tag{1.20}$$

$$= x_n^* \tag{1.21}$$

For a real periodic signal, there are two symmetry properties. They are as follows.

$$|x_{-n}| = |x_n| \tag{1.22}$$

$$\varphi_{-n} = -\varphi_n \tag{1.23}$$

That is, if $x_p(t)$ is real, then Equations (1.22) and (1.23) hold and if it holds then $x_p(t)$ has to be real. This is because the complex exponentials at nf_0 and $-nf_0$ can be combined into a cosine term. As an example, only nonzero coefficients of a periodic signal are considered as $x_{\pm 2}$, $x_{\pm 1}$, $x_0.x_0 = x_0^*$ implies, x_0 is real and let

$$x_{-2} = 2e^{j\frac{\pi}{4}} = x_2^* \quad \text{and} \tag{1.24}$$

$$x_{-1} = 3e^{j\frac{\pi}{3}} = x_1^* \quad \text{and} \tag{1.25}$$

$$x_0 = 1 \tag{1.26}$$

Then

$$x_p(t) = 2e^{j\frac{\pi}{4}} e^{-j4\pi f_0 t} + 3e^{j\frac{\pi}{3}} e^{-j2\pi f_0 t} + 1 + 3e^{-j\frac{\pi}{3}} e^{j2\pi f_0 t} + 2e^{-j\frac{\pi}{4}} e^{j4\pi f_0 t} \tag{1.27}$$

Combining the appropriate terms gives

$$x_p(t) = 4\cos\left[4\pi f_0 t - \frac{\pi}{4}\right] + 6\cos\left[2\pi f_0 t - \frac{\pi}{3}\right] + 1 \tag{1.28}$$

which is a real signal. The above form of representing $x_p(t)$ in terms of cosines is called the **trigonometric form of the Fourier series**.

The Fourier series exists only when the function $x(t)$ satisfies the following three conditions.

1. $x(t)$ is well defined and a single-valued function.
2. It must possess only a finite number of discontinuities in the period T.
3. It must have a finite number of positive and negative maxima in the period T.

A periodic function $x(t)$ may be expressed in the form of trigonometric Fourier series comprising the following sine and cosine terms.

$$\begin{aligned} x(t) = a_0 + a_1 \cos \omega_0 t + a_2 \cos 2\omega_0 t + \ldots + a_n \cos n\omega_0 t + \ldots \\ + b_1 \sin \omega_0 t + b_2 \sin 2\omega_0 t + \ldots + b_n \sin n\omega_0 t + \ldots \end{aligned} \tag{1.29}$$

$$x(t) = a_0 + \sum_{n=1}^{\infty} (a_n \cos n\omega_0 t + b_n \sin n\omega_0 t) \quad (t_0 \le t \le t_0 + T) \tag{1.30}$$

where $T = \dfrac{2\pi}{\omega}$ and a_n and b_n are coefficients.

The above equation is the trigonometric Fourier-series representation of the function $x(t)$ over an interval $(t_0, t_0 + T)$

The constant $b_0 = 0$ because $\sin n\omega_0 t = 0$ for n.

The constant a_0 is given by

$$a_0 = \frac{1}{T}\int_0^T x(t).dt \tag{1.31}$$

The constant a_n is given by

$$a_n = \frac{2}{T}\int_0^T x(t)\cos n\omega_0 t.dt \tag{1.32}$$

The constant b_n is given by

$$b_n = \frac{2}{T}\int_0^T x(t)\sin n\omega_0 t.dt \tag{1.33}$$

The Fourier series for a periodic signal is same for the entire interval (–∞, ∞) as for the interval (t_0, $t_0 + T$).

The constant a_0 is given by

$$a_0 = \frac{1}{T}\int_0^T x(t).dt \tag{1.34}$$

The constant a_n is given by

$$a_n = \frac{2}{T}\int_0^T x(t)\cos n\omega_0 t.dt \tag{1.35}$$

The constant b_n is given by

$$b_n = \frac{2}{T}\int_0^T x(t)\sin n\omega_0 t.dt \tag{1.36}$$

EXAMPLE 1.12

For the periodic waveform given in Figure 1.22, obtain the Fourier-series representation.

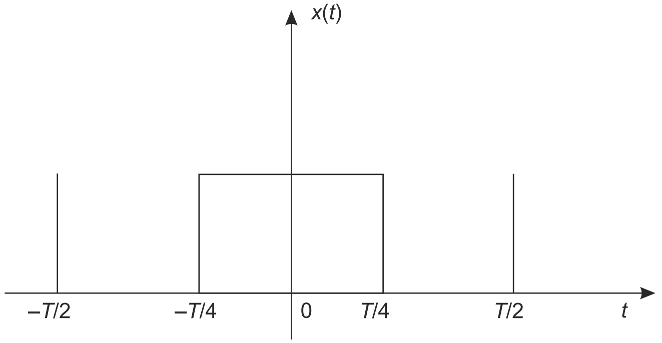


Fig. 1.22

Solution

The given waveform for one period can be written as

$$x(t) = \begin{cases} 0 & \text{for} \quad -T/2 < t < -T/4 \\ A & \text{for} \quad -T/4 < t < T/4 \\ 0 & \text{for} \quad T/4 < t < T/2 \end{cases}$$

From the figure, it is clear that the given signal is an even function.

$$x(t) = x(-t) \text{ and } \quad \therefore b_n = 0$$

The Fourier-series expression is given by

$$x(t) = a_0 + \sum_{n=1}^{\infty} (a_\text{n} \cos n\omega_0 t + b_n \sin n\omega_0 t)$$

$$\therefore b_n = 0$$

$$x(t) = a_0 + \sum_{n=1}^{\infty} (a_n \cos n\omega_0 t)$$

$$x(t) = a_0 + a_1 \cos \omega_0 t + a_2 \cos 2\omega_0 t + a_3 \cos 3\omega_0 t + ...$$

$$a_0 = \frac{1}{T} \int_{-T/2}^{T/2} x(t).dt$$

$$a_0 = \frac{1}{T} \int_{-T/4}^{T/4} A.dt$$

$$= \frac{1}{T} [At]_{-T/4}^{T/4}$$

$$= \frac{A}{T} [T/4 - (-T/4)]$$

$$= \frac{A}{T} . \frac{T}{2} = \frac{A}{2}$$

The constant a_n is given by

$$a_n = \frac{2}{T} \int_{-T/2}^{T/2} x(t) \cos n\omega_0 t.dt$$

$$a_n = \frac{2}{T} \int_{-T/4}^{T/4} A \cos n\omega_0 t.dt$$

$$= \frac{2A}{T} \left[\frac{\sin n\omega_0 t}{n\omega_0} \right]_{-T/4}^{T/4}$$

$$= \frac{2A}{n\omega_0 T}\left[\sin n\omega_0 t\frac{T}{4} - \sin\left(-n\omega_0 t\frac{T}{4}\right)\right]$$

$$= \frac{2A}{n\omega_0 T}\left[2\sin\left(n\omega_0 t\frac{T}{4}\right)\right]$$

$$= \frac{4A}{n\omega_0 T}\left[\sin\left(n\omega_0 t\frac{T}{4}\right)\right]$$

But $\omega_0 = \frac{2\pi}{T} \Rightarrow \omega_0 T = 2\pi$

$\therefore$ $$a_\text{n} = \frac{4A}{2\pi n}\left[\sin\left(\frac{2\pi n}{4}\right)\right]$$

$$= \frac{2A}{\pi n}\left[\sin\left(\frac{\pi n}{2}\right)\right]$$

Putting $n = 1, 2, 3, ...$

$$a_1 = \frac{2A}{\pi}\left[\sin\left(\frac{\pi}{2}\right)\right] = \frac{2A}{\pi}.1 = \frac{2A}{\pi}$$

$$a_2 = \frac{2A}{2\pi}\left[\sin\left(\frac{2\pi}{2}\right)\right] = \frac{2A}{\pi}.0 = 0$$

$$a_3 = \frac{2A}{3\pi}\left[\sin\left(\frac{3\pi}{2}\right)\right] = \frac{2A}{3\pi}.(-1) = -\frac{2A}{3\pi}$$

Putting all these values in the following equation,

$$x(t) = a_0 + a_1 \cos \omega_0 t + a_2 \cos 2\omega_0 t + a_3 \cos 3\omega_0 t + ...$$

$$x(t) = \frac{A}{2} + \frac{2A}{\pi}\left(\cos\omega_0 t - \frac{1}{3}\cos 3\omega_0 t + \frac{1}{5}\cos 5\omega_0 t + ...\right)$$

1.10 FOURIER TRANSFORM

Similar to periodic signals, aperiodic signals can also be represented in the frequency domain. In an aperiodic signal, the period of the signal approaches infinity, Fourier transform is an approach to develop the frequency-domain representation of an aperiodic signal over an entire interval.

However, unlike the discrete spectrum of the periodic case, there is a continuous spectrum for the aperiodic case. That is, the frequency components constituting a given signal $x(t)$ lie in a continuous range and quite often this range could be $(-\infty, \infty)$. Equation (1.16) expresses $x(t)$ as a sum over a discrete set of frequencies. Its counterpart for the aperiodic case is an integral relationship given by

$$x(t) = \int_{-\infty}^{\infty} X(\omega) e^{j2\pi ft} .dt \tag{1.37}$$

where $X(\omega)$ is the Fourier transform of $x(t)$.

Let the integral be treated as a sum over incremental frequency ranges of width Δf. Let $X(\omega).\Delta f$ be the incremental complex amplitude of $e^{j2\pi f_i t}$ at the frequency $f = f_i$. If we sum a large number of such complex exponentials, the resulting signal should be a very good approximation to $x(t)$. This leads to signal representation with a sum of complex exponentials replaced by an integral, where a continuous range of frequencies, with the appropriate complex amplitude distribution will synthesise the given signal $x(t)$.

Equation (1.37) is called the **synthesis relation** or **Inverse Fourier Transform (IFT)** relation. In relation with IFT equation, the Fourier transform is expressed as follows.

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} .dt \tag{1.38}$$

So simply it can be expressed as

$$X(\omega) = F[x(t)] \tag{1.39}$$

$$x(t) = F^{-1}[X(\omega)] \tag{1.40}$$

Symbolically, it may be expressed as

$$x(t) \leftrightarrow X(\omega) \tag{1.41}$$

In general, Fourier transform $X(\omega)$ is a complex function of ω and may be expressed as

$$X(\omega) = |X(\omega)| e^{j\theta(\omega)} \tag{1.42}$$

where $|X(\omega)|$ is called the **amplitude spectrum** and $e^{j\theta(\omega)}$ is called the **phase spectrum**.

For a function $x(t)$ to be Fourier transformable, it is necessary to satisfy the following conditions.

1. The function $x(t)$ is a single-valued function with a finite number of maxima and minima and a finite number of discontinuities in any finite time interval.
2. The function $x(t)$ is absolutely integrable. Mathematically,

$$\int_{-\infty}^{\infty} x(t).dt < \infty \tag{1.43}$$

1.10.1 Properties of Fourier Transform

The following are the properties of Fourier transform.

1. Time-scaling Property

If $x(t) \leftrightarrow X(\omega)$ then for any real constant a,

$$x(at) \leftrightarrow \frac{1}{|a|} X\left(\frac{\omega}{a}\right) \tag{1.44}$$

2. Linearity Property

If $x_1(t) \leftrightarrow X_1(\omega)$, and

$x_2(t) \leftrightarrow X_2(\omega)$

then $$a_1x_1(t) + a_2x_2(t) \leftrightarrow a_1X_1(\omega) + a_2X_2(\omega) \tag{1.45}$$

3. Duality or Symmetry Property

If $x(t) \leftrightarrow X(\omega)$

then $$X(\omega) \leftrightarrow 2\pi x(w) \tag{1.46}$$

4. Time-shifting Property

If $x(t) \leftrightarrow X(\omega)$

then $$x(t–b) \leftrightarrow X(\omega).e^{-jwb} \tag{1.47}$$

5. Frequency-shifting Property

If $x(t) \leftrightarrow X(\omega)$

then $$e^{j\omega_0 t}.x(t) \leftrightarrow X(\omega - \omega_0) \tag{1.48}$$

6. Time Differentiation Property

If $x(t) \leftrightarrow X(\omega)$

then $$\frac{dx(t)}{dt} \leftrightarrow j\omega.X(\omega) \tag{1.49}$$

1.10.2 Bandlimited Signals

If the limiting of a deterministic or a stochastic signal's Fourier transform or power spectral density is zero above a certain finite frequency then the signal is said to be bandlimited.

A bandlimited signal can be fully reconstructed from its samples, provided that the sampling rate exceeds twice the maximum frequency in the bandlimited signal. This minimum sampling frequency is called **Nyquist rate**.

For an example, a simple sinusoidal signal is considered as a deterministic bandlimited signal which can be expressed as follows:

$$x(t) = \sin(2\pi ft + \theta)$$

This signal is sampled at a rate $f_s = \frac{1}{T} > 2f$ so that with the samples $x(nT)$ for all integers n, it is possible to recover $x(t)$ completely from the samples $x(nT)$. Similarly, sum of sinusoids with different frequencies and phases are also bandlimited to the highest of their frequencies. Figure 1.23 shows a bandlimited signal

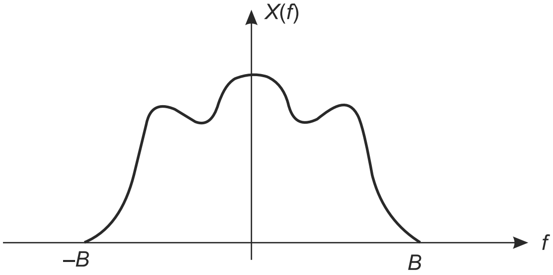


Fig. 1.23 A bandlimited signal

For the signal $x(t)$, the Fourier transform is $X(f)$. From Figure 1.20, the highest frequency component in $x(t)$ is B. As a result, the Nyquist rate is twice the highest frequency component in the signal and it is expressed as

Nyquist rate = $2B$

1.10.3 Bandpass Signals

A bandpass signal is a signal containing a band of frequencies away from zero frequency, such as a signal that comes out of a bandpass filter. It refers to a specific range of frequencies required for transmission of a signal over an appreciable distance for effective communication by applying special modulation techniques for effective transmission of a signal.

All the Radio-Frequency (RF) signals transmitted and received in wireless systems are bandpass signals. The spectrum of a bandpass signal is confined to a band not including 0 Hz in the frequency domain. An example of the Fourier transform of such a signal is shown in Figure 1.24.

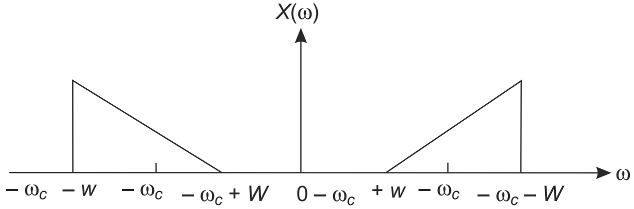


Fig. 1.24 Fourier transform of a bandpass signal

The frequency ω_c radians per second or $f_c = \dfrac{\omega_c}{2\pi}$ is called the **centre frequency** for the signal and the signal bandwidth is $B = 2W$. When a low-frequency message is intentionally translated up to a band centred on a frequency ω_c, this frequency is called the **carrier frequency**. The spectrum of the bandpass signal is negligible outside intervals surrounding $\pm\omega_c$. If the intervals or bandwidths around $\pm\,\omega_c$ are 'relatively small' then the signal is said to be narrowband.

EXAMPLE 1.13

Find the Fourier transform of the following rectangular function shown in Figure 1.25.

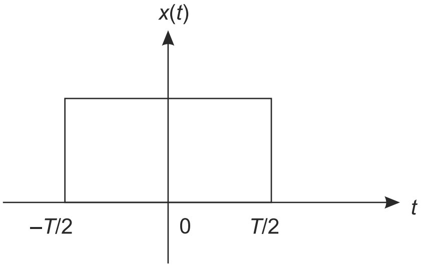


Fig. 1.25

Solution

$$x(t) = \text{rect}\left(\frac{t}{T}\right)$$

$$= \begin{cases} 1 & \text{for} \quad -T/2 < t < T/2 \\ 0 & \text{otherwise} \end{cases}$$

The Fourier transform is given by

$$X(\omega) = F[x(t)] = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}\,.dt$$

$$X(\omega) = \int_{-\infty}^{\infty} \text{rect}\left(\frac{t}{T}\right) e^{-j\omega t}\,.dt$$

$$= \int_{-T/2}^{T/2} 1.e^{-j\omega t}\,.dt$$

$$= \left[\frac{e^{-j\omega t}}{-j\omega}\right]_{-T/2}^{T/2}$$

$$= -\frac{1}{j\omega}\left[e^{-j\omega\frac{T}{2}} - e^{j\omega\frac{T}{2}}\right]$$

$$= \frac{1}{j\omega}\left[e^{j\omega\frac{T}{2}} - e^{-j\omega\frac{T}{2}}\right]$$

It is known that

$$e^{j\theta} = \cos\theta + j\sin\theta \quad \text{and} \quad e^{-j\theta} = \cos\theta - j\sin\theta$$

Hence, $2\cos\theta = e^{j\theta} + e^{j\theta}$ and $2j\sin\theta = e^{j\theta} - e^{-j\theta}$

Putting $\theta = \frac{\omega T}{2}$

$$2j\sin\frac{\omega T}{2} = e^{j\omega\frac{T}{2}} - e^{-j\omega\frac{T}{2}}$$

$$\therefore \qquad X(\omega) = \frac{1}{j\omega}\left[2j\sin\frac{\omega T}{2}\right]$$

$$= \frac{2\sin\omega\frac{T}{2}}{\omega}$$

EXAMPLE 1.14

Find the Fourier-transform pair of δ(ω).

Solution

Inverse Fourier-transform is expressed as

$$x(t) = F^{-1}(X(\omega)) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)e^{j\omega t}.d\omega$$

$$x(t) = F^{-1}(\delta(\omega)) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \delta(\omega)e^{j\omega t}\,d\omega$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{j\omega t}\,d\omega$$

$$= \frac{1}{2\pi}[e^{j\omega t}] \text{ at } \omega = 0$$

$$= \frac{1}{2\pi}\times e^{0}$$

$$= \frac{1}{2\pi}$$

$$F^{-1}(\delta(\omega)) = \frac{1}{2\pi}$$

$$F\left(\frac{1}{2\pi}\right)=\delta(\omega)$$

$$\frac{1}{2\pi}\leftrightarrow\delta(\omega)$$

This is the Fourier-transform pair.

EXAMPLE 1.15

Find the Fourier transform of the following rectangular function shown in Figure 1.26.

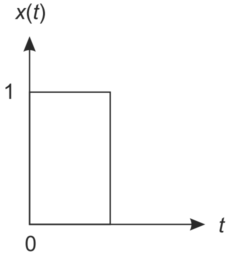


Fig. 1.26

Solution

The Fourier transform can be expressed as

$$F[x(t)]=X(\omega)=\frac{1}{2\pi}\int_{-\infty}^{\infty}x(t)e^{-j\omega t}\,.dt$$

$$\therefore \qquad X(\omega)=\int_0^T 1.e^{-j\omega t}\,.dt$$

$$=\int_0^T e^{-j\omega t}\,.dt$$

$$=\left(\frac{e^{-j\omega t}}{-j\omega}\right)_0^T$$

$$=-\frac{1}{j\omega}(-e^{-j\omega t}-e^0)$$

$$=-\frac{1}{j\omega}(e^{-j\omega T}-1)$$

$$=-\frac{1}{j\omega}\left(\frac{e^{-j\omega T/2}}{e^{j\omega T/2}}-1\right)$$

$$= -\frac{1}{j\omega}\left(\frac{e^{-j\omega T/2} - e^{j\omega T/2}}{e^{j\omega T/2}}\right)$$

$$= \frac{1}{j\omega}\frac{1}{e^{\mathrm{j\omega T/2}}}\left(e^{j\omega T/2} - e^{j\omega T/2}\right)$$

$$= \frac{2e^{-j\omega T/2}}{\omega}\left(\frac{e^{j\omega T/2} - e^{-j\omega T/2}}{2j}\right)$$

$$= Te^{-j\omega T/2}\left(\frac{\sin\left(\frac{\omega T}{2}\right)}{\left(\frac{\omega T}{2}\right)}\right)$$

$$= Te^{-j\omega T/2}\,\sin c\left(\frac{\omega T}{2}\right)$$

EXAMPLE 1.16

Find the Fourier transform pair of the following signal represented in Figure 1.27.

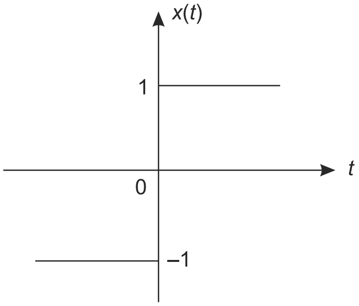


Fig. 1.27

Solution

$x(t)$ is a signum function and it is expressed as

$$x(t) = \mathrm{sgn}(t) = \begin{cases} -1 & \text{for} \quad t < 0 \\ 0 & \text{for} \quad t = 0 \\ 1 & \text{for} \quad t > 0 \end{cases}$$

$\mathrm{sgn}(t) = u(t) - u(-t)$

The Fourier transform can be expressed as

$$F[x(t)] = X(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} x(t)e^{-j\omega t}.dt$$

$$\therefore\ X(\omega) = \int_{-\infty}^{0}(-1).e^{-j\omega t}.dt + \int_{0}^{\infty}1.e^{-j\omega t}.dt$$

$$= -\int_{-\infty}^{0} e^{-j\omega t}.dt + \int_{0}^{\infty} e^{-j\omega t}.dt$$

$$= \left(\frac{e^{-j\omega t}}{-j\omega}\right)_{-\infty}^{0} + \left(\frac{e^{-j\omega t}}{-j\omega}\right)_{0}^{\infty}$$

$$= \frac{1}{j\omega}(e^{0} - e^{-\infty}) - \frac{1}{j\omega}(e^{-\infty} - e^{0})$$

$$= \frac{1}{j\omega}(1-0) - \frac{1}{j\omega}(0-1)$$

$$= \frac{1}{j\omega} + \frac{1}{j\omega}$$

$$= \frac{2}{j\omega}$$

$\therefore$ sgn(t) $\leftrightarrow \frac{2}{j\omega}$ is the Fourier-transform pair.

1.11 NOISE IN COMMUNICATION SYSTEMS

Noise is defined as an extraneous form of energy with random frequency and amplitude which tends to interfere with reception of a signal from a distant station. Noise may be picked up by a signal during its transmission from a transmitter to a receiver, which is commonly termed **external noise**. Alternately, noise may be produced within receiving equipment while it is receiving a signal. This type of noise is termed **internal noise**. It is to be noted that transmitting equipment do not produce noise in general. In fact, the signal level is raised to such a high magnitude in the transmitting equipment that any noise existing in the transmitting system can be easily ignored in comparison to the signals.

1.11.1 External Noise

There are two types of external noises.

1. Atmospheric noise
2. Human-made noise

1. Atmospheric Noise

These are electrical disturbances in the lower air layers. These disturbances are caused by lightning discharges. These noises are also termed **static noises** and these are in the form of impulses. Atmospheric disturbance in a communication channel is independent of the intensity of any single atmospheric factor as on the frequency with which they occur.

Power due to the lightning discharges is so large that it causes serious interference. This interference can be reduced by fitting a capacitor or a choke-capacitor combination of a sufficiently less HF resistance across the motor terminals that reduces these radiations to a negligible level.

The actual amount of noise present depends on the bandwidth of the receiver. So, it is desirable to limit the bandwidth of the receiver to a frequency range, which will be sufficient to receive the signal. If the noise is of large amplitude, a circuit with a low bandwidth will tend to distortion.

2. Human-made Noise

Human made noise results from the radiations produced by electrical machinery, ignition system of automobiles, fluorescent tubes, etc. which produce sparks. The radiations produced by sparking contain a wide band of frequencies, whose mean frequency varies with respect to the inductance and capacitance associated with the connecting leads of the apparatus providing disturbance.

Interference due to sparking in DC motors, dynamos, etc. can be reduced by fitting a capacitor or a choke-capacitor combination of a sufficiently less HF resistance across the motor terminals that reduces these radiations to a negligible level.

Ignition systems of automobiles/aircrafts are other powerful sources of interference, which are severe on 6–7 metres. In aircraft, this problem is solved by shielding the ignition leads up to engines and spark plugs.

Interference may also be caused by insulators used on power-transmission lines when such insulators are dirty or coated with salts. Leakage currents flow through these leakage paths. These are unsteady currents and probably travel by small arcs between deposited particles. RF components in these currents are radiated by the line conductors acting as aerials causing disturbance in communication services.

Interference may reach a receiver in a number of ways.

(a) Direct radiations from a noise source to the receiver aerial or to the receiver itself.

(b) Due to RF, current in the main lines produced by these disturbances may reach either directly or through inductive or capacitive coupling to receiver aerials.

(c) Interference may be caused by stray coupling between the two supply systems.

These interferences can be reduced placing receiver aerials in a high, open position and connecting them to the receivers through screened cables. Large ships contain a heavy number of electrical machines and it becomes necessary to use radio transmission while signals are received at other frequencies and it becomes necessary to shield the receiving room completely. Power leads entering the receiving room are fitted with filters.

1.11.2 Internal Noise

Even though all forms of external noise may be eliminated, there would still be noise generated in the receiver itself, which is completely internal to the system. Proper designing of the entire system can reduce these noises.

Internal noise can be classified into three major types:

1. Thermal agitation noise
2. Shot noise
3. Partition noise

1. Thermal Agitation Noise

Thermal noise constitutes the most important source of noise, which is generated by the random motion of electrons in a conductor. The free electrons in a conductor are in continuous motion with a velocity that depends on the temperature of the conductor. The motion of free electrons contribute the minimum amount of current. The sum of the total current is zero over a long time interval.

At any particular instant, there may be a current either in one direction or in the other direction. This flow of charge develops a random voltage at the terminals of the conductor and the effect is said to be thermal agitation noise.

The maximum thermal noise which a resistance R can deliver in the frequency band B Hz at absolute temperature T is given by

$$P_n = kTB, \tag{1.50}$$

where k is the Boltzmann's constant (1.38×10^{-23} J/K)

Figure 1.28 gives the Thevenin's equivalent circuit of a resistor as a noise source.

If R_L is the effective load and is noise-free while R is the noise-producing resistance then under matched load conditions, noise power across R is

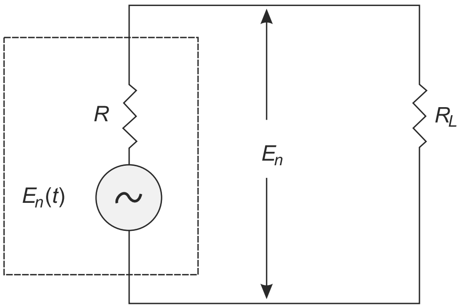


Fig. 1.28 Thevenin's equivalent circuit of a noise source

$$P_n = \left(\frac{E_n}{R+R_{\mathrm{L}}}\right)^2 \times R_{\mathrm{L}} \tag{1.51}$$

$$P_n = \frac{E_n^2}{4R} \quad (\because R = R_{\mathrm{L}}\ ) \tag{1.52}$$

$$E_n^2 = 4.\ R.\ P_n = 4RKTB \tag{1.53}$$

$$\therefore E_n\ (t) = \sqrt{4\,RKTB}\ = \text{Equivalent noise voltage} \tag{1.54}$$

Thermal agitation noise is produced in all the resistances, but under normal conditions, the thermal noise produced across the input-tuned circuit is of importance since this noise component gets sufficient amplification to become comparable to the signal level. Noise components present in the input of subsequent stages do not get enough amplification to cause disturbance in reception.

Thermal agitation voltage is used in some measuring apparatus as a standard input voltage for adjusting an equipment to give standard amplification.

2. Shot Noise

These noises arise due to the fact that the current in a tube or semiconductor device is due to the movement of discrete electric charges carried by electrons or holes, and this current develops a minute noise voltage across the anode or collector load. Noise at the receiver output depends on the passband of the succeeding receiver circuits. Usually, the shot noise of the first stage contributes the maximum noise to the receiver output.

Sometimes saturated diodes can be used as sources of known noise in which the magnitude of shot noise can be determined. The current in a diode is given by

$$I_{\mathrm{sn}} = \sqrt{2.q.I_{\mathrm{b}}B} \tag{1.55}$$

where

I_{sn} is shot noise current,

q is charge of the electron (1.6×10^{-19} coloumb),

I_{b} is the current flowing from anode and cathode, and

B is bandwidth.

3. Partition Noise

The partition of emission current between the anode and screen grid produces random fluctuations commonly termed partition noise. Partition noise is very less in diodes than transistors. Due to this, many microwave receivers offer diode circuits. Recently, for the purpose of low-noise microwave amplification, gallium arsenide FET has been used. The mean squared value of partition noise in transistor is given as

$$I_{\text{pn}} = 2I_{\text{C}}\left[1 - \frac{|\alpha|^2}{\alpha_0}\right] \tag{1.56}$$

where

I_{pn} is partition noise current,

I_{C} is collector current,

α is current amplification factor, and

α_0 is current amplification factor at low frequencies.

1.11.3 Noise Figure

In communication systems, the received signals are of low power and accompanied by noise. Further amplification becomes necessary for satisfactory reception but this amplifies the noise also. Since noise cannot be completely eliminated, the performance of a system is measured in terms of the ratio of signal power to noise power at different points in the receiving system.

To measure the 'noiseness' of a circuit, it is usual to compare the signal-to-noise ratio at the input of the network to that of the signal-to-noise ratio at its output. This ratio is known as 'noise figure'. It can be represented by *F*.

$$F = \frac{\text{SNR power at input}}{\text{SNR power at output}} \tag{1.57}$$

$$F = \frac{\left(\dfrac{P_{\text{SI}}}{P_{\text{NI}}}\right)}{\left(\dfrac{P_{\text{SO}}}{P_{\text{NO}}}\right)}$$

$$F = \frac{P_{\text{SI}}}{P_{\text{NI}}} \times \frac{P_{\text{NO}}}{P_{\text{SO}}} \tag{1.58}$$

where

P_{SI} is signal power at input,

P_{SO} is signal power at output,

P_{NI} is noise power at input, and

P_{NO} is noise power at output.

For comparing the performance of radio receivers working at different impedance levels, the use of equivalent noise resistance is very difficult. In that case, noise figure is used to calculate the noise level present in its output.

Noise figure in dB $= F_{\text{dB}} = 10 \log_{10} F$ (1.59)

Noise figure is unity for a noiseless ideal receiver which introduces no noise of its own.

1.11.4 Noise Temperature

The noise power is directly proportional to temperature and is independent of resistor value. This provides a convenient way of expressing available noise power from different noise sources. The power specified in terms of temperature is called noise temperature.

The noise power is

$$P_n = K.T_N.B \tag{1.60}$$

From which,

Noise temperature $T_n = \dfrac{P_N}{K.B}$ (1.61)

1.12 MATHEMATICAL MODELS FOR COMMUNICATION CHANNELS

Generally, a communication channel is a physical medium which is used to send a signal from a transmitter to the receiver, or it can be any physical medium which transmits information or used for transmitting information like optical fibre, storage media, etc.

When a signal transmits through the channel, the channel inherently adds some noise to the signal. This noise is known as **thermal noise**. The noise generated by the components is categorised as thermal noise, also known as **additive noise**. The effect of the noise on the signal can be reduced by increasing the power in the transmitted signal but there will be more power consumption. Because of the limitations of the channel, it is necessary to design the communication system such that more data can be transmitted as possible without getting corrupted, i.e. these constraints are to be considered while deriving new algorithms for reliable transmission of information over a channel with high data rates.

1.12.1 Types and Characteristics of Channels

There are many types of channels available with their operating frequencies.

1. **Wireline Channel** It operates at frequencies of few kHz to several hundreds of kHz.
2. **Fibre Optical Channel** It provides bandwidth in magnitudes several times higher than that of wireline channels.
3. **Wireless Electromagnetic Channel** It operates in the range of 10 kHz to $\cong$ 100 GHz. This is further categorised as long-wave radio, short-wave radio and microwave radio as they operate in radio frequency, they are also known as *radio channels*.
4. **Underwater Acoustic Channel** It is operated at extremely low frequencies.
5. **Storage Channel** like magnetic tapes, magnetic disks, etc.

1.12.2 Mathematical Models

Any mathematical model in a communication system can reflect the most important characteristics of the transmission medium. This mathematical model of the transmission medium is used to design the channel encoder, modulator at the transmitting end, demodulator and channel decoder at the receiving end. The following are the channel models frequently used to characterise many of the physical channels.

1. Additive Noise Channel

The simplest mathematical model for a communication channel is the additive noise channel which is illustrated in Figure 1.29. In this model, the transmitted signal $s(t)$ is corrupted by an additive random noise process $n(t)$.

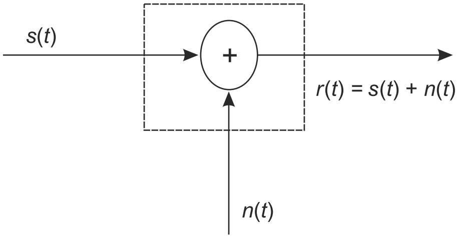


Fig. 1.29 Additive noise channel

Physically, the additive noise process may arise from electronic components and amplifiers at the receiver of the communication system or from interference encountered in transmission similar to radio-signal transmission.

If the noise is introduced primarily by electronic components and amplifiers at the receiver, it may be characterised as thermal noise. This type of noise is characterised statistically as a **Gaussian noise process**. Hence, the resulting mathematical model for the channel is usually called the **additive Gaussian noise channel**. Because this channel model applies to a broad class of physical communication channels and because of its mathematical tractability, this is the predominant channel model used in our communication-system analysis and design. Channel attenuation is easily incorporated into the model. When the signal undergoes attenuation in transmission through the channel, the received signal is

$$r(t) = as\ (t) + n(t) \tag{1.62}$$

where a represents the attenuation factor.

2. The Linear Filter Channel

In some physical channels such as wireline telephone channels, filters are used to ensure that the transmitted signals do not exceed specified bandwidth limitations and thus they do not interfere with one another. Such channels are generally characterized mathematically as linear filter channels with additive noise. Figure 1.30 shows a linear filter channel.

Hence, if the channel input is the signal $s(t)$, the channel output is the signal with additive noise.

$$r(t) = s(t)*h(t) + n(t) \tag{1.63}$$

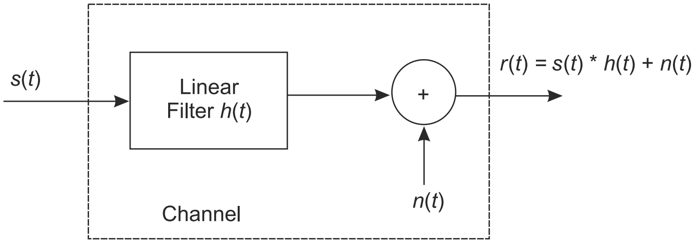


Fig. 1.30 A linear filter channel

$$= \int_{-\infty}^{\infty} h(\tau)s(t-\tau)d\tau + n(t) \tag{1.64}$$

where $h(t)$ is the impulse response of the linear filter and * denotes convolution.

3. The Linear Time-Variant Filter Channel

Some of the physical channels like ionospheric radio channels which result in time-variant multipath propagation of the transmitted signal may be characterised mathematically as time-variant linear filters. Such linear filters are characterised by the time-variant channel impulse response $h(\tau; t)$ where $h(\tau; t)$ is the response of the channel at time t, due to an impulse applied at time $t - \tau$. The linear time-variant filter channel with additive noise is illustrated in Figure 1.31.

For an input signal $s(t)$, the channel output signal is

$$r(t) = s(t) * h(\tau, t) + n(t) \tag{1.65}$$

$$= \int_{-\infty}^{\infty} h(\tau;t)s(t-\tau)d\tau + n(t) \tag{1.66}$$

A good model for multipath signal propagation through physical channels, such as the ionosphere and mobile cellular radio channels in which the time-variant impulse response has the form

$$h(\tau;t) = \sum_{k=1}^{L} a_k\,(t)s(t - T_k\,) \tag{1.67}$$

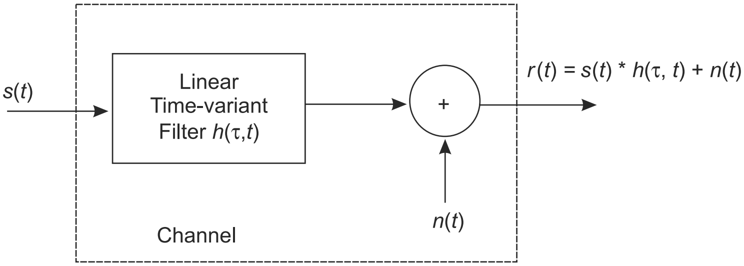


Fig. 1.31 A linear time-variant filter channel

where the $a_k(t)$ represent the possibly time-variant attenuation factors for the L multipath propagation paths. If Equation (1.65) is substituted into Equation (1.67), the received signal will be

$$r(t) = \sum_{k=1}^{L} a_k(t)s(t - \tau_k) + n(t) \tag{1.68}$$

Hence, the received signal consists of L multipath components, where each component is attenuated by $a_k(t)$ and delayed by t_k.

1.13 ELECTROMAGNETIC SPECTRUM

The electromagnetic spectrum is the range of all possible frequencies of electromagnetic radiation. The electromagnetic spectrum of an object is the characteristic distribution of electromagnetic radiation emitted or absorbed by that particular object. It extends from low frequencies used for modern radio to γ-radiation at the short-wavelength end, covering wavelengths from thousands of kilometres down to a fraction of the size of an atom.

1.13.1 Range of the Spectrum

EM waves are typically described by any of the following three physical properties:

1. The frequency f
2. Wavelength λ
3. Photon energy E

The frequency range of EM waves is ranging from 2.4×10^{23} Hz (1 GeV gamma rays) down to the ~1 kHz, since wavelength is inversely proportional to the wave frequency. Gamma rays have very short wavelengths and highest energy and radio waves have very low energy.

$$f = \frac{c}{\lambda} \text{ or} \tag{1.69}$$

$$f = \frac{E}{h} \text{ or} \tag{1.70}$$

$$E = \frac{hc}{\lambda} \tag{1.71}$$

where

c = 299,792,458 m/s is the speed of light in vacuum, and

h is Planck's constant = $4.13566733(10)\times10^{-15}$ eV

Table 1.1 shows electromagnetic spectrum with different physical properties.

Table 1.1 Electromagnetic spectrum with different physical properties

Class	Frequency	Wavelength	Energy
Gamma rays (Y)	(300–30) EHz (EHz: Exahertz)	(1–10) pm	1.24 MeV–124 keV
Hard X-rays (HX)	(30–3) EHz	(10–100) pm	(124–12.4) keV
Soft X-rays (SX)	(300–30) PHz (PHz: Petahertz)	(100 pm–1 nm)	(12.4–1.24) keV
Extreme Ultraviolet (EUV)	(30–3 PHz)	(1–10) nm	1.24 keV–124 eV
Near Ultraviolet (NUV)	(3 PHz–300 THz) (THz: Terahertz)	(10–100) nm	(12.4–1.24) eV
Near Infrared (NIR)	(300–30) THz	100 nm–1 µm	1.24 eV–124 MeV
Mid Infrared (MIR)	(30–3) THz	(1–10) µm	(124–12.4) MeV
Far Infrared (FIR)	3 THz–300 GHz (GHz: Gigahertz)	(10–100) µm	(12.4–1.24) MeV
Extreme High Frequency (EHF)	(300–30) GHz	100 µm–1 mm	1.24 MeV–124 µeV
Super High Frequency (SHF)	(30–3) GHz	1 mm–1 cm	(124–12.4) µeV
Ultra High Frequency (UHF)	3 GHz–300 MHz (MHz: Megahertz)	1 cm–1 dm	(12.4–1.24) µeV
Very High Frequency (VHF)	(300 –30) MHz	1 dm–1 m	1.24 µeV–124 neV
High Frequency (HF)	(30–3) MHz	(1–10) m	(124–12.4) neV
Medium Frequency (MF)	3 MHz –300 kHz (kHz: kilohertz)	(10–100) m	(12.4–1.24) neV
Low Frequency (LF)	(300–30) kHz	(1–10) km	1.24 neV–124 peV
Very Low Frequency (VLF)	(30 –3) kHz	(10–100) km	(124–12.4) peV
Voice Frequency (VF)	3 kHz–300 Hz (Hz: Hertz)	100 km–1 Mm	(12.4–1.24) peV
Super Low Frequency (SLF)	(300–30) Hz	(1–10) Mm	1.24 peV–124 feV
Extremely Low Frequency (ELF)	(30–3) Hz	(10–100) Mm	(124–12.4) feV

Whenever electromagnetic waves exist in a medium with matter, their wavelength is decreased. Generally, electromagnetic radiation is classified by wavelength into

1. Radio wave,
2. Microwave,
3. Infrared, and
4. The visible region which is perceived as light, ultraviolet, X-rays and gamma rays.

The behavior of electromagnetic radiation depends on its wavelength. When electromagnetic radiation interacts with single atoms and molecules, its behaviour also depends on the amount of energy per photon it carries. Spectroscopy can detect a much wider region of the electromagnetic spectrum than the visible range of 400 nm to 700 nm.

1.13.2 Types of Radiation

The electromagnetic spectrum is divided into five major types of radiation. Figure 1.32 illustrates an electromagnetic spectrum which includes radio waves (including microwaves), light (including ultraviolet, visible and infrared), heat radiation, X-rays, gamma rays and cosmic rays.

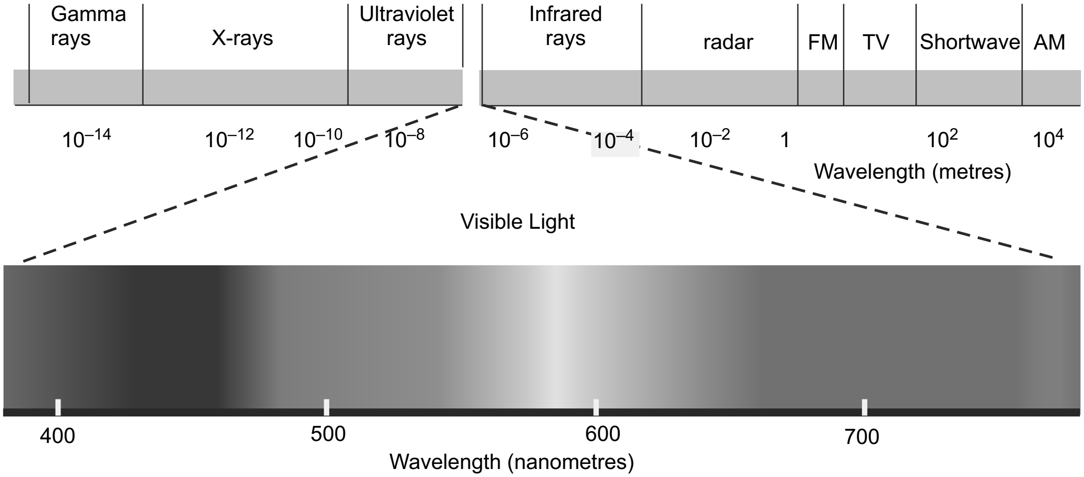


Fig. 1.32 Electromagnetic spectrum

1. Radio Frequency

Radio waves are generally available with wavelengths ranging from hundreds of metres to about one millimetre. They are used for transmission of data via modulation, Television, mobile phones, etc. all use radio waves. Radio waves can be made to carry information by varying a combination of the amplitude, frequency and phase of the wave within a frequency band.

2. Microwaves

The Super High Frequency (SHF) and Extreme High Frequency (EHF) of microwaves will be next in the frequency scale. They are electromagnetic waves with wavelengths ranging from as long as one metre to as short as one millimetre, or equivalently, with frequencies between 300 MHz (0.3 GHz) and 300 GHz. They are typically short enough to employ tubular metal waveguides of reasonable diameter.

3. Infrared Radiation

The infrared part of the electromagnetic spectrum covers the range from roughly 300 GHz (1 mm) to 400 THz (750 nm). It can be divided into three parts.

(a) **Far-infrared region**, from 300 GHz (1 mm) to 30 THz (10 μm). The lower part of this range may also be called **microwave**. This radiation is typically absorbed by so-called rotational modes in gas-phase molecules, by molecular motions in liquids and by phonons in solids. The water in the earth's atmosphere absorbs so strongly in this range that it renders the atmosphere effectively opaque.

(b) **Mid-infrared region**, from 30 to 120 THz (10 to 2.5 μm). Hot objects such as black-body radiators can radiate strongly in this range. It is absorbed by molecular vibrations, where the different atoms in a molecule vibrate around their equilibrium positions. This range is sometimes called the **fingerprint region** since the mid-infrared absorption spectrum of a compound is very specific for that compound.

(c) **Near-infrared region**, from 120 to 400 THz (2,500 to 750 nm). Physical processes that are relevant for this range are similar to those for visible light.

4. Visible Radiation (Light)

Above the infrared region in frequency spectrum will be visible light. This is the range in which the sun and stars emit most of their radiation. Visible light and near-infrared light are typically absorbed and emitted by electrons in molecules and atoms that move from one energy level to another. The light we see with our eyes is really a very small portion of the electromagnetic spectrum. A rainbow shows the visible part of the electromagnetic spectrum and the infrared region would be located just beyond the red side of the rainbow with ultraviolet appearing just beyond the violet end.

At most wavelengths, the information carried by electromagnetic radiation is not directly detected by human senses. Electromagnetic radiation with a wavelength between 380 nm and 760 nm (790–400 terahertz) is detected by the human eye and perceived as visible light. Other wavelengths, especially near infrared (longer than 760 nm) and ultraviolet (shorter than 380 nm), are also sometimes referred to as light, especially when the visibility to humans is not relevant.

5. Ultraviolet Light

This is radiation whose wavelength is shorter than the violet end of the visible spectrum and is longer than that of an X-ray. Ultraviolet rays can break chemical bonds, making molecules unusually reactive or ionising them. The sun emits a large amount of UV radiation, which could quickly turn the earth into a barren desert. However, most of it is absorbed by the atmosphere's ozone layer before reaching the surface.

6. X-Rays

X-rays are the next to ultraviolet rays which are also ionising, but due to their higher energies they can also interact with matter. Hard X-rays have shorter wavelengths than soft X-rays. As they can pass through most substances, X-rays can be used to 'see through' objects, most notably diagnostic X-ray images in medicine, which is known as **radiography**, as well as for high-energy physics and astronomy.

7. Gamma Rays

Next to hard X-rays are gamma rays, which were discovered by Paul Villard in 1900. These are the most energetic photons having no lower limit to their wavelength. Gamma rays are useful for the irradiation of food and seeds for sterilisation. In the medical field, they are used in radiation cancer therapy and some kinds of diagnostic imaging such as PET Scans. The wavelengths of gamma rays can be measured with high accuracy by means of Compton scattering.

Summary

The process of establishing a connection or link between two points for information exchange is called communication. Simply, it is the process of conveying messages at a distance. A communication system is an assembly of different electronic equipments mainly used for communication purpose.

A communication system involves three major components such as process of information, transmission and reception.

- Process of information includes
- Collection of data
- Processing of data
- Storage of the information

Next, transmission includes the further processing of data such as encoding. Finally, reception includes

- Decoding
- Storage of the data
- Interception

The electrical signals produced by transducers at transmitting section are of two types.

- Analog signals that continuously vary with time
- Digital signals which are not continuous

Signals can be further classified under the following categories.

- Continuous-time and discrete-time signals
- Real and complex signals
- Deterministic and random signals
- Periodic and aperiodic signals
- Even and odd signals
- Energy and power signals

Signals contain information about a variety of things and activities in the physical world. Broadly, there are two types of signal representation. They are:

- Time-domain representation
- Frequency-domain representation

There are a number of different mathematical transforms which are used to analyse time functions and are called frequency-domain methods. The following methods are the most common and they are in use with different fields.

- Fourier series is specifically useful for repetitive signals
- Fourier transform is specifically useful for nonrepetitive signals
- Laplace transform is useful for control systems, etc.
- Z-transform is specifically useful for discrete signals
- Wavelet transform is specifically useful for signal compression, etc.

Noise is defined as an extraneous form of energy with random frequency and amplitude which tends to interfere with the reception of a signal from a distant station. Noise may be picked up by a signal during its transmission from a transmitter to a receiver, which is commonly termed external noise. Alternately, noise may be produced within receiving equipment while it is receiving a signal. This type of noise is termed internal noise.

A communication channel is a physical medium which is used to send the signal from the transmitter to the receiver or it can be of any physical medium which transmits information or used for transmitting information like optical fibre, storage media, etc. By considering the limitations of the channel, it is necessary to design the communication system such that more data can be transmitted as possible without getting corrupted.

The electromagnetic spectrum is the range of all possible frequencies of electromagnetic radiation. The electromagnetic spectrum of an object is the characteristic distribution of electromagnetic radiation emitted or absorbed by that particular object. It extends from low frequencies used for modern radio to γ-radiation at the short-wavelength end, covering wavelengths from thousands of kilometres down to a fraction of the size of an atom.

REVIEW QUESTIONS

PART-A

1. What is the significance of electronic communication?
2. What are the major components of communication?
3. Mention the major types of communication.
4. Classify an electrical signal.
5. Differentiate between continuous-time signal and discrete-time signal.
6. What are real and complex signals?
7. Differentiate between deterministic and random signals.
8. What do you mean by an aperiodic signal? Give an example.
9. What is an odd signal? Give an example.
10. Differentiate between odd and even signals.
11. When is a signal said to be an energy signal? Give one example.
12. When is a signal said to be a power signal? Give one example.
13. How will you represent signals?
14. State the significance of Fourier series and give Fourier-series pair.
15. What is the purpose of Fourier transform? Give Fourier-transform pair.
16. Give any two properties of Fourier transform.
17. What do you mean by bandlimited signals?
18. Define band pass signal.
19. Mention the causes of noise in a communication channel.
20. What are the possible types of noise in a communication channel?
21. What is thermal agitation noise?
22. What do you mean by shot noise?
23. Define noise figure
24. Why are mathematical models for communication channels necessary?
25. Draw the mathematical model of an additive noise channel.
26. Differentiate between linear and a linear time-variant filter channel.
27. State the physical properties of an electromagnetic spectrum.
28. What are the different wavelength regions of the electromagnetic spectrum?

PART-B

1. Draw the block diagram of an electronic communication system and explain the functioning of each block.
2. What are the constituents of a communication system?
3. How will you classify electrical signals produced by transducers at the transmitting section in a communication system? Explain them.
4. Explain odd and even signals.
5. Describe real and complex signals with examples.
6. Differentiate between energy and power signals. Give examples
7. What are the differences between periodic and aperiodic signals? Give examples.
8. When will you prefer Fourier transform? State and prove its important properties.
9. State the sources of noise in a communication channel. Explain them in detail.
10. Why are mathematical models for communication channels necessary? How will you model them?
11. Explain the modelling of physical communication channels with neat sketches.
12. With a neat sketch, explain the different regions of an electromagnetic spectrum.

2

AMPLITUDE MODULATION

Objectives

- To know the need for modulation in a communication system
- To discuss the different types of modulation in detail
- To provide the process of Amplitude Modulation (AM) and its representation in detail.
- To provide the process of Double Side-Band Suppressed Carrier—Amplitude Modulation (DSB-SC-AM) and its representation in detail
- To provide the process of Single Side-Band Suppressed Carrier—Amplitude Modulation (SSB-SC-AM) and its representation in detail
- To know about the purpose of Vestigial Side Band (VSB) and its representation
- To discuss various methods of generation of AM and also about the functioning of AM transmitters in detail

2.1 INTRODUCTION

Modulation is the process by which some character of a high-frequency carrier signal is varied in accordance with the instantaneous value of another signal called modulating or message signal.

The signal containing information to be transmitted is known as **modulating** or **message signal**. It is also known as **baseband signal**. The term "baseband" means the band of frequencies representing the signal supplied by the source of information. Usually, the frequency of the carrier is greater than the modulating signal. The signal resulting from the process of modulation is called **modulated signal**.

2.2 NEED FOR MODULATION

2.2.1 For Easy Transmission

By considering that the communication medium is free space, antennas are needed to transmit and receive the signal. The antenna radiates effectively when its height is of the order of wavelength of the signal to be transmitted.

For example, for a frequency of 1 kHz, the height of the antenna required for effective radiation would be half of the wavelength, i.e.

$$\text{Antenna height} = \frac{\lambda}{2} = \frac{C}{2f} \tag{2.1}$$

$$= \frac{3\times 10^8}{2\times 1\times 10^6} = 150 \text{ km} \tag{2.2}$$

where

λ is the wavelength of the signal to be fixed,

C is the velocity of light, and

f is the frequency of the signal to be transmitted.

But it is highly impractical to construct and install such an antenna. However, the height of the antenna can be reduced by modulation techniques and it achieves effective radiation.

The process of modulation provides frequency translation, i.e. Audio Frequency (AF) signals are translated into Radio Frequency (RF) signals. These RF signals act as carrier signals and AF signals act as message signals. Hence, the height of the antenna required is very much reduced. Here, 1 kHz baseband signal is translated into a high-frequency signal of 1 MHz.

$$\text{Antenna height} = \frac{\lambda}{2} = \frac{C}{2f}$$

$$= \frac{3\times 10^8}{2\times 1\times 10^6} = 150 \text{ m} \tag{2.3}$$

This height of the antenna is practically achievable.

2.2.2 Narrow Banding

Assume that the baseband signal in a broadcast system redirected directly with the frequency range extending from 50 Hz to 10 kHz . If an antenna is designed for 50 Hz, it will be too long for 10 kHz and vice versa. Hence, it is impossible to provide a wideband antenna.

However, if an audio signal is modulated to radio frequency range of 1 MHz then the ratio of lowest to highest frequency will be

$$\frac{10^6 + 50}{10^6 + 10^4} = \frac{1}{1.01} \cong 1 \tag{2.4}$$

Therefore, the same antenna will be suitable for the entire band extending from $(10^6 + 50)$ Hz to $(10^6 + 10^4)$ Hz. Thus, modulation converts a wideband signal to narrow band. This is called narrow banding.

2.2.3 Multiplexing

If more than one signal uses a single channel then modulation may be used to translate different signals to different special locations, thus enabling the receiver to select the desired signal. Application of multiplexing includes data telemetry, FM stereophonic broadcasting and long-distance telephones.

2.2.4 To Overcome Equipment Limitations

Occasionally, in signal-processing applications, the frequency of the signal to be processed and the frequency range of the processing apparatus does not match. If the equipment is elaborately complex it is necessary to fix some frequency range in the equipment, and translate the frequency range of the signal corresponding to the fixed range of the equipment. The modulation can be used to accomplish this frequency translation.

2.2.5 Modulation for Frequency Assignment

Modulation allows several radios and TV stations for broadcasting simultaneously at different carrier frequencies and allows different receivers to be tuned to select different stations.

2.2.6 Modulation to Reduce Noise and Interferences

Even though the effect of noise and interferences cannot be completely eliminated in a communication system, it is possible to minimise the effect, by using certain types of modulation schemes. They require a transmission BW larger than the BW of the message signal.

2.3 TYPES OF MODULATION

Since it is impractical to propagate information signals over standard transmission media, it is necessary to modulate the source information onto a higher frequency analog signal, a carrier, which carries the information through the system. The information signal modulates the carrier by changing either its amplitude, frequency or phase. Modulation is simply the process of changing one or more properties of the analog carrier in proportion with the information signal. Depending on the types of signals involved in the modulation process, modulation is basically of two types:

a) Analog modulation

b) Digital modulation

If the information signal is of analog type, the modulation is further classified as follows:

1. Amplitude Modulation (AM)

It is a process in which the information is analog and the amplitude (V) of the carrier is varied in accordance with the instantaneous value of the information signal.

2. Frequency Modulation(FM)

It is a process in which the information is analog and the frequency (f) of the carrier is varied in accordance with the instantaneous value of the information signal.

3. Phase Modulation (PM)

It is a process in which the information is analog and the phase (θ) of the carrier is varied in accordance with the instantaneous value of the information signal.

Let a sinusoidal carrier wave in analog modulation be given by

$$V_c(t) = V_c \sin(\omega_c t + \theta) \tag{2.5}$$

$$= A \sin(\omega_c t + \theta) \tag{2.6}$$

where

A is the amplitude of the carrier signal,

ω_c is the angular frequency, and

θ is the phase angle.

Any of these parameters can be varied in accordance with the baseband or message signal. Among AM, FM and PM, FM and PM are combinedly called **angle modulation**.

If the information signal is of digital type, the modulation is further classified as follows:

1. Amplitude Shift Keying (ASK)

If the information signal is digital and the amplitude (V) of the carrier is varied proportional to the information signal, a digitally modulated signal known as Amplitude Shift Keying (ASK) is produced.

2. Frequency Shift Keying (FSK)

If the information signal is digital and the frequency (f) of the carrier is varied proportional to the information signal, a digitally modulated signal known as Frequency Shift Keying (FSK) is produced.

3. Phase Shift Keying (PSK)

If the information signal is digital and the phase (θ) of the carrier is varied proportional to the information signal, a digitally modulated signal known as Phase Shift Keying (PSK) is produced.

If both the amplitude and the phase are varied proportional to the information signal then the resultant is Quadrature Amplitude Modulation (QAM). ASK, FSK, PSK and QAM are forms of digital modulation.

2.4 AMPLITUDE MODULATION

2.4.1 Definition and Representation

Amplitude modulation is the process of changing the amplitude of the carrier in accordance with the amplitude of the message signal. Frequency and phase of the carrier signal are not altered during the process. It is a low-quality form of modulation and often used for commercial broadcasting of both audio and video signals.

Let

$$V_{\mathrm{m}}(t) = V_{\mathrm{m}} \sin \omega_{\mathrm{m}} t \tag{2.7}$$

$$V_{\mathrm{c}}(t) = V_{\mathrm{c}} \sin \omega_{\mathrm{c}} t \tag{2.8}$$

where

V_{m} is the maximum amplitude of the modulating signal,

V_{c} is the maximum amplitude of the carrier signal,

ω_{m} is the angular frequency of the modulating signal, and

ω_{c} is the angular frequency of the carrier signal.

According to the definition,

$$V_{\mathrm{AM}} = V_{\mathrm{c}} + V_{\mathrm{m}} \sin \omega_{\mathrm{m}} t \tag{2.9}$$

$$= V_{\mathrm{c}} \left[1 + \frac{V_{\mathrm{m}}}{V_{\mathrm{c}}} \sin \omega_{\mathrm{m}} t \right] \tag{2.10}$$

$$= V_{\mathrm{c}} \left[1 + m_{\mathrm{a}} \sin \omega_{\mathrm{m}} t \right] \tag{2.11}$$

where

$$m_{\mathrm{a}} = \frac{V_{\mathrm{m}}}{V_{\mathrm{c}}} = \text{Modulation index or depth of modulation}$$

Modulation index or **coefficient** is an indicator to describe the amount of amplitude change (modulation) present in an AM waveform (depth of modulation), basically stated in the form of percentage.

$$m_a = \frac{V_m}{V_c} \times 100 \tag{2.12}$$

The instantaneous amplitude of the modulated signal is

$$\begin{aligned} V_{AM}(t) &= V_{AM} \sin \omega_c t \\ &= V_c\,[1 + m_a \sin \omega_m t] \sin \omega_c t \\ &= V_c \sin \omega_c t + m_a\, V_c \sin \omega_m t \sin \omega_c t \\ &= V_c \sin \omega_c t + \frac{m_a V_c}{2}\,[\cos(\omega_c - \omega_m)t - \cos(\omega_c + \omega_m)t] \\ &= V_c \sin \omega_c t + \frac{m_a V_c}{2} \cos(\omega_c - \omega_m)t - \frac{m_a V_c}{2} \cos(\omega_c + \omega_m)t] \end{aligned} \tag{2.13}$$

2.4.2 Frequency Spectrum of AM wave

In Equation (2.13), the first term of RHS represents the carrier wave. The second and third terms are identical, called as Lower Side Band (LSB) and Upper Side Band (USB) respectively. Figure 2.1 shows the frequency spectrum of an AM wave.

The above figure shows the side-band terms lying on either side of the carrier term which are separated by $\omega_{m.}$ The frequency of LSB is $\omega_c - \omega_m$ and that of USB is $\omega_c + \omega_m$. The bandwidth (BW) of AM can be determined by using these side bands. Hence, BW is twice the frequency of the modulating signal.

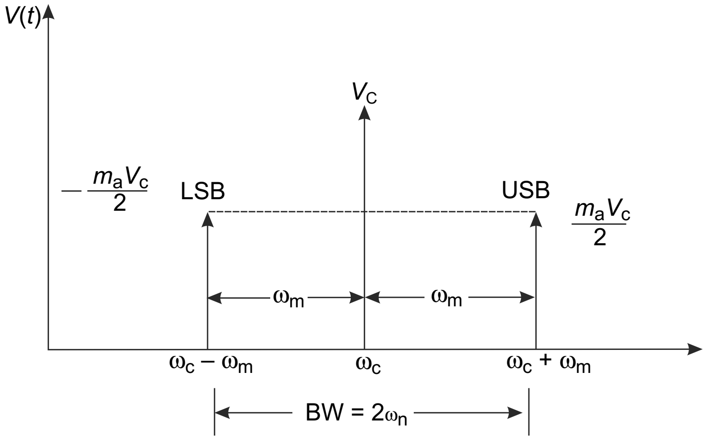


Fig 2.1 Frequency spectrum of AM wave

EXAMPLE 2.1

A sinusoidal carrier voltage of 500 kHz frequency and 200 V amplitude is modulated by a sinusoidal voltage of 10 kHz frequency producing 50% modulation. Calculate the frequency and amplitude of USB and LSB.

Solution

Frequency of LSB = (500 – 10) kHz = 490 kHz

Frequency of LSB = (500 + 10) kHz = 510 kHz

$$\text{Amplitude} = \frac{m_a V_c}{2}$$

$$= \frac{0.5 \times 200}{2} = 50 \text{ V}$$

2.4.3 Graphical Representation of an AM Wave

Graphical representation of an AM wave is illustrated in Figure 2.2.

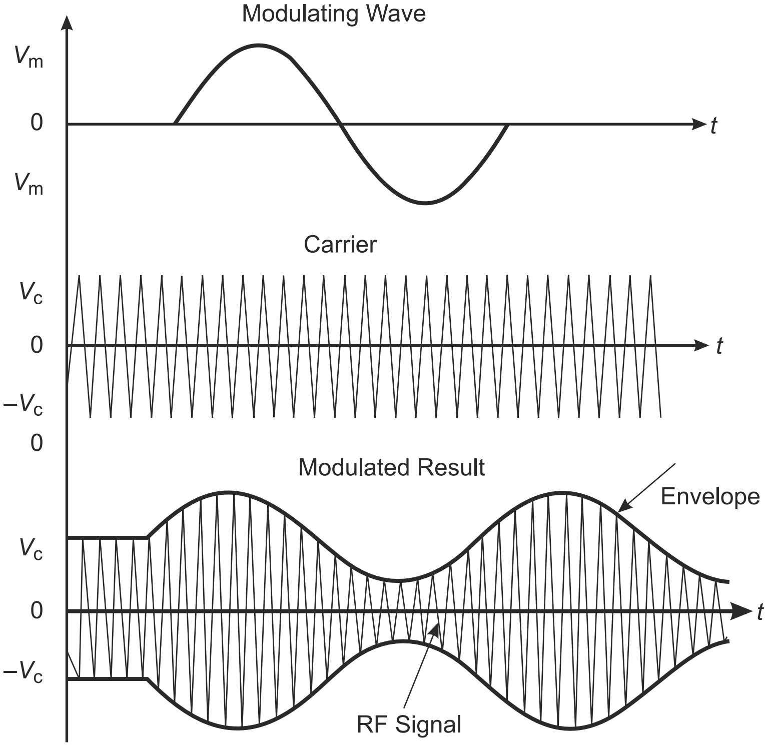


Fig. 2.2 Graphical representation of an AM wave

From the above figure, the amplitude of the carrier is varied in accordance with the modulating signal while frequency remains the same. The AM waveform reaches minimum value when the modulating signal amplitude is at maximum negative. The repetition rate of the envelope is equal to the frequency of the modulating signal, and the shape of the envelope is identical to the shape of the modulating signal.

From Figure 2.2,

$$2\,V_{\text{modulating}}(\text{max}) = V_{\text{max}} - V_{\text{min}} \tag{2.14}$$

$$V_{\text{m}} = V_{\text{modulating}}(\text{max}) = \frac{V_{\text{max}} - V_{\text{min}}}{2} \tag{2.15}$$

and

$$V_{\text{Carrier}}(\text{max}) = V_{\text{max}} - V_{\text{m}} \tag{2.16}$$

$$= V_{\text{max}} - \frac{V_{\text{max}} - V_{\text{min}}}{2}$$

$$= \frac{V_{\text{max}} + V_{\text{min}}}{2} \tag{2.17}$$

$$\therefore\ m_{\text{a}} = \frac{V_{\text{m}}}{V_{\text{c}}} = \frac{(V_{\text{max}} - V_{\text{min}})/2}{(V_{\text{max}} + V_{\text{min}})/2}$$

$$= \frac{V_{\text{max}} - V_{\text{min}}}{V_{\text{max}} + V_{\text{min}}} \tag{2.18}$$

Using the equation relating maximum peak-to-peak amplitude and minimum peak-to-peak amplitude to modulation factor,

$$m_{\text{a}} = \frac{\max pp - \min pp}{\max pp + \min pp} \tag{2.19}$$

Percent modulation can be calculated as

Percent modulation = $M = m_{\text{a}} \times 100$

EXAMPLE 2.2

Determine the modulation factor and percent modulation of the signal shown as Figure 2.3.

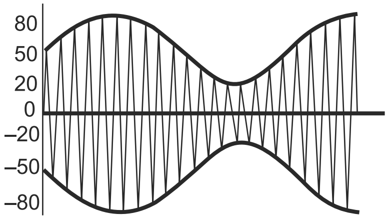


Fig. 2.3

Solution

$$m = \frac{\max pp - \min pp}{\max pp + \min pp}$$

From the given figure,

$$\max pp = 2(80) = 160$$

$$\min pp = 2(20) = 40$$

$$m = \frac{160 - 40}{160 + 40}$$

$$\therefore\ m = 0.6$$

$\therefore$ percent modulation = $M = m \times 100 = 60\%$

EXAMPLE 2.3

Determine the modulation factor and percent modulation of the signal shown as Figure 2.4.

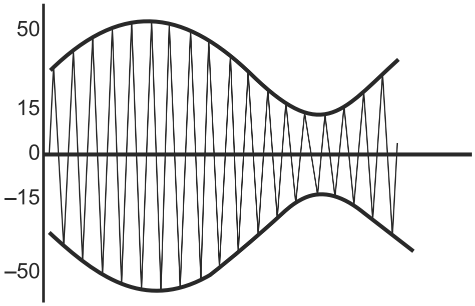


Fig. 2.4

Solution

$$m = \frac{\max pp - \min pp}{\max pp + \min pp}$$

From the given figure,

$$\max pp = 2(50) = 100$$

$$\min pp = 2(15) = 30$$

$$m = \frac{100 - 30}{100 + 30}$$

$\therefore$ $m = 0.538$

$\therefore$ percent modulation = $M = m \times 100 = 53.8\%$

2.4.4 Phasor Representation of AM with Carrier

Figure 2.5 shows the phasor representation of an AM with carrier.

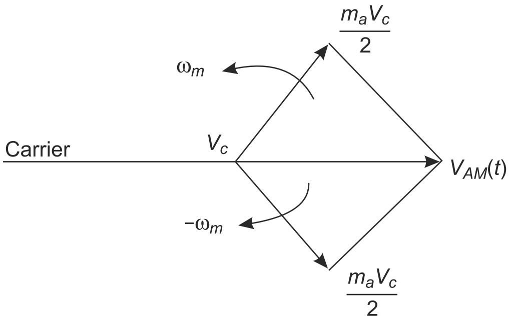


Fig. 2.5 Phasor representation of AM with carrier

It is the easy way of representation of an AM wave, where V_c is the carrier wave phasor, taken as reference phasor. The two side bands having a frequency of ($\omega_c + \omega_m$) and ($\omega_c - \omega_m$) are represented by two phasors rotating in opposite directions with angular frequency of ω_m. The resultant phasor is $V(t)$. It depends on the position of the side-band phasor and carrier wave phasor.

2.4.5 Degrees of Modulation

The modulating signal is preserved in the envelope of amplitude-modulated signal. Only if $V_m \leq V_c$ then, $m_a < 1$.

where V_m = Maximum amplitude of modulating signal, and

V_c = Maximum amplitude of carrier signal.

There are three degrees of modulation depending upon the amplitude of the message signal relative to carrier amplitude.

1. Under-modulation
2. Critical modulation
3. Overmodulation

1. Under-modulation

In this case, the modulation index is $m_a < 1$, i.e. $V_m < V_c$. It is shown in Figure 2.6.

Here, the envelope of an AM signal does not reach the zero-amplitude axis. Hence, the message signal is fully preserved in the envelope of an AM wave. This is known as under-modulation. An envelope detector can recover the message signal without any distortion.

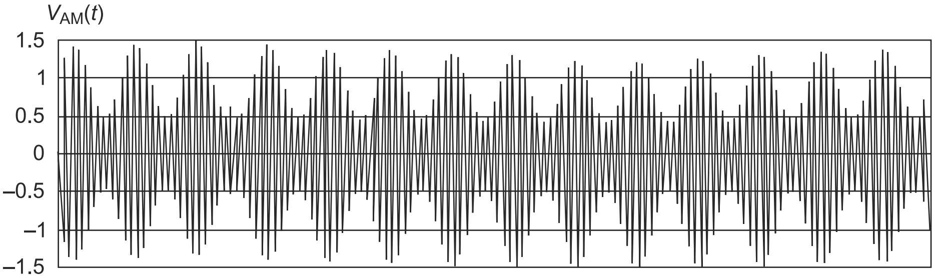


Fig. 2.6 Under-modulation

2. Critical Modulation

In this case, the modulation index is $m_a = 1$, i.e. $V_m = V_c$. It is shown in Figure 2.7.

Here, the envelope of the modulating signal just reaches the zero-amplitude axis. The message signal remains preserved. This is known as critical modulation.

In this case also, the modulated signal can be recovered by using an envelope detector without any distortion.

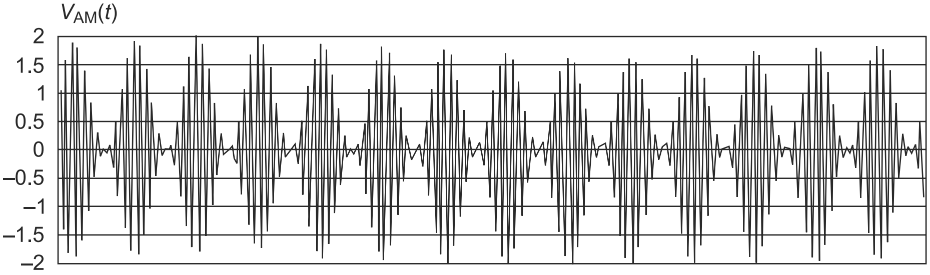


Fig. 2.7 Critical modulation

3. Overmodulation

In this case, the modulation index is $m_d > 1$, i.e. $V_m > V_c$. It is shown in Figure 2.8.

In overmodulation, the message signal cannot be fully recovered without any distortion.

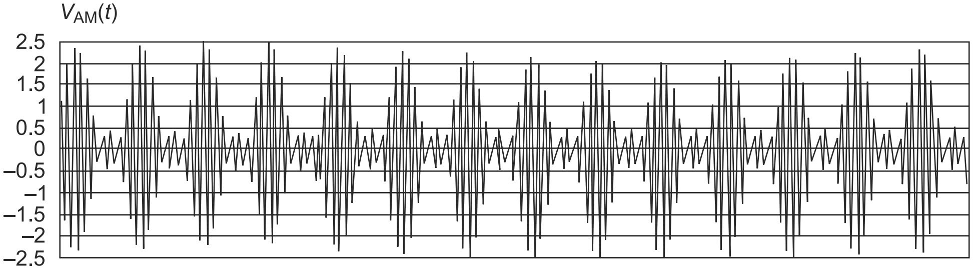


Fig. 2.8 Over-modulation

2.4.6 Power Relation in AM

The modulated wave contains three terms such as carrier wave, Lower Side Band (LSB) and Upper Side Band (USB). Therefore, the modulated wave contains more power than the carrier had before modulation took place. Since the amplitude of side bands depend on the modulation index, it is preserved that the total power in the modulated wave depends on the modulation index.

The total power in the modulated wave will be

$$P_{\mathrm{T}} = P_{\mathrm{C}} + P_{\mathrm{LSB}} + P_{\mathrm{USM}}$$

$$= \frac{V_{\mathrm{Carrier}}^2}{R} + \frac{V_{\mathrm{LSB}}^2}{R} + \frac{V_{\mathrm{USB}}^2}{R} \tag{2.20}$$

where

V_{Carrier} = rms carrier voltage

V_{LSB} = rms value of lower side-band voltage

V_{USB} = rms value of upper side-band voltage

R = Resistance in which power is dissipated

$$P_{\mathrm{Carrier}} = \frac{V_{\mathrm{Carrier}}^2}{R} = \frac{\left(V_{\mathrm{C}} / \sqrt{2}\right)^2}{R} + \frac{V_{\mathrm{C}}^2}{2R} \tag{2.21}$$

where

V_{C} = Maximum amplitude of carrier wave

V_{m} = Maximum amplitude of modulating wave

Similarly,

$$P_{\mathrm{LSB}} = \frac{V_{\mathrm{LSB}}^2}{R} \tag{2.22}$$

$$= \frac{\left(\dfrac{\dfrac{m_{\mathrm{a}} V_{\mathrm{C}}}{2}}{\sqrt{2}}\right)^2}{R}$$

$$= \frac{m_{\mathrm{a}}^2 V_{\mathrm{C}}^2}{8R} \quad \left[\because V_{\mathrm{LSB}} = V_{\mathrm{USB}} = \frac{m_{\mathrm{a}} V_{\mathrm{c}}}{2}\right] \tag{2.23}$$

$$\therefore\ P_{\mathrm{T}} = P_{\mathrm{c}} + P_{\mathrm{LSB}} + P_{\mathrm{USB}}$$

$$= \frac{V_{\mathrm{C}}^2}{2R} + \frac{m_{\mathrm{a}}^2\, V_{\mathrm{C}}^2}{8R} + \frac{m_{\mathrm{a}}^2\, V_{\mathrm{C}}^2}{8R}$$

$$= \frac{V_C^2}{2R} + 2\left[\frac{m_a^2 V_C^2}{8R}\right]$$

$$= \frac{V_C^2}{2R} + \frac{m_a^2 V_C^2}{4R}$$

$$= \frac{V_C^2}{2R}\left[1 + \frac{m_a^2}{2}\right] \tag{2.24}$$

It is known that,

$$P_C = \frac{V_C^2}{2R} \tag{2.25}$$

Substitute it in Equation (2.18).

$$\therefore\ P_T = P_C\left[1 + \frac{m_a^2}{2}\right] \tag{2.26}$$

If $m_a = 1$, i.e. 100% modulation

$$\frac{P_T}{P_C} = 1.5$$

$\therefore$

$$P_T = 1.5\ P_C \tag{2.27}$$

2.4.7 Current Relation in AM

From Equation (2.20),

$$P_T = P_c\left[1 + \frac{m_a^2}{2}\right]$$

We know that,

$$P_T = I_T^2 \,.\, \mathrm{R} \tag{2.28}$$

$$I_T^2 \,.\, R = I_C^2 \,.\, R \tag{2.29}$$

Hence,

$$I_T^2 = I_C^2\left[1 + \frac{m_a^2}{2}\right]$$

$$\therefore\ I_T = I_C\sqrt{\left[1 + \frac{m_a^2}{2}\right]} \tag{2.30}$$

where

I is the total current, and

I_C is the carrier current.

EXAMPLE 2.4

The total power content of an AM signal is 1000 W. Determine the power being transmitted at the carrier frequency and at each of the side bands when the % modulation is 100%.

Solution

The total power consists of the power at the carrier frequency, that at the upper side band and that at the lower side band. Since the % modulation is 100%, m_a=1.

$$P_T = P_C + P_{LSB} + P_{USB}$$

$$P_T = P_C + \frac{m_a^2 P_C}{4} + \frac{m_a^2 P_C}{4}$$

$$= P_C + \frac{m_a^2 P_C}{2}$$

$$1000 = P_C + \frac{(1.0)^2\ P_C}{2}$$

$$= P_C + 0.5P_C$$

$$= 1.5P_C$$

$$\frac{1000}{1.5} = P_C$$

$$P_C = 666.67 \text{ W}$$

This leaves 1000 – 666.67 = 333.33 W to be shared equally between upper and lower side bands.

$$P_{USB} + P_{LSB} = 333.33 \text{ W}$$

In AM, $P_{USB} = P_{LSB}$

$$\therefore\ 2P_{LSB} = 333.33$$

$$\therefore\ P_{USB} = P_{LSB} = \frac{333.33}{2}$$

$$= 166.66 \text{ W}$$

EXAMPLE 2.5

Determine the power content of the carrier and each of the side bands for an AM signal having a percent modulation of 80% and a total power of 2500 W.

Solution

The total power of an AM signal is the sum of the power at the carrier frequency and the power contained in the side bands.

$$P_T = P_C + P_{LSB} + P_{USB}$$

$$P_T = P_C + \frac{m_a^2 P_C}{4} + \frac{m_a^2 P_C}{4}$$

$$P_T = P_C + 2\left[\frac{m_a^2 P_C}{4}\right]$$

$$P_T = P_C + \frac{m_a^2 P_C}{2}$$

$$2500 = P_C + \frac{(0.8)^2\, P_C}{2}$$

$$= P_C + \frac{0.64\, P_C}{2}$$

$$= 1.32\, P_C$$

$$2500 = 1.32\, P_C$$

$$\therefore\ P_C = \frac{2500}{1.32} = 1893.9 \text{ W}$$

The power in the two side bands is the difference between the total power and the carrier power.

$$P_{LSB} = P_{USB} + 2500 - 1893.9$$

$$P_{LSB} + P_{USB} = 606.1 \text{ W}$$

$$\therefore\ P_{LSB} = P_{USB} = \frac{606.1}{2} \text{ W} = 303.5 \text{ W}$$

EXAMPLE 2.6

The power content of the carrier of an AM wave is 5 kW. Determine the power content of each of the side bands and the total power transmitted when the carrier is modulated as 75%

Solution

In an AM wave, the power in each of the side bands is equal.

$$P_{LSB} = P_{USB} = \frac{m_a^2 P_c}{4}$$

$$= \frac{(0.75)^2\,(5000)}{4}$$

$$\therefore\ P_{\text{LSB}} + P_{\text{USB}} = 703.13\text{ W}$$

The total power is the sum of the carrier power and the power in the two side bands.

$$P_{\text{T}} = P_{\text{C}} + P_{\text{LSB}} + P_{\text{USB}}$$

$$= 5000 + 703.13 + 703.13$$

$$= 6406.26\text{ W}$$

EXAMPLE 2.7

An AM wave has a power content of 800 W at its carrier frequency. Determine the power content of each side band for a 90% modulation.

Solution

$$P_{\text{LSB}} = P_{\text{USB}} = \frac{m_{\text{a}}^2 P_{\text{c}}}{4}$$

$$= \frac{(0.9)^2\,(800)}{4} = 162\text{ W}$$

EXAMPLE 2.8

Determine the percent modulation of an AM wave which has power content at the carrier of 8 kW and 2 kW in each of its side bands when the carrier is modulated by a simple audio tone.

Solution

With the power content of the side bands and the carrier, the relationship of side-band power can be used to determine the modulation factor. After multiplying it with 100, it provides percent modulation.

$$P_{\text{LSB}} = P_{\text{USB}} = \frac{m_{\text{a}}^2 P_{\text{c}}}{4}$$

$$2 \times 10^3 = \frac{m_{\text{a}}^2\,(8\times10^3)}{4}$$

$$m_{\text{a}} = \frac{4\times2\times10^3}{8\times10^3}$$

$$\therefore\ m_{\text{a}} = 1$$

Percent modulation $M = m_{\text{a}} \times 100$

$$\therefore\ M = 100\%$$

EXAMPLE 2.9

The total power content of an AM wave is 600 W. Determine the percent modulation of the signal if each of the side bands contains 75 W.

Solution

To determine M, the carrier power is first determined. Once P_C is found, the relationship of side-band power can be used to determine the modulation factor. After multiplying it with 100, it provides percent modulation.

$$P_T = P_C + P_{LSB} + P_{USB}$$

$$600 = P_C + 75 + 75$$

$$P_C = 600 - 150$$

$$\therefore\ P_C = 450$$

Now, using the relationship between the carrier power and the side-band power,

$$P_{LSB} = P_{USB} = \frac{m_a^2 P_c}{4}$$

$$75 = \frac{m_a^2(450)}{4}$$

$$m_a^2 = \frac{4(75)}{450} = 0.667$$

$$\therefore\ m_a = 0.816$$

The percent modulation is calculated as

$$M = 0.816 \times 100$$

$$\therefore\ M = 81.6\%$$

EXAMPLE 2.10

Find the percent modulation of an AM wave whose total power content is 2500 W and whose side bands each contain 400 W.

Solution

To determine M, the carrier power is first determined. Once P_C is found, the relationship of side-band power can be used to determine the modulation factor. After multiplying it with 100, it provides percent modulation.

$$P_T = P_C + P_{LSB} + P_{USB}$$

$$2500 = P_{\mathrm{C}} + 400 + 400$$

$$P_{\mathrm{C}} = 2500 - 800 = 1700 \text{ W}$$

$$P_{\mathrm{LSB}} = P_{\mathrm{USB}} = \frac{m_{\mathrm{a}}^2 P_{\mathrm{c}}}{4}$$

$$400 = \frac{m_{\mathrm{a}}^2\,(1700)}{4}$$

$$m_{\mathrm{a}}^2 = \frac{400(4)}{1700}$$

$$m_{\mathrm{a}}^2 = \frac{1600}{1700} = 0.941$$

$$\therefore\ m_{\mathrm{a}} = 0.970$$

The percent modulation is calculated as

$$M = 0.970 \times 100$$

$$\therefore\ M = 97\%$$

EXAMPLE 2.11

Determine the power content of each of the side bands and of the carrier of an AM signal that has a percent modulation of 85% and contains 1200 W total power.

Solution

$$P_{\mathrm{T}} = P_{\mathrm{C}} = \left(1 + \frac{m_{\mathrm{a}}^2}{2}\right)$$

$$1200 = P_{\mathrm{C}}\left(1 + \frac{(0.85)^2}{2}\right)$$

$$1200 = P_{\mathrm{C}}\left(1 + \frac{0.7225}{2}\right)$$

$$1200 = P_{\mathrm{C}}\,(1 + 0.3613)$$

$$1200 = P_{\mathrm{C}}\,(1.3613)$$

$$P_{\mathrm{C}} = \frac{1200}{1.3613}$$

$$\therefore\ P_{\mathrm{C}} = 881.5 \text{ W}$$

Total power is the sum of carrier power and power in side bands.

$$P_T = P_C + P_{SB}$$

$$881.5 + P_{SB} = 1200$$

$$P_{SB} = 1200 - 881.5 = 318.5$$

$$P_{SB} = P_{LSB} + P_{USB} = 318.5$$

$$P_{LSB} = P_{USB} = \frac{P_{SB}}{2} = \frac{318.5}{2}$$

$$\therefore\ P_{LSB} = P_{USB} = 159.25\text{ W}$$

EXAMPLE 2.12

An AM signal in which the carrier is 70% modulated contains 1500 W at the carrier frequency. Determine the power content of the upper and lower side bands for this percent modulation. Calculate the power at the carrier and the power content of each of the side bands when the percent modulation drops to 50%.

Solution

For 70% modulation,

$$P_{LSB} = P_{USB} = \frac{m_a^2 P_c}{4}$$

$$= \frac{(0.7)^2\,(1500)}{4} = 183.75$$

$$\therefore\ P_{LSB} = P_{USB} = 183.75\text{ W}$$

In standard AM transmission, carrier power remains the same, regardless of percent modulation.

$$P_{C_{50}} = P_{C_{70}} = 1500\text{ W}$$

Power in side bands for 50% modulation is,

$$P_{LSB} = P_{USB} = \frac{m_a^2 P_c}{4}$$

$$= \frac{(0.5)^2\,(1500)}{4} = 93.75$$

$$\therefore\ P_{LSB} = P_{USB} = 93.75\text{ W}$$

EXAMPLE 2.13

The percent modulation of an AM wave changes from 40% to 60%. Originally, the power content at the carrier frequency was 900 W. Determine the power content at the carrier frequency and within each of the side bands after the percent modulation has risen to 60%.

Solution

In standard AM transmission, carrier power remains the same, regardless of percent modulation.

$$P_{\mathrm{C60}} = P_{\mathrm{C40}} = 900 \text{ W}$$

Power in side bands for 60% modulation is,

$$P_{\mathrm{LSB}} = P_{\mathrm{USB}} = \frac{m_{\mathrm{a}}^2 P_{\mathrm{c}}}{4}$$

$$= \frac{(0.60)^2\,(900)}{4} = 81.0$$

$$\therefore\ P_{\mathrm{LSB}} = P_{\mathrm{USB}} = 81.0 \text{ W}$$

2.4.8 Percentage Efficiency

It can be defined as the ratio of power in side bands to total power because side bands only contain the useful information.

$$\%\,\eta = \frac{\text{Total power in side bands}}{\text{Total power}} \times 100 \tag{2.31}$$

$$= \frac{P_{\mathrm{LSB}} + P_{\mathrm{USB}}}{P_{\mathrm{T}}} \times 100$$

$$= \frac{\dfrac{m_{\mathrm{a}}^2 V_{\mathrm{C}}^2}{8R} + \dfrac{m_{\mathrm{a}}^2 V_{\mathrm{C}}^2}{8R}}{\dfrac{V_{\mathrm{C}}^2}{2R}\left[1 + \dfrac{m_{\mathrm{a}}^2}{2}\right]} \times 100$$

$$= \frac{\dfrac{m_{\mathrm{a}}^2 V_{\mathrm{C}}^2}{4R}}{\dfrac{V_{\mathrm{C}}^2}{2R}\left[1 + \dfrac{m_{\mathrm{a}}^2}{2}\right]} \times 100$$

$$= \frac{\dfrac{m_{\mathrm{a}}^2}{2}}{1 + \dfrac{m_{\mathrm{a}}^2}{2}} \times 100$$

$$\%\,\eta = \frac{m_{\mathrm{a}}^2}{2 + m_{\mathrm{a}}^2} \times 100 \tag{2.32}$$

If $m_a = 1,$

$$\% \ \eta = \frac{1}{3} \times 100 = 33.3\% \tag{2.33}$$

From this, it is calculated that only 33.3% of energy is used and the remaining power is wasted by the carrier transmission along with the side bands.

EXAMPLE 2.14

Determine percentage efficiency and percentage of the total power carried by the side bands of the AM wave when $m_a = 0.5$ and $m_a = 0.3$.

Solution

$$\% \ \eta = \frac{\frac{m_a^2}{2}}{1+\frac{m_a^2}{2}} \times 100$$

When $m_a = 0.5,$

$$\% \ \eta = \frac{\frac{(0.5)^2}{2}}{1+\frac{(0.5)^2}{2}} \times 100 = 11.11\%$$

When $m_a = 0.3,$

$$\% \ \eta = \frac{\frac{(0.3)^2}{2}}{1+\frac{(0.3)^2}{2}} \times 100 = 43\%$$

2.4.9 Drawbacks of AM

In conventional AM double side-band system, the carrier signal does not carry information; the information is contained in the side bands.

Due to the nature of this system, the drawbacks are as follows:

1. Carrier power constitutes two-thirds or more of the total transmitted power.
2. Both side bands contain the same information. Transmitting both side bands is redundant and thus causes it to utilise twice as much bandwidth as needed with single side-band system.
3. Conventional AM is both power- and bandwidth-inefficient.

2.5 MODULATION BY SEVERAL SINE WAVES

Let V_1, V_2, V_3,..., etc. be the simultaneous modulation voltages. Then the total modulating voltage V_T will be equal to the square root of the sum of the square of individual voltages.

$$V_T = \sqrt{V_1^2 + V_2^2 + V_3^2 + \ldots} \tag{2.34}$$

Dividing both sides by V_c, we get

$$\frac{V_T}{V_c} = \sqrt{\frac{V_1^2 + V_2^2 + V_3^2 + \ldots}{V_c^2}}$$

$$= \sqrt{\frac{V_1^2}{V_c^2} + \frac{V_2^2}{V_c^2} + \frac{V_3^2}{V_c^2} + \ldots}$$

$$= \sqrt{m_1^2 + m_2^2 + m_3^2 + \ldots} \tag{2.35}$$

Modulation index can also be found out by using the principle that while modulating the carrier simultaneously using several sine waves, the carrier power will be unaffected. The total side-band power will be the sum of individual side-band power.

$$P_{SB} = P_{SB1} + P_{SB2} + P_{SB3} + \ldots \tag{2.36}$$

We know that,

$$P_{SB} = \frac{P_C m_a^2}{2}$$

Substituting this in Equation (2.30),

$$\frac{P_C m_T^2}{2} = \frac{P_C m_1^2}{2} + \frac{P_C m_2^2}{2} + \frac{P_C m_3^2}{2} + \ldots \tag{2.37}$$

By cancelling $P_c/2$ on both sides,

$$m_T^2 = m_1^2 + m_2^2 + m_3^2 + \ldots$$

$$m_T^2 = \sqrt{m_1^2 + m_2^2 + m_3^2 + \ldots} \tag{2.38}$$

The total modulation must be less than unity. If it is greater than unity, over modulation will occur and it will result in distorted output.

2.6 DOUBLE-SIDE-BAND SUPPRESSED-CARRIER AMPLITUDE MODULATION (DSB-SC-AM)

The major considerable parameters of a communication system are transmitting power and bandwidth. Hence, it is very much necessary to save the power and bandwidth in a communication system.

In AM with carrier, from the calculation of efficiency, it is found that only 33.3% of energy is used and the remaining power is wasted by the carrier transmission along with the side bands.

In order to save the power in amplitude modulation, the carrier is suppressed, because it does not contain any useful information. This scheme is called the Double-Side-Band Suppressed-Carrier Amplitude Modulation (DSB-SC AM).

It contains only LSB and USB terms, resulting in a transmission bandwidth that is twice the bandwidth of the message signal.

Let the modulating signal be

$$V_{\mathrm{m}}(t) = V_{\mathrm{m}} \sin \omega_{\mathrm{m}} t \tag{2.39}$$

and the carrier signal be

$$V_{\mathrm{c}}(t) = V_{\mathrm{c}} \sin \omega_{\mathrm{c}} t \tag{2.40}$$

When multiplying both the carrier and modulating signals, the product which is obtained is the DSB-SC AM signal.

$$\begin{aligned} V(t)_{\mathrm{DSB-SC}} &= V_{\mathrm{m}} \sin \omega_{\mathrm{m}} t + V_{\mathrm{c}} \sin \omega_{\mathrm{c}} t \\ &= V_{\mathrm{m}} V_{\mathrm{c}} \sin \omega_{\mathrm{m}} t \sin \omega_{\mathrm{c}} t \\ &= \frac{V_{\mathrm{m}} V_{\mathrm{c}}}{2} \left[\cos(\omega_{\mathrm{c}} - \omega_{\mathrm{m}})t - \cos(\omega_{\mathrm{c}} + \omega_{\mathrm{m}})t \right] \end{aligned} \tag{2.41}$$

We know that

$$V(t)_{\mathrm{AM}} = V_{\mathrm{c}} \sin \omega_{\mathrm{c}} t + \frac{m_{\mathrm{a}} V_{\mathrm{c}}}{2} \cos(\omega_{\mathrm{c}} - \omega_{\mathrm{m}})t - \frac{m_{\mathrm{a}} V_{\mathrm{c}}}{2} \cos(\omega_{\mathrm{c}} + \omega_{\mathrm{m}})t \tag{2.42}$$

By comparing Equation (2.34) and Equation (2.35), the carrier term $V_{\mathrm{c}} \sin w_{\mathrm{c}} t$ is missing and only two side bands are present in Equation (2.35). Hence, it is said to be DSB-SC AM.

Figure 2.9 shows the frequency spectrum of DSB-SC AM, from which it is clear that the carrier term ω_{c} is suppressed. It contains only two side-band terms having the frequency of $(\omega_{\mathrm{c}} - \omega_{\mathrm{m}})$ and $(\omega_{\mathrm{c}} + \omega_{\mathrm{m}})$ Hence, this scheme is called DSB-SC AM. Figure 2.10 shows the graphical representation of DSB-SC AM. It exhibits the phase reversal at zero crossing.

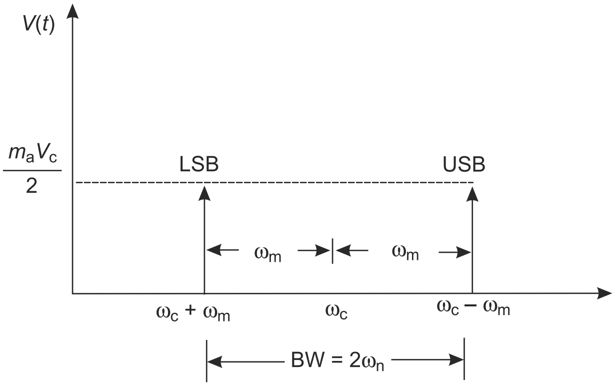


Fig. 2.9 Frequency Spectrum of DSB-SC AM

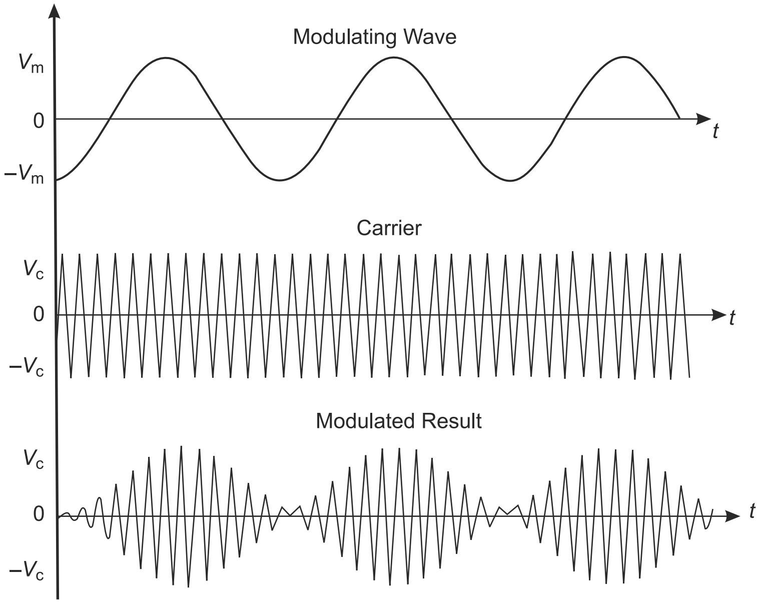


Fig. 2.10 Graphical representation of DSB-SC AM

2.6.1 Phasor Representation of DSB-SC AM

Let us assume that the carrier phase is a reference phasor and oriented in the horizontal direction as shown in Figure 2.11 by the dotted line since it is suppressed after modulation.

The USB term $\frac{m_a V_c}{2}\cos(\omega_c + \omega_m)t$ rotates at an angular frequency of ω_m in anticlockwise direction and the LSB term $\frac{m_a V_c}{2}\cos(\omega_c + \omega_m)t$ rotates at an angular frequency of ω_m in

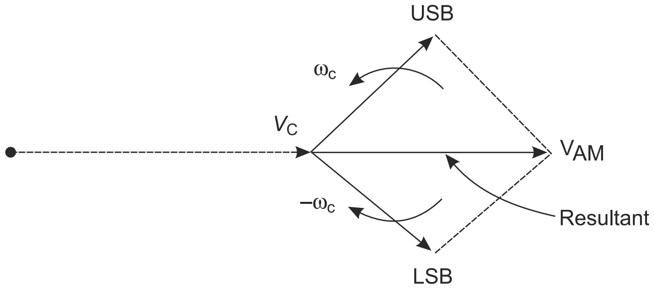


Fig. 2.11 Phasor representation of DSB-SC AM

clockwise direction. Hence, the resultant amplitude of the modulated wave at any point is the vector sum of the two side bands.

2.6.2 Power Calculation in DSB-SC AM

The total power transmitted in AM is

$$
\begin{aligned}
P_T &= P_C + P_{LSB} + P_{USB} \\
&= \frac{V_C^2}{2R} + \frac{m_a^2 V_C^2}{8R} + \frac{m_a^2 V_C^2}{8R} \\
&= \frac{V_C^2}{2R} + 2\left[\frac{m_a^2 V_C^2}{8R}\right] \\
&= \frac{V_C^2}{2R} + \frac{m_a^2 V_C^2}{4R} \\
&= \frac{V_C^2}{2R}\left[1 + \frac{m_a^2}{2}\right] \qquad (2.43)
\end{aligned}
$$

where $P_C = \frac{V_C^2}{2R}$

If the carrier is suppressed then the total power transmitted is

$$
\begin{aligned}
P_T' &= P_{LSB} + P_{USB} \qquad (2.44) \\
&= \frac{m_a^2 V_C^2}{8R} + \frac{m_a^2 V_C^2}{8R} = \frac{m_a^2 V_C^2}{4R} \\
&= \frac{V_C^2}{2R}\left[\frac{m_a^2}{2}\right] \qquad (2.45)
\end{aligned}
$$

$$\therefore\ P'_{\text{T}} = P_{\text{LSB}} + P_{\text{USB}}$$

$$= \frac{m_{\text{a}}^2}{2} \times P_{\text{C}} \qquad (2.46)$$

$$\text{Power saving} = \frac{P_{\text{T}} - P'_{\text{T}}}{P_{\text{T}}} \qquad (2.47)$$

$$= \frac{\left[1+\frac{m_{\text{a}}^2}{2}\right]P_{\text{C}} - \left[\frac{m_{\text{a}}^2}{2}.P_{\text{C}}\right]}{\left[1+\frac{m_{\text{a}}^2}{2}\right]P_{\text{C}}}$$

$$= \frac{P_{\text{C}} + \cancel{\frac{m_a^2}{2}P_{\text{C}}} - \cancel{\frac{m_{\text{a}}^2}{2}P_{\text{C}}}}{\left[1+\frac{m_{\text{a}}^2}{2}\right]P_{\text{C}}}$$

$$= \frac{\cancel{P_C}}{\left[1+\frac{m_{\text{a}}^2}{2}\right]\cancel{P_C}} \qquad (2.48)$$

$$\%\ \text{power saving} = \frac{2}{2+m_{\text{a}}^2} \times 100 \qquad (2.49)$$

If $m_{\text{a}} = 1,$

$$\%\ \text{power saving} = \frac{2}{2+1} \times 100$$

$$= \frac{2}{3} \times 100 = 66.7\% \qquad (2.50)$$

$\therefore$ 66.7% of power is saved in DSB-SC AM.

2.6.3 Advantage and Disadvantage of DSB-SC AM

Advantage of DSB-SC AM

1. Efficient in terms of power usage
2. Low power consumption
3. 100% modulation efficiency
4. Large bandwidth

Disadvantage of DSB-SC AM

1. Product detector required for demodulation of DSB signal which is quite expensive
2. Complex detection
3. Signal rarely used because the signal is difficult to recover at the receiver

2.6.4 Applications of DSB-SC AM

1. Used in analog TV systems to transmit the colour information
2. For transmitting stereo information in FM sound broadcast at VHF

2.7 SINGLE SIDE-BAND SUPPRESSED-CARRIER AMPLITUDE MODULATION (SSB-SC AM)

In AM with carrier, both the transmitting power and bandwidth are wasted. Hence, the DSB-SC AM scheme has been introduced in which power is saved by suppressing the carrier component, but the bandwidth remains the same.

Increase in the saving of power is possible by eliminating one side band in addition to the carrier component, because the USB and LSB are related by symmetry about the carrier frequency so either one side band is enough for transmitting as well as recovering the useful message. In addition to that, transmission bandwidth can be cut into half if one side band is suppressed along with the carrier. This scheme is known as Single-Side-Band Suppressed-Carrier Amplitude Modulation (SSB-SC AM).

The block diagram of SSB-SC AM is shown in Figure 2.12.

The SSB-SC AM can be obtained as follows:

In order to suppress one of the side bands, the input signal fed to the modulator 1 is 90° out of phase with that of the signal fed to the modulator 2.

Let $$V_1(t) = V_m \sin \omega_m t \,.\, V_c \sin \omega_c t \tag{2.51}$$

$$V_2(t) = V_m \sin (\omega_m t + 90°).\, V_c \sin (\omega_c t + 90°) \tag{2.52}$$

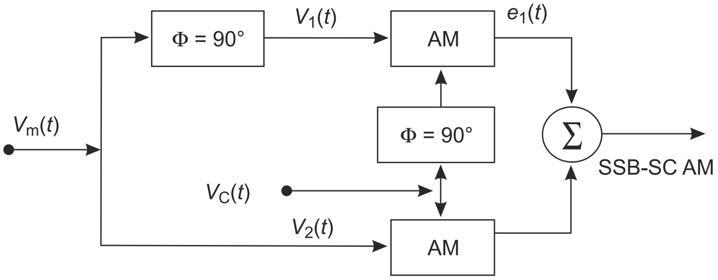


Fig. 2.12 Block diagram of SSB-SC AM

$$V_2(t) = V_{\mathrm{m}} \cos \omega_{\mathrm{m}} t.\ V_{\mathrm{c}} \cos \omega_{\mathrm{c}} t$$

$$\therefore\ V(t)_{\mathrm{SSB}} = V_1(t) + V_2(t) \tag{2.53}$$

$$= V_{\mathrm{m}}\ V_{\mathrm{c}}[\sin \omega_{\mathrm{m}} t.\ \sin \omega_{\mathrm{c}} t + \cos \omega_{\mathrm{m}} t.\ \cos \omega_{\mathrm{c}} t] \tag{2.54}$$

$$= \frac{V_{\mathrm{m}} V_{\mathrm{c}}}{2} \cos(\omega_{\mathrm{c}} - \omega_{\mathrm{m}})t \tag{2.55}$$

We know that for DSB-SC AM,

$$V(t)_{\mathrm{DSB}} = \frac{V_{\mathrm{m}} V_{\mathrm{c}}}{2}[\cos(\omega_{\mathrm{c}} - \omega_{\mathrm{m}})t - \cos(\omega_{\mathrm{c}} + \omega_{\mathrm{m}})t] \tag{Eq. 2.41}$$

By comparing Equation (2.55) with Equation (2.41), one of the side bands is suppressed. Hence, this scheme is known as SSB-SC AM.

The frequency spectrum of SSB-SC AM is shown in Figure 2.13.

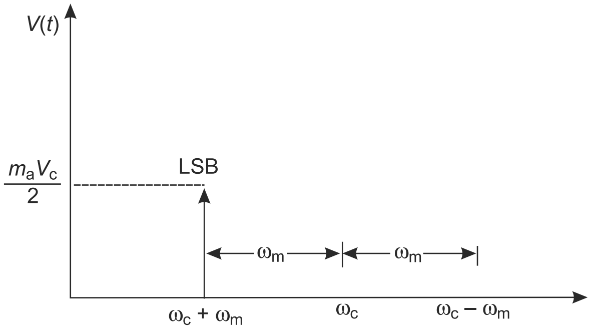


Fig. 2.13 Frequency spectrum of SSB-SC AM

It shows that only one side-band signal is present, the carrier and the other (upper) side-band signals are suppressed. Thus, the bandwidth required reduces from $2\omega_{\mathrm{m}}$ to ω_{m}. That means, the bandwidth requirement is reduced to half compared to AM and DSB-SC signals.

The graphical representation and phasor representation of SSB-SC AM system is shown in Figure 2.14 and Figure 2.15.

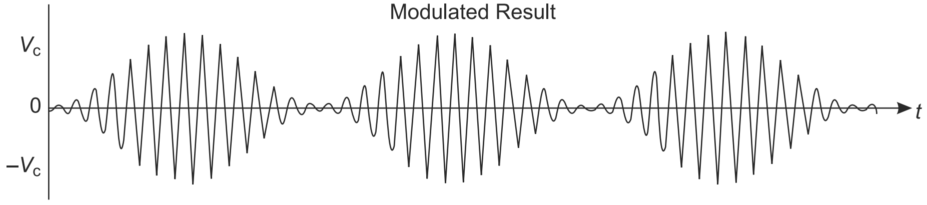


Fig. 2.14 Graphical representation of SSB-SC AM

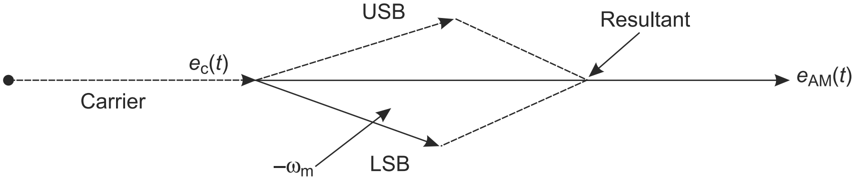


Fig. 2.15 Phasor representation of SSB-SC AM

2.7.1 Power Calculation in SSB-SC AM

Power in side bands $P_T'' = P_{SB} = \frac{1}{4} m_a^2 P_c$

Power saving with respect to AM with carrier = $\frac{P_T - P_T''}{P_T}$ (2.56)

$$= \frac{\left[1 + \frac{m_a^2}{2}\right] P_C - \left[\frac{m_a^2}{4} \cdot P_C\right]}{\left[1 + \frac{m_a^2}{2}\right] P_C}$$

$$= \frac{P_C + \frac{m_a^2}{2} P_C - \frac{m_a^2}{4} P_C}{\left[1 + \frac{m_a^2}{2}\right] P_C} = \frac{\left[1 + \frac{m_a^2}{4}\right]}{\left[1 + \frac{m_a^2}{2}\right]}$$

$$= \frac{4 + \frac{m_a^2}{4}}{2 + \frac{m_a^2}{2}}$$

$$= \frac{4 + m_a^2}{2 + 2m_a^2} \tag{2.57}$$

For $m_a = 1$,

% power saving = $\frac{5}{6} \times 100 = 83.3\%$ (2.58)

2.7.2 Saving of Power in SSB-SC AM with respect to DSB-SC AM

Power saving with respect to AM with carrier = $\frac{P_T - P_T''}{P_T}$ (2.59)

$$= \frac{\frac{1}{2}m_a^2 P_C - \frac{1}{4}m_a^2 P_C}{\frac{1}{2}m_a^2 P_C} = \frac{\frac{1}{4}m_a^2 P_C}{\frac{1}{2}m_a^2 P_C}$$

$$= \frac{1}{2} \times 100 = 50\% \tag{2.60}$$

From the above analysis, in AM with carrier, the total power is $\left[1 + \frac{m_a^2}{2}\right]$ times the carrier power.

If the carrier is suppressed and only the side bands are transmitted then 66.67% of power is saved. If in addition to carrier, one of the side bands is also suppressed, the power saving is 83.35 over AM with carrier.

EXAMPLE 2.15

A single side-band signal contains 1 kW. How much power is contained in the side bands and how much at the carrier frequency?

Solution

In a single side-band transmission, the carrier and one of the side bands have been eliminated. Therefore, all the transmitted power is transmitted at one of the side bands regardless of percent modulation. Thus,

$$P_{SB} = 1 \text{ kW}$$

$$P_C = 0 \text{ W}$$

EXAMPLE 2.16

An SSB transmission contains 10 kW. This transmission is to be replaced by a standard AM signal with the same power content. Determine the power content of the carrier and each of the side bands when the percent modulation is 80%.

Solution

The total power content of the new AM signal is to be the same as the total power content of the SSB signal.

$$P_T = P_{SSB} = 10 \text{ kW}$$

$$P_T = P_C + P_{LSB} + P_{USB}$$

$$P_T = P_C + \frac{m_a^2 P_C}{4} + \frac{m_a^2 P_C}{4}$$

$$10{,}000 = P_C + \frac{(0.8)^2\, P_C}{4} + \frac{(0.8)^2\, P_C}{4}$$

$$= P_C + \frac{0.64\, P_C}{4} = 1.32\, P_C$$

$$P_C = \frac{10{,}000}{1.32} = 7575.76$$

$$\therefore\ P_C = 7575.76 \text{ W}$$

The power content of the side bands is equal to the difference between the total power and the carrier power.

$$P_{LSB} + P_{USB} = 10{,}000 - 7575.76$$

$$= 2424.24$$

$$P_{LSB} = P_{USB} = \frac{2424.24}{2}$$

$$\therefore\ P_{LSB} = P_{USB} = 1212.12 \text{ W}$$

2.7.3 SSB-SC over AM with Carrier

The main reason SSB is superior to AM, and most other forms of modulation, is due to the following:

1. Since the carrier is not transmitted, there is a reduction by 50% of the transmitted power (–3 dBm). In AM with 100% modulation, half the power is comprised of the carrier; with the remaining half power in both side bands.
2. Because in SSB, only one side band is transmitted, there is a further reduction by 50% in transmitted power (–3 dBm (+) –3 dBm = –6 dBm).
3. Finally, because only one side band is received, the receiver's needed bandwidth is reduced by one half—thus effectively reducing the required power by the transmitter another 50% (–3 dBm (+) –3 dBm (+) –3 dBm = –9 dBm).

It is to be noted that if a receiver's bandwidth can be reduced by 50%, the needed transmitter power is also reduced by 50%, i.e. the receiver's Signal-to-Noise Ratio (SNR) is improved as the receiver bandwidth is reduced. This implies that the signal containing the information is not lost, which is the case in this instance.

2.7.4 Advantages and Disadvantages of SSB-SC AM

Advantages of SSB

1. SSB amplitude modulation is widely used by military or radio amateurs in high-frequency communication. It is because the bandwidth is the same as the bandwidth of modulating signals.

2. Occupies one half of the spectrum space
3. Efficient in terms of power usage
4. Less noise on the signal

Disadvantages of SSB

1. When no information or modulating signal is present, no RF signal is transmitted.
2. Most information signals transmitted by SSB are not pure sine waves.
3. A voice signal will create a complex SSB signal.

2.7.5 Applications of SSB-SC AM

1. Two-way radio communications
2. Frequency-division multiplexing
3. Up conversion in numerous telecommunication systems

2.8 VESTIGIAL SIDE-BAND AM (VSB AM)

Some of the modulating signals of very large bandwidth such as video signals, TV and high-speed data signals have very low-frequency components along with the rest of the signal. These components give rise to side bands, very close to the carrier frequency, which are difficult to remove by using filters. So it is not possible to fully suppress one complete side band in case of television signals. Similarly, the low video frequencies contain the most important information of the picture. If any of its side bands (LSB/USB) is suppressed, it would result in phase distortion at low frequencies. Therefore, a compromise has been made to suppress the part of the lower side band. Then, the radiated signal consists of full USB together with the carrier and vestige of the (partially suppressed) lower side band. This kind of modulation is known as vestigial side-band modulation (VSB).

A VSB AM system utilises the advantages of DSB-SC AM and SSB-SC AM and avoids their disadvantages. VSB transmission is similar to SSB transmission, in which one of the side bands is completely removed. In VSB transmission, however, the second side-band is not completely removed, but is filtered to remove all but the desired range of frequencies.

VSB signals are very easy to generate and their bandwidth is slightly greater than SSB-SC AM but less than DSB-SC AM.

Figure 2.16 shows the block diagram of VSB modulation and Figure 2.17 shows its frequency response.

In VSB, instead of rejecting one side band completely, a gradual cut of one side band is acceptable. The VSB filter in the block diagram gives the partial suppression of the transmission side band (VSB), and the neighbourhood of the carrier is exactly compensated by the partial transmission of the corresponding part of suppressed side band (LSB).

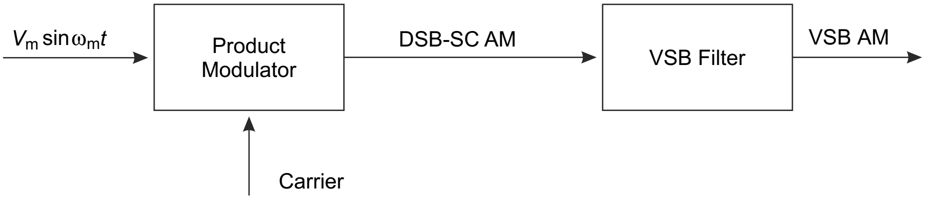


Fig. 2.16 Block diagram of VSB AM

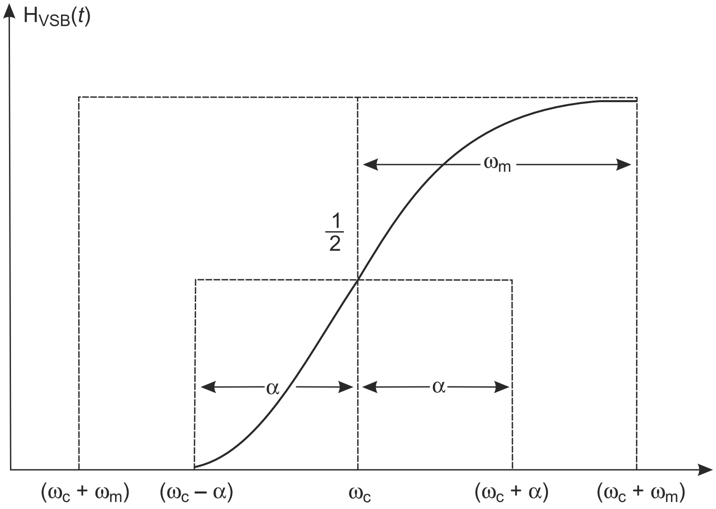


Fig. 2.17 Frequency spectrum of VSB AM

The VSB side-band filters have a transition width of 2α Hz and transmission width of $B_T = \omega_m + \alpha$ for α is 1< $\omega_{m.}$

The frequency transform of a VSB signal is

$$H_{\text{VSB}}(f) = H_{\text{DSB}}(f) - H\alpha(f) \tag{2.61}$$

where

$H_{\text{DSB}}(f)$ is magnitude transfer function of DSB system

$H_\alpha(f)$ is magnitude transfer function of VSB filter

$$V_{\text{VSB}}(t) = \underbrace{\frac{V_c}{2}\sin\omega_c t + \frac{V_m V_c}{2}\left[\cos(\omega_c - \omega_m)t - \cos(\omega_c + \omega_m)t\right]}_{\text{Response of AM}} - \underbrace{\frac{\left[V_m + V_\alpha(t)\right]V_c}{2}\cos\omega_c t}_{\text{Response of VSB filter}}$$

$$= \frac{V_c}{2}\sin\omega_c t + \left[1 + V_m(t)\right] - \frac{\gamma(t)V_c}{2}\cos\omega_c t \tag{2.62}$$

where $\gamma(t) = V_{\mathrm{m}}(t) + V_{\alpha}(t)$

If $\gamma(t) = 0$ then the circuit will be acting as AM modulator with carrier. If $\gamma(t) = V_{\mathrm{m}}(t)$ then the output will be SSB with carrier.

Advantages of VSB AM

1. Bandwidth of VSB is greater than SSB but less than DSB system.
2. Power transmission is greater than DSB but less than SSB system
3. No loss in frequency component and, hence, no phase distortion.

2.9 APPLICATIONS OF AM

1. The AM SSB is used in telephone systems, two-way radio and also in military communication.
2. The AM DSB is used in FM and TV broadcasting

2.10 GENERATION OF AM WAVES

The methods for generation of AM waves are classified into two major types:

1. Nonlinear modulators
2. Linear modulators

2.10.1 Nonlinear AM modulators

The relation between the amplitude of the modulating signal and the resulting depth of modulation is nonlinear in this type of modulators. In general, any device operated in nonlinear region of its output characteristic is capable of producing AM waves when the carrier and the modulating signals are fed at the input. The devices making use of nonlinear *V-I* characteristics are diodes, triode tubes, Bipolar Junction Transistors (BJTs) and Field Effect Transistors (FETs), which are also called **square-law modulators**. They are suitable for use at low operating voltages.

In such a nonlinear modulator, the output current flowing through the load is given by the power series

$$i = a_0 + a_1 e_1 + a_2 e_2^2 + \ldots \tag{2.63}$$

where $a_0 + a_1$, a_2, ..., etc, are constants and e_1 is the input voltage to the device.
Figure 2.18 shows a basic nonlinear modulator.

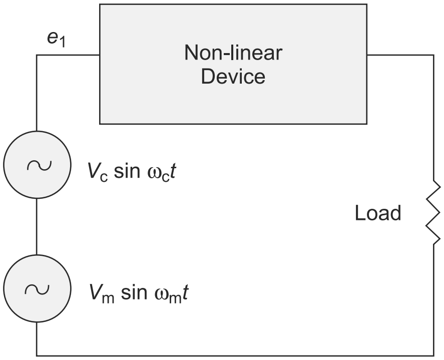


Fig. 2.18 Basic nonlinear modulator

By considering the modulator circuit shown above,

$$e_1 = V_c \sin \omega_c t + V_m \sin \omega_m t \tag{2.64}$$

$$\begin{aligned} \therefore\ i &= a_0 + a_1(V_c \sin \omega_m t + V_m \sin \omega_m t) + a_2(V_c \sin \omega_c t + V_m \sin \omega_m t)^2 + \ldots \\ &= a_0 + a_1 V_c \sin \omega_c t + a_1 V_m \sin \omega_m t \\ &\quad + a_2 V_c^2 \sin^2 \omega_c t + a_2 V_m^2 \sin^2 \omega_m t \\ &\quad + 2a_2 V_c V_m \sin \omega_c t \sin \omega_m t \end{aligned}$$

$$\begin{aligned} &= a_0 + a_1 V_c \sin \omega_c t + a_1 V_m \sin \omega_m t \\ &\quad + a_2 V_c^2 \sin^2 \omega_c t + a_2 V_m^2 \sin^2 \omega_m t \\ &\quad + 2a_2 V_c V_m (\cos (\omega_c - \omega_m)t - \cos (\omega_c + \omega_m)t) \end{aligned} \tag{2.65}$$

The last term of Equation (2.66) gives the upper and lower side bands while the second term gives the carrier. If the load is a resonant circuit, side bands and carrier may be selected giving the AM output. As a_1 is considerably larger than a_2, the depth of modulation that is available without distortion is low. Also note that the circuit efficiency is quite low because a sufficient number of components are filtered out from the plate current. Thus, this type of circuit is essentially a low-level circuit. A common triode-tube circuit using the square-law modulation technique is shown in Figure 2.19.

Figure 2.20 shows square-law modulator using diode. The operation of the diode square-law modulator is as follows. Due to the nonlinearity of transfer characteristics of the diode, the magnitude of the carrier component is greater during the positive half cycle of the modulating voltage and lesser in magnitude during the negative half cycle of the modulating signal.

The diode modulator does not provide amplification and a single diode is unable to balance out the undesired frequency completely. These limitations can be eliminated by using

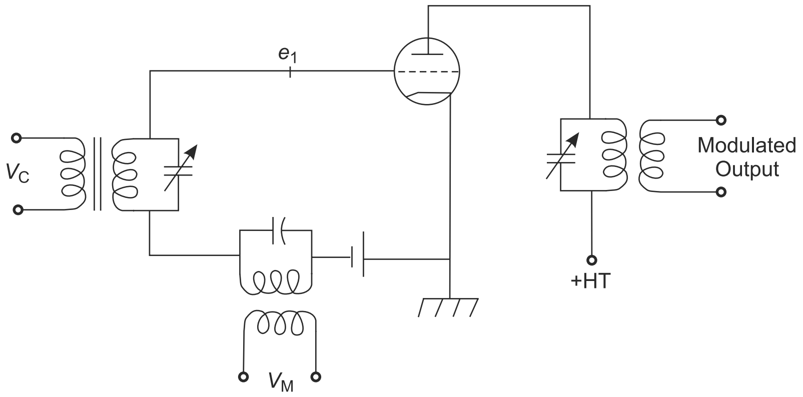


Fig. 2.19 Square-law modulator using triode tube

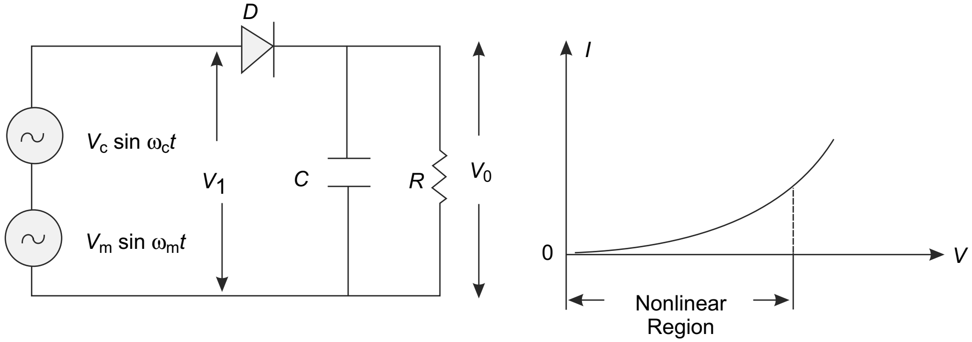


Fig. 2.20 Square-law modulator using diode

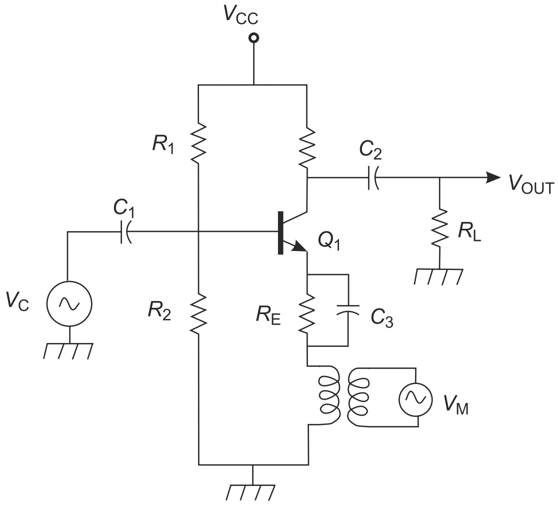


Fig. 2.21 Low-level modulator using transistor

amplifying devices like transistor and FET in a balanced mode. Figure 2.21 shows the low-level modulator using a transistor.

Advantages of a Low-level Modulator

1. Less modulating signal power is required to achieve high percentage of modulation
2. The advantage of using a linear RF amplifier is that the smaller early stages can be modulated, which only requires a small audio amplifier to drive the modulator.

Disadvantages of a Low-level Modulator

1. The great disadvantage of this system is that the amplifier chain is less efficient**,** because it has to be linear to preserve the modulation.

2.10.2 Linear AM Modulators

For radio transmitters requiring good linearity and high power output, a linear modulator circuit is commonly employed. Figure 2.22 shows the AM modulator using transistor in which the modulation takes place in the collector, the output element of the transistor. Therefore, if this is the final active stage of the transmitter, it is called a high-level modulator.

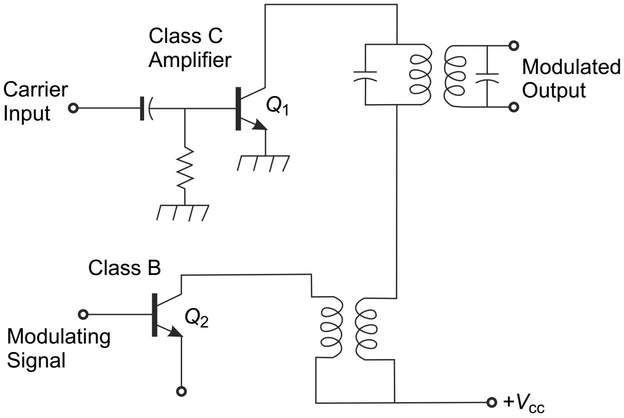


Fig. 2.22 AM modulator using transistor

From the above figure, the output stage of the transmitter is a high-power Class C amplifier. Class C amplifiers conduct for only a portion of the positive half cycle of their input signal. The collector current pulses make the tuned circuit oscillate at the desired output frequency. Then, the tuned circuit reproduces the negative portion of the carrier signal.

The modulator is a linear power amplifier that takes the low-level modulating signal and amplifies it to a high power level. The modulating output signal is coupled through a modulation transformer T_1 to the Class C amplifier. The secondary winding of the modulation transformer is connected in series with the collector supply voltage +V_{CC} of the Class C amplifier.

With no modulating signal as input, there is no modulation voltage, i.e. zero modulation voltage across the secondary of T_1, the collector supply voltage is applied directly to the Class C amplifier, and the output is a steady sinusoidal wave.

When there is a modulating signal present, the ac voltage of the modulating signal across the secondary of the modulation transformer is added to and subtracted from the dc collector supply voltage. This varying supply voltage is then applied to the Class C amplifier, causing the amplitude of the current pulses through the transistor Q_1 to vary. Due to this, the amplitude of the carrier sine wave varies in accordance with the modulated signal.

When the modulation signal goes positive, it adds to the collector supply voltage, thereby increasing its value and causing higher current pulses and a higher amplitude carrier. When the modulating signal goes negative, it subtracts from the collector supply voltage, decreasing it.

For 100% modulation, the peak of the modulating signal across the secondary of T_1 must be equal to the supply voltage. When the positive peak occurs, the voltage applied to the collector is twice the collector supply voltage. When the modulating signal goes negative, it subtracts from the collector supply voltage. When the negative peak is equal to the supply voltage, the effective voltage applied to the collector of Q_1 is zero, producing zero carrier output.

One disadvantage of collector modulators is the need for a modulation transformer that connects the audio amplifier to the Class C amplifier in the transmitter. For higher power, the transformer is larger and more expensive. For very high-power applications, the transformer is eliminated and the modulation is accomplished at a lower level with a suitable modulator circuit. The resulting AM signal is amplified by a high-power linear amplifier.

To replace the transformer, it is preferred to use a transistorised version of a collector modulator, which replaces the transformer with an emitter follower as shown in Figure 2.23.

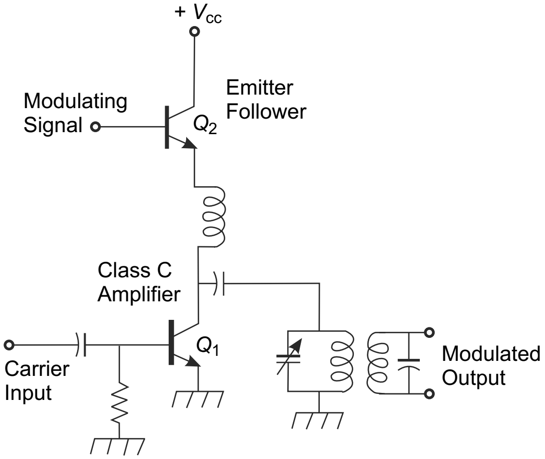


Fig. 2.23 High-level modulator with emitter follower

From the above figure, the modulating signal is applied to the emitter follower Q_2, which is an audio amplifier. Since the emitter follower is connected in series with the collector supply voltage +V_{CC}, this causes the amplified modulating signal to vary the collector supply voltage to the Class C amplifier. Q_2 varies the supply voltage to the transistor Q_1.

If the modulating voltage goes positive, the supply voltage to Q_1 increases and thus the carrier amplitude also increases in proportion to the modulating signal. If the modulating signal goes negative, the supply voltage to Q_1 decreases, thereby decreasing the carrier amplitude in proportion to the modulating signal.

Advantages of a High-level Modulator

1. Eliminates the need for a large, heavy and expensive transformer
2. Improves the frequency response
3. Best suitable for power levels below about 100 W

Disadvantages of a High-level Modulator

1. The emitter follower must dissipate more power as a Class C amplifier.
2. Very inefficient.

2.10.3 Generation of Double Side-Band Suppressed-Carrier AM (DSB-SC AM)

In AM, the carrier of the modulated wave does not contain any information and a considerable part of the total power in the wave goes waste. If the carrier is suppressed and the side bands are only transmitted, a considerable amount of power saving is done.

However, the suppression of the carrier from a transmitted wave leads to a complicated design at the receiving end. As a result, the receiver becomes costly. Therefore, it is necessary to employ DSB-SC AM for broadcast service and utilise the suppressed carrier service.

There are two ways of generating DSB-SC AM.

1. Balanced modulators
2. Ring modulators

1. Balanced Modulators

The commonly used circuit for the suppression of the carrier is termed the balanced modulator, which employs devices with nonlinear characteristic. Figure 2.24 shows the push-pull balance modulator.

From the figure, the carrier is applied to the centre-tapped transformer and is in phase at the two bases of two transistors. The modulated signal is antiphase at the two bases of transistors. The output currents are given as follows:

$$i_1 = K(1 + m_a \sin \omega_m t) \sin \omega_c t$$

$$i_2 = K(1 - m_a \sin \omega_m t) \sin \omega_c t$$

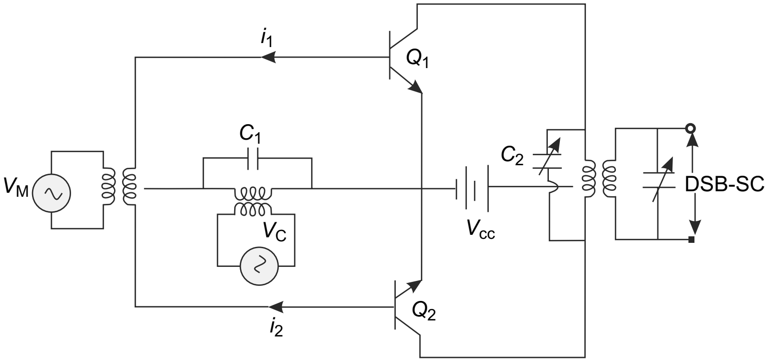


Fig. 2.24 Balanced modulator

where K is the constant of proportionality.

The output developing across the secondary of the output transformer is proportional to the difference of the currents i_1 and i_2.

$$|i_1 - i_2| = 2Km_a \sin \omega_m t \sin \omega_c t$$

$$= Km_a [\cos (\omega_c - \omega_m)t - \cos (\omega_c + \omega_m)t]$$

From the above equation, it is noted that the output contains only the two side bands and there is no carrier in the wave. The carrier is suppressed. The tank circuit is tuned to the carrier frequency and it responds to a band of frequencies centred at ω_c. Thus, a DSB-SC wave is generated at the output side. For complete suppression of the carrier signal, a band pass filter is required at the output of the circuit to remove the undesired frequency terms.

Advantages

1. Suppression of carrier results in considerably amount of power saving.
2. Used in carrier-current telephony system, in which one side band is suppressed out to reduce the bandwidth of the channel required for transmission.
3. Security is maintained.
4. Increases the overall efficiency.

2. Ring Modulators

This is another type of modulator used to produce DSB-SC AM waves. Figure 2.25 shows the circuit arrangement of a ring modulator. It employs four diodes as nonlinear devices and the carrier signal is connected between the centre taps of the input and output transformers.

By considering that there is no carrier and the modulating signal is present, diodes D_1 and D_2 or D_3 and D_4 will conduct depending on the signal polarity and will provide an effective short circuit, thereby restricting the signal from reaching the output. When the carrier alone is present, the flow of current in the two halves of the output transformer is equal and opposite and no output can be developed across the output.

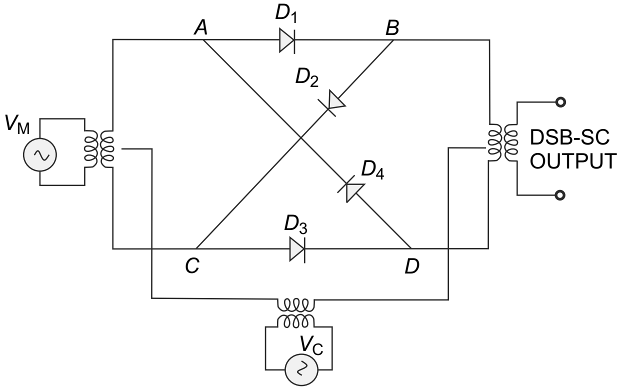


Fig. 2.25 Ring modulator

By considering both the carrier and the modulating signals to be present, diodes D_1 and D_3 conduct during the positive half cycle of the carrier, while diodes D_2 and D_4 do not conduct. During the negative half cycle of the carrier, diodes D_2 and D_4 conduct and diodes D_1 and D_3 do not conduct.

When both the carrier and the modulating signals are present, the resultant potential in one half of the output transformer becomes larger than the other and output is obtained.

Consider the modulating signal V_m and carrier signal V_c such that

$$V_m(t) = V_m \sin \omega_m t$$

$$V_c(t) = V_c \sin \omega_c t$$

The output voltage equals the product of two signals, which is given as

$$V_0(t) = V_m V_c \sin \omega_c t \sin \omega_m t$$

$$= \frac{V_m V}{2}\left[\cos(\omega_c - \omega_m)t - \cos(\omega_c + \omega_m)t\right]$$

From the above equation, it is noted that the output is free from the carrier and other terms, and it contains only side-band terms.

2.10.4 Generation of Single Side-Band Suppressed-Carrier AM (SSB-SC AM)

A single side band is sufficient to convey information to the distant receiver. This reduces the power and bandwidth requirements considerably. Single side-band suppressed carrier AM waves can be generated in two ways:

1. Filter method
2. Phase-shift method

1. Filter Method

A simple arrangement to produce SSB-SC signals is the use of a filter for filtering out one of the side bands leaving at the output of the other side band alone. The block diagram of the filter method is shown in Figure 2.26(a).

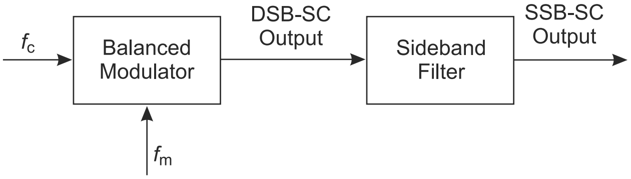


Fig. 2.26(a) Filter method of SSB-SC generation

This method involves generation of DSB-SC signals followed by extraction of the desired side band using the appropriate filter. This method can be used to generate SSB-SC waves. The baseband is restricted and appropriately related to the carrier frequency. Under this condition, the desired side band is selected by an appropriate filter.

There are two methods of side-band selection. Many transmitters consist of two filters, one that will pass the upper side band and the second that will pass the lower side band. A switch is used to select the desired side band. Figure 2.26(b) shows the filter method with selection of lower and upper side bands.

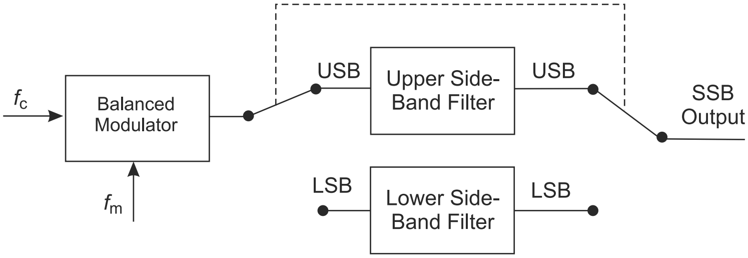


Fig. 2.26(b) Selection of side bands with two filters

With the modulating signal and carrier signal as inputs to the balanced modulator, the outputs from the lower side-band filter and upper side-band filter are as follows:

Lower side-band $f_{LSB} = f_c - f_m$

Upper side-band $f_{USB} = f_c + f_m$

Those outputs are based on the selection of lower or upper side-band filters by using a selection switch.

Since it is difficult to design a filter in an effort to eliminate the unwanted side band, such a filter will introduce attenuation in the wanted side band also. Increasing the bandwidth may result in passing some of the unwanted side bands to the output.

For satisfactory performance of the system, the following two requirements have to be satisfied.

1. The passband of the filter should be same as that of the desired side band.
2. The separation region between passband and stopband should not exceed twice the maximum frequency component present in the baseband.
3. The unwanted side band and whose nearest frequency component is separated from the desired side band by twice the lowest frequency component of the modulating signal.

Advantages

1. Small size
2. Good bandpass characteristics
3. Good attenuation characteristics
4. Adequate frequency limit
5. Use of crystal filters will make the method cheaper
6. Flat and wider bandwidth

Disadvantages

1. The maximum operating frequencies are lower than the transmitting frequency.
2. Transmission of very high (like 10 MHz) frequency and very low (like 50 Hz) frequency through the filter method is not possible because the filter is not able to follow such a steep response.

2. Phase-Shift Method

This method employs two balanced modulators as shown in Figure 2.27. The balanced modulator 1 is given the modulating signal and carrier signal in the usual manner. The balanced modulator 2 is given these signals after a phase shift of 90°.

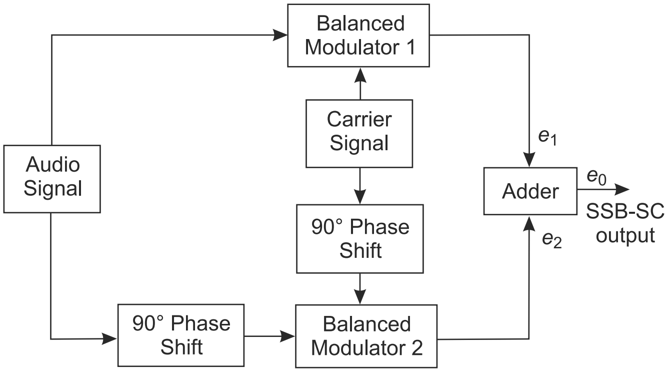


Fig. 2.27 Phase shift method of SSB-SC generation

Consider the modulating signal V_m and carrier signal V_c for the balanced modulator 1 such that

$$V_m(t) = V_m \sin \omega_m t$$

$$V_c(t) = V_c \sin \omega_c t$$

Then the output of the balanced modulator 1 is given as follows:

$$V_1 = \frac{V_m . V_c}{2}\left[\cos(\omega_c - \omega_m)t + \cos(\omega_c + \omega_m)t\right]$$

Similarly, the modulating signal V_m and carrier signal V_c for the balanced modulator 2 are

$$V_m(t) = V_m \cos \omega_m t$$

$$V_c(t) = V_c \cos \omega_c t$$

Then the output of the balanced modulator 2 is given as follows:

$$V_2 = \frac{V_m . V_c}{2}\left[\cos(\omega_c - \omega_m)t + \cos(\omega_c + \omega_m)t\right]$$

When two voltages are added in the adder circuit, the upper side band vanishes and the lower side band alone is the result of the adder circuit. This is the SSB-SC output from the phase-shift method.

Advantages

1. It can switch from one side band to another, i.e. LSB or USB
2. It is able to generate SSB at any frequency, thereby eliminating the need of frequency mixer.
3. Low audio modulating frequency may also be used.

Disadvantages

1. Need of critical phase-shift network
2. Needs two balanced modulators which must both give exactly the same output
3. Complex system

Applications

1. Police wireless communication
2. SSB telegraphy system
3. Point-to-point radio telephone communication
4. VHF and UHF communication systems

2.10.5 Generation of Vestigial Side-Band AM (VSB AM)

AM vestigial side band is a form of amplitude modulation in which the carrier and one complete side band are transmitted, but only part of the second side band is transmitted.

The carrier is transmitted at full power. In VSB, the lower modulating signal frequencies are transmitted double side band and the higher modulating signal frequencies are transmitted single side band. Consequently, the lower frequencies can appreciate the benefit of 100% modulation, whereas the higher frequencies cannot achieve more than the effect of 50% modulation. Consequently, the low-frequency modulating signals are emphasised and produce large-amplitude signals in the demodulator than the high frequencies.

A VSB AM can be generated by passing a DSB-SC signal through an appropriate filter as shown in Figure 2.28. The system is similar to DSB-SC except that the filter characteristics should appropriate.

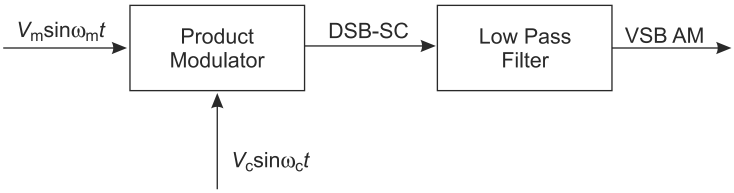


Fig. 2.28 VSB-AM modulator

2.11 AM TRANSMITTERS

2.11.1 Basic Radio Transmission System

A radio transmitter is a device that transmits information by means of radio waves. The signal is translated in terms of a high-frequency wave commonly termed carrier and the process of translation into high frequency is termed **modulation**. All radio transmitters use one form of modulation or the other for transmission of intelligence.

All radio-transmitting systems must have a section for generation of high-frequency carrier wave, a section for converting information into electrical impulses and amplifying them to the required level, a section for modulating the carrier with signal intelligence, amplification stages for increasing the level of the modulated wave to the desired power and antenna system for transmitting these signals into free space. Figure 2.29 shows a basic radio-transmission system.

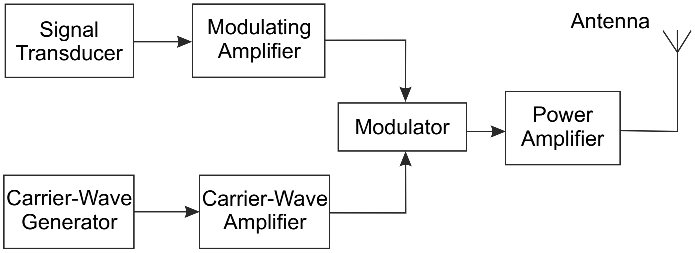


Fig. 2.29 A basic radio-transmission system

Names of the transmitters are depending on the type of signal to be transmitted, type of modulation employed, the carrier frequency used or the type of radio waves radiated by the system.

1. Depending on the type of signals used such as speech or code signal or a picture signal, a transmitter may be termed broadcast transmitter, telephony transmitter or telegraphy transmitter.
2. Depending on the modulation process employed, it may be termed AM or FE transmitter.
3. Depending on the frequency of the carrier employed, it may be termed medium-wave, short-wave, VHF, UHF or microwave transmitter.
4. A transmitter may be termed long-distance transmitter or a line- of-sight transmitter depending on whether the transmission is by sky waves or space waves.

2.11.2 Modulation Systems in AM Transmitters

AM transmitters are generally used for radio broadcasts over long, medium or short waves, point-to-point communication systems using radio telephony/telegraphy signals over short waves or VHF waves and picture signal transmission over the VHF or UHF ranges. The depth of modulation is directly proportional to the magnitude of the modulation signal.

The modulation may take place at an early stage of the transmitter where the carrier level is low. Such a system would require a modulating signal of lower magnitude. For this reason, this type of modulating circuit is termed **low-level modulation**. Since the power contained by modulated waves produced by a low-level modulation circuit is quite low, it is necessary to provide one or more stages of power amplification to increase the power level of the modulated wave to the requisite level. Since these power-amplifier stages are required to amplify the carrier as well as the side bands equally, they must possess sufficient bandwidth to accommodate these frequencies. If not, side-band elimination will occur. Such a circuit operated with low-power gain and has low efficiency. Class B linear RF power amplifiers are usually employed for this purpose. A low-level modulation system is shown in Figure 2.30.

When a large modulated power is to be transmitted, the system described above is found to be quite unsuitable because of low efficiency. In such systems, the RF carrier and the modulating signal are amplified to the desired levels before modulation takes place. After

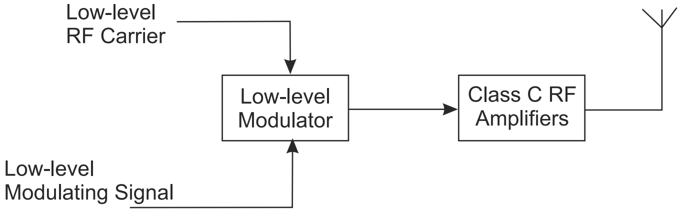


Fig. 2.30 A low-level modulation system

modulation, which is usually carried out by using a Class C plate-modulated amplifier, the output is fed to the antenna without further amplification. Such a system is termed **high-level modulation**. This system has more power-conversion efficiency than the low-level modulation system. A high-level modulation system is shown in Figure 2.31.

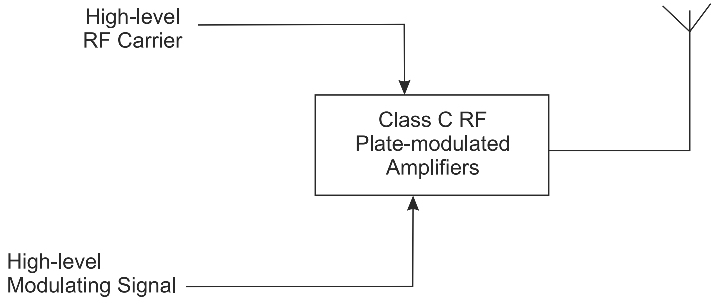


Fig. 2.31 A high-level modulation system

2.11.3 AM Broadcast Transmitters

AM transmitters are used in large numbers for transmission of music, speech, vice or light entertainment programmes for the general public. These transmitters have power output in the range of 1 kW to 100 kW or more and operate on long, medium or short waves.

The AM transmitters usually consist of four stages in the RF chain and two stages in the AF chain excluding the plate-modulated Class C power amplifier which is common to both the chains. The modulated output is fed to the antenna for radiation. Broadcast transmitters use high-level modulation because a large amount of power is dissipated in the power amplification stages. Figure 2.32 shows the block diagram of an AM transmitter from which the functioning of each block is as follows:

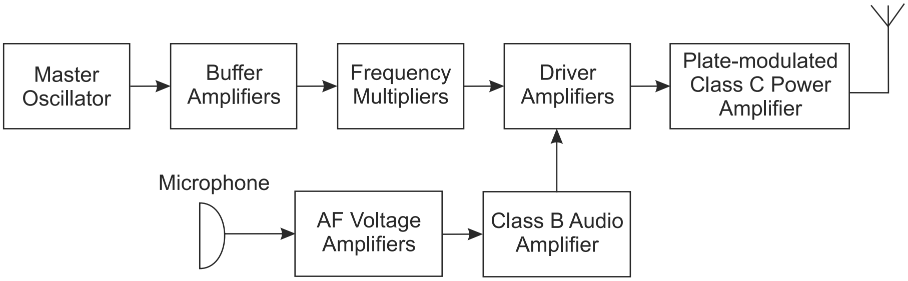


Fig. 2.32 Block diagram of an AM transmitter

1. Master Oscillator

The master oscillator generates high-frequency waves which are subsequently used as the carrier. Any oscillator can be used for this purpose but it should fulfill the following requirements.

(a) The circuit must provide a carrier of specified frequency. The master oscillator frequency can be adjusted to any desired value by suitable selection of frequency-determining components in the tank circuit of the master oscillator. This generates only the subharmonics of the final carrier frequency and the required frequency is brought with the help of harmonic generators.

(b) Frequency should be drift-free. A slow variation of frequency with time is termed **frequency drift**. Frequency drift in an oscillator takes place due to variation in tank-circuit inductance and capacitance with temperature variations. In addition, inter-electrode capacitances also vary with temperature. Changes in the tank-circuit inductance or capacitance due to temperature variations because of device heating while operating can be avoided to a large extent by enclosing the components in the constant-temperature channels.

Drift in frequency due to changes in inter-electrode capacitance can be minimised by keeping the value of frequency-determining components larger in comparison with the inter-electrode capacitance. To meet this requirement, the frequency of the master oscillator must be kept low and the desired carrier frequency signal is then obtained by the use of frequency-multiplier stages.

In addition to these slow frequency drifts, abrupt frequency changes may take place due to abrupt changes in load. Such a frequency change is termed **frequency scintillation** and it can be avoided by operating the master oscillator at lighter loads.

(c) Master oscillator frequency should be easily adjustable. In most transmitters, the master oscillators are operated as crystal-controlled oscillators in order to achieve a high degree of frequency stability. The frequency of an oscillator can be changed only by changing the crystal. In transmitters employing *LC* circuits in the master oscillator, the frequency can be readily adjusted by variation of load.

(d) Changes in supply voltages. Changes in supply voltages produce changes in the operating currents resulting in a change in the master-oscillator frequency, this type of frequency change can be avoded by using a high Q tank circuit.

2. Buffer Amplifier

The master oscillator is followed by a buffer amplifier. For high stability, the load of the oscillator should be light and constant. When the buffer amplifier is absent then the master oscillator will be directly connected to the power amplifier which would draw grid current and hence draw power from the master oscillator resulting in frequency variations.

To isolate the oscillator from the power amplifier stages and minimise the changes in the oscillator frequency due to variation in coupling and antenna loading, a buffer amplifier is used in between the oscillator and an intermediate amplifier.

3. Frequency Multipliers

When the transmitter has to be operated at a high frequency band such as (3–30) MHz or in a VHF band such as 300 MHz to 3 GHz, frequency multipliers are used to obtain the required output frequency. This is done due to frequency stability of oscillators becoming poor as their frequency is increased. Thus, frequency of an oscillator is increased with the help of multiplier stages.

4. Power Amplifier

The RF voltage generated by the master oscillator has usually very small power. This power level is raised to the required value by a chain of amplifiers. The number of stages depends on the power required to draw the final power amplifier. Class C amplifiers having high collector efficiency of the order of 75% and are used to deliver an appreciable amount of power to the succeeding stages. For noiseless transmission, the pick-up device would be a microphone and the audio amplifier. For a television transmitter, the pick-up device is a camera and so a video amplifier is required.

5. Modulated Amplifiers

This is generally a push-pull Class B or Class C tuned amplifier and the modulating signal is fed into the conductor of the amplifier. Generally, the plate-modulator method is used in a high-power radio broadcast and radio-telephone transmitters. For low-power transistorised radio transmitters, collector, base or emitter modulation may be used.

6. Antenna Matching Network

The output of the final amplifier is fed to an impedance-matching network which includes the tank circuit in the collector of the final amplifier. The Q of this circuit is kept low so as to pass the side bands without amplitude or frequency distortion.

EXAMPLE 2.17

A transmitter supplies 10 kW to the antenna when unmodulated. Find the total power radiated when modulated to 30%.

Solution

$$\text{Total power } P_\mathrm{T} = P_\mathrm{c}\left(1+\frac{m_\mathrm{a}^2}{2}\right)$$

$$P_\mathrm{T} = 10\times 10^3\left(1+\frac{(0.3)^2}{2}\right)$$

$$P_\mathrm{T} = 10.45 \text{ kW}$$

EXAMPLE 2.18

The rms values of the antenna currents before and after modulation are 10 A and 12 A respectively. Determine the modulation factor.

Solution

$$I_T = I_0\sqrt{1+\frac{m_a^2}{2}}$$

$$12 = 10\sqrt{1+\frac{m_a^2}{2}}$$

$$\therefore\ m_a = 0.938$$

EXAMPLE 2.19

The antenna current of an AM transmitter is 10 A when only carrier is sent, but increases to 10.93 A when the carrier is modulated. Find the percent modulation. Determine the antenna current and the depth of modulation.

Solution

$$\frac{I_T}{I_C} = \sqrt{1+\frac{m_a^2}{2}}$$

$$m_a = \sqrt{2\left[\frac{I_T}{I_C}\right]^2 - 1}$$

$$m_a = \sqrt{2\left[\frac{10}{10.93}\right]^2 - 1} = 0.821$$

$$I_T = I_C\sqrt{1+\frac{m_a^2}{2}}$$

$$I_T = 10.93\sqrt{1+\frac{(0.821)^2}{2}}$$

$$I_T = 12.63\text{ A}$$

EXAMPLE 2.20

An AM transmitter radiates 10 kW with carrier unmodulated and 11.5 kW, when the carrier is modulated. Calculate the modulation factor if any sine wave, corresponding to 40% modulation, is transmitted simultaneously. Determine the total power.

Solution

$$\frac{P_T}{P_C} = 1 + \frac{m_a^2}{2}$$

$$\frac{m_a^2}{2} = \frac{P_T}{P_C} - 1 = 0.15$$

$$m_a = 0.547$$

If the two signals are simultaneously modulated,

$$m_T = \sqrt{m_1^2 + m_2^2}$$

$$m_T = \sqrt{0.547^2 + 0.4^2}$$

$$m_T = \sqrt{0.46} = 0.678$$

$$P_T = P_c\left(1 + \frac{m_T^2}{2}\right)$$

$$P_T = 10 \times 10^3 \left(1 + \frac{(0.678)^2}{2}\right)$$

$$P_T = 12.298 \text{ kW}$$

2.11.4 SSB-AM Transmitters

Single side-band AM transmission consists of transmitting only one side band with the suppression of the other side band along with the carrier and it is preferred for AM radio telephony. The major advantages of SSB AM transmitters are as follows:

1. From SSB AM transmitters, there is an improvement in signal-to-noise ratio of 9 to 12 dB over the Double Side-Band (DSB) AM system.
2. In comparison between DSB AM, SSB AM occupies only half the frequency band as compared to AM-DSB system, which improves signal-to-noise ratio and also allows two channels to be transmitted within the same frequency band.
3. Due to suppression of the carrier, it provides a considerable reduction in the operation and maintenance cost of the overall transmitter.
4. SSB signals cannot be received by ordinary broadcast receivers. Thus, a certain amount of privacy is automatically included in SSB system.

The main drawback of the SSB AM system is that the transmitter and receiver equipments are costlier.

Figure 2.33 shows the block diagram of an SSB AM transmitter.

Audio signal with a frequency of (0–3) kHz is fed to a balanced modulator which is also fed with a carrier of 100 kHz. The carrier is suppressed in the balanced modulator and the two side bands are passed to the output. The lower side band is filtered out while the upper

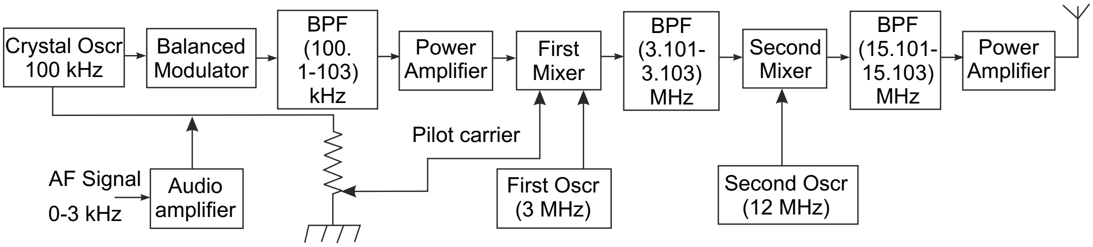


Fig. 2.33 Block diagram of SSB AM transmitter

side band (100.1–103) kHz is fed to the first mixer after amplification at the power-amplifier stage. The sum and difference frequencies are produced at the mixer output of which sum frequencies are selected by the bandpass filter to the second mixer.

The second mixer mixes the signal frequencies (3.101–3.103) MHz with the second oscillator output at 12 MHz and produces components at different frequencies of which the bandpass filter selects only (15.101–15.103) MHz. This signal is amplified in the final amplifier output. Another feature of SSB transmission is that the carrier is not totally absent but is transmitted at a considerably reduced level, which helps in saving considerable amount of power and simplifies the receiver equipment design.

EXAMPLE 2.21

An SSB transmitter using the filter method operates at a frequency of 4.5 MHz. The voice frequency range is 250 to 3000 Hz. Calculate the upper and lower side-band ranges. What should be the approximate centre frequency of a BPF to select the LSB?

Solution

(i) *For upper side-band:*

Lower limit $f_{\text{LL}} = f_c + 250$

$= 4.5 \times 10^6 + 250 = 45{,}00{,}250$ Hz

Upper side band $f_{\text{UL}} = f_c + 3000$

$= 4.5 \times 10^6 + 3000 = 45{,}03{,}000$ Hz

Range of upper side band is from 45,00,250 to 45,03,000 Hz.

(ii) *For lower side band:*

Lower limit $f_{\text{LL}} = f_c - 250$

$= 4.5 \times 10^6 - 250 = 44{,}99{,}750$ Hz

Upper side band $f_{UL} = f_c - 3000$

$$= 4.5 \times 10^6 - 3000 = 44{,}97{,}000 \text{ Hz}$$

Range of upper side band is from 44,99,750 to 44,97,000 Hz.

Summary

In general, AM is very simple. But, it is not the most efficient modulation to use, both in terms of the amount of spectrum it takes up, and the way in which it uses the power that is transmitted. This is the reason why it is not widely used these days either for broadcasting and or two-way radio communication. Even the long, medium and short-wave broadcasts will ultimately change because of the fact that amplitude modulation, AM, is subject to much higher levels of noise than are other modes.

In order to overcome the wastage of power in AM, the best method is to suppress the carrier. When the carrier is suppressed, both the upper and lower side bands are left, resulting Double Side-Band Suppressed-Carrier (DSB SC). It is also not necessary to use both the side bands for transmitting the desired message and so one of the side bands can be suppressed, resulting in Single Side-Band-Suppressed-Carrier (SSB SC).

Both DSB-SC and SSB-SC modulation methods are now most popular in communication. Their applications include two-way radio communications, transmission of stereo information in FM sound broadcast at VHF, marine applications and frequency division multiplexing.

REVIEW QUESTIONS

PART-A

1. State the needs of modulation.
2. Define modulation.
3. What are the types of modulation?
4. Define amplitude modulation.
5. Draw the frequency spectrum of an AM.
6. Draw the phasor representation of an AM wave
7. Define modulation index.

8. Give the equation to calculate the power in AM.
9. Give the relationship between the carrier power and the side-band power.
10. What are the three different degrees of freedom?
11. Draw the graphical representation of under-modulation in an AM wave.
12. Draw the graphical representation of over-modulation in an AM wave.
13. Draw the graphical representation of critical modulation in an AM wave.
14. State the drawbacks of AM.
15. What is DSB SC AM?
16. Draw the frequency spectrum of DSB SC AM.
17. Draw the phasor representation of DSB SC AM.
18. What is the amount of power saving in DSB SC AM?
19. List the advantages and disadvantages of DSB SC AM.
20. List the applications of DSB SC AM.
21. Define Single Side-Band Suppressed-Carrier (SSB SC).
22. State the advantages and disadvantages of SSB SC.
23. What is the amount of power saving in SSB SC AM?
24. What is the saving of power in SSB SC AM with respect to DSB SC AM?
25. What are the applications of SSB SC AM?
26. Compare over SSB SC AM over AM with carrier.
27. What is vestigial side-band modulation?
28. Mention the advantages of VSB AM.

PART-B

1. What is amplitude modulation? Draw its graphical and phasor representations. Derive its power calculation. Mention its advantages and disadvantages.
2. What is double side-band suppressed carrier AM? Compare it with AM with carrier with respect to frequency spectrum, phasor representation and power calculations.
3. Mention the purpose of single side-band suppressed carrier AM. Compare it with AM with carrier with respect to frequency spectrum, phasor representation and power calculations.
4. What is vestigial side-band modulation? Explain.
5. Give the comparison between AM with carrier, DSB-SC and SSB-SC modulation methods in detail.

PROBLEMS

1. An antenna transmits an AM signal having a total power content of 15 kW. Determine the power being transmitted at the carrier frequency and at each of the side bands when the percent modulation is 85%.
2. Calculate the power content of the carrier and of each of the side bands of an AM signal whose total broadcast power is 40 kW when the percent modulation is 60%.
3. Determine the power contained at the carrier frequency and within each of the side bands for an AM signal whose total power content is 15 kW when the modulation factor is 0.70.
4. An amplitude-modulated signal contains a total power of 6 kW. Calculate the power being transmitted at the carrier frequency and at each of the side bands when the percent modulation is 100%.
5. An AM wave has a power content of 1800 W at its carrier frequency. What is the power content of each of the side bands when the carrier is modulated to 85%?
6. An AM signal contains 500 W at its carrier frequency and 100 W in each of its side bands.
   (a) Determine the percent modulation of the AM signal.
   (b) Find the allocation of power if the percent modulation is changed to 60%.
7. 1200 W is contained at the carrier frequency of an AM signal. Determine the power content of each of the side bands for each of the following percent modulations.
   (a) 40% (b) 50% (c) 75% (d) 100%
8. An AM wave has a total transmitted power of 4 kW when modulated 85%. How much total power should an SSB wave contain in order to have the same power content as that contained in the two side bands?
9. An SSB transmission contains 800 W. This transmission is to be replaced by a standard AM signal with the same power content. Determine the power content of the carrier and each of the side bands when the percent modulation is 85%.
10. A commercial AM station is broadcasting wit has average transmitted carrier power of 10 kW. The modulation index is set at 0.707 for a sinusoidal modulating signal. Determine the transmission power, efficiency and the average power in the carrier component of the transmitted signal.

3

AMPLITUDE DEMODULATION

Objectives

- To know the purpose of amplitude demodulation in a communication system
- To discuss the different types of AM detectors with their detection procedure in detail
- To provide various AM receivers with their block diagram and functioning in detail
- To know the significance of AGC and its different types

3.1 INTRODUCTION

The reverse process of amplitude modulation is called **AM demodulation** from which the original modulating signal is recovered from a modulated wave. AM demodulators, or detectors, are circuits that accept amplitude-modulated signals and recover the original modulating signal or message signal.

An ideal demodulator should produce at its output a demodulated signal that resembles the original modulating signal in all respects. If there is any deviation from the wave shape of the modulating signal, it is termed **distortion**.

3.2 TYPES OF AM DETECTORS

3.2.1 Series Diode Detector

A simple series diode detector for AM demodulation purpose is shown in Figure 3.1. This series diode detector is the basic type of diode receiver, which is also referred to as a **voltage**

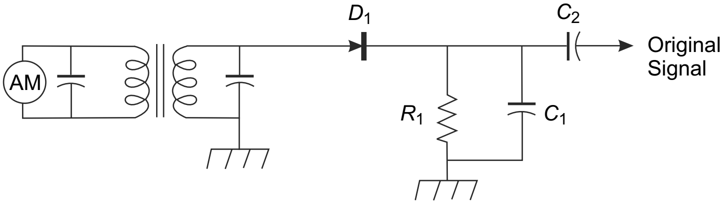


Fig. 3.1 A series diode detector

diode detector. Since the semiconductor diode is in series with both the input voltage and the load impedance, this is called a series-diode detector.

From the above figure, the AM signal is usually a transformer coupled and applied to a basic half-wave rectifier circuit consisting of a diode D_1 and R_1. The diode conducts when the positive half cycles of the AM signals occur. During the negative half cycles, the diode is reverse biased and no current flows through it. Hence, it results in the voltage across R_1 that is a series of positive pulses whose amplitude varies with the modulating signal. A capacitor C_1 is connected across the resistor R_1, effectively filtering out the carrier and thus recovering the original message signal.

During each positive cycle of the AM signal, the capacitor charges quickly to the peak value of the pulses passed by the diode. When the pulse voltage drops to zero, the capacitor discharges into the resistor R_1. The time constant R_1C_1 is chosen to be long compared to the carrier period. As a result, the capacitor discharges only during the time the diode is not conducting. When the next pulse comes along, the capacitor again charges to its peak value and during reverse biasing of the diode, the capacitor again discharges a small amount into the resistor. The resulting waveform across the capacitor is a close approximation to the original modulating signal. Figure 3.2 shows the input and output waveforms of the series diode detector.

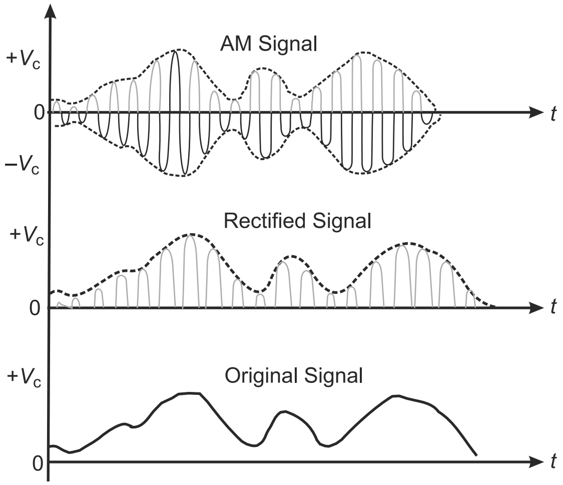


Fig. 3.2 Input–output waveforms

Because the diode detector recovers the envelope of the AM signal, which is the original information signal, the circuit is also called an **envelope detector**. The purpose of a blocking capacitor C_2 is to completely filter out the carrier signal.

3.2.2 Shunt Diode Detector

The circuit diagram of a shunt diode detector is shown in Figure 3.3.

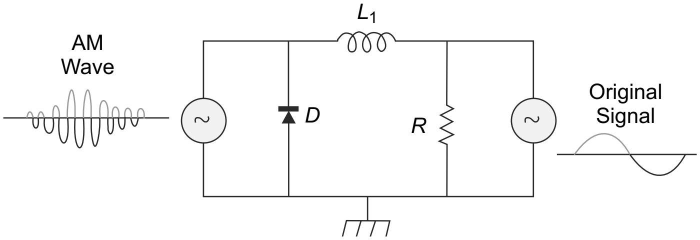


Fig. 3.3 Shunt diode detector

The shunt diode detector is similar to the series diode detector except that the output variations are current pulses rather than voltage pulses. This current passed through a shunt resistor develops the voltage output.

The input is an AM modulated envelope. On the negative half cycles of the AM wave, the diode D_1 is forward biased and shunts the signal to ground. On the positive half cycles, current flows from the output through L_1 to the input. A field is built up around L_1 that tends to keep the current flowing. This action integrates the RF current pulses and causes the output to follow the modulation envelope closely. Shunt resistor R_1 develops the output voltage from this current flow. Although the shunt detector operates on the principle of current flow, it is the output voltage across the shunt resistor that is used to reproduce the original modulation signal. The shunt diode detector is easily identified by noting that the detector diode is in parallel with both the input and load impedance. The waveforms associated with this detector are identical to those obtained from a series diode detector.

The shunt diode detector is used where the voltage variations are too small to produce a full output from audio amplifier stages. Additional current amplifiers are required to bring the output to a usable level.

3.2.3 Common-Emitter Detector

The main purpose of a common-emitter detector is to supply an amplified detected output in receivers. A typical transistor common-emitter detector is shown in Figure 3.4. Input transformer T_1 acts as a frequency-selective device and L_2 inductively couples the input modulation envelope to the base of the transistor Q_1. Resistors R_1 and R_2 are fixed-bias voltage dividers that set the bias levels for Q_1. To eliminate carrier signal resistor R_1 is bypassed by the

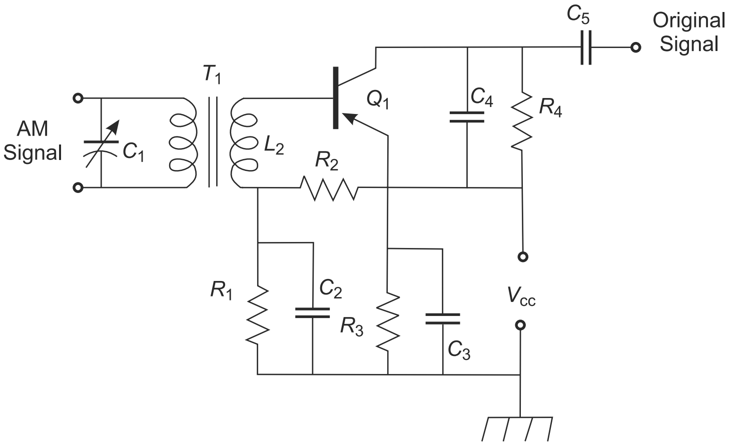


Fig. 3.4 Common-emitter detector

capacitor C_2. This *RC* combination also acts as the load for the diode detector. The detected output is in series with the biasing voltage and controls collector current. The output is developed across R_4 which is also bypassed to remove the carrier by the capacitor C_4. Resistor R_3 is used for temperature stabilisation and C_3 bypasses it for both carrier and modulating signals.

When there is no input signal, there is less conduction in the transistor Q_1. When an input signal is applied to the base of Q_1, it is rectified by the emitter-base junction, which is operating as a diode. It is developed across R_1 as a dc bias voltage which controls bias and collector current for Q_1. The output collector current flows through the resistor R_4. Any carrier signal in the output is bypassed across the collector load resistor by capacitor C_4 without bypassing the message signal. After the modulation envelope is detected in the base circuit, it is amplified in the output circuit to provide the original output. Due to the amplification in this circuit, weaker signals can be detected than with a simple detector.

3.2.4 Common-Base Detector

Figure 3.5 is a circuit diagram of a typical common-base detector. In this circuit, detection occurs in the emitter-base junction and amplification occurs at the output of the collector junction. Transformer T_1 is tuned by the capacitor C_3 to the frequency of the incoming modulated envelope. Resistor R_1 and the capacitor C_1 combination forms a self-biasing network which sets the dc operating point of the emitter junction. The original message output is taken from the collector circuit through audio transformer T_2. The primary of T_2 forms the detector output load and the carrier is bypassed by the capacitor C_2.

When the capacitor C_3 is tuned to the proper frequency, the signal is passed to the emitter of Q_1. When no input signal is present, bias is determined by the resistor R_1. When the input signal becomes positive, current flows through the emitter-base junction, causing it to be

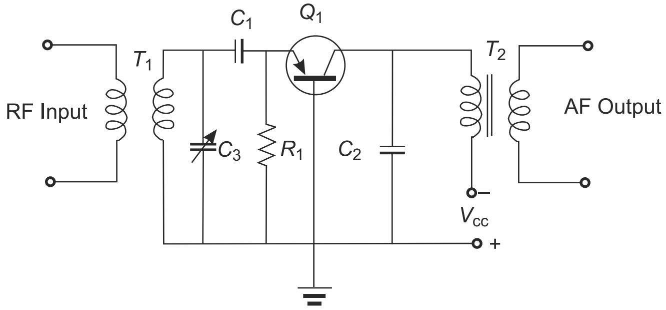


Fig. 3.5 Common-base detector

forward biased. The C_1 and R_1 combination acts as a filter network. This action provides a varying dc voltage that follows the peaks of the carrier-modulated envelope.

The varying dc voltage on the emitter changes the bias on Q_1 and causes collector current to vary in accordance with the detected voltage. Transformer T_2 couples these modulating signal current changes to the output. Thus, Q_1 detects the AM wave and then provides amplification for the detected waveform.

3.2.5 Synchronous or Coherent Detector

The synchronous detector uses carrier synchronisation for retrieving the message signal. These detectors are mainly used for detecting DSB SC or SSB SC signals because of their empirical nature. Figure 3.6 shows the block diagram of a synchronous detector.

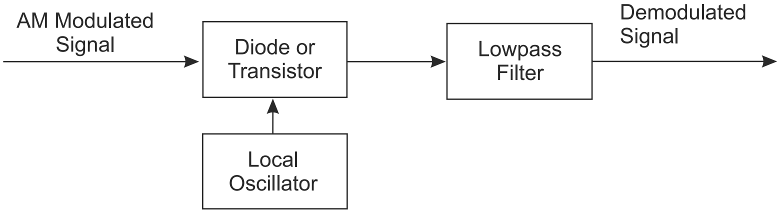


Fig. 3.6 Block diagram of a synchronous detector

The incoming AM modulated signal is first multiplied with a locally generated carrier signal and then entered through a lowpass filter. The filter bandwidth is same as the message bandwidth or sometimes larger. It is assumed that the local oscillator is exactly synchronised with the carrier in both phase and velocity, hence the name synchronous detector.

Case 1

If the input to the detector is a DSB SC signal then

$$V_1 = \frac{m_a V_c}{2} \sin \omega_c t \sin \omega_m t \tag{3.1}$$

$$V_2 = V \sin \omega_c t \tag{3.2}$$

Then the output is

$$y(t) = V_1 . V_2 = \frac{m_a V_c V}{2} \sin^2 \omega_c t \sin \omega_m t$$

$$= \frac{m_a V_c V}{2} \frac{(1 - \cos 2\omega_c t)}{2} \sin \omega_m t \tag{3.3}$$

After passing through LPF,

$$V_0 = \frac{m_a V_c V}{4} \sin \omega_m t \tag{3.4}$$

Case 2

If the input to the detector is an SSB SC signal then

$$V_1 = \frac{m_a V_c}{2} \sin \omega_c t \sin \omega_m t \tag{3.5}$$

$$V_2 = V \sin \omega_c t \tag{3.6}$$

Then the output is

$$y(t) = V_1 . V_2 = \frac{m_a V_c V}{2} \cos (\omega_c - \omega_m)t \cdot \sin \omega_m t$$

$$= \frac{m_a V_c V}{2} \left[\frac{\sin \omega_m t - \sin(2\omega_c t - \omega_m t)}{2} \right] \tag{3.7}$$

The first term of the output is the modulating signal and is passed through the output. The second term is an RF ripple component and is attenuated by the low pass filter. Then the final recovered information signal is

$$V_0 = \frac{m_a V_c V}{4} \sin \omega_m t \tag{3.8}$$

Thus, the synchronous detector is capable of demodulating DSB-SC and SSB SC AM. However, the synchronous detector is effective only when the locally generated signal is properly synchronised with the modulating signal. Any shift in phase or frequency of the carrier from the local oscillator results in phase or delay distortion.

A practical synchronous detector is shown in Figure 3.7, in which a centre-tapped transformer provides the two equal but inverted signals. The carrier signal is applied to the centre tap.

Diode D_2 is placed in the direction opposite to that of the diode D_1. These diodes are used as switches, which are turned off and on by the clock, which is used as the bias voltage. Carrier is derived by amplifying the AM signal.

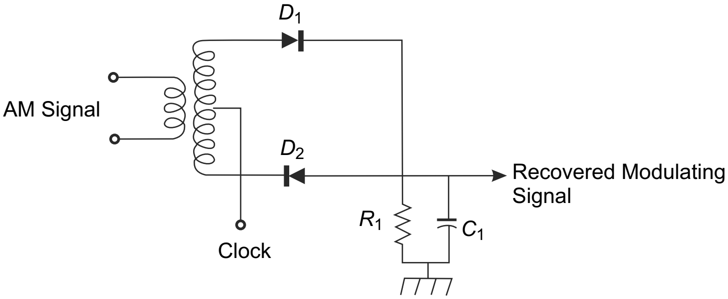


Fig. 3.7 A practical synchronous detector

When the clock is positive, D_1 is forward biased and it acts as a short circuit and connects the AM signal to the load resistor. As a result, positive half cycles appear across the load. When clock is negative, D_2 is forward biased. During this time, the negative cycles of the AM signal are occurring, which makes the lower output of the secondary winding positive. With D_2 conducting, the positive half cycles are passed to the load and the circuit performs full-wave rectification. As a result, the capacitor across the load filters out the carrier, leaving the original modulating signal across the load.

3.2.6 Quadrature PLL Detector

Quadrature or Costas detectors consist of two synchronous detectors. One detector is supplied with DSB-SC AM and a locally generated carrier which is in phase with the transmitted carrier. This detector is said to be the in-phase detector or *I*-channel. The second detector is supplied with DSB-SC AM and a locally generated carrier which is in quadrature phase with the carrier. This detector is said to be quadrature detector or *Q*-channel. These two detectors are coupled together to form a negative feedback system designed in such a way to maintain the local oscillator synchronous with the carrier. The block diagram of such a detector is shown in Figure 3.8.

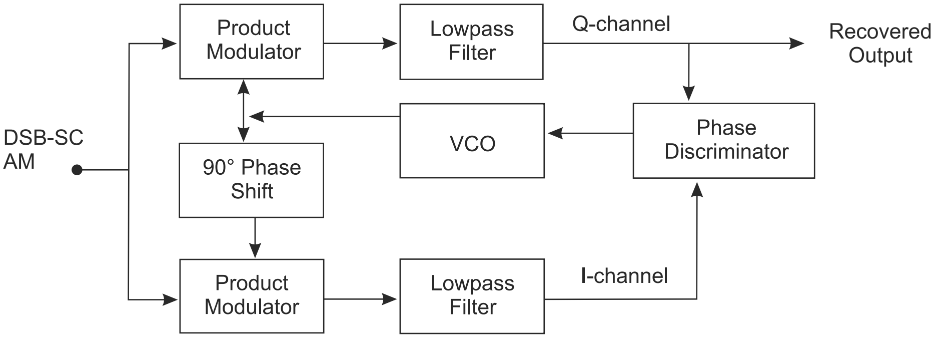


Fig. 3.8 Costas detector for DSB-SC AM

By assuming that the phase of the local oscillator signal to be same as that of the carrier, *I*-channel output contains the desired demodulated signal whereas *Q*-channel output is zero due to the quadrature null effect of *Q*-channel. It indicates that the local oscillator carrier signal is properly synchronised with the transmitting carrier.

If there is a phase shift of φ between local oscillator carrier and the transmitting carrier then *I*-channel output will remain in the same value. But *Q*-channel output contains some signal which is proportional to sin φ. This *Q*-channel output will have the same polarity as the *I*-channel output for one direction of local oscillator phase shift and opposite polarity for opposite direction of local oscillator-phase shift.

By combining the *I* and *Q*-channel outputs in phase discrimination, a dc signal is obtained that automatically corrects the phase errors in VCO. It is apparent that phase control in the Costas detector ceases with modulation and that phase lock has to be reestablished.

3.2.7 Vestigial Side-Band Demodulation

Envelope detectors are also suitable for demodulation of vestigial side-band signals. A VSB signal may be considered a normal AM signal to which a pair of side bands having smaller amplitude than side bands of the signals are added in phase quadrature with the carrier as shown in Figure 3.9. Figure 3.9(a) shows a DSB AM signal which may be produced by usual modulation methods. In Figure 3.9(b) the same signal is shown, to which are added side bands having magnitude smaller than the magnitude of the side bands of the AM signal but in phase quadrature with the carrier. As a result of these quadrature components, the USB is reduced while the LSB is increased as shown in Figure 3.9(c), which is the resultant VSB signal.

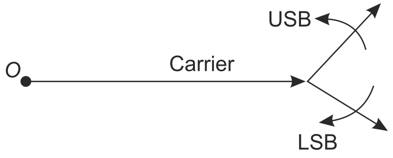


Fig. 3.9(a) DSB AM signal

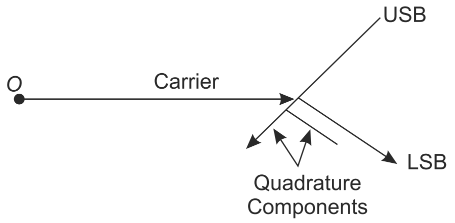


Fig. 3.9(b) Signal with quadrature components

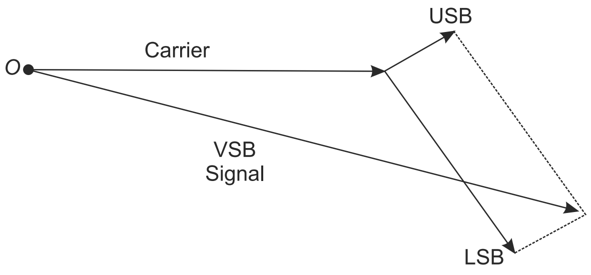


Fig. 3.9(c) Resultant VSB signal

If the DSB AM signal is represented by the expression,

$$e(t) = V_{\mathrm{c}}(1 + m_{\mathrm{a}} \sin \omega_{\mathrm{m}}t) \sin \omega_{\mathrm{c}}t$$

$$= V_{\mathrm{c}} \sin \omega_{\mathrm{c}}t + m_{\mathrm{a}} V_{\mathrm{c}} \sin \omega_{\mathrm{m}}t \sin \omega_{\mathrm{c}}t \qquad (3.9)$$

Then the quadrature pair of side bands is given as $\alpha\, m_{\mathrm{a}} V_{\mathrm{c}} \cos \omega_{\mathrm{c}}t \cos \omega_{\mathrm{m}}t$, where the value of α lies between 0 and 1. Therefore, the VSB signal given to the detector input is given by

$$V_{\mathrm{t}}(t) = V_{\mathrm{c}}\, [(1 + m_{\mathrm{a}} \sin \omega_{\mathrm{m}}t) \sin \omega_{\mathrm{c}}t + \alpha\, m_{\mathrm{a}}V_{\mathrm{c}} \cos \omega_{\mathrm{c}}t \cos \omega_{\mathrm{m}}t]$$

$$= V_{\mathrm{c}}\, \{[1 + P(t)] \sin \omega_{\mathrm{c}}t + Q(t) \cos \omega_{\mathrm{c}}t\} \qquad (3.10)$$

where

$$P(t) = m_{\mathrm{a}} \sin \omega_{\mathrm{m}}t \text{ and } Q(t) = \alpha\, m_{\mathrm{a}}\, \omega_{\mathrm{m}}t \qquad (3.11)$$

$$\therefore \quad V_{\mathrm{i}}(t) = V_{\mathrm{c}}\, [A \sin \omega_{\mathrm{c}}t + B \cos \omega_{\mathrm{c}}t] \qquad (3.12)$$

where

$$A = 1 + P(t) \text{ and } B = Q(t) \qquad (3.13)$$

Or alternatively,

$$V_{\mathrm{i}}(t) = V_{\mathrm{c}}\left[\sqrt{A^2 + B^2} \cdot \sin(\omega_{\mathrm{c}}t + \phi)\right] \qquad (3.14)$$

where

$$\phi = \tan^{-1}(B/A) \qquad (3.15)$$

It can be seen that the VSB signal input to the detector is equivalent to an angle-modulated wave whose amplitude also changes with the modulating signal. The detector does not respond to angle modulation but gives an output that is proportional to the amplitude changes.

The detector output,

$$e_0(t) = KV_{\mathrm{c}}\left(\sqrt{A^2 + B^2}\right) \qquad (3.16)$$

where K is the rectification efficiency.

Equation (3.16) may be rewritten as

$$e_0(t) = AKV_{\mathrm{c}}\left(\sqrt{1 + (B/A)^2}\right)$$

$$\cong AKV_{\mathrm{c}}\, (1 + m_{\mathrm{a}} \sin \omega_{\mathrm{m}}t) \text{ if } B < A \qquad (3.17)$$

The expression for the detector output described by Equation (3.17) shows that there is a modulating signal present at the detector output provided the quadrature component is smaller than the original side bands, i.e. $B < A$. As the quadrature component is proportional to $\alpha.m_{\mathrm{a}}$, the side-band power and, hence, the demodulated output may be increased by increasing m_{a} but at the same time keeping the product $\alpha.m_{\mathrm{a}}$ small.

3.3 AM RECEIVERS

3.3.1 Functions of a Radio Receiver

In general, a radio receiver is a device that picks up the desired signal from the various signals propagating at that time through the atmosphere, amplifies the desired signal to the requisite level, recovers from it the original modulating signal and displays it in the desired manner. The difference between receivers of various types arise from the way in which the receivers demodulate the signal. This would also depend on the modulation employed at the transmitter side.

3.3.2 Classification of Radio Receivers

Radio receivers are classified based on the following criteria:

1. Based on the Principle of Operation

(a) Straight or Tuned Radio Frequency Receivers These receivers are operated in a straightforward manner without frequency conversion or mixing.

(b) Super Heterodyne Receivers In these receivers, the incoming RF signal is converted to standard Intermediate Frequency (IF) before detection takes place. This is done with the help of frequency converters.

2. Based on the Application

(a) AM Receivers These receivers can be used to receive the speech or music radiated from AM broadcast transmitter operating on long-wave, medium-wave or short-wave band.

(b) FM Receivers These receivers are used for receiving broadcast programmes from FM broadcast in VHF and UHF bands.

(c) TV Receivers For receiving television broadcast in VHF or UHF bands, TV receivers are used.

(d) Communication Receivers These are superheterodyne receivers used for the reception of telegraph and short-wave telephone signals. These receivers are more costly and complicated due to the use of more components used in the receiver design.

In addition to the above types, code receivers are simple superheterodyne receivers with the addition of an IF beat frequency oscillator to produce audio beat notes with IF signal and radar receivers used for receiving radar signals.

3.3.3 Features of Broadcast Receivers

The broadcast receivers must possess the following features.

1. Simplicity in Operation

These receivers are required to be handled by all kinds of users and it is essential for it to be simple to operate. These receivers basically have three controls like frequency-band selection switch, tuning control and volume control.

2. Good Fidelity

These receivers should have a reasonably large and uniform frequency response over the entire audio-frequency band.

3. Good Selectivity

These receivers should have the ability to discriminate the desired signals from the unwanted signals at other frequencies, preferably the side bands of adjacent channels in the frequency spectrum.

4. Adaptability to Different Types of Antenna

A broadcast receiver should be designed to operate satisfactorily with any type of antenna.

Though various forms of receiver circuits have been developed, two types of receivers have real significance.

1. Tuned Radio Frequency (TRF) receivers
2. Super heterodyne receivers

3.3.4 Tuned Radio Frequency Receivers

The TRF receiver is a simple receiver employing straightforward circuit arrangements. It uses two or three stages of RF amplification all tuned simultaneously to the desired signal frequency so that these stages provide selection as well as amplification to the signal. The amplified signal is then demodulated in a detector stage. The demodulated signal is amplified by the AF amplifier stages and fed to the loudspeaker. Figure 3.10 shows the block diagram of such a receiver.

A TRF receiver employs two identical stages of RF amplifiers tuned together. Their purpose is to select and amplify the incoming frequency and reject the unwanted frequencies. After the signal was amplified to a suitable level, it was demodulated using a suitable detector and then

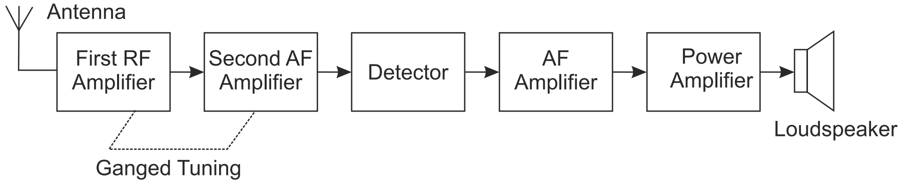


Fig. 3.10 Block diagram of a TRF receiver

finally fed to the loudspeaker. Before passing it to the loudspeaker, it was amplified by using AF and power amplifiers.

These receivers are easy to design but work satisfactorily only at medium-wave frequencies. At high frequencies, it is difficult to design the receiver. This limitation is mainly due to the inability associated with the high gain achieved at the desired frequency by a multistage amplifier. If each RF stage has a gain of 300, the overall gain will be 90,000 and $\frac{1}{90{,}000}$ of the output of the last stage is needed, which is to be passed back to the input of the first stage with correct polarity causing oscillations.

Another drawback in TRF receivers is the wide variation in the Q-factor and the bandwidth of the tuned circuits employed in RF amplifiers at different frequencies of the frequency band. Consider a tuned circuit required to have a bandwidth of 10 kHz at a frequency of 550 kHz. The Q of the circuit must be

$$Q = \frac{f_0}{BW} = \frac{550}{10} = 55 \tag{3.18}$$

At the higher end of the band, i.e. 1650 kHz, the inductive reactance and, hence, the circuit Q of the coil will increase by a factor of $\frac{1650}{550} = 165$. In practice, the circuit losses increase with frequency and the Q does not rise directly as frequency increases; the practical values of the Q lying in the range of 100–140. Taking an average value of 120 for the Q would give a bandwidth of $\frac{1650}{120} = 13.75\text{ kHz}$. This increased bandwidth results in a pick-up of stations of adjacent frequencies.

If such a receiver is to be used at short waves, satisfactory reception at 20 MHz would require the tuned circuit to have a $Q = \frac{20\text{ MHz}}{10\text{ kHz}} = 2000$. Since this value of Q cannot be obtained with ordinary tuned circuits; selectivity at this frequency will be very poor.

Due to the drawbacks such as inability in gain, reception of stations of frequencies adjacent to the desired signal and variation in the bandwidth over the band, the TRF receiver needs to be replaced by another suitable receiver called superheterodyne receiver.

EXAMPLE 3.1

A TRF receiver is to be tuned with a single tuned circuit using a 10 mH inductor. The ideal to 10 kHz bandwidth occurs at 500 kHz.(a). Calculate the capacitance range of the variable capacitance in the LC tank circuit to tune the receiver from 600 to 1500 kHz. (b). Calculate the bandwidth of the receiver at 600 to 1500 kHz.

Solution

$$f = \frac{1}{2\pi\sqrt{LC}}$$

$$C_{\max} = \frac{1}{(2\pi f)^2 L}$$

$$C_{\max} = \frac{1}{(2\pi \times 600 \times 10^3)^2 \times 10 \times 10^{-6}} = 7.04\,nf$$

$$C_{\min} = \frac{1}{(2\pi \times 1500 \times 10^3)^2 \times 10 \times 10^{-6}} = 1.126\,nf$$

$$\text{Bandwidth } Q = \frac{f_0}{BW}$$

$$= \frac{600 \times 10^3}{100} = 6000 \text{ Hz} = 6 \text{ kHz}$$

$$\text{Bandwidth } Q = \frac{f_0}{BW}$$

$$= \frac{1500 \times 10^3}{100} = 15000 \text{ Hz} = 15 \text{ kHz}$$

EXAMPLE 3.2

An AM commercial broadcast receiver operating in a frequency band of (535 to 1605) kHz with an input filter factor of 54. Determine the bandwidth at the low and high ends of RF spectrum.

Solution

The bandwidth at the low frequency end of the AM spectrum is centreed around a carrier frequency of 540 kHz and is

$$\text{Bandwidth } Q = \frac{f_0}{BW}$$

$$= \frac{540 \times 10^3}{54} = 10000 \text{ Hz} = 10 \text{ kHz}$$

The bandwidth at the high frequency end of the AM spectrum is centreed around a carrier frequency of 1600 kHz and is

$$\text{Bandwidth } Q = \frac{f_0}{BW}$$

$$= \frac{1600 \times 10^3}{54} = 29{,}630 \text{ Hz} = 29.63 \text{ kHz}$$

3.3.5 AM Superheterodyne Receivers

Superheterodyne principle is the process of operation on modulated radio waves to obtain similarly modulated waves of different frequencies. This process includes the use of an input signal with the local oscillator signal which determines the change of frequency.

A superheterodyne receiver may be defined as one in which one or more changes of frequency take place before the AF signal is extracted from the modulated wave. A receiver in which the change of frequency takes place twice before detection is usually called a **double superheterodyne receiver**.

In superheterodyne receivers, the modulated signal of the carrier frequency (f_m) is fed to a circuit called **mixer** to which is also fed the voltage at frequency (f_0) generated by a local oscillator. As a result, the output of the mixer stage is a voltage of frequency (f_{IF}), which is the difference of the signal frequency f_m and the local oscillator frequency f_0. This difference frequency is called **Intermediate Frequency (IF)**. The signal frequency and the local oscillator frequency can be varied by using ganged tuned capacitors in these stages. This results in a mixer output that has a constant frequency irrespective of the frequency to which the receiver may be tuned. Thus, IF is fixed for a receiver. It should be noted that the IF signal is exactly similar to the modulated signal and the only difference is in their carrier frequencies.

The IF amplifiers, being tuned voltage amplifiers, use transformers in the input and output circuits. Each of these transformers consists of a pair of mutually coupled tuned circuits. With these fixed-frequency tuned circuits as plate load, the IF amplifiers provide most of the gain and selectivity to the receiver. As the gain and selectivity of IF amplifiers remain constant at all incoming signal frequencies, the sensitivity and selectivity of the receiver is fairly uniform over the entire frequency range.

The block diagram of a superheterodyne receiver is shown in Figure 3.11.

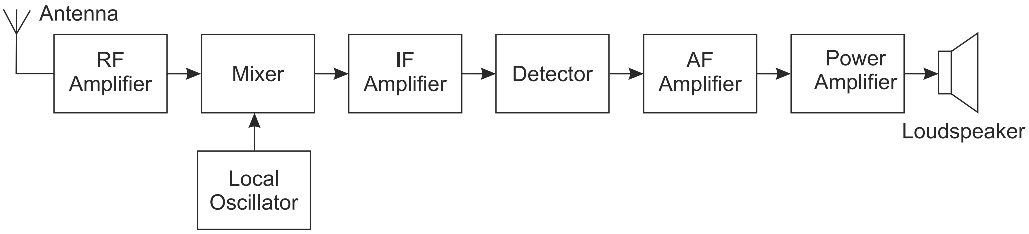


Fig. 3.11 Block diagram of superheterodyne receiver

In a broadcasting system, the receiver performs the following functions other than just demodulating the incoming signal.

(a) Carrier Frequency Tuning This is mainly done in order to select the desired signal, i.e. TV or radio signal.

(b) Filtering Filtering is used to separate the desired signal from the other modulated signals that may be picked up along the way.

(c) Amplification This is done in order to compensate the loss of signal power during transmission.

Basically, the receiver consists of an RF section, a mixer and a local oscillator, one IF section, a detector and a power amplifier.

1. RF Amplifier

The incoming AM signal is picked up by the receiving antenna first and is passed to the RF amplifier. The RF amplifier is a tuned voltage amplifier and couples the antenna to the mixer. It selects the desired signals from the antenna and amplifies the signals to the requisite level.

The advantages of having an RF amplifier stage in the receiver are listed as follows:

(a) The RF amplifier amplifies the incoming signal voltage to a high level before feeding it to the mixer. This increases the overall sensitivity of the receiver.

(b) It provides discrimination or selectivity against signals of unwanted frequencies. This results in improved image signal and adjacent channel selectivity.

(c) It provides better coupling between the antenna and the mixer so that energy extracted from the EM waves is effectively utilised. This is important at VHF and higher frequencies.

(d) The increased level of output signal at the mixer input increases the noise performance of the receiver. This results because the magnitude of the signal at the mixer input is brought to a level quite higher than the noise magnitude. Thus, signal-to-noise ratio is increased.

(e) It isolates the local oscillator circuit from antenna, thereby preventing radiation of local oscillator energy.

Figure 3.12 shows the circuit diagram of a one-stage RF amplifier. It is a small-signal amplifier using parallel tuned circuit as the load impedance. This parallel-output tuned circuit is tuned to the incoming desired signal frequency. The output from the receiving antenna is transformer coupled to the base of the transistor. The secondary coil of the input tuned circuit is tuned to the incoming desired signal frequency with the help of a ganged tuning capacitor. In fact, tuning capacitors in the input side and the output side of the RF amplifier are ganged together. In addition to this, small trimming capacitors are connected in shunt with these tuning capacitors for the purpose of RF alignment.

A self-bias is provided with the help of resistors R_1 and R_2 and R_E-C_E combination. A decoupling network consisting of the resistor R_b and the capacitor C_b is placed in the collector supply lead.

The amplified RF signal developed across the collector tuned circuit is coupled through a step-down transformer to the input of the frequency mixer. This step-down transformer provides the impedance matching between the high impedance of the RF amplifier collector circuit and the low impedance of the base-to-emitter circuit of the following stage. Also, the

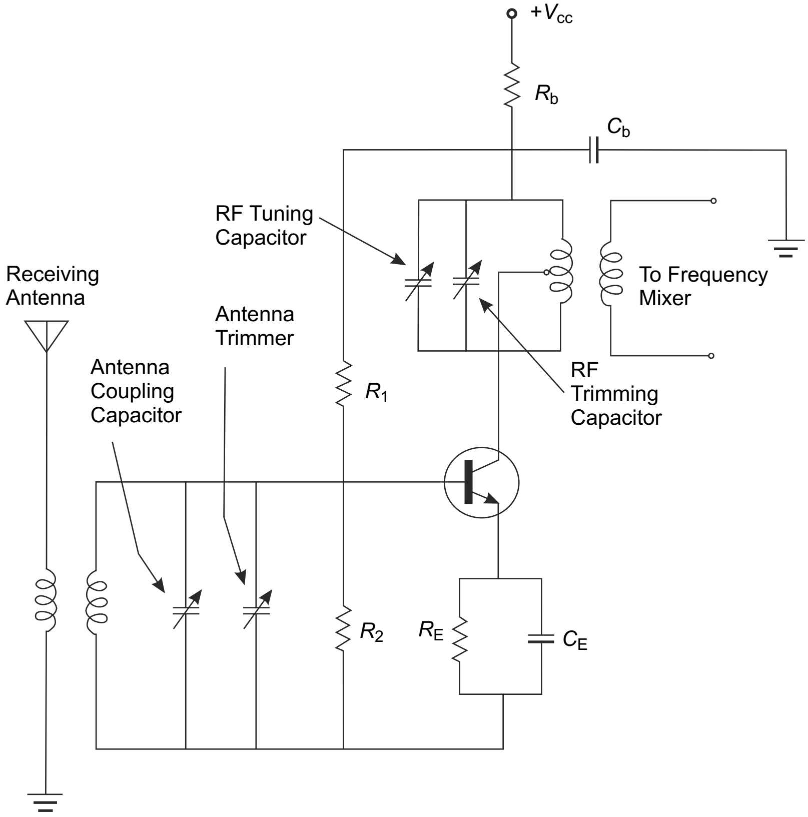


Fig. 3.12 Circuit diagram of RF amplifier

collector is connected to the suitable point on the primary of the output transformer so that load impedance on top of the collector is optimum.

2. Local Oscillator

All local oscillators are *LC* oscillators and use a single tuned circuit to determine the frequency of oscillation. Figure 3.13 shows a typical local oscillator circuit used in a superheterodyne receiver along with the mixer circuit.

The local oscillator frequency of the standard broadcast receiver is usually made higher than the incoming signal frequency by an amount equal to intermediate frequency (IF) of 455 kHz. If the standard AM broadcast band is assumed to extend from 540 kHz to 1650 kHz, the local oscillator frequency is made smaller than the incoming signal frequency. In such a case, the local oscillator should be capable of varying frequency in the range of 85 kHz to 1195 kHz. This gives a ratio of maximum to minimum frequency equal to 14:1. However, this ratio cannot be achieved by normal tunable capacitance. The capacitance of normal tunable capacitance equal to 10:1 gives a maximum frequency ratio of 3.2:1. This means that if the

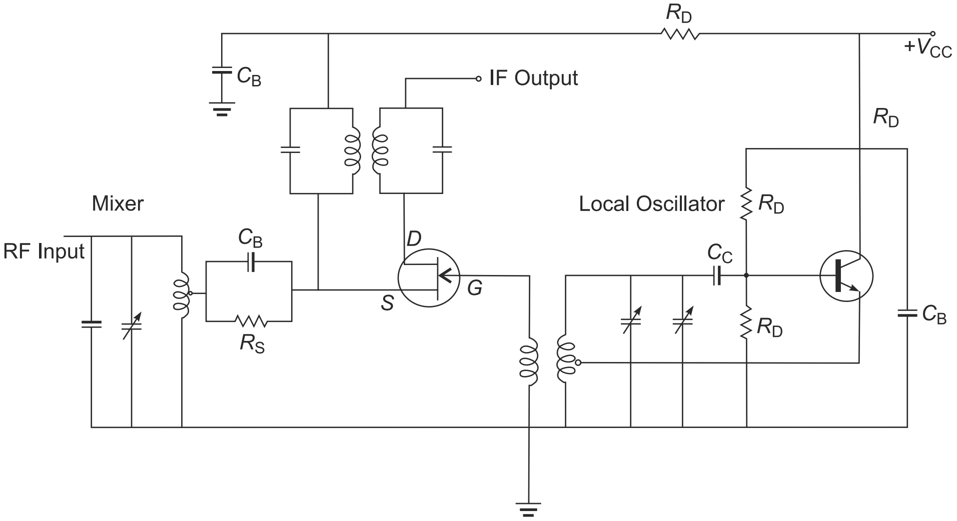


Fig. 3.13 A local oscillator circuit

local oscillator frequency is kept lower than the signal frequency by an amount of 455 kHz, the normal tunable capacitors cannot be used.

3. Frequency Changers

The combination of a mixer and local oscillator constitute the frequency changer. Both of them provide 'heterodyne' function, where the incoming signal is converted to a predetermined fixed frequency called the **intermediate frequency**. This intermediate frequency is lower than the incoming carrier frequency. The result of heterodyning is

$$f_{\text{IF}} = f_0 - f_{\text{m}} \tag{3.19}$$

Since the output of the frequency changer is neither the original input frequency nor the output baseband frequency, it is called intermediate frequency. Sometimes the frequency-changer circuits are referred to as **first detector**. In case of double frequency conversion, the demodulator becomes the third detector.

4. IF Section

Intermediate Frequency (IF) amplifiers are tuned voltage amplifiers that are operated in Class A with a fixed resonant load. The IF section has the bandwidth corresponding to the required signal that the receiver is intended to handle. This section provides most of the amplification and selectivity of the receiver.

The intermediate frequency of a receiver is always a compromise between various factors as described below:

(a) If the intermediate frequency is made too high, adjacent channel rejection as well as selectivity becomes poor.

(b) A high value of intermediate frequency makes the difference between signal and local oscillator frequency large and as a result, tracking becomes difficult.

(c) If the intermediate frequency is lowered, the difference between a signal frequency and its image frequency is reduced; this results in a poorer image signal rejection. Thus, intermediate frequency must be made high if image signals are to be completely rejected.

(d) A low intermediate frequency makes the selectivity sharp, thereby increasing the adjacent channel rejection. Too low an IF makes the selectivity too sharp that may result in cutting off side bands. To avoid this, magnification factor Q of the IF circuits has to be lowered which results in low-stage gain of IF circuits.

(e) If a low If is to be used then a high stability of the local oscillator frequency must be maintained because any drift in the local oscillator frequency results in large percentage IF drift.

(f) Finally, the IF of a receiver should be selected as to be lower than the lowest signal frequency to be received by the receiver otherwise signal frequencies close to the intermediate frequency will be difficult to receive and heterodyne whistles will be heard in the receiver output.

5. Image Frequency

An image frequency is any frequency other than the selected radio frequency carrier that, if allowed to enter and mix with the local oscillator, will produce a cross-product frequency that is equal to the intermediate frequency. An image frequency is equivalent to a second radio frequency that will produce an IF that will interfere with the IF from the desired radio frequency.

Once an image frequency has been mixed to IF, it cannot be filtered out or suppressed. If the selected RF carrier and its image frequency enter a receiver at the same time, they both mix with the local oscillator frequency and produce difference frequencies that are equal to the IF. Consequently, two different stations are received and demodulated simultaneously, producing two sets of frequencies.

For a radio frequency to produce a cross product equal to the IF, it must be replaced from the local oscillator frequency by a value equal to the IF. With high side injection, the selected RF is below the local oscillator by an amount equal to the IF. Therefore, the image frequency is the radio frequency that is located in the IF frequency above the local oscillator. Mathematically, for high side injection, the image frequency (f_{im}) is

$$f_{\text{im}} = f_c + f_{\text{IF}} \tag{3.20}$$

Because the desired RF equals the local oscillator frequency minus IF,

$$f_{\text{im}} = f_{\text{RF}} + 2f_{\text{IF}} \tag{3.21}$$

For a superheterodyne receiver using high side injection, the frequency spectrum is shown in Figure 3.14.

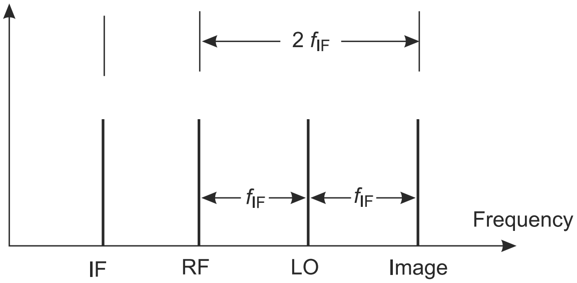


Fig. 3.14 Frequency spectrum with image frequency

From the above figure, it is noted that the higher the IF, the farther away in the frequency spectrum the image frequency is from the desired RF. Therefore, for better image-frequency rejection, a high intermediate frequency is preferred. However, the higher the IF, the more difficult it is to build stable amplifiers with high gain. Therefore, there is trade off when selecting the IF for a radio receiver between image frequency rejection and IF gain and stability.

3.3.6 Image Frequency Rejection Ratio (IFRR)

The image frequency rejection ratio (IFRR) is a numerical measure of the ability of a preselector to reject the image frequency. For a single tuned preselector, the ratio of its gain at the desired RF to the gain at the image frequency is the IFRR. Mathematically, IFRR is,

$$\text{IFRR} = \sqrt{(1+Q^2\rho^2)} \tag{3.22}$$

where

$$\rho = \left({f_{\text{im}}}/{f_{\text{RF}}} \right) - \left({f_{\text{RF}}}/{f_{\text{im}}} \right) \tag{3.23}$$

If there is more than one tuned circuit in the front end of a receiver, the total IFRR is simply the product of the two ratios.

Once an image frequency has been down0converted to IF, it cannot be removed. Therefore, to reject the image frequency, it has to be blocked prior to the mixer/converter stage. Image-frequency rejection is the primary purpose for the RF preselector. If the bandwidth is sufficiently low, the image frequency is prevented from entering the receiver.

The ratio of the RF to the IF is also an important consideration for image frequency rejection. The closer the RF is to the IF, the closer the RF is to the image frequency.

7. Demodulator or Detector

The output of the IF amplifier section is fed to a demodulator which recovers the baseband or message signal. Diode detector is most common choice in radio receivers. If coherent detection is used, then a coherent signal source must be provided in the receiver. The detector also supplies dc bias voltage to RF and IF stages in the form of an AGC circuit. Finally, the recovered signal is power amplified en-route to the loudspeaker.

EXAMPLE 3.3

An AM standard broadcast receiver is to be designed having an intermediate frequency of 455 kHz. Calculate the required frequency that the local oscillator should be at when the receiver is tuned to 540 kHz if the local oscillator tracks above the frequency of the received signal.

Solution

The intermediate frequency is generated by producing a difference frequency between the carrier and the local oscillator.

$$f_{\mathrm{IF}} = f_0 - f_{\mathrm{m}}$$

or

$$f_{\mathrm{IF}} = f_{\mathrm{m}} - f_0$$

To find the local oscillator frequency,

$$f_0 = f_{\mathrm{IF}} + f_{\mathrm{m}}$$

$$= (455 \times 10^3) + (540 \times 10^3) = 995 \text{ kHz}$$

EXAMPLE 3.4

Repeat the Example 3.3, if the local oscillator tracks below the frequency of the received signal.

Solution

To find the local oscillator frequency,

$$f_0 = f_{\mathrm{m}} + f_{\mathrm{IF}}$$

$$= (540 \times 10^3) - (455 \times 10^3) = 85 \text{ kHz}$$

EXAMPLE 3.5

An AM superheterodyne receiver uses high side injection and has a carrier frequency of 1355 kHz. Determine the IF carrier, USB and LSB for an RF wave that is made up of a carrier frequency of 900 kHz, USB of 905 kHz and LSB of 895 kHz.

Solution

$$f_{\mathrm{IF}} = f_{\mathrm{C}} + f_{\mathrm{RF}}$$

$$= 1355 \times 10^3 - 900 \times 10^3 = 455 \text{ kHz}$$

The upper and lower IFs are

$$f_{\text{IF(USF)}} = f_{\text{C}} - f_{\text{RF(LSF)}}$$

$$= 1355 \times 10^3 - 895 \times 10^3 = 460 \text{ kHz}$$

$$f_{\text{IF(LSF)}} = f_{\text{C}} - f_{\text{RF(USF)}}$$

$$= 1355 \times 10^3 - 905 \times 10^3 = 450 \text{ kHz}$$

EXAMPLE 3.6

An AM superheterodyne receiver has IF of 455 kHz, RF of 600 kHz and local oscillator frequency of 1055 kHz. Determine the image frequency and IFRR for a preselector Q of 100.

Solution

To find the image frequency:

$$f_{\text{im}} = f_{\text{C}} + f_{\text{IF}}$$

$$= 1055 \times 10^3 + 455 \times 10^3 = 1510 \text{ kHz}$$

Or

$$f_{\text{im}} = f_{\text{RF}} + 2f_{\text{IF}}$$

$$= 600 \times 10^3 + 2(455 \times 10^3) = 1510 \text{ kHz}$$

To find the IFRR:

$$\rho = \frac{1510 \times 10^3}{600 \times 10^3} - \frac{600 \times 10^3}{1510 \times 10^3}$$

$$= 2.51 - 0.397 = 2.113$$

$$\text{IFRR} = \sqrt{1 + (100)^2 \ (2.113)^2} = 211.3$$

EXAMPLE 3.7

A receiver using high side injection has an RF carrier of 27 MHz and an IF centre frequency of 455 kHz. Determine the local oscillator frequency, image frequency, IFRR for a preselector Q of 100 and preselector Q required to achieve the same IFRR as that achieved for an RF carrier of 600 kHz.

Solution

To find the local oscillator frequency:

$$f_{\text{c}} = f_{\text{RF}} + f_{\text{IF}}$$

$$= 27 \times 10^6 + 455 \times 10^3 = 27.455 \text{ MHz}$$

To find the image frequency:

$$f_{\text{im}} = f_{\text{c}} + f_{\text{IF}}$$

$$= 27.455 \times 10^6 + 455 \times 10^3 = 27.91 \text{ MHz}$$

To find the IFRR:

$$\rho = \frac{27.91 \times 10^6}{27 \times 10^6} - \frac{27 \times 10^6}{27.91 \times 10^6}$$

$$\text{IFRR} = \sqrt{1 + (100)^2\,(0.066)^2} = 6.71$$

To find the preselector Q:

$$Q = \sqrt{\frac{(\text{IFRR}^2 - 1)}{\rho^2}} = 3167$$

EXAMPLE 3.8

For an AM receiver with RF amplifier loaded to an antenna, the coupling circuit is 100. If IF is 455 kHz, find the image frequency and its rejection ratio at 1000 kHz and at 25 MHz. Also, find the IF to make the image rejection as good as 25 MHz as it would be at 1000 kHz.

Solution

At 1000 kHz:

To find the image frequency:

$$\text{Image frequency } f_{\text{im}} = f_{\text{RF}} + 2f_{\text{IF}}$$

$$= 1000 \times 10^3 + 2\,(455 \times 10^3) = 1.91 \text{ MHz}$$

To find the image frequency rejection ratio:

$$\rho = \frac{1.91 \times 10^6}{1000 \times 10^3} - \frac{1000 \times 10^3}{1.91 \times 10^6}$$

$$= 1.91 - 0.523 = 1.3864$$

$$\text{IFRR} = \sqrt{1 + Q^2 \rho^2}$$

$$= \sqrt{1 + (100)^2\,(1.3864)^2} = 138.64$$

At 25 MHz:

To find the image frequency:

$$\text{Image frequency } f_{\text{im}} = f_{\text{RF}} + 2f_{\text{IF}}$$

$$= 25 \times 10^6 + 2(455 \times 10^3) = 25.91 \text{ MHz}$$

To find the image frequency rejection ratio:

$$\rho = \frac{25.91 \times 10^6}{25 \times 10^6} - \frac{25 \times 10^6}{25.91 \times 10^6}$$

$$= 1.0364 - 0.9648 = 0.0715$$

$$\text{IFRR} = \sqrt{1 + Q^2 \rho^2}$$

$$= \sqrt{1 + (100)^2\,(0.0715)^2} = 7.219$$

To find the IF to make IFRR as good at 25 MHz as it would be at 1000 kHz:

$$\text{IFRR} = \sqrt{1 + Q^2 \rho^2} = 138.64$$

$$\sqrt{1 + (100)^2\,\rho^2} = 138.64$$

$$1 + 100\,\rho^2 = 19209.96$$

$$\rho = 1.386$$

We know that,

$$\rho = \left(f_{\text{im}} \big/ f_{\text{RF}}\right) - \left(f_{\text{RF}} \big/ f_{\text{im}}\right)$$

$$\therefore \qquad \left(f_{\text{im}} \big/ f_{\text{RF}}\right) - \left(f_{\text{RF}} \big/ f_{\text{im}}\right) = 1.386$$

$$f_{\text{im}}{}^2 - f_{\text{RF}}{}^2 = 1.386 \times f_{\text{im}} \cdot f_{\text{RF}}$$

$$f_{\text{im}}{}^2 - (25 \times 10^6)^2 - (1.386 \times 25 \times 10^6)\, f_{\text{im}} = 0$$

$$f_{\text{im}}{}^2 - (34.65 \times 10^6)\, f_{\text{im}} - (6.25 \times 10^{14}) = 0$$

By solving the equation,

$$f_{\text{im}} = \frac{34.65 \times 10^6 \pm \sqrt{(34.65 \times 10^6)^2 - 4 \times 1 \times 6.25 \times 10^{14}}}{2} = 95.483 \text{ MHz}$$

$$\therefore \qquad f_{\text{im}} = f_{\text{RF}} + 2f_{\text{IF}} = 95.483 \text{ MHz}$$

$$25 \times 10^6 + 2\, f_{\text{IF}} = 95.483 \times 10^6$$

$$2f_{\text{IF}} = 95.483 \times 10^6 - 25 \times 10^6 = 70.483 \times 10^6$$

$$\therefore \qquad f_{\text{IF}} = \frac{70.483 \times 10^6}{2} = 35.24 \text{ MHz}$$

EXAMPLE 3.9

Calculate the image-frequency rejection in dB of a double superheterodyne receiver which has a first IF at 2 MHz and a second IF at 200 kHz. An RF amplifier is tuned to the circuit which has Q = 75 and which is turned to 30 MHz.

Solution

$$f_{\text{IF1}} = 2 \text{ MHz}$$

$$f_{\text{IF2}} = 200 \text{ MHz}$$

$$\text{IFRR} = \sqrt{1+Q^2\rho^2}$$

$$\rho_1 = \left(f_{\text{im}}/f_{\text{RF}}\right)-\left(f_{\text{RF}}/f_{\text{im}}\right)$$

$$\rho_1 = \frac{34\times10^6}{30\times10^6}-\frac{30\times10^6}{34\times10^6}=0.251$$

$$\rho_2 = \frac{2.4\times10^6}{2\times10^6}-\frac{2\times10^6}{2.4\times10^6}=0.366$$

$$\text{IFRR} = \sqrt{1+Q^2\,(\rho_1^2+\rho_2^2)}$$

$$\text{IFRR} = \sqrt{1+(75)^2\,(0.25^2+0.366^2)}=33.257$$

In dB,

$$\text{IFRR}_{\text{dB}} = 20 \log_{10} (33.257) = 30.467$$

EXAMPLE 3.10

A superheterodyne receiver having an RF amplifier is tuned to 15 MHz. The IF is 455 kHz. The RF amplifier and preselector are equal and are such that the image rejection is 41.58 dB. Calculate Q.

Solution

$$f_{\text{IF}} = 455 \text{ kHz}$$

$$\text{IFRR} = 41.58 \text{ dB} = 119.9$$

$$\rho_1 = \left(f_{\text{im}}/f_{\text{RF}}\right)-\left(f_{\text{RF}}/f_{\text{im}}\right)$$

$$f_{\text{im}} = f_{\text{RF}} + 2f_{\text{IF}}$$

$$= 15 \times 10^6 + 2(455 \times 10^3) = 15.455 \text{ MHz}$$

$$\therefore \qquad \rho = \frac{15.455 \times 10^6}{15 \times 10^6} - \frac{15 \times 10^6}{15.455 \times 10^6}$$

$$= 1.030 - 0.970 = 0.0594$$

$$\text{IFRR} = \sqrt{1 + Q^2 \rho^2}$$

$$119.9 = \sqrt{1 + Q^2\ 0.0594^2}$$

$$119.9^2 = 1 + Q^2\ 0.0594^2$$

$$14387.9 = 1 + 0.00353\ Q^2$$

$$0.00353\ Q^2 = 14386.9$$

$$\therefore \qquad Q = \sqrt{\frac{14386.9}{0.00353}} = 2018.81$$

8. Advantages of Superheterodyne Receivers over TRF Receivers

1. Improved selectivity in terms of adjacent channels
2. More uniform selectivity over the entire frequency range
3. Improved receiver stability
4. Higher gain per stage because IF amplifiers are operated at a lower frequency
5. Uniform bandwidth because of fixed intermediate frequency

9. Applications of Superheterodyne Receivers

Suitable for radio receiver applications like AM, FM, SSB, communication, television and radar receivers.

3.3.7 Automatic Gain Control (AGC)

Generally, it is observed that the amplitude of the IF carrier signal at the detector input may vary as much as 30 or 40 dB. This results in corresponding variations in reproduction at the receiver output. At a minimum carrier the loudspeaker output becomes inaudible and gets mixed in noise. On the other hand, at the carrier maximum, the output of the loudspeaker becomes very large. The effective and simplest means adopted to overcome this drawback is Automatic Gain Control (AGC). It is also called **Automatic Volume Control (AVC)**. It has been observed that a properly designed AGC system reduces the amplitude variation due to fading from a high value of 30 to 40 dB to a small value of 3 to 4 dB.

1. Principle of AGC

The principle of AGC operation consists of the following two steps.

(a) Derivation of rectification of carrier voltage in a linear diode detector with a dc voltage proportional to the carrier amplitude.

(b) Application of this dc voltage as a reverse biased voltage at the input of the RF amplifier, frequency mixer and the IF amplifier.

Hence if the carrier signal amplitude increases, the AGC bias increases and the gains of all the tuned stages preceding the detector decrease resulting in decrease in carrier amplitude at the input of the detector bringing it back to its normal value. Now if the carrier amplitude decreases due to some reason then the reverse action takes place. Hence, the AGC smoothens out the variations in the carrier amplitude to a very large extent.

2. Simple AGC Circuit

Figure 3.15 shows the circuit of a linear diode detector with simple AGC. In this circuit, the half-wave rectified voltage is developed across the load resistor R. Capacitor C filters the RF components due to which only the dc and the modulating frequency voltage are obtained across the load resistor R.

The dc component is removed from the output by the use of coupling capacitor C_C. AGC is picked up from the diode end of the load resistor R. But since this voltage consists of modulating frequency component as well, therefore an AGC filter consisting of a series resistor R_A and a shunt capacitor C_A is used to remove the modulating frequency component and thus leaving only a positive dc voltage as the required AGC bias.

The time constant of this AGC filter is suitably selected to remove all modulating-frequency components. This time constant R_AC_A must be large enough to remove even the

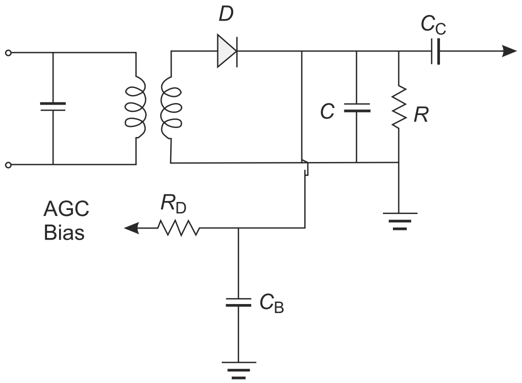


Fig. 3.15 Simple AGC circuit

lowest modulating-frequency component from the AGC bias to follow the change in carrier amplitude. A typical time constant of AGC filter is in the range of 0.1 to 0.2 second. Now this positive AGC bias is applied at the base of *PNP* transistors of preceding tuned stages. This bias then reduces the net forward bias at the emitter junction, thereby reducing the gain of the amplifier. However, in the case of *NPN* transistors, a negative AGC bias is applied at the bases of transistors of preceding tuned stages. In this case, the detector circuit is similar to that shown in Figure 3.11 except that the polarity of the diode is reversed.

3. AGC Circuit with π-Filter

To provide the better removal of RF components from the modulating-frequency output, a π-filter is used in place of a simple capacitor filter. Figure 3.16 shows the circuit diagram of a linear diode detector with π-filter and simple AGC. A manual volume control is also provided by using a variable resistor at the input of the first audio amplifier.

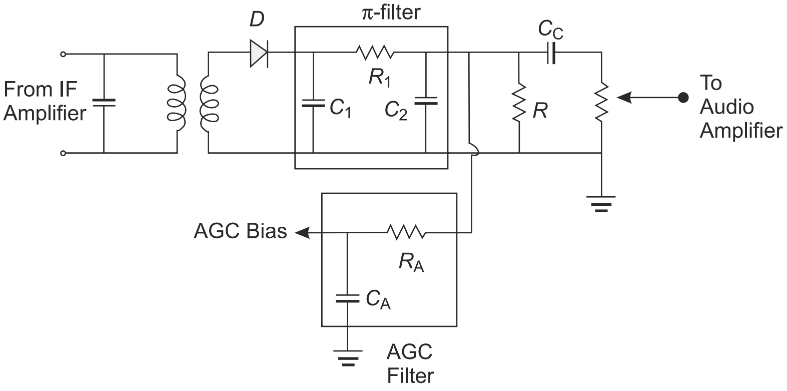


Fig. 3.16 AGC circuit with π-Filter

4. Delayed AGC

The above simple AGC systems suffer from a drawback that the AGC becomes operative even for very weak signals. The resultant is that the receiver gain starts decreasing as soon as detector diode starts producing the output. On the other hand, an ideal AGC system must remain inoperative till the input carrier voltage reaches a reasonable large predetermined voltage. Subsequently, the AGC must come into 0 (zero) to maintain output level constant instead of variation in input level of carrier voltage. If a delay is produced in AGC operation then it serves some purpose. Figure 3.17 shows an arrangement to produce delay in AGC operation. With zero and small signal voltages, the diode D_2 conducts due to which the AGC bias just equals the potential of cathode of this diode. Hence, AGC remains fixed at a low positive value. As the input carrier voltage increases, the AGC bias produced due to rectification of carrier voltage in the detector diode D_1 increases. Also, when this rectified

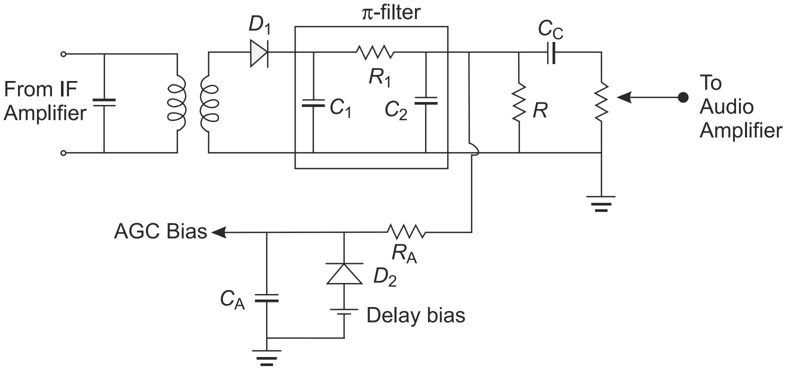


Fig. 3.17 Linear diode detector with delayed AGC

bias magnitude exceeds the magnitude of the positive cathode voltage of the diode D_2, then D_2 stops to conduct and the AGC system works normally.

5. Amplified and Delayed AGC

If the delayed AGC bias is amplified before application as reverse bias to the tuned amplifiers, the AGC behaviour on characteristics closely approaches the ideal delayed AGC. Figure 3.18 shows the circuit diagram of delayed AGC with additional IF amplifier.

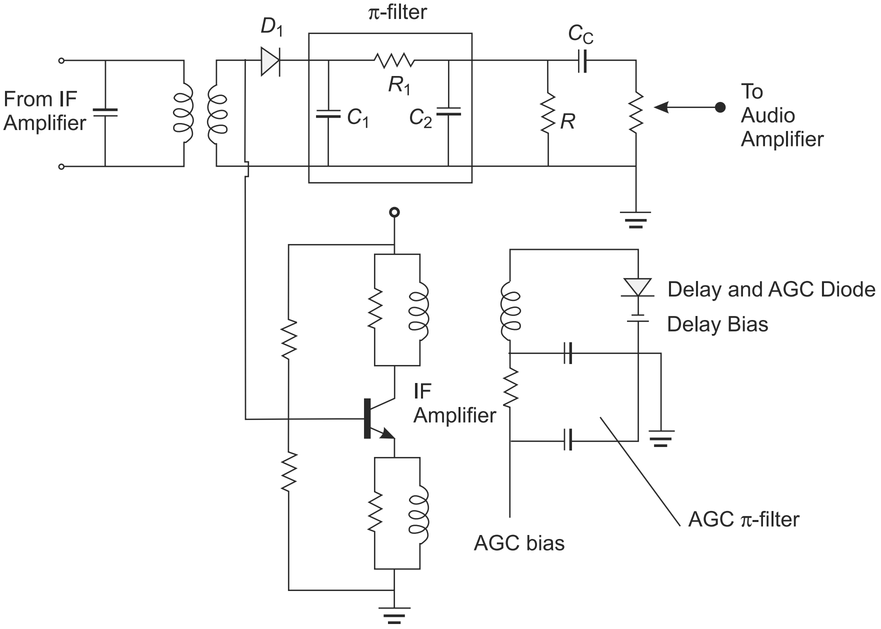


Fig. 3.18 Linear diode detector with amplified and delayed AGC

6. AGC Characteristics

The following graph shown in Figure 3.19 shows the characteristics of various categories of AGC.

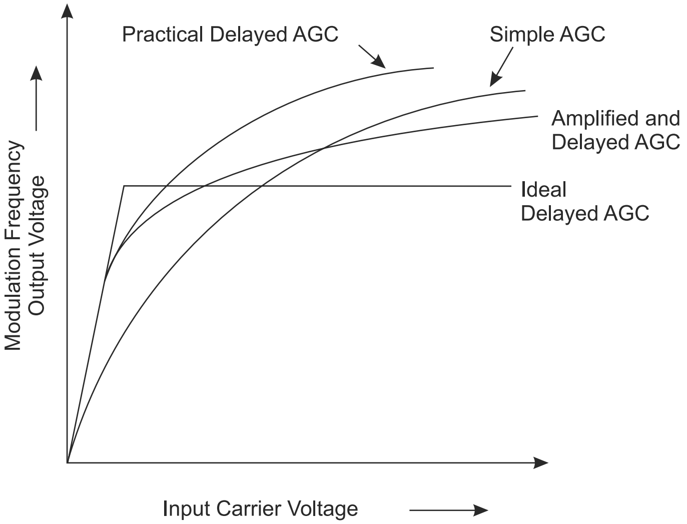


Fig. 3.19 AGC characteristics

Summary

AM demodulators are circuits that accept amplitude-modulated signals and recover the original modulating signal or message signal. The basic detector is the diode detector. Because the diode detector recovers the envelope of the AM signal, which is the original information signal, the circuit is also called an envelope detector. It is also useful in demodulation of VSB signals.

In a rectifier detector, the percentage efficiency is 31.8% and it also produces only one third of the input signal at the output, the rectifier detector is not very efficient. It also produces distortions at the output.

In a Costas detector for DSB-SC AM, by combining the *I* and *Q* channel outputs in phase discrimination, a dc signal is obtained that automatically corrects the phase errors in VCO. It is apparent that phase control in the Costas detector ceases with modulation and that phase lock has to be reestablished.

Though various forms of receiver circuits have been developed, two types of receivers have real significance.

1. Tuned Radio Frequency (TRF) receivers
2. Superheterodyne receivers

Due to the drawbacks such as inability in gain, reception of stations of frequencies adjacent to the desired signal and variation in the bandwidth over the band, the TRF receiver needs to be replaced by a superheterodyne receiver.

REVIEW QUESTIONS

PART-A

1. What is demodulation?
2. State the principle of a synchronous detector.
3. List any two AM detectors.
4. Define heterodyning principle.
5. Mention the type of detector used for DSB-SC demodulation.
6. What are the functions of a radio receiver?
7. How will you classify receivers based on principle?
8. How will you classify receivers based on application?
9. List out the features of AM broadcast receivers.
10. What do you mean by fidelity of a receiver?
11. What do you mean by selectivity of a receiver?
12. What are the major functions of AM broadcast receivers?
13. What are the drawbacks of TRF receivers?
14. How will you overcome the drawbacks of TRF receivers?
15. What are the functions of RF amplifiers in a receiver?
16. What is frequency conversion in a superheterodyne receiver?
17. Define intermediate frequency.
18. What is image frequency?
19. Define image-frequency rejection ratio.
20. List the advantages of superheterodyne receivers.
21. List the applications of superheterodyne receivers.
22. Define automatic gain control and mention its importance.
23. Why do you prefer delayed AGC?

PART-B

1. With a neat diagram, explain the operation of an envelope detector.
2. Explain the operation of a diode rectifier detector with a neat diagram.
3. Draw the block diagram of a synchronous detector and explain. Give the details on a practical synchronous diode detector.
4. Explain the functioning of a Costas detector for demodulation of DSB-SC AM.
5. Draw the block diagram of a tuned radio-frequency receiver and explain the functioning of each block. State its drawbacks also.
6. Explain the functioning of a superheterodyne receiver with a neat block diagram.
7. What do you mean by image frequency? How will you perform image-frequency rejection ratio in an AM superheterodyne receiver?
8. What function is served by having an intermediate frequency section in an AM superheterodyne receiver? What is a local oscillator and what is its function in an AM receiver?
9. An AM receiver is tuned to a station whose carrier frequency is 750 kHz. What frequency should the local oscillator be set to provide an intermediate frequency of 455 kHz if the local oscillator frequency tracks below the received frequency? If it tracks above?
10. Repeat the above problem for a station having a carrier frequency of 1200 kHz.
11. With a neat circuit diagram, explain the functioning of a linear diode detector.
12. Describe the functioning of delayed AGC with a neat circuit diagram.

4

ANGLE MODULATION

Objectives

- To know about frequency modulation and phase modulation with their frequency spectrum, bandwidth calculation, power calculation and vector representation
- To discuss the differences between AM and FM, and FM and PM
- To provide the process of FM generation by two methods such as direct method and indirect methods in detail.
- To discuss about two basic types of FM transmitters such as directly modulated FM transmitter and indirectly modulated FM transmitter using PM

4.1 INTRODUCTION

Angle modulation is modulation in which the angle of a sine-wave carrier is varied by a modulating wave. There are two types of angle modulation.

1. Frequency Modulation (FM)
2. Phase Modulation (PM)

In **frequency modulation**, the modulating signal causes the carrier frequency to vary. These variations are controlled by both the frequency and the amplitude of the modulating wave. In **phase modulation,** the phase of the carrier is controlled by the modulating waveform.

The amplitude of the carrier wave remains constant in an FM process. This is an advantage over amplitude modulation, because all natural, internal and external noises consist of amplitude variations. The receiver cannot distinguish between the amplitude variations that represent noise and those that represent the desired signal. So amplitude modulation is generally noisy than frequency modulation. Since the amplitude of the wave remains constant, the power associated with an FM wave is constant.

4.2 FREQUENCY MODULATION

4.2.1 Definition and Representation

Frequency modulation is the process by which the frequency of the carrier signal is changed in accordance with the instantaneous amplitude of the modulating signal. The amplitude of the carrier wave is maintained constant in this process.

The number of times per second that the instantaneous frequency is varied from the carrier frequency is controlled by the frequency of the modulating signal. The amount by which the frequency departs from the average is controlled by the amplitude of the modulating signal. This variation is referred to as the frequency deviation of the frequency-modulated wave. Two important points to be noted are

1. Amount of frequency shift is proportional to the amplitude of the modulating signal
2. Rate of frequency shift is proportional to the frequency of the modulating signal

Let the message signal and carrier signal be,

$$V_{\mathrm{m}}(t) = V_{\mathrm{m}} \sin \omega_{\mathrm{m}} t, \text{ and} \tag{4.1}$$

$$V_{\mathrm{c}}(t) = V_{\mathrm{c}} \sin (\omega_{\mathrm{c}} t + \theta) \tag{4.2}$$

V_{m} is the maximum amplitude of the modulating signal.

V_{c} is the maximum amplitude of the carrier signal.

ω_{m} is the angular frequency of modulating signal.

ω_{c} is the angular frequency of carrier signal.

φ is the total instantaneous phase angle of carrier $[\phi = (\omega_{\mathrm{c}} t + \theta)]$.

θ is the initial phase angle.

$$\therefore \quad V_{\mathrm{c}}(t) = V_{\mathrm{c}} \sin \phi \tag{4.3}$$

The angular velocity may be determined by finding the rate of change of this phase angle. That is,

$$\omega_{\mathrm{c}} = \frac{d\phi}{dt} \tag{4.4}$$

During the process of frequency modulation, the frequency of the carrier signal is changed in accordance with the instantaneous amplitude of the modulating signal. Then the frequency of the carrier after modulation is

$$\omega_{\mathrm{i}} = \omega_{\mathrm{c}} + KV_{\mathrm{m}}(t) = \omega_{\mathrm{c}} + KV_{\mathrm{m}} \cos \omega_{\mathrm{m}} t \tag{4.5}$$

where K is a constant of proportionality.

Maximum frequency deviation occurs when the cosine terms in Equation (4.5) has a value ±1.

Under this condition, the instantaneous angular velocity is given by

$$\omega_i = \omega_c \pm KV_m \tag{4.6}$$

So that the maximum frequency deviation Δf is given by

$$\Delta f = \frac{KV_m}{2\pi} \tag{4.7}$$

This gives $2\pi\ \Delta f = KV_m$

Equation (4.6) may be rewritten as

$$\omega_i = \omega_c + 2\pi\ \Delta f \cos \omega_m t \tag{4.8}$$

Integration of the above equation gives the phase angle of the frequency-modulated wave.

$$\begin{aligned} \therefore\quad \phi_i &= \int \omega_i\, dt \\ &= \int [\omega_c + 2\pi\ \Delta f \cos \omega_m t]\, dt \\ &= \omega_c t + \frac{2\pi\ \Delta f}{\omega_m} \sin \omega_m t + \theta_i \end{aligned} \tag{4.9}$$

where θ_i is a constant of integration representing a constant phase angle and may be neglected since it has no importance in the modulation process.

The instantaneous amplitude of the modulating signal is given by

$$\begin{aligned} V(t)_{\text{FM}} &= V_c \sin \phi_i \\ &= V_c \sin \left(\omega_c t + \frac{2\pi\ \Delta f}{\omega_m} \sin \omega_m t \right) \\ &= V_c \sin (\omega_c t + m_f \sin \omega_m t) \end{aligned} \tag{4.10}$$

where

$$m_f = \frac{2\pi\ \Delta f}{\omega_m} = \frac{KV_m}{\omega_m} = \text{Modulation index of FM}$$

$$\begin{aligned} V(t)_{\text{FM}} &= V_c \sin (\omega_c t + m_f \sin \omega_m t) \\ &= V_c [\sin \omega_c t \,.\, \cos (m_f \sin \omega_m t) + \cos \omega_c t \sin (m_f \sin \omega_m t)] \end{aligned} \tag{4.11}$$

The graphical representation of an FM wave is shown in Figure 4.1.

4.2.2 Frequency Spectrum of an FM Wave

Different frequency components in an FM wave can be determined in the same way as followed in AM, i.e. by expanding the expression for the waves, which involves the use of Bessel functions.

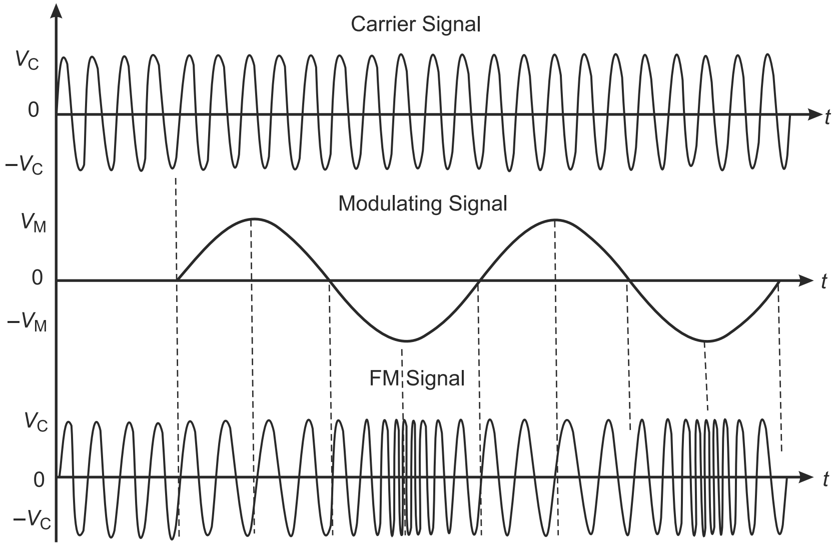


Fig. 4.1 Graphical representation of FM wave

$$V(t)_{FM} = V_c\,[\sin \omega_c t\,.\,\cos(m_f \sin \omega_m t) + \cos \omega_c t \sin (m_f \sin \omega_m t)]$$

$$\begin{aligned} &= V_c\,[J_0\,(m_f) \sin \omega_c t + J_1\,(m_f)\,\{\sin (\omega_c + \omega_m)t - \sin (\omega_c - \omega_m)t\} \\ &\quad + J_2\,(m_f)\,\{\sin (\omega_c + 2\omega_m)t - \sin (\omega_c - 2\omega_m)t\} \\ &\quad + J_3\,(m_f)\,\{\sin (\omega_c + 3\omega_m)t - \sin (\omega_c - 3\omega_m)t\} \\ &\quad + J_4\,(m_f)\,\{\sin (\omega_c + 4\omega_m)t - \sin (\omega_c - 4\omega_m)t\} + ...] \end{aligned} \tag{4.12}$$

where J_0, J_1, J_2, J_3... are the coefficients of zero, first, second and higher orders for the Bessel function and the order of the coefficient is denoted by the subscript m_f.

The Bessel function can be written as

$$J_n(m_f) = \left(\frac{m_f}{2}\right)^n \left[\frac{1}{n!} - \frac{\left({}^{m_f}\!/\!_{2}\right)^2}{1!(n+1)!} + \frac{\left({}^{m_f}\!/\!_{2}\right)^4}{2!(n+2)!} - ...\right] \tag{4.13}$$

The plot of a Bessel function is shown in Figure 4.2.

In order to know the amplitude of the carrier and the amplitude of the side bands, it is necessary to know the value of corresponding Bessel function $J_n(m_f)$ and multiply it with V_C. For example, an FM wave with a maximum deviation $\Delta f = \pm 75$ kHz and a maximum audio frequency of 15 kHz has a modulation index $m_f = 5$. The wave has a total of 8 upper side bands and an equal number of lower side bands. The magnitude of the eigth side band is only 2% of the carrier amplitude.

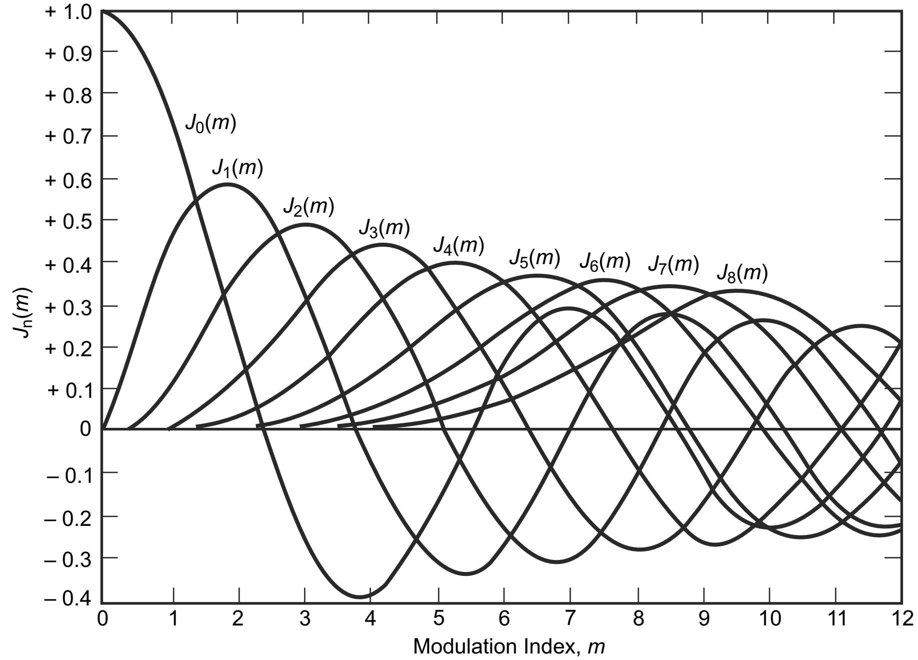


Fig. 4.2 A plot of Bessel function

The FM wave contains an infinite number of side bands; each side band is separated from the next by the modulating frequency f_m. However, out of these, there are only a few side bands which carry a significant amount of power. The remaining side bands have such a low power that they get attenuated during propagation and do not convey any message to the receiver. In practice, it is to be notified that all the side bands having amplitudes greater than 55 of the carrier as significant side bands and neglect all the remaining side bands with amplitude less than 5% of the carrier. An increase in the modulating-frequency signal amplitude at the transmitter results in larger frequency deviation and as a consequence in a larger bandwidth of a modulated signal.

The amplitude of FM remains constant; hence the power of the FM is same as that of the unmodulated carrier. The total power of the FM signal depends upon the power of the unmodulated carrier, whereas in AM, the total power depends on the modulation index. In AM, the increased modulation index increases the side-band power. But in FM, the total power remains constant with increased modulation index and only the bandwidth is increased.

The frequency spectrum of an FM wave for $m_f = 0.5$ and $m_f = 5$ is shown in Figure 4.3.
The upper frequency reached is equal to the rest or carrier frequency plus the frequency deviation.

$$f_H = f_c + \Delta f \tag{4.14}$$

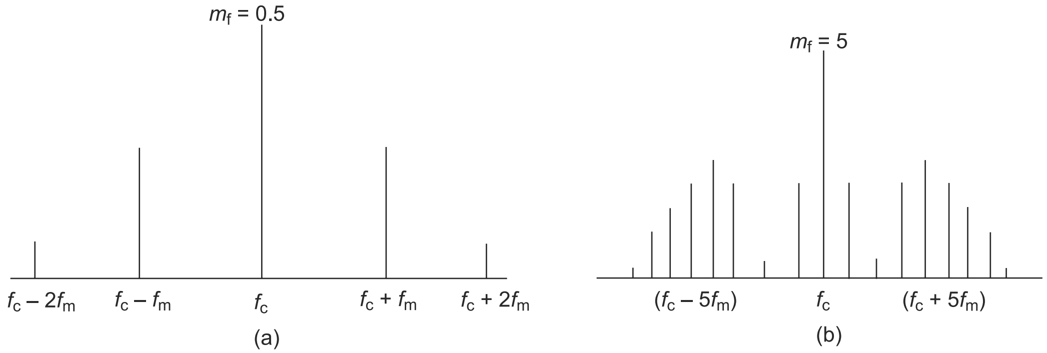


Fig. 4.3 Frequency Spectrum of FM wave with (a) $m_f = 0.5$, and (b) $m_f = 5$

The lower frequency reached by the modulated wave is equal to the rest or carrier frequency minus the frequency deviation.

$$f_L = f_c - \Delta f \tag{4.15}$$

The modulation index is now calculated by

$$m_f = \frac{\Delta f}{f_m} \tag{4.16}$$

The percent modulation for an FM wave is then calculated as

$$M = \frac{\Delta f_{\text{actual}}}{\Delta f_{\text{max}}} \times 100 \tag{4.17}$$

EXAMPLE 4.1

A 107.6 MHz carrier is frequency modulated by a 7 kHz sine wave. The resultant FM signal has a frequency deviation of 50 kHz. a). Find the carrier swing of the FM signal. b) Determine the highest and lowest frequencies attained by the modulated signal. c). What is the modulation index of FM?

Solution

a). *Relating carrier swing to frequency deviation*

$$\text{Carrier swing} = 2\Delta f$$

$$= 2 \times 50 \times 10^3 = 100 \text{ kHz}$$

b). *The upper frequency reached is equal to the rest or carrier frequency plus the frequency deviation.*

$$f_H = f_c + \Delta f$$

$$= 107.6 \times 10^6 + 50 \times 10^3 = 107.65 \text{ MHz}$$

The lower frequency reached by the modulated wave is equal to the rest or carrier frequency minus the frequency deviation.

$$f_L = f_c + \Delta f$$

$$= 107.6 \times 10^6 - 50 \times 10^3 = 107.55 \text{ MHz}$$

c). The modulation index

$$m_f = \frac{\Delta f}{f_m}$$

$$= \frac{50 \times 10^3}{7 \times 10^3} = 7.143$$

EXAMPLE 4.2

Determine the frequency deviation and carrier swing for a frequency-modulated signal which has a resting frequency of 105.000 MHz and whose upper frequency is 105.007 MHz when modulated by a particular wave. Find the lowest frequency reached by the FM wave.

Solution

Frequency deviation is the maximum change in frequency of the modulated signal away from the rest or carrier frequency.

$$\Delta f = (105.007 - 105.000) \times 10^6$$

$$= 0.007 \times 10^6 = 7 \text{ kHz}$$

$$\text{Carrier swing} = 2\ \Delta f$$

$$= 2 \times 7 \times 10^3 = 14 \times 10^3 = 14 \text{ kHz}$$

The lowest frequency can be determined as

$$f_L = f_c - \Delta f$$

$$= (105.000 - 0.007) \times 10^6 = 104.993 \text{ MHz}$$

EXAMPLE 4.3

What is the modulation index of an FM signal having a carrier swing of 100 kHz when the modulating signal has a frequency of 8 kHz?

Solution

$$\text{Frequency deviation } \Delta f = \frac{\text{Carrier swing}}{2}$$

$$= \frac{100 \times 10^3}{2} = 50 \text{ kHz}$$

$$\text{Modulation index } m_{\text{f}} = \frac{\Delta f}{f_{\text{c}}}$$

$$= \frac{50 \times 10^3}{8 \times 10^3} = 6.25$$

EXAMPLE 4.4

A frequency-modulated signal which is modulated by a 3 kHz sine wave reaches a maximum frequency of 100.02 MHz and minimum frequency of 99.98 MHz. a). Determine the carrier swing b). Find the carrier frequency. c). Calculate the frequency deviation of the signal. d). What is the modulation index of the signal?

Solution

$$\text{Carrier swing} = f_{\text{max}} - f_{\text{min}}$$

$$= 100.02 \times 10^6 - 99.98 \times 10^6$$

$$= 0.04 \times 10^6 = 40 \text{ kHz}$$

$$\text{Carrier frequency } f_{\text{c}} = \frac{f_{\text{max}} + f_{\text{min}}}{2}$$

$$f_{\text{c}} = \frac{100.02 \times 10^6 + 99.98 \times 10^6}{2}$$

$$= 100 \times 10^6 = 100 \text{ MHz}$$

$$\text{Frequency deviation } \Delta f = \frac{\text{Carrier swing}}{2}$$

$$\Delta f = \frac{40 \times 10^3}{2}$$

$$\Delta f = 20 \text{ kHZ}$$

$$\text{Modulation index} \quad m_{\text{f}} = \frac{\Delta f}{f_{\text{c}}}$$

$$= \frac{20 \times 10^3}{3 \times 10^3} = 6.667$$

EXAMPLE 4.5

An FM transmission has a frequency deviation of 20 kHz. (a) Determine the percent modulation of this signal if it is broadcast in the 88–108 MHz band. (b) Calculate the percent modulation if this signal were broadcast as the audio portion of a television broadcast.

Solution

(a) The percent modulation for an FM wave if the maximum frequency deviation in the FM broadcast band is 75 kHz.

$$M = \frac{\Delta f_{\text{actual}}}{\Delta f_{\text{max}}} \times 100$$

$$= \frac{20 \times 10^3}{75 \times 10^3} \times 100 = 26.67\%$$

(b) The percent modulation for an FM wave if the maximum frequency deviation in the FM broadcast band is 25 kHz.

$$M = \frac{20 \times 10^3}{25 \times 10^3} \times 100 = 80\%$$

EXAMPLE 4.6

What is the frequency deviation and carrier swing necessary to provide 75% modulation in the FM broadcast band?

Solution

Percent modulation of FM signal

$$M = \frac{\Delta f_{\text{actual}}}{\Delta f_{\text{max}}} \times 100$$

The maximum frequency deviation permitted in the FM broadcast is 75 kHz.

$$75 = \frac{\Delta f_{\text{FM}}}{75 \times 10^3} \times 100$$

$$\Delta f_{\text{FM}} = \frac{75 \times 75 \times 10^3}{100} = 56.25 \text{ kHz}$$

Carrier swing is related to frequency deviation by

$$\text{Carrier swing} = 2\Delta f$$

$$= 2 \times 56.25 \times 10^3 = 112.5 \text{ kHz}$$

EXAMPLE 4.7

Determine the percent modulation of an FM signal which is being broadcast in the 88–108 MHz band, having a carrier swing of 125 kHz.

Solution

$$\Delta f = \frac{\text{Carrier swing}}{2}$$

$$= \frac{125 \times 10^3}{2} = 62.5 \text{ kHz}$$

The percent modulation for an FM wave is

$$M = \frac{\Delta f_{\text{actual}}}{\Delta f_{\text{max}}} \times 100$$

$$M = \frac{62.5 \times 10^3}{75 \times 10^3} \times 100 = 83.3\%$$

EXAMPLE 4.8

The percent modulation of the sound portion of a TV signal is 80%. Determine the frequency deviation and the carrier swing of the signal.

Solution Percent modulation of FM signal

$$M = \frac{\Delta f_{\text{actual}}}{\Delta f_{\text{max}}} \times 100$$

$$80 = \frac{\Delta f_{\text{actual}}}{25 \times 10^3} \times 100$$

$$\Delta f_{\text{actual}} = \frac{80 \times 25 \times 10^3}{100} = 20 \text{ kHz}$$

Carrier swing is related to frequency deviation by

$$\text{Carrier swing} = 2\Delta f$$

$$= 2 \times 20 \times 10^3 = 40 \text{ kHz}$$

4.2.3 Bandwidth Calculation in FM

The bandwidth of frequency modulation can be determined by using Bessel table and is defined as

$$B = 2(n \times f_{\text{m}}) \text{ Hz} \tag{4.18}$$

where

n is the number of significant sidebands, and

f_{m} is the modulating signal frequency (Hz).

By using Carson's rule, the bandwidth of frequency modulation is defined as

$$B = 2(\Delta f + f_{\text{m}}) \text{ Hz} \tag{4.19}$$

and carrier swing is denoted as $2\Delta f$

where

Δf is the Peak frequency deviation (Hz), and

f_m is the Modulating signal frequency (Hz).

Such a system is termed **narrowband FM**. It finds its use in police, defence, fire services, etc. with frequency deviation lying in the range of 15–25 kHz.

For values of modulation index $m_f < 0.6$, the side bands produce a wide frequency spectrum but their amplitudes decrease. In such cases, the number of significant side bands increases and the system is referred as wideband FM.

Wideband FM requires a considerably larger bandwidth as compared to the corresponding AM wave. When $m_f > 5$, the Bessel coefficients $J_0(m_f)$ diminish rapidly and the system bandwidth to an approximation may be given by

$$B = 2 \,.\, m_\mathrm{f} \,.\, f_\mathrm{m} \text{ Hz} \tag{4.20}$$

Such a large bandwidth leads to improvement in the quality of reception, with very low noise as compared to AM. Narrowband FM does not possess this characteristic due to small bandwidth, frequency deviation, bandwidth, etc.

EXAMPLE 4.9

If a 6 MHz band were being considered for use with the same standards that apply to the 88–108 MHz band, how many FM stations could be accommodated?

Solution

Each station requires a total bandwidth of 400 kHz.

$$\text{Number of channels} = \frac{6\times10^6}{400\times10^3} = 15$$

EXAMPLE 4.10

Determine the bandwidth of a narrowband FM signal which is generated by a 4 kHz audio signal modulating a 125 MHz carrier.

Solution

$$\begin{aligned}\text{Bandwidth} &= 2f_\mathrm{c}\\ &= 2\times4\times10^3 = 8 \text{ kHz}\end{aligned}$$

EXAMPLE 4.11

A 2 kHz audio signal modulates a 50 MHz carrier, causing a frequency deviation of 2.5 kHz. Determine the bandwidth of the FM signal.

Solution

$$\text{Modulation index} \quad m_f = \frac{\Delta f}{f_c}$$

$$= \frac{2.5 \times 10^3}{2 \times 10^3} = 1.25$$

Since this is less than π/2, a narrowband signal is to be dealt. Thus,

$$\text{Bandwidth} = 2f_c$$

$$= 2 \times 2 \times 10^3 = 4 \text{ kHz}$$

EXAMPLE 4.12

In a FM wave, the frequency deviation is 25 kHz. What is the phase deviation when the modulating signal frequency is 10 kHz?

Solution

$$\text{Modulation index} \quad m_f = \frac{\Delta f}{f_c}$$

$$= \frac{25 \times 10^3}{10 \times 10^3} = 2.5$$

EXAMPLE 4.13

The carrier frequency of a broadcast signal is 100 MHz. The maximum frequency deviation is 75 kHz. If the highest audio frequency modulating the carrier is limited to 15 kHz, what is the approximate bandwidth?

Solution

$$\text{Bandwidth} = 2m_f f_m$$

$$\text{Modulation index} = m_f = \frac{\Delta f}{f_c}$$

$$= \frac{75 \times 10^3}{15 \times 10^3} = 5$$

$$\text{Bandwidth} = 2 \times 5 \times 15 \text{ kHz} = 150 \text{ kHz}$$

$$\text{Overall bandwidth} = (150 + 25 + 25) \text{ kHz} = 200 \text{ kHz}$$

EXAMPLE 4.14

The maximum deviation allowed in an FM broadcast system is 75 kHz. If the modulating signal is of 10 kHz, find the bandwidth of the FM signal. Find the bandwidth when the modulating frequency is doubled.

Solution

Given that

$$\Delta f = 75 \text{ kHz and } f_m = 10 \text{ kHz}$$

$$BW = 2(\Delta f + f_m)$$

$$= 2(75 + 10) = 170 \text{ kHz}$$

When the modulating frequency is doubled,

$$\text{Bandwidth} = 2(75 + 20) = 190 \text{ kHz}$$

4.2.4 Power Calculation in FM

In frequency modulation, power of an angle-modulated wave is equal to the power of unmodulated carrier. The power from an unmodulated carrier signal is redistributed among carrier and side bands.

The average power of angle modulation is defined as

$$P_C = \frac{V_C^2}{2R} \tag{4.21}$$

where

V_c is the peak voltage of the unmodulated carrier, and

R is the load resistance.

The instantaneous power for angle modulation is defined as

$$P_C = \frac{m(t)^2}{R}$$

$$P_C = \frac{V_C^2}{R}\left[\frac{1}{2} + \frac{1}{2}\cos(2\omega_c t + 2\theta(t)\right] \tag{4.22}$$

Average power of the second term is zero. Thus,

$$P_C = \frac{V_C^2}{2R} \tag{4.23}$$

The total power for a modulated wave is defined as follows.

$$P_t = P_0 + P_1 + P_2 + \ldots + P_n \tag{4.24}$$

$$P_t = \frac{V_c^2}{2R} + \frac{2(V_1)^2}{2R} + \frac{2(V_2)^2}{2R} + \cdots + \frac{2(V_n)^2}{2R} \tag{4.25}$$

where

P_0 is the modulated carrier power,

P_1 is the power in the first set of sidebands,

P_2 is the power in the second set of sidebands, and

P_n is the power in the nth set of sidebands.

EXAMPLE 4.15

Determine the unmodulated carrier power for the FM modulator. Assume a load resistance of R_L= 50 Ω. Determine the total power in the angle-modulated wave.

Solution

Unmodulated carrier power $P_C = \frac{10^2}{2(50)} = 1$ W

Total power in the angle-modulated wave is

$$P_T = \frac{7.7^2}{2(50)} + \frac{2(4.4)^2}{2(50)} + \frac{2(1.1)^2}{2(50)} + \frac{2(0.2)^2}{2(50)}$$

$$= 0.5929 + 0.3872 + 0.0242 + 0.0008$$

$$= 1.0051 \text{ W}$$

4.2.5 Vector Representation of FM Wave

Any alternating wave having a frequency f_c and amplitude V may be represented as a rotating vector **OA** with **OA** representing magnitude and rotating as a constant angular velocity $\omega_{c.}$ If its frequency is increased or decreased, the rotation of the phasor will be faster or slower than ω_c and the phasor would advance or retard to position **OB** or **OC** from the position **OA**. However, it will maintain a constant magnitude. Such a situation is depicted in Figure 4.4. The result is variation in phase angle φ of the wave.

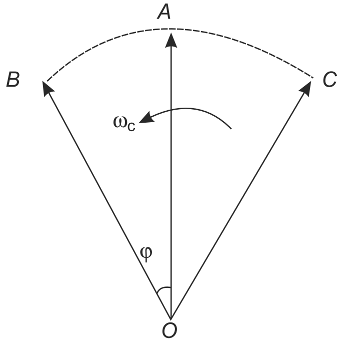


Fig. 4.4 Vector representation of an FM wave

4.2.6 Comparison between Amplitude Modulation and Frequency Modulation

Table 4.1 shows the differences between Amplitude Modulation (AM) and Frequency Modulation (FM).

Table 4.1 Differences between AM and FM

S. No.	Amplitude Modulation (AM)	Frequency Modulation (FM)
1	Amplitude of the carrier is varied according to amplitude of modulating signal	Frequency of the carrier is varied according to amplitude of modulating signal
2	Poor fidelity due to narrow bandwidth	Due to the large bandwidth, fidelity is better
3	Most of the power is in the carrier, hence it is less efficient	All the transmitted power is useful, hence it is more efficient
4	More noise interference	Less noise interference
5	Adjacent channel interference present	Adjacent channel interference is avoided due to guard bands
6	AM broadcast operates in MF and HF ranges	FM broadcast operates in VHF and UHF ranges
7	In AM, only carrier and two side bands are present	Infinite number of side bands are present
8	Transmission equipment is simple	Transmission equipment is complex
9	Transmitted power varies according to modulation index	Transmitted power remains constant irrespective of modulation index
10	Depth of modulation cannot be increased above 1	Depth of modulation has no limitation, it can be increased by increasing frequency deviation.

4.3 PHASE MODULATION

4.3.1 Definition and Representation

Phase modulation is defined as the process by which the phase of the carrier signal can be changed in accordance with the instantaneous amplitude of the modulating signal. The frequency and amplitude remains constant after the modulation process.

Let the modulating signal and the carrier signal are

$$V_{\mathrm{m}}(t) = V_{\mathrm{m}} \cos \omega_{\mathrm{m}} t \tag{4.26}$$

$$V_{\mathrm{c}}(t) = V_{\mathrm{c}} \sin (\omega_{\mathrm{c}} t + \theta) \tag{4.27}$$

where θ is the phase angle of the carrier signal.

$$\theta = KV_{\mathrm{m}}(t) = KV_{\mathrm{m}} \cos \omega_{\mathrm{m}} t \tag{4.28}$$

If the carrier signal varies sinusoidally with the modulating signal then the phase-modulated wave is represented by

$$V(t)_{\text{PM}} = V_c \sin(\omega_c t + \theta) \tag{4.29}$$

$$V(t)_{\text{PM}} = V_c \sin(\omega_c t + KV_m \cos \omega_m t)$$

$$V(t)_{\text{PM}} = V_c \sin(\omega_c t + m_p \cos \omega_m t) \tag{4.30}$$

where $m_p = KV_m$ = Modulation index of phase modualtion

Equation (4.25) representing the phase-modulated wave may be expanded in a way similar to FM. This gives

$$V_0 = V[J_0(m_p) \sin \omega_c t + J_1(m_p) \sin(\omega_c + \omega_m)t - \sin(\omega_c - \omega_m)t$$
$$+ J_2(m_p) \sin(\omega_c + 2\omega_m)t + \sin(\omega_c - 2\omega_m)t \tag{4.31}$$
$$+ \ldots$$

where m_p is termed **modulation index** for phase modulation and equals $\Delta\theta$. The PM waves, like FM waves, have identical frequency spectrum. Figure 4.5 shows the graphical representation of a PM wave.

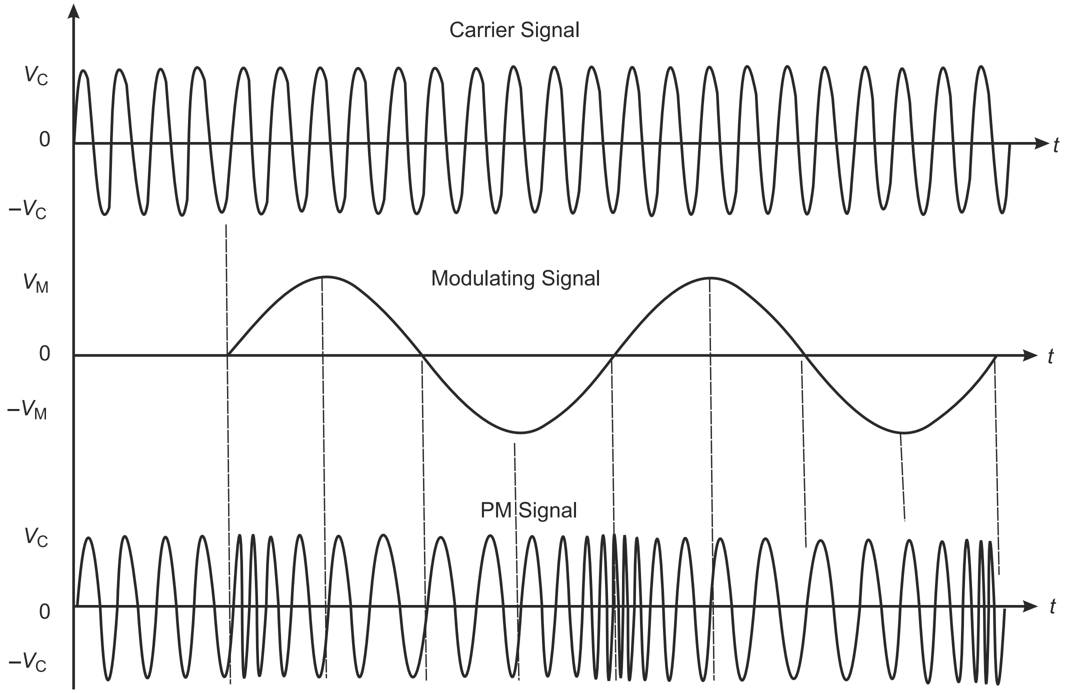


Fig. 4.5 Graphical representation of a PM wave

In PM, $\Delta\theta$ is given a fixed maximum value so that as the modulating frequency f_m varies, the frequency deviation Δf also varies and $\Delta\theta = m_{\text{p}} = \dfrac{\Delta f}{f_{\text{m}}}$ remains constant. This is different from FM in which Δf is constant and can be given a large value.

It is to be noted that FM and PM are closely related and may be termed angle modulation, because in both cases, the changes in phase angle as well as frequency of the modulated carrier take place.

4.3.2 Conversion of PM to FM

A frequency-modulated wave can be obtained from phase modulation. This is done by integrating the modulating signal before applying it to the phase modulator, which is shown in Figure 4.6.

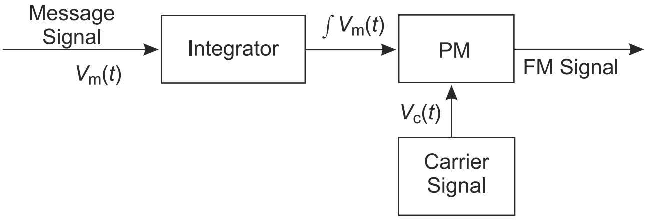


Fig. 4.6 Conversion of PM to FM

The modulating signal in PM is

$$V_{\text{m}}(t) = V_{\text{m}} \cos \omega_{\text{m}} t \tag{4.32}$$

This signal is fed to an integrator and after integration of the modulating signal,

$$V_{\text{m}}(t) = \int V_{\text{m}} \cos \omega_{\text{m}} t \,.\, dt$$

$$= \frac{V_{\text{m}}}{\omega_{\text{m}}} \sin \omega_{\text{m}} t \tag{4.33}$$

In phase modulation,

$$\theta \propto V_{\text{m}}(t)$$

$$\theta = KV_{\text{m}}(t) = \frac{KV_{\text{m}}}{\omega_{\text{m}}} \sin \omega_{\text{m}} t$$

The instantaneous value of the modulated voltage is given by

$$V(t)_{\text{FM}} = V_{\text{c}} \sin (\omega_{\text{c}} t + \theta) \tag{4.34}$$

$$V(t)_{\text{FM}} = V_{\text{c}} \sin \left(\omega_{\text{c}} t + \frac{KV_{\text{m}}}{\omega_{\text{m}}} \sin \omega_{\text{m}} t \right)$$

$$m_\mathrm{f} = \frac{\Delta f}{f_\mathrm{m}} = \frac{KV_\mathrm{m}}{\omega_\mathrm{m}} \sin \omega_\mathrm{m} t$$

$$\therefore \qquad V(t)_\mathrm{FM} = V_\mathrm{c} \sin \left(\omega_\mathrm{c} t + \frac{KV_\mathrm{m}}{\omega_\mathrm{m}} \sin \omega_\mathrm{m} t \right) \tag{4.35}$$

This equation is the expression for FM wave.

4.3.3 Conversion of FM to PM

The PM wave can be obtained from FM by differentiating the modulating signal before applying it to the frequency modulator circuit, which is shown in Figure 4.7.

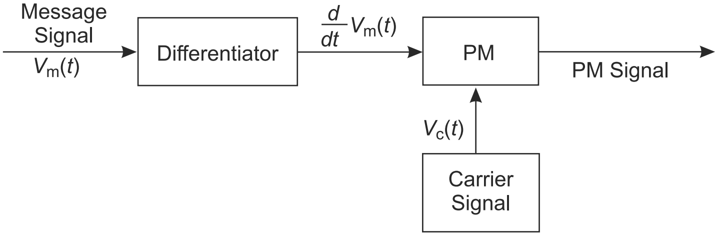


Fig. 4.7 Conversion of FM to PM

The modulating signal in PM is $V_\mathrm{m}(t) = V_\mathrm{m} \cos \omega_\mathrm{m} t$

This signal is fed to a differentiator and after differentiation of the modulating signal,

$$V_\mathrm{m}(t) = \frac{d}{dt} V_\mathrm{m} \cos \omega_\mathrm{m} t$$

$$= -\,\omega_\mathrm{m} \,.\, V_\mathrm{m} \sin \omega_\mathrm{m} t \tag{4.36}$$

After frequency modulation,

$$\omega_\mathrm{i} = \omega_\mathrm{c} + K.V_\mathrm{m}(t)$$

$$\omega_\mathrm{i} = \omega_\mathrm{c} + K\,(-\omega_\mathrm{m} . V_\mathrm{m} \sin \omega_\mathrm{m} t)$$

$$\omega_\mathrm{i} = \omega_\mathrm{c} - K\omega_\mathrm{m} . V_\mathrm{m} \sin \omega_\mathrm{m} t \tag{4.37}$$

The instantaneous phase angle of an FM signal is

$$\phi_\mathrm{i} = \int \omega_\mathrm{i}\, dt \tag{4.38}$$

$$\phi_\mathrm{i} = \int (\omega_\mathrm{c} - K\omega_\mathrm{m} . V_\mathrm{m} \sin \omega_\mathrm{m} t) dt$$

$$= \omega_\mathrm{c} t + \frac{K\omega_\mathrm{m} . V_\mathrm{m}}{\omega_\mathrm{m}} \cos \omega_\mathrm{m} t$$

$$= \omega_\mathrm{c} t + K \,.\, V_\mathrm{m} \cos \omega_\mathrm{m} t \tag{4.39}$$

The instantaneous voltage after modulation is given by

$$V(t)_{\mathrm{PM}} = V_{\mathrm{c}} \sin \phi_{\mathrm{i}} \tag{4.40}$$

$$= V_{\mathrm{c}} \sin(\omega_{\mathrm{c}}t + K.V_{\mathrm{m}} \cos \omega_{\mathrm{m}}t)$$

$$= V_{\mathrm{c}} \sin(\omega_{\mathrm{c}}t + m_{\mathrm{p}} \cos \omega_{\mathrm{m}}t) \tag{4.41}$$

This is the expression for a phase-modulated wave. The process of integration and differentiation are linear. Therefore, no new frequencies are generated.

4.3.4 Comparison between Frequency Modulation (FM) and Phase Modulation (PM)

Table 4.2 shows the differences between Frequency Modulation (FM) and Phase Modulation (PM).

Table 4.2 Differences between FM and PM

S.No	Frequency Modulation (FM)	Phase Modulation (PM)
1	The maximum frequency deviation depends upon amplitude of modulating voltage and modulating frequency.	The maximum phase deviation depends only upon the amplitude of the modulating voltage.
2	Frequency of the carrier is modulated by the modulating signal.	Phase of the carrier is modulated by the modulating signal.
3	Modulation index is increased as modulation frequency is reduced and vice versa.	Modulation index remains same if the modulating frequency is changed.

4.4 GENERATION OF FM

Frequency-modulated signals can be generated by two methods:

1. Direct method
2. Indirect method

4.4.1 Direct Method of FM Generation

In the direct method of FM generation, the modulating signal directly varies the instantaneous frequency of a carrier signal by means of a voltage-controlled device. For such a device, a sinusoidal oscillator having highly selective frequency-determining network is to be used and the frequency can be controlled by reactive components of this network.

Another method of FM generation is that the FM is generated by a tank circuit consisting of *L* or *C*. If the variation of FM wave can be made proportional to the voltage supplied by the modulation circuits then a complete FM wave is obtained.

1. Semiconductor Reactance Modulator

To frequency-modulate the low-power semiconductor transmitter, the semiconductor-reactance modulator is used. A typical frequency-modulated oscillator stage operated as a reactance modulator is shown in Figure 4.8. Transistor Q_1 along with its associated circuitry forms an oscillator and Q_2 is the modulator which is connected to the circuit so that its collector-to-emitter capacitance (C_{CE}) is in parallel with a portion of the RF oscillator coil L_1. Due to the modulator operation, the output capacitance of Q_2 is varied. Thus, the frequency of the oscillator is shifted in accordance with the modulation, the same way as if C_1 were varied.

When the modulating signal is applied to the base of Q_2, the emitter-to-base bias varies at the modulation rate which causes the collector voltage of Q_2 to vary at the same modulating rate. When the collector voltage increases, output capacitance C_{CE} decreases and when the collector voltage decreases, C_{CE} increases.

An increase in the collector voltage has the effect of spreading the plates of C_{CE} farther apart by increasing the width of the barrier. A decrease of collector voltage reduces the width of the *PN* junction and has the same effect as pushing the capacitor plates together to provide more capacitance.

When the output capacitance decreases, there is an increase in instantaneous frequency of the oscillator tank circuit. When the output capacitance increases, there is a decrease in

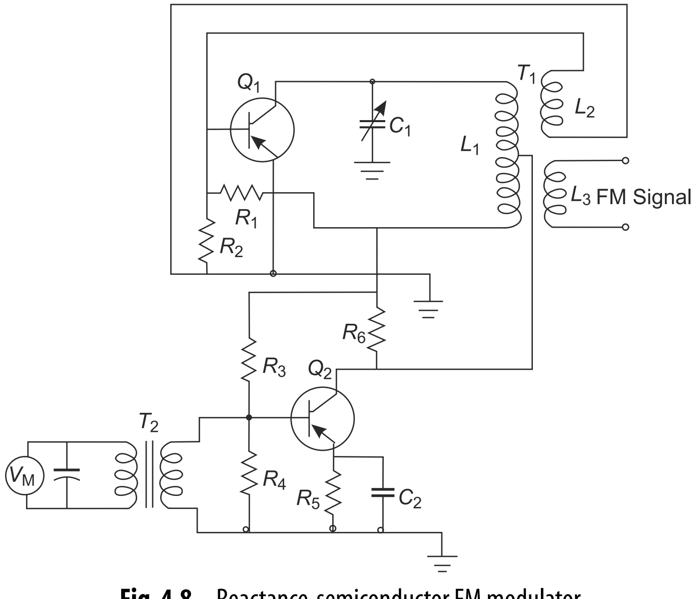


Fig. 4.8 Reactance-semiconductor FM modulator

instantaneous frequency of the oscillator tank circuit. This decrease in frequency produces a lower frequency in the output because of the shunting effect of C_{CE}. Thus, the frequency of the oscillator tank circuit increases and decreases at an audio-frequency-modulating rate. The output of the oscillator, therefore, is a frequency-modulated RF signal.

Since the audio modulation causes the collector voltage to increase and decrease, an AM component is induced into the output. This produces both an FM and AM output. The amplitude variations are then removed by placing a limiter stage after the reactance modulator and only the frequency modulation remains. Frequency multipliers or mixers are used to increase the oscillator frequency to the desired output frequency.

For high-power applications, linear RF amplifiers are used to increase the steady-amplitude signal to a higher-power output. With the initial modulation occurring at low levels, FM represents a savings of power when compared to conventional AM. This is because FM noise-reducing properties provide a better signal-to-noise ratio than is possible with AM.

2. Multivibrator Modulator

Astable multivibrator is another type of frequency modulator which is illustrated in Figure 4.9.

The fundamental frequency of the multivibrator is varied by giving the modulating voltage in series with the base terminals of the multivibrator transistors. The amount of variation is proportional to the amplitude of the modulating voltage. For this purpose, the fundamental frequency of the multivibrator is to be high in relation to the highest modulating frequencies.

The purpose of using a symmetrical multivibrator is to eliminate the unwanted even harmonics since a multivibrator output consists of the fundamental frequency and its entire

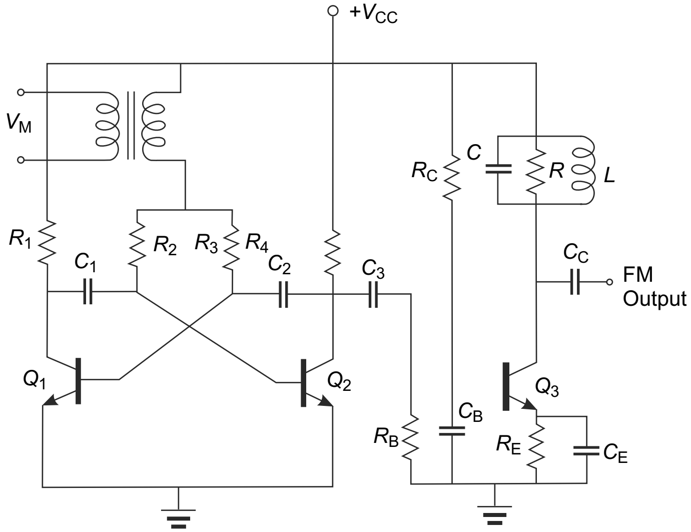


Fig. 4.9 Multivibrator modulator

harmonics. The desired fundamental frequency can be amplified after all other odd harmonics are eliminated in the *LCR* filter section. A single frequency-modulated carrier is then made available for further amplification and transmission. Proper design of the multivibrator will cause the frequency deviation of the carrier to faithfully follow the modulating voltage.

It is important to consider that the *RC* coupling from one multivibrator transistor base to the collector of the other has a time constant which is greater than the actual gate length by a factor of 10 or more. Under these conditions, a rise in base voltage in each transistor is essentially linear from cut-off to the bias at which the transistor is switched on. Since this rise in base voltage is a linear function of time, the gate length will change as an inverse function of the modulating voltage. This action will cause the frequency to change as a linear function of the modulating voltage.

3. *Varactor FM Modulator*

The varactor is simply a diode, or *PN* junction, that is designed to have a certain amount of capacitance between junctions. The capacitance of a varactor, as with regular capacitors, is determined by the area of the capacitor plates and the distance between the plates.

The depletion region in the varactor is the dielectric and is located between the *P* and *N* elements, which serve as the plates. Capacitance is varied in the varactor by varying the reverse bias which controls the thickness of the depletion region. The varactor is so designed that the change in capacitance is linear with the change in the applied voltage. This is a special design characteristic of the varactor diode. Figure 4.10 shows a varactor modulator for FM generation.

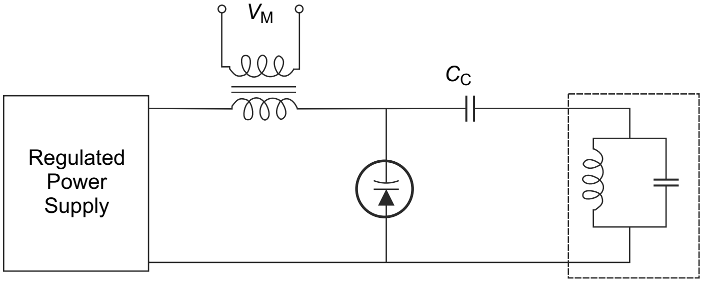


Fig. 4.10 Varactor modulator

If the modulating signal is applied to the input of the varactor diode, during the positive signal, reverse bias increases and the depletion region width increases and this decreased capacitance increases the frequency of the oscillator. During the negative half of the signal, the reverse bias decreases which results in a decrease in oscillator frequency.

4.4.2 Indirect Method of FM Generation

This method is also referred to as **Armstrong method of FM generation**, in which the modulating signal is integrated and then phase-modulated with the carrier signal. The result is

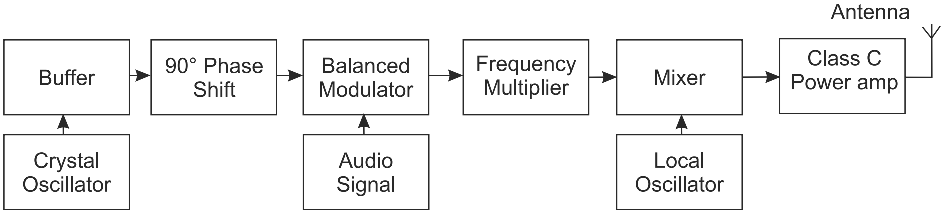


Fig. 4.11 Armstrong method of FM generation

an FM signal and to get the desired form of the output, frequency multipliers are used. Figure 4.11 shows the block diagram of the Armstrong method of FM generation.

From the above block diagram, it is to be noted that the carrier is not directly involved in producing FM signal, but it is injected. The use of a crystal-controlled oscillator is for the generation of carrier frequency. The effect of a mixer is to change the centre frequency only, whereas the effect of a frequency multiplier is to multiply centre frequency and frequency deviation equally.

Figure 4.12 shows the phasor diagram of an amplitude-modulated signal.

$$V_{\mathrm{AM}}(t) = V_{\mathrm{c}} \sin \omega_{\mathrm{c}} t + \frac{m_{\mathrm{a}} V_{\mathrm{c}}}{2} \cos(\omega_{\mathrm{c}} - \omega_{\mathrm{m}})t - \frac{m_{\mathrm{a}} V_{\mathrm{c}}}{2} \cos(\omega_{\mathrm{c}} + \omega_{\mathrm{m}})t \qquad (4.42)$$

It is to be noted that the resultant of the two side-band frequency vectors is always in phase with the unmodulated carrier so that there is amplitude variation but no phase variation.

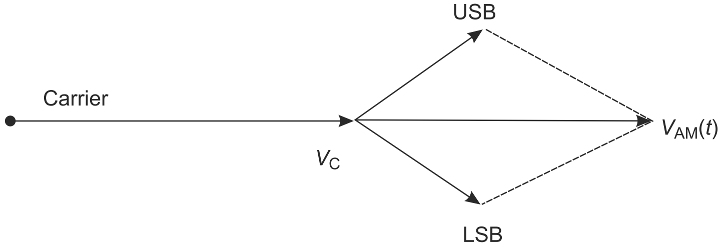


Fig. 4.12 Phasor diagram of AM

In order to generate FM through PM, phase shift is needed between the modulated and unmodulated carrier. Hence, an AM voltage is added to an unmodulated voltage at 90° out of phase. The result is PM, which is shown in Figure 4.13.

From the AM signal, only the two side bands are added to unmodulated voltage and the result is the two side-band voltages are always in quadrature with the carrier voltage and the modulation increases. Hence, some form of phase modulation will be obtained. For complete removal of AM present in it, an amplitude limiter is preferred.

The output of the amplitude limiter is a phase-modulated output, which is shown in Figure 4.14.

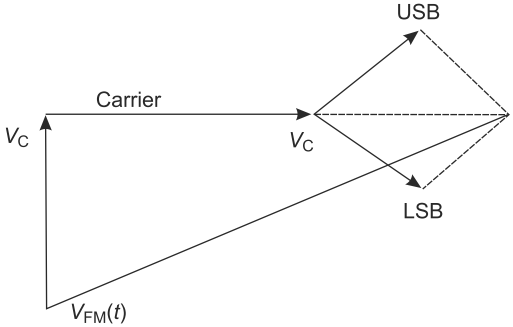


Fig. 4.13 Phasor diagram of FM

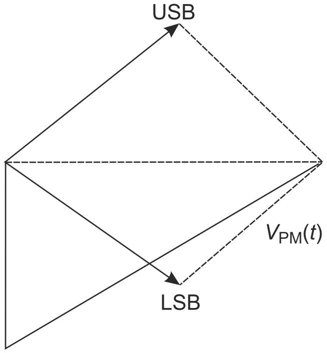


Fig. 4.14 Phasor diagram of PM

4.5 FM TRANSMITTERS

A frequency-modulated transmitter may consist of a modulating system that can directly produce frequency-modulated waves by varying the master oscillator frequency. Such circuits employ *LC* circuits in master oscillator circuits.

Alternately, the transmitting equipment may contain a crystal oscillator which is phase-modulated by audio signals. The PM wave is then converted into a frequency-modulated wave. The basic difference between the two circuits is that while the first circuit employs an *LC* circuit in the master oscillator and its frequency is likely to change with changes in circuit parameters, the second circuit gives a drift-free frequency. As such, the first circuit always employs some form of Automatic Frequency Control (AFC) circuit. Another difference is that the first method produces more frequency deviation and requires less number of stages of frequency multipliers whereas the phase-modulator circuit produces smaller frequency deviations and requires more stages of frequency multiplication.

There are two basic types of FM transmitters:

1. Directly modulated FM transmitter
2. Indirectly modulated FM transmitter using PM

4.5.1 Directly Modulated FM Transmitter

Figure 4.15 shows the block diagram of a directly modulated FM transmitter which uses a reactance-tube modulator. It produces frequency deviation in proportion to the signal amplitude. A varactor diode modulator can also be used alternately for this purpose. The resulting FM signal is passed through a number of frequency-multiplier stages. These stages not only raise the centre frequency of the signal but the frequency deviation is multiplied by the same factor as well. The modulated wave is then amplified to the required power level by Class C power amplifier stages and transmitted.

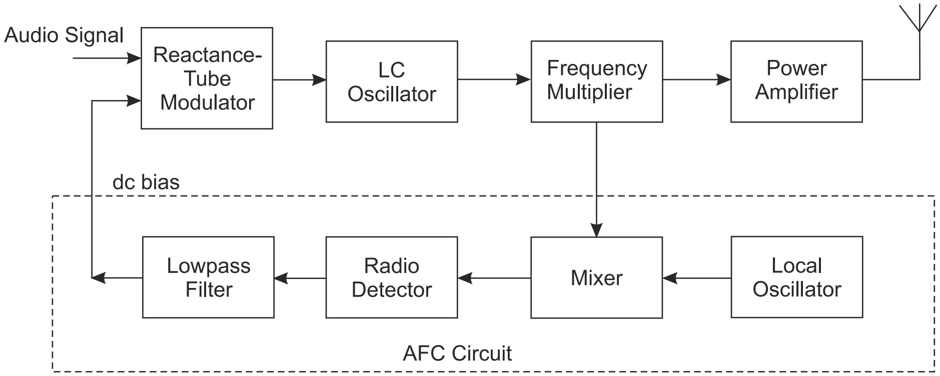


Fig. 4.15 Block diagram of a directly modulated FM transmitter

A part of the output of the frequency multiplier stages is passed to AFC circuit shown in dotted lines. The purpose of this circuit is to make correction in the centre frequency of the transmitter should any drift in it take place due to changes in circuit parameters. The signal from the frequency multiplier is mixed with the crystal oscillator output in a mixer and the difference frequency is fed to a discriminator which gives dc output according to frequency shift with respect to centre frequency.

When the frequency of the transmitter is exactly equal to centre frequency, the discriminator output is zero and there is no dc correcting bias. Any positive or negative drift in the frequency produces a corresponding correction bias at the discriminator which when applied to the reactance-tube modulator brings the *LC* master oscillator frequency back to its centre value.

If a modulating signal is present, the resulting AF signal produced at the discriminator output is not allowed to reach the reactance-tube modulator because of the lowpass filter which has a cut-off lower than the signal. Frequency deviation resulting from audio signal as well as shifts due to change in circuit parameters causes the discriminator output to contain AF output along with slowly varying dc bias. While the former is usually taken to AF stages and produces a sidetone, it is the latter component which reaches the modulator and results in frequency correction.

4.5.2 Indirectly Modulated FM Transmitter

The block diagram of the indirect method of an FM transmitter is shown in Figure 4.16. It comprises of a crystal oscillator, the output of which is fed to a phase modulator. The audio signal is integrated and also applied to the phase modulator. The resultant wave at the phase-modulator output is passed through several stages of frequency multipliers to obtain the desired frequency deviation and increase the central frequency. The signal is amplified by the power-amplifier stage to the required power level.

To understand the circuit action in production of FM waves, assume an audio signal $V_m \sin \omega_m t$ at the input of the transmitter. The input to the phase modulator is given by the integrator

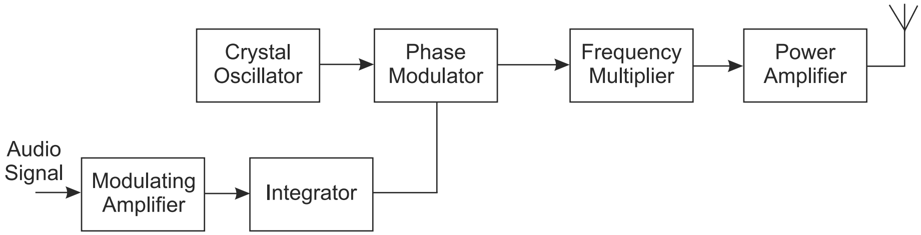


Fig. 4.16 Block diagram of indirect method of FM transmitter

$$\int V_{\mathrm{m}} \sin \omega_{\mathrm{m}} t \, dt = -\frac{V_{\mathrm{m}} \cos \omega_{\mathrm{m}} t}{\omega_{\mathrm{m}}} \tag{4.43}$$

The phase shift produced by this signal at the modulator output is given by

$$\theta = -\frac{K V_{\mathrm{m}} \cos \omega_{\mathrm{m}} t}{\omega_{\mathrm{m}}} \tag{4.44}$$

$$\therefore \quad \omega_{\mathrm{i}} = \frac{d\theta}{dt} = -\frac{d}{dt}\left[\frac{K V_{\mathrm{m}} \cos \omega_{\mathrm{m}} t}{\omega_{\mathrm{m}}}\right] = K V_{\mathrm{m}} \sin \omega_{\mathrm{m}} t \tag{4.45}$$

Frequency deviation $\Delta f = K V_{\mathrm{m}}$ (4.46)

Thus, the circuit results in frequency modulation with deviation proportional to the peak amplitude of the modulating signal.

There is another method of producing FM waves from a crystal oscillator. This method employs a balance modulator and is called Armstrong method. Figure 4.17 shows the block diagram of Armstrong method of FM transmission.

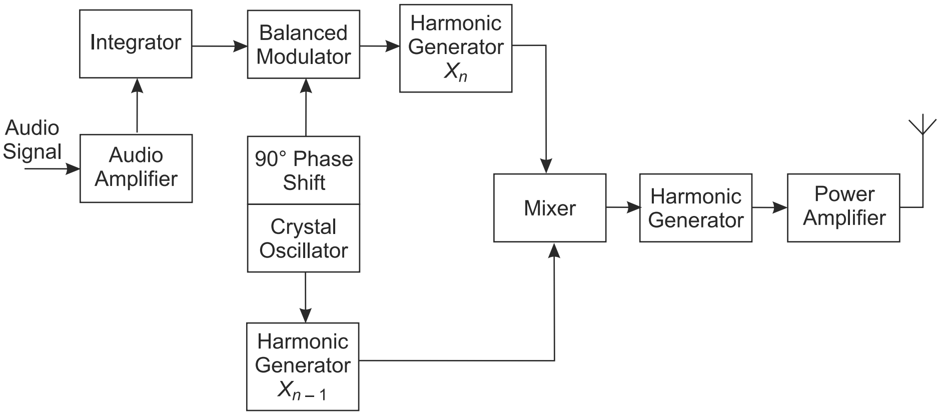


Fig. 4.17 Block diagram of Armstrong method of FM transmission

The Armstrong method utilises a balanced modulator with audio and carrier signals after 90° phase shift. The balanced modulator output gives the upper and lower side bands only with carrier suppressed completely. The frequency of the side bands is increased in a harmonic generator stage and fed to a mixer stage, the other input to this stage being the carrier after passing through another harmonic generator stage. The different frequency components at the mixer are the carrier frequency and the side-band frequencies. This output is again multiplied by a number of frequency multiplier stages, raised to the required power level and transmitted.

To understand the circuit action, assume the modulating-signal amplitude to be $V_m \sin \omega_m t$ and the carrier amplitude to be $V_c \sin \omega_c t$. Input to the balanced modulator is given by

$$\text{Modulating signal} = -\frac{V_m \cos \omega_m t}{\omega_m}$$

$$\text{Carrier signal} = V_c \cos \omega_c t$$

Output of the balanced modulator is given by

$$= 2\frac{V_m}{\omega_m} \cos \omega_c t \cos \omega_m t$$

$$= \frac{V_m}{\omega_m} [\cos (\omega_c + \omega_m)t + \cos(\omega_c - \omega_m)t] \tag{4.47}$$

At the mixer, the output is the selected sum of the carrier and the balanced modulator output.

$$\text{Mixer output} = V_c \sin \omega_c t + \frac{V_m}{\omega_m} \cos(\omega_c + \omega_m)t + \frac{V_m}{\omega_m} \cos(\omega_c - \omega_m)t \tag{4.48}$$

The phasor diagram of the above equation is shown in Figure 4.18, which shows that the mixer output contains a phase deviation ϕ_p.

$$\phi_p = \tan^{-1}\left(\frac{2V_m}{\omega_m V_c}\right) \tag{4.49}$$

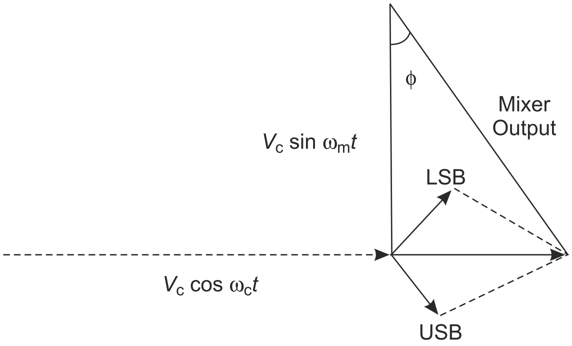


Fig. 4.18 Phasor diagram of Equation (4.44)

In addition to this, it also contains an amplitude-modulation component. However, if the side-band components have very low amplitude then amplitude modulation is negligible. Then for small angles,

$$\tan \phi_p \cong \phi_p = \frac{2V_m}{\omega_m V_c} \tag{4.50}$$

The frequency deviation produced by the system equals Δf such that

$$\Delta f = \frac{2V_m f_m}{\omega_m V_c} = \frac{V_m}{\pi V_c} \tag{4.51}$$

This equation shows that the frequency deviation produced by the system is directly proportional to the magnitude of the modulating signal.

4.6 ADVANTAGES AND DISADVANTAGES OF FM

Advantages

1. FM system has infinite number of side bands in addition to a single carrier. Each side band is separated by a frequency f_m. Hence, its bandwidth is infinite.
2. In FM, the side bands at equal distances from f_c have equal amplitudes. The side-band distribution is symmetrical about the carrier frequency.
3. The amplitude of a frequency-modulated wave in FM is independent of modulation index.
4. In FM, the total transmitted power always remains constant but an increase in the modulation index increases the bandwidth of the system whereas in AM, increased modulation increases the side-band power and which then increases the total transmitted power.
5. In FM, all the transmitted power is useful whereas in AM, most of the transmitted power is used by the carrier. But the carrier usually does not contain any useful information. Hence, the power is wasted.
6. Noise is very less in FM, hence there is an increase in the signal-to-noise ratio. This feature does not exist in AM.
7. FM system operates in UHF and VHF range of frequencies and at these frequencies the space wave is used for propagation, so that the radius of reception is limited slightly more than the line of sight. It is thus possible to operate several independent transmitters on the same frequency with considerably less interference than would be possible with AM.

Disadvantages

1. FM requires much wider channels.

2. Transmission and reception equipment for FM are more complex and hence more expensive.
3. Since reception is limited in line of sight, the area of reception for FM is much smaller than that for AM.

Summary

In frequency modulation, the modulating signal causes the carrier frequency to vary. These variations are controlled by both the frequency and the amplitude of the modulating wave. In phase modulation, the phase of the carrier is controlled by the modulating waveform. The amplitude of the carrier wave remains constant in the FM process. Since the amplitude of the wave remains constant, the power associated with an FM wave is constant.

Frequency-modulated wave can be obtained from phase modulation. This is done by integrating the modulating signal before applying it to the phase modulator. Similarly, the PM wave can also be obtained from FM by differentiating the modulating signal before applying it to the frequency-modulator circuit.

Frequency-modulated signals can be generated in two ways:

- Direct method of FM
- Indirect method of FM

The primary requirement of FM generation is a variable output frequency. The frequency is directly proportional to the instantaneous amplitude of the modulating signal. Another requirement of FM generation is that the frequency deviation is independent of modulating frequency.

A frequency-modulated transmitter may consist of a modulating system that can directly produce FM waves by varying the master-oscillator frequency. Such circuits employ *LC* circuits in master-oscillator circuits. Alternately, the transmitter equipment may contain a crystal oscillator which is phase-modulated by the audio signals. The PM wave is then converted into FM wave.

REVIEW QUESTIONS

PART-A

1. Define frequency modulation.
2. Define phase modulation.
3. What is angle modulation?

4. Define frequency deviation.
5. Define phase deviation.
6. Give the expression for a frequency-modulated wave.
7. Describe the relationship between the instantaneous carrier frequency and modulating signal for FM.
8. Draw the frequency spectrum of FM wave for $m_{\mathrm{f}} = 0.5$ and $m_{\mathrm{f}} = 5$.
9. How will you calculate the bandwidth of FM wave?
10. What is the instantaneous power of angle modulation?
11. Draw the vector representation of an FM wave.
12. Give the expression of a phase-modulated signal.
13. What are the methods of FM generation?
14. What are the two types of FM transmitters?
15. State the advantages and disadvantages of FM.

PART-B

1. Explain the process of frequency modulation and derive its frequency spectrum.
2. Explain the process of phase modulation and draw its graphical representation.
3. How will you convert FM to PM, and PM to FM?
4. With neat block diagrams, describe the conversion process from frequency-modulated wave to phase-modulated wave.
5. With neat block diagrams, describe the conversion process from phase-modulated wave to frequency-modulated wave.
6. With neat circuit diagrams, explain any direct methods of FM generation.
7. With neat circuit diagrams, explain any indirect methods of FM generation.
8. Draw the block diagram of a directly modulated FM transmitter and explain its principle of operation.
9. Draw the block diagram of an indirectly modulated FM transmitter and explain its principle of operation.
10. List the salient features of frequency modulation in comparison with amplitude modulation.

5

FREQUENCY DEMODULATION

Objectives

- To know about the process of FM detection and methods for FM detection such as balanced-slope detector, Foster–Seeley discriminator and ratio detector, etc. To discuss the differences between AM and FM, and FM and PM
- To provide the process of FM demodulation of phase-modulated waves
- To discuss about different FM receivers used at the receiver side of the communication system

5.1 INTRODUCTION

The process of extracting the modulating signal from a frequency-modulated carrier is known as **frequency demodulation**. The electronic circuits that perform the frequency demodulation process are called **FM detectors**.

FM detectors perform the detection in two steps.

1. The original modulating signal is recovered from its AM signal by using a linear envelope detector.
2. It converts the frequency-modulated signal into its corresponding amplitude-modulated signal by using frequency-dependent circuits whose output voltage depends on input frequency from which the original modulating signal is detected. Such circuits are called **frequency discriminators**.

5.2 TYPES OF FM DETECTORS

FM detectors can be divided into two types:

1. Slope detectors
2. Phase discriminators

5.2.1 Single-Tuned Discriminator or Slope Detectors

The principle of operation depends on the slope of the frequency-response characteristics of a frequency-selective network. Figure 5.1 shows the circuit diagram of a slope detector.

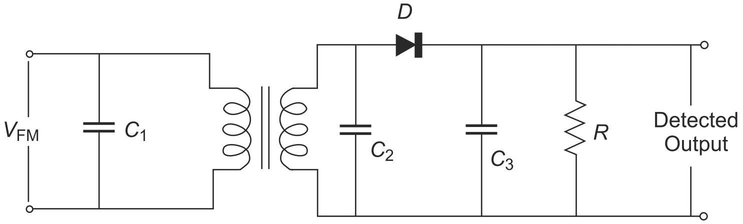


Fig. 5.1 Slope detector

The slope detector consists of a parallel *LC* tuned circuit which acts as a frequency discriminator. It is slightly detuned from the carrier frequency ω_c. A low-frequency deviation produces small amplitude variation, while a high-frequency deviation produces large amplitude variation. Through this action, the FM signal is changed to AM signal. Thus, the AM signal is detected by a diode detector followed by the discriminator.

This detector is simple in construction and is economically cheap. The main disadvantages are the following:

1. The nonlinear characteristic of the circuit causes a harmonic distortion.
2. It does not eliminate the amplitude variation and the output is sensitive to any amplitude variation in the input FM signal which is not a desirable feature.

5.2.2 Stagger-Tuned Discriminator or Balanced-Slope Detector

To overcome the limitations of a slope detector, balanced-slope detector is used. The circuit diagram of a balanced-slope detector is shown in Figure 5.2, which consists of two *LC* circuits.

From the above circuit diagram, the circuit I is tuned to $f_c + \Delta f$, i.e. above IF by an amount Δf and the circuit II is tuned to $f_c - \Delta f$. The output taken across two *RC* loads, when added up gives the total output.

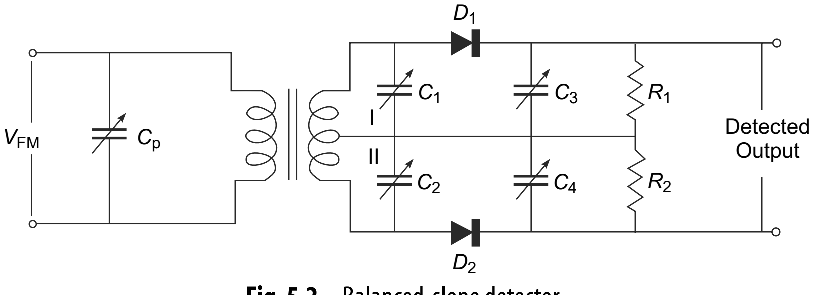


Fig. 5.2 Balanced-slope detector

When the input frequency is f_c, the voltage across the upper half of the secondary coil will be less than the maximum value. This is because maximum voltage occurs at resonance which will happen only at $f_c + \Delta f$. The voltage across the lower half of the secondary coil will be identical to that of the upper half. Therefore, the output of the diode D_1 is positive and that of the diode D_2 is negative. Hence, the detector output will be zero.

When the input frequency is $f_c + \Delta f$, the circuit I will be at resonance. But in the circuit II, the input is far away from the resonant frequency $f_c - \Delta f$. The output now will be positive and maximum.

When the input frequency is $f_c - \Delta f$, the circuit II will be at resonance. The output of the diode D_2 will be maximum and that of the diode D_1 will be negative and maximum.

When the input frequency is between $f_c + \Delta f$ and $f_c - \Delta f$, the output will lie between the above two extremes. It will be positive or negative depending on which side of f_c the input lies.

When the input frequency is beyond $f_c + \Delta f$ and $f_c - \Delta f$, the tuned circuit response will make the output fall. Thus, the input frequency characteristics of the detector will follow S-shaped curve, which is shown in Figure 5.3.

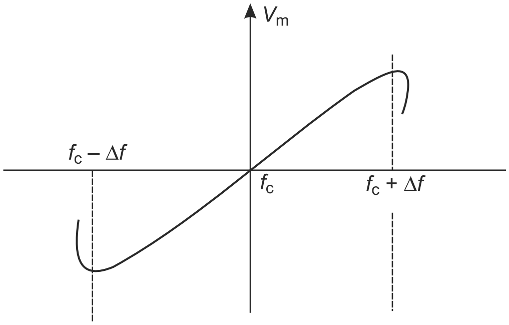


Fig. 5.3 S-shaped frequency response

A balanced slope detector exhibits some drawbacks:

1. Amplitude limiting cannot be provided.
2. Linearity is not good when compared to a single-tuned slope detector.
3. It is very difficult to align because of three different frequencies to which various tuned circuits are to be tuned.
4. Distortion in the output across *RC* filter is possible.

5.2.3 Foster–Seeley Discriminator

The circuit diagram of a Foster–Seeley discriminator is shown in Figure 5.4. Due to its circuit configuration and operation, it is also called **centre-tuned discriminator**.

It is possible to obtain the same S-shaped response curve from a circuit in which the primary and secondary windings are both tuned to the centre frequency of the incoming signals. This

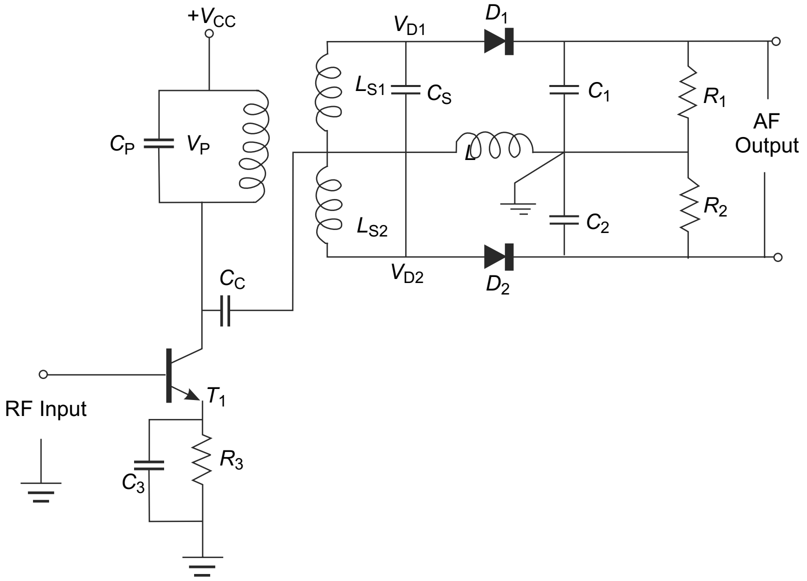


Fig. 5.4 Foster–Seeley discriminator

is desirable because it greatly simplifies alignment and also the process yields better linearity than slope detection. In this discriminator, the same diode and load arrangement is used as in the balanced-slope detection. But the method of ensuring that voltages fed to the diodes vary linearly with the deviation of the input signal has been changed completely.

From the diagram, $C_{1,}$ C_3 and C_4 have low reactance at the carrier frequency. Consequently, the RF voltage appearing across the tuned circuit also appears across the RF choke. The centre-tapped tuned transformer is placed so that the vector sum of the carrier and induced carrier in the secondary appears across the upper diode, which is $V_{ao} = (V_1 + V_2)$. The vector difference of the carrier and the induced carrier in the secondary appears across the lower diode $V_{bo} = (V_1–V_2)$.

Assume that the instantaneous frequency applied to primary is above the resonant frequency. The tuned circuit acts as the capacitive circuit and voltage across both secondaries will lag the applied voltage. The magnitude of this voltage is identical to that of the voltage across the RF choke $L_{3.}$ The vector sum across the upper diode will produce a greater volt than the vector difference across the lower diode. The diode rectifiers detect the peak value of these voltages, which charge the capacitors C_3 and C_4 to the peak of their respective values. Because of the placement of the two diodes in the circuit, the output of the detector is equal to the difference of the voltage across each capacitor.

If the instantaneous voltage is equal to centre frequency or resonance frequency, the vector sum and vector-difference voltages are identical and the detector voltage across the capacitor is zero. Thus, if the frequency deviation of the modulated carrier is small, the output amplitude will follow this deviation and produce a low-amplitude signal at a frequency equal to the deviation rate. Therefore, the unit will function as an angle-frequency-modulated detector.

A mathematical analysis is given to show that the voltage applied to each diode is the sum of primary voltage and corresponding half of the secondary voltage. It will be shown that the primary and secondary voltages are

1. 90° out of phase when input frequency is f_c
2. Less than 90° out of phase when $f_{in} > f_c$
3. More than 90° out of phase when $f_{in} < f_c$

If $f < f_{0,}$ the circuit is acting as capacitive circuit. If $f = f_0$, the circuit is acting as a resistive circuit and if $f > f_0$, the circuit is acting as an inductive circuit. The frequency response of Foster–Seeley discriminator is shown in Figure 5.5.

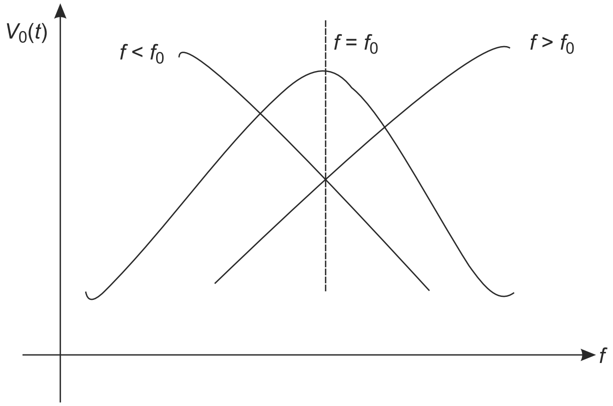


Fig. 5.5 Frequency response

The resistances forming the load are made much larger than the capacitive reactances. It can thus be seen that the circuit composed of C_C, L_3 and C_4 is effectively placed across the primary winding. This is shown in Figure 5.6. L_3 is an RF choke and has a high reactance compared to C and C_C.

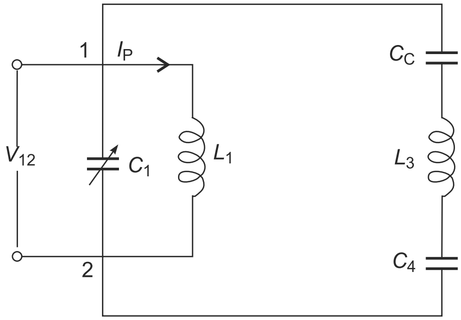


Fig. 5.6 Primary circuit of the discriminator

The voltage across L_3 is

$$V_{L_3} = \frac{V_{12} \,. Z_{L3}}{Z_{L_3} + Z_{C_c} + Z_{C_4}} \tag{5.1}$$

$$V_{L3} = \frac{V_{12} \cdot jX_{L3}}{jX_{L_3} - j(X_{C_c} + X_{C_4})} \tag{5.2}$$

As L_3 is assigned a very large reactance,

i.e. $$X_{\mathrm{L3}} >> X_{Cc} + X_{\mathrm{C4}} \tag{5.3}$$

hence, $$V_{12} \cong \mathrm{V}_{\mathrm{L3}} \tag{5.4}$$

It is concluded that the voltage across the inductance is equal to the applied primary voltage. The primary current by neglecting the impedance coupled in secondary and the primary resistance is given by

$$I_{\mathrm{P}} = \frac{V_{12}}{jX_{L3}} = \frac{V_{12}}{jX_{L_1}} \tag{5.5}$$

The mutually induced emf in the secondary winding is

$$V_{\mathrm{s}} = jX_{\mathrm{m}} I_{\mathrm{p}} = -jX_{\mathrm{m}} I_{\mathrm{p}} \tag{5.6}$$

The secondary circuit of the discriminator is shown in Figure 5.7.

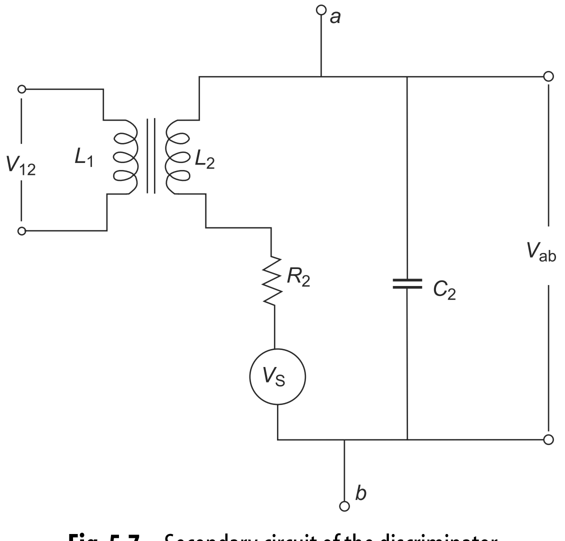


Fig. 5.7 Secondary circuit of the discriminator

The voltage across the secondary winding V_{ab} can be calculated with respect to the secondary circuit of the discriminator.

$$V_{ab} = \frac{-jX_{\text{m}} \, . V_{12} \, . jX_{C_2}}{jX_{L_1} \, (R_2 + jX_2)} \tag{5.7}$$

where $$X_2 = X_{L_2} + X_{C_2} \tag{5.8}$$

Total voltage applied to D_1 will be

$$V_{a0} = V_{ac} + V_L = \frac{1}{2} V_{ab} + V_{12} \tag{5.9}$$

Total voltage applied to D_2 will be

$$V_{b0} = V_{bc} + V_L = -\frac{1}{2} V_{ab} + V_{12} \tag{5.10}$$

The dc output voltage will be calculated as

$$V_{a'b'} = V_{a'0} - V_{b'0} \tag{5.11}$$

Since the output voltage is proportional to the peak value of the RF voltage applied to the respective diode,

$$V_{a'b'} \propto V_{a0} - V_{b0} \tag{5.12}$$

Case 1

Consider $f_{\text{in}} = f_{\text{c}}$; then $X_2 = 0$, i.e. reactance is zero at resonance.

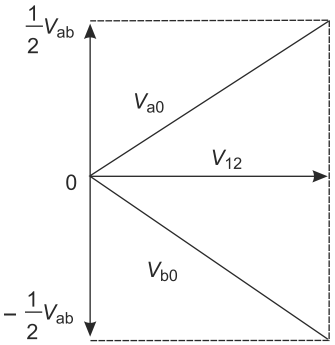


Fig. 5.8(a) Case 1: $f_{\text{in}} = f_{\text{c}}$

$$\text{Output voltage } V_{ab} = \frac{jX_{\text{m}}}{L_1} . \frac{. V_{12} \, . X_{C_2}}{R_2}$$

$$= \frac{V_{12} \, . X_{C_2} \, M}{L_1 R_2} \angle 90° \tag{5.13}$$

Thus, the secondary voltage leads the applied primary voltage by 90°. Therefore, $\frac{1}{2}V_{ab}$ will lead V_{12} by 90° and $-\frac{1}{2}V_{ab}$ will lag V_{12} by 90°. It is possible to add the diode input voltages vectorially. This is shown in Figure 5.8 (a). In this case, $V_{a0} = V_{b0}$ and hence the discriminator output is zero.

Case 2

Consider $f_{in} > f_{c,}$ reactance is inductive.

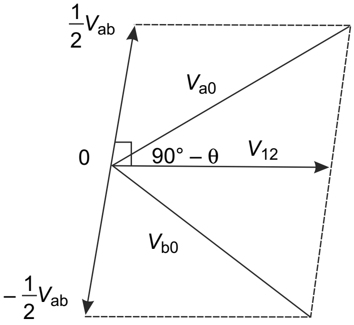


Fig. 5.8(b) Case 2: $f_{in} > f_c$

Then $$V_{ab} = \frac{jX_m \, .V_{12} \, .X_{C_2}}{L_1\,(R_2 + jX_2\,)} \tag{5.14}$$

where $$X_2 = X_{L_2} - X_{C_2} = \text{Positive value} \tag{5.15}$$

$$|Z_2| = \sqrt{R_2^2 + X_2^2} \text{ and } \theta = \tan^{-1}\left(\frac{X_2}{R_2}\right) \tag{5.16}$$

Hence, $$V_{ab} = \frac{X_m \, .V_{12} \, .X_{C_2} \angle 90°}{L_1 |Z_2| \angle\theta}$$

$$V_{ab} = \frac{X_m \, .V_{12} \, .X_{C_2} \angle 90 - \theta}{L_1 |Z_2| \angle\theta} \tag{5.17}$$

Thus, V_{ab} leads V_{12} by less than 90° so that $-\frac{1}{2}V_{ab}$ must lag V_{12} by more than 90°. In this case, $V_{a0} > V_{b0}$. Thus, the discriminator output will be positive when f_{in} is greater than f_c.

Case 3

Consider $f_{in} < f_c$, the reactance X_2 will be capacitive, i.e. negative, and the angle of impedance Z_2 will also be negative.

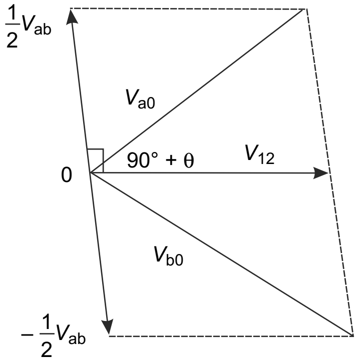


Fig. 5.8(c) Case 3: $f_{in} < f_c$

$$X_2 = X_{L_2} - X_{C_2} = \text{Negative value} \tag{5.18}$$

$$|Z_2| = \sqrt{R_2^2 + X_2^2} \text{ and } \theta = \tan^{-1}\left(\frac{-X_2}{R_2}\right) \tag{5.19}$$

Hence, $$V_{ab} = \frac{X_m \cdot V_{12} \cdot X_{C_2} \angle 90°}{L_1\,(R_2 - jX_2\,)}$$

$$V_{ab} = \frac{X_m \cdot V_{12} \cdot X_{C_2} \angle 90 + \theta}{L_1 |Z_2| \angle\theta} \tag{5.20}$$

Thus, V_{ab} will lead V_{12} by more than 90° so that $V_{a0} < V_{b0}$. Thus, the discriminator output will be negative when f_{in} is greater than f_c.

If the frequency response is plotted for the phase discriminator, it will follow the required 'S' shape. As the input frequency moves farther and farther away from the centre frequency, the disparity between the two diode-input voltages becomes greater. The output of the discriminator will increase up to the limits of the useful range. Beyond these limits, the diode-input voltages are reduced because of the frequency response of the transformer, so that the overall output falls.

Advantages of Foster–Seeley Discriminator

1. The phase discriminator is much easier to align than the balanced-slope detector.
2. Only two tuned circuits are necessary and both are tuned to the same frequency.
3. Linearity is better, because the device relies less upon the frequency response and more on the primary-secondary phase relation which is quite linear.

Disadvantage of Foster–Seeley Discriminator Need of a separate amplitude limiting circuit

5.2.4 Ratio Detector

In a phase discriminator, change in the magnitude of the input signal will give rise to amplitude change in the resulting output voltage. This makes prior limiting circuit necessary. It is possible to modify the discriminator circuit to provide limiting so that the amplitude limiter may be dispensed with. The modified circuit is now called ratio detector.

A ratio detector is closely resembles the Foster–Seeley discriminator in certain aspects. The input circuits of the ratio detector and the Foster–Seeley discriminator are similar but the connections of the diode D_2 are reversed in a ratio detector. As a result, the rectified voltages of D_1 and D_2 become additive. A large capacitor C_0 is placed across resistors R_1 and R_2. The output is taken in between the resistor-capacitor network. The schematic diagram of such a modified ratio detector is shown in Figure 5.9.

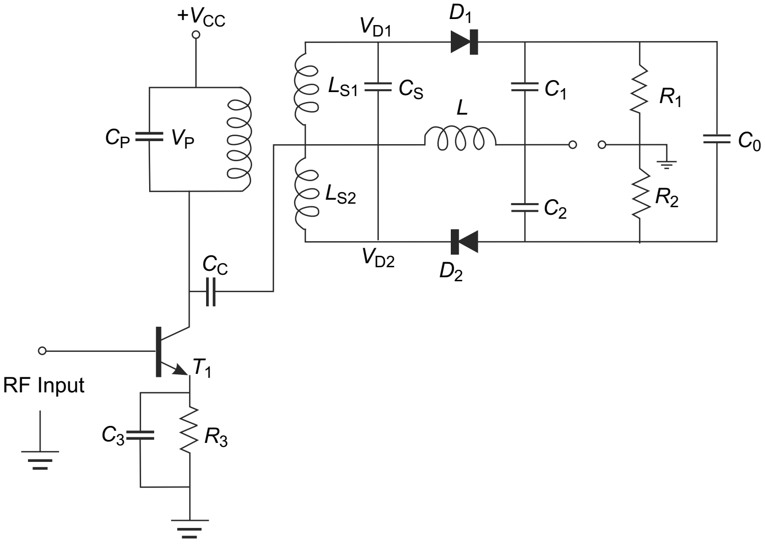


Fig. 5.9 Ratio detector

1. Operation

With the diode D_2 reversed, the polarity at a, 0 and b changes. So $V_{a'b'}$ is the sum voltage rather than difference. It is now possible to connect a large capacitor between a_1 and b_1 to maintain this sum voltage constant. Once C_0 is connected, then $V_{a'b'}$ is not the output voltage in this case and the output is taken across O and O'.

The capacitor charges to a voltage $V = (V_1 + V_2)$. Normally, R_1 and R_2 are made equal and the voltage at the junction of R_1 and R_2 measured with respect to the bottom point equals $\left(\frac{V_1+V_2}{2}\right)$. This voltage remains fixed because of large value of the capacitor C_0 and the resulting high time constant.

Case 1 When signal frequency equals f_c, both the diodes will ne equal. As C_1 is always kept equal to C_2, V_1 and V_2 are equal. The potential difference across the output terminals equals zero.

Case 2 When signal frequency is greater than f_c ($f_{in} < f_c$) input for D_1 exceeds input for D_2. As a result, V_1 is increased to $V_1 + \Delta V$ and V_2 becomes $V_2 + \Delta V$.

Voltage at R_1R_2 junction is given by

$$V_{R_1R_2} = \frac{V_1 + \Delta V + V_2 + \Delta V}{2} = \frac{V_1 + V_2}{2} \tag{5.21}$$

Output voltage V_0 is given by

$$V_0 = \frac{V_1 + V_2}{2} - (V_1 + \Delta V)$$

$$= -\Delta V \tag{5.22}$$

Case 3 For signal frequencies smaller than f_c ($f_{in} > f_c$), voltage across diode D_1, i.e. V_1 is reduced to $V_1 - \Delta V$ and the voltage through D_2 increases to $V_2 + \Delta V$. The potential at the junction of R_1R_2 is still constant and equals $\left(\frac{V_1 + V_2}{2}\right)$, while the potential at the junction of C_1C_2 measured with respect to bottom becomes $V_2 + \Delta V$.

Therefore, the output voltage V_0 is given by

$$V_0 = \frac{V_1 + V_2}{2} - (V_1 + \Delta V)$$

$$= \Delta V \tag{5.23}$$

The above three cases show that the output voltage follows deviation in the frequency of the signal. Hence, this circuit will convert frequency-modulated signal into original modulating signal in the same way as a Foster–Seeley discriminator.

2. Amplitude Limiting by Ratio Detector

The main advantage of using a ratio detector is that there is no need for separate amplitude-limiting circuitry. This limiting action is achieved with the help of the capacitor C_0 connected across R_1 and R_2. The following explanation gives the behaviour of the detector to amplitude changes.

(a) If the input voltage V_{12} is constant, C_0 will charge up-to potential existing between *a* and *b*. This will be a dc voltage and no current will flow through the capacitor, i.e. the input impedance of C_0 is infinite. The total impedance for two diodes is, therefore, the sum of R_1 and R_2. The output will remain the same.

(b) For an increase in input due to noise, the output voltage follows the input. The capacitor C_0 must get charged to that level. To charge itself to the increased input level, the capacitor C_0 draws charging current from the input resonant circuit, thereby loading to a great extent. As a result, the magnification factor Q, the circuit is lowered reducing the input signal level.

(c) If there is a decrease in the input-signal level, the capacitor must discharge to the input level. Thus, there is a discharging current through R_1 and R_2. This reduces loading upon the input resonant circuit causing its magnification factor Q to increase. As a result, the input-signal level is increased.

Thus, the ratio detector output remains free from amplitude fluctuations in the signal input and converts frequency changes into amplitude changes.

5.3 DEMODULATION OF PHASE-MODULATED WAVES

It is known that a phase-modulated wave with phase deviation $\Delta\phi$ produces a frequency deviation f such that $\Delta f = \Delta\phi.\ f_m$. Thus, if a circuit capable of providing division by $\frac{1}{f_m}$ is incorporated in the output of a Foster–Seeley discriminator or ratio detector, the output obtained from this arrangement will be proportional to phase deviation $\Delta\phi$. A commonly employed technique is illustrated by a block diagram shown in Figure 5.10.

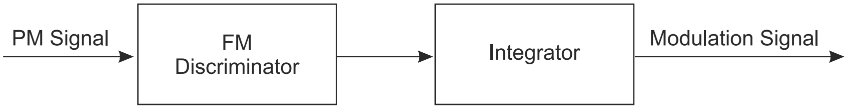


Fig. 5.10 Block diagram of a PM demodulator

The PM modulator is not much used in communication but is very commonly employed for providing automatic frequency control of an oscillator and in servo systems. Figure 5.11 shows the basic arrangement of a phase-sensitive error-control circuit.

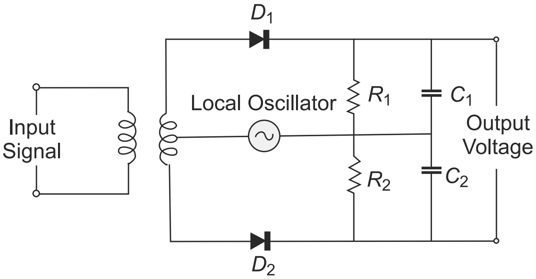


Fig. 5.11 A phase-sensitive error-control circuit

Here, the input signal is compared with a very stable local oscillator and the output V_0 is proportional to the phase error of the input signal. The output voltage is fed via a feedback servo loop to reduce the error signal.

5.4 FM RECEIVERS

5.4.1 FM Superheterodyne Receiver

Frequency-modulated receivers are also superheterodyne receivers employing double-frequency conversion. The RF, mixer and IF stages are similar to that employed in AM receivers except that the device selection and the circuit design are done with a view to VHF or UHF operation. AGC is not normally employed in these receivers. The resonant circuits employed in various RF and IF stages are designed to have adequate bandwidth to accommodate frequency-modulated signals. Figure 5.12 shows the block diagram of an FM superheterodyne receiver.

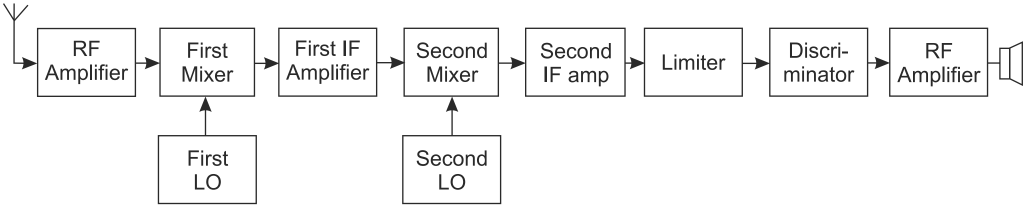


Fig. 5.12 Block diagram of an FM superheterodyne receiver

1. RF Section

The RF section is similar to that employed in AM receivers. They select the desired signal from many modulated signals and amplify them to the requisite level. The only difference between AM and FM is that the RF sections in FM receivers are designed to operate at VHF and UHF ranges. Since the receivers are to operate in VHF or UHF ranges with space-wave propagation, such signals are not prone to selective or general fading. These FM receiver circuits do not require AGC bias.

2. Frequency Changer

It is the combination of the mixer and the local oscillator. They perform the same function as in AM receivers. By the heterodyne operation using mixer and local oscillator, the input and the local-oscillator frequency can be used to produce the difference frequency which is called the intermediate frequency. The frequency-changer circuit is also called a **detector circuit**. Therefore, an FM receiver employing double-frequency conversion is also called **triple-detection receiver**.

3. IF Section

It consists of one or more stages of tuned amplifiers which have an adequate bandwidth to accommodate the frequency-modulated signals. The main purpose of an IF amplifier is to amplify the intermediate frequency present at the output of the frequency-changer circuit and to provide most of the gain and selectivity of the receiver.

4. Limiter

The amplified IF signals at the second IF amplifier are passed through the limiter stage before being passed on to the discriminator circuit for conversion into audio signals. A limiter stage is not required if a ratio detector is employed instead of a Foster–Seeley discriminator. The purpose of a limiter stage is to provide a constant-amplitude IF signal with the same frequency deviation as produced by the modulating signal at the input of the discriminator so that the amplitude variations in FM due to external and internal noises do not reach the receiver output.

The limiter stage commonly employed is similar to the IF amplifier. But it is operated in such a way that during positive half cycle of the input signal, the amplifier goes into saturation and during negative half cycle, the amplifier goes to cut-off.

Another common feature of FM reception is that there is unbearable noise at the receiver output when there is no input signal. This could be prevented by using a squelch circuit.

5. De-emphasis Circuit

It is always employed in FM receiver circuits to restore relative magnitudes of different components of AF signals as in the original modulating signal.

6. Detector

The FM detector, usually a Foster–Seeley discriminator or ratio detector, can be used to recover the original modulating signal from the modulated signal. The output of the detector is passed to an audio amplifier to amplify the detected or baseband signal to the required level.

5.4.2 Characteristics of FM Receivers

The overall performance of radio receivers is measured from the characteristics of the receiver. It helps to evaluate the relative merits of a particular circuit design and also helps to find out the usefulness of a receiver under special operating conditions. The various characteristics of the receiver are as follows.

1. Sensitivity

Sensitivity of a receiver is defined as a measure of its ability to receive weak signals. The sensitivity curve of a receiver is shown in Figure 5.13.

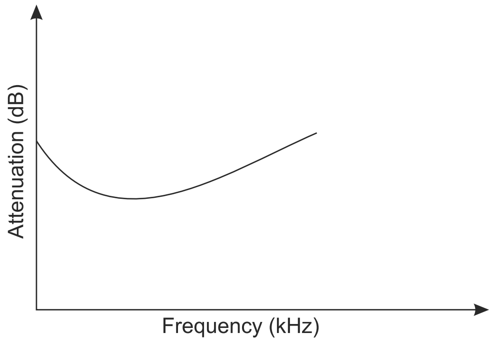


Fig. 5.13 Sensitivity curve of a receiver

2. Selectivity

Selectivity of a receiver is defined as its ability to select the desired signal among the various signals present and rejecting all the other unwanted signals. As the selection is done by resonant circuits, the selectivity is entirely dependent upon the frequency response of these resonant circuits. A resonant circuit with high quality factor (Q) is more selective than a circuit with low Q. Figure 5.14 shows the selectivity curve of an FM receiver.

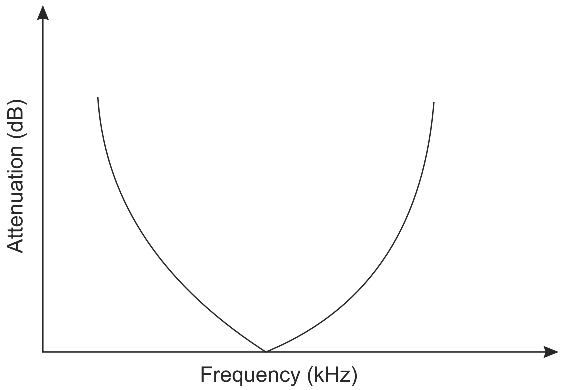


Fig. 5.14 Selectivity curve

3. Stability

It is the ability of a receiver to deliver a constant amount of output for a given period of time, when the receiver is supplied with a signal of constant amplitude and frequency.

4. Fidelity

This may be defined as the degree to which a system accurately reproduces at the output the essential characteristics of signals that are impressed upon the output. Figure 5.15 shows the overall fidelity curve of a receiver.

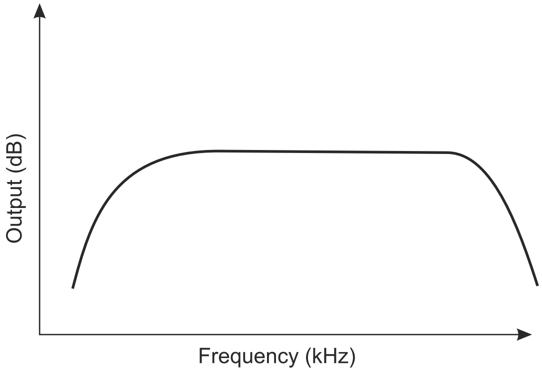


Fig. 5.15 Overall fidelity curve for a receiver

5. Signal-to-Noise Ratio

Signal-to-noise ratio may be defined as the ratio of signal power to noise power at the receiver output. A good receiver should have high Signal-to-Noise Ratio (SNR) which indicates negligible noise present at the output. Signal-to-noise ratio at the receiver output may be expressed as

$$\text{SNR} = \frac{m_a E_0}{E_n} \tag{5.24}$$

where E_0 and m_a are the carrier amplitude and degree of modulation respectively, of the input signal frequency for which SNR is desired.

5.4.3 Automatic Frequency Control

The main purpose of Automatic Frequency Control (AFC) is to lock the receiver to any valid RF signal. The basic block diagram of a basic FM receiver is shown in Figure 5.16.

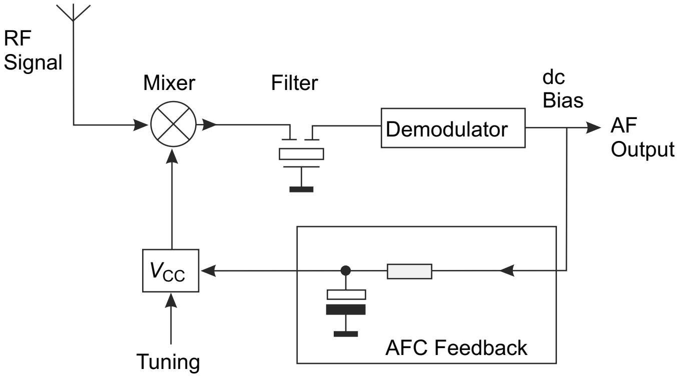


Fig. 5.16 Block diagram of a basic FM receiver

The RF signal enters a mixer and mixes with the VCO which can be a coil and a tunable capacitor. The product from the mixer is then filtered and entered into the demodulator to provide the sound output. The modulator is assumed to work at a frequency of 455 kHz. If the receiver is tuned perfectly to the RF component, 455 kHz will be received out of the mixer. The demodulator will then have a constant bias of a few volts and an ac component which will be the audio.

If the VCO frequency is changed to a few kilohertz, the demodulator will now have a different bias voltage depending on the frequency error of the VCO. The ac component will still be there, but a bit distorted and the dc bias from the demodulator is filtered by a resistor and a capacitor. The AFC feedback will, after the filtering, regulate the VCO to correct it so that the frequency slides back to 455 kHz and the reception will be a perfect one.

5.4.4 Pre-Emphasis

The noise has greater effect on higher modulating frequencies than on the lower ones. The effect of noise on higher frequencies can be reduced by artificially boosting them at the receiver. This boosting of higher modulating frequencies at the transmitter is called pre-emphasis. The pre-emphasis circuit will be placed at the transmitting side of the frequency modulator. It is used to increase the gain of the higher frequency component as the input signal frequency increases, the impendence of the collector voltage increases. If the signal frequency is lesser then the impendence decreases which increases the collector current and hence decreases the voltage. Figure 5.17 shows the circuit diagram of pre-emphasis.

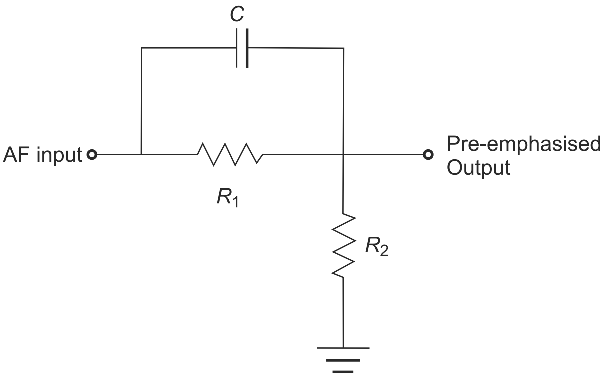


Fig. 5.17 Pre-emphasis circuit

5.4.5 De-Emphasis

The de-emphasis circuit will be placed at the receiving side. It acts as a lowpass filter. The boosting gain for higher frequency signal in the transmitting side is done by the pre-emphasis circuit is filtered to the same value by the lowpass filter. Figure 5.18 shows the circuit diagram of de-emphasis.

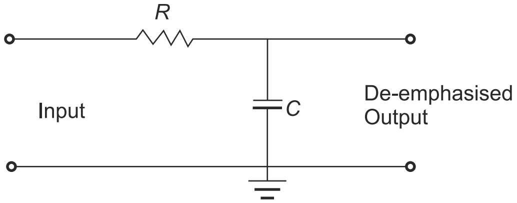


Fig. 5.18 De-emphasis circuit

The cut-off frequency is given as follows:

$$f_c = \frac{1}{2\pi RC} \tag{5.25}$$

The characteristic curve showing both pre-emphasis and de-emphasis is shown in Figure 5.19.

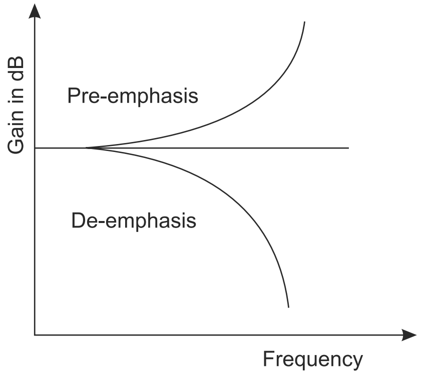


Fig. 5.19 Pre-emphasis and de-emphasis gain in dB

EXAMPLE 5.1

For a cut-off frequency of 2 kHz, give the design considerations for a de-emphasis circuit.

Solution

The cut-off frequency for de-emphasis circuit is expressed as

$$f_c = \frac{1}{2\pi RC}$$

It is given that cut-off frequency = 2 kHz

Assume C = 0.1μF

$$2 \times 10^3 = \frac{1}{2\pi R \times 0.1 \times 10^{-6}}$$

$$\therefore\ R = \frac{1}{2\pi f C}$$

$$= \frac{1}{2\pi \times 2 \times 10^3 \times 0.1 \times 10^{-6}} = 796 \text{ Hz}$$

EXAMPLE 5.2

For the de-emphasis circuit shown in Figure 5.20, determine the cut-off frequency.

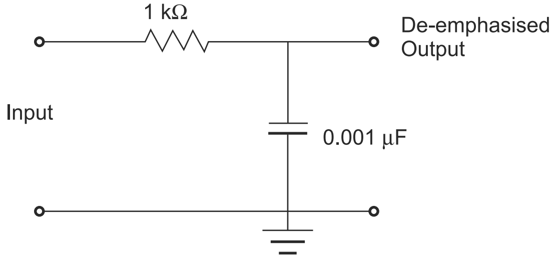


Fig. 5.20

Solution

The cut-off frequency for a de-emphasis circuit is expressed as

$$f_c = \frac{1}{2\pi RC}$$

It is given that $R = 1$ kΩ and $C = 0.001\mu$F

$$\therefore f_c = \frac{1}{2\pi \times 1 \times 10^3 \times 0.001 \times 10^{-6}} = 159 \text{ kHz}$$

Summary

Frequency demodulation is the process of extracting the modulating signal from a frequency-modulated carrier.

FM detectors can be divided into two types:

- Slope detectors
- Phase discriminators

In a slope detector, the principle of operation is dependent on the slope of the frequency-response characteristics of a frequency-selective network. To overcome the limitations of a slope detector, a balanced-slope detector is used.

The Foster–Seeley discriminator is much easier to align than the balanced-slope detector and in which only two tuned circuits are necessary and both are tuned to the same frequency. Linearity is also better in it. In a Foster–Seeley discriminator, change in the magnitude of the input signal will give rise to amplitude change in the resulting output voltage. This makes prior limiting circuit necessary.

It is possible to modify the discriminator circuit to provide limiting so that the amplitude limiter may be dispensed with. The modified circuit is now called ratio detector. The main advantage of using a ratio detector is that there is no need for separate amplitude-limiting circuitry.

Frequency-modulated receivers are also superheterodyne receivers employing double-frequency conversion. The RF, mixer and IF stages are similar to that employed in AM receivers except for the device selection and the circuit design. The main purpose of Automatic Frequency Control (AFC) is to lock the receiver to any valid RF signal.

The pre-emphasis circuit will be placed at the transmitting side of the frequency modulator. It is used to increase the gain of the higher frequency component as when the input signal frequency increases, the impedance of the collector voltage increase. If the signal frequency is lesser then the impedance decreases which increase the collector current and hence decreases the voltage. The de-emphasis circuit will be placed at the receiving side. It acts as a lowpass filter. The boosting gain for higher frequency signal in the transmitting side is done by the pre-emphasis circuit filtered to the same value by the lowpass filter.

REVIEW QUESTIONS

PART-A

1. Define Frequency demodulation.
2. How will you perform FM detection?
3. What are the methods of FM detection?
4. State the principle of slope detection.
5. What are the disadvantages of a slope detector?
6. State the drawbacks of a balanced-slope detector.
7. What are the advantages of a Foster–Seeley discriminator?
8. Mention the differences between the phase discriminator and ratio detector.
9. What are the advantages of a ratio detector?
10. Draw the block diagram of a PM demodulator.
11. What is the purpose of a de-emphasis circuit in FM receiver?
12. List out the desirable characteristics of an FM receiver.
13. Define sensitivity of a receiver.
14. Draw the sensitivity curve of a receiver.

15. What do you mean by selectivity of a receiver?
16. Draw the selectivity curve of a receiver.
17. What is stability?
18. Define fidelity of a receiver.
19. Draw the overall fidelity curve of an FM receiver.
20. Give the expression of SNR of a receiver.
21. Define pre-emphasis
22. What is de-emphasis?
23. What is the significance of AFC?

PART-B

1. What are the different methods of detection of FM signal? What are their advantages and disadvantages?
2. With a neat diagram, explain the operation of a slope detector for FM demodulation.
3. Explain the principle of operation of a balanced-slope detector and state its drawbacks.
4. With a suitable diagram, describe how FM demodulation can be performed using a Foster–Seeley discriminator.
5. Explain the principle of operation of a ratio detector for FM detection with a neat sketch. Compare its merits over Foster–Seeley discriminator.
6. Explain about the amplitude limiting achieved in a ratio detector.
7. Draw the sketch of an FM slope detector and explain its operation. Why is this method not often used in practice?
8. With a neat block diagram, explain the principle of operation of an FM receiver.
9. Explain the principle of operation of double superheterodyne FM receiver with a neat block diagram. What are the reasons for the choice of IF frequency?
10. Explain the salient characteristics of a receiver in detail.

6

PULSE MODULATION

Objectives

- To know the purpose of pulse modulation in communication systems
- To discuss the different types of pulse modulations
- To provide the details about analog pulse modulation in detail
- To discuss the process of pulse amplitude modulation with its mathematical representation, graphical form, different generation methods, PAM transmission and reception in detail
- To provide the details about pulse duration modulation and pulse position modulation methods with their mathematical representation, graphical form, different generation methods, transmission and reception in detail
- To understand the concepts behind digital pulse modulation such as pulse-code modulation and delta modulation in detail

6.1 INTRODUCTION

Pulse-modulation schemes aim at transferring a narrowband analog signal over an analog baseband channel as a two-level signal by modulating a pulse wave. Some pulse-modulation schemes also allow the narrowband analog signal to be transferred as a digital signal with a fixed bit rate, which can be transferred over a digital transmission system.

6.2 DEFINITION

Pulse modulation provides an alternate form of communication system that finds an attractive application in communication links using time-division multiplexing with better signal-to-noise ratios. The characteristic of a train of rectangular pulses is varied in accordance with the modulating signal.

The aim of pulse-modulation methods is to transfer a narrowband analog signal, for example a phone call, over a wideband baseband channel or, in some of the schemes, as a bit stream over another digital transmission system.

6.3 TYPES OF PULSE MODULATION

Pulse modulation can be divided into two major types.

1. Analog pulse modulation
2. Digital pulse modulation

6.4 ANALOG PULSE MODULATION

If a message is adequately described by its sample values, it can be transmitted by analog pulse modulation by which the sample values directly modulate a periodic pulse train with one pulse for each sample.

Figure 6.1 shows an information signal and the resulting pulse-modulated signals. It makes several items apparent.

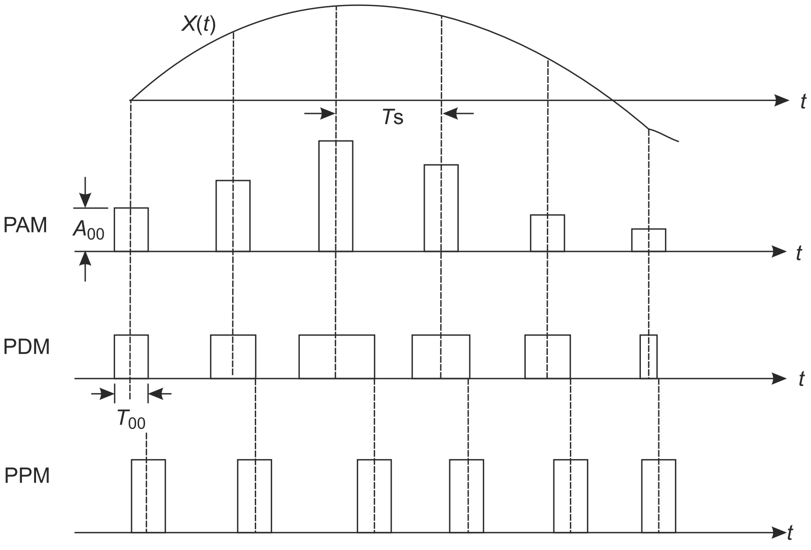


Fig. 6.1 Types of analog pulse modulation

1. If a rectangular pulse is to accommodate the information signal, a method must be established by which the information signal is sampled. It is obvious that if the repetition frequency of the rectangular pulse is extremely small, many variations of the information signal could take place and thereby go undetected. This is a loss of information.
2. If more samples are taken then more faithful representation of the information is obtained. If an infinite number of samples are taken, the PAM will exactly duplicate the information signal. However, due to short time, the bandwidth required to represent the amplitude-modulated pulse would become infinite.
3. The process of sampling must offer some advantage to justify the added complexity. The justification is that between periodic samples of one signal, other information signals may also be sampled. Since the samples are taken in a time sequence, several information signals can be accommodated in the same channel allocation.

There are three types of analog pulse modulation systems.

1. Pulse Amplitude Modulation (PAM)
2. Pulse Duration Modulation (PDM)
3. Pulse Position Modulation (PPM)

6.4.1 Pulse Amplitude Modulation (PAM)

If the amplitude of a pulse is altered in accordance to that of the amplitude of the modulating signal to make it accommodate the information signal, this modulation process is called Pulse Amplitude Modulation (PAM).

6.4.2 Pulse Duration Modulation (PDM)

In this modulation, pulse duration is altered in accordance with the amplitude of the modulating signal to accommodate the information signal. It is apparent that the duration of the pulse can be controlled only to a certain degree. Faithful transmission of a very narrow pulse requires a large bandwidth and a pulse that is too wide will run into the next pulse in sequence.

6.4.3 Pulse Position Modulation (PPM)

In this modulation, pulse position is altered in accordance with the amplitude of the modulating signal to accommodate the information signal. In this case, the position can vary over specified regions without loss of information.

In regard to these three techniques, the following is to be noted:

(a) If pulse duration or pulse position modulation is employed, it is imperative that the leading edge and trailing edge be sufficiently defined, since their position determines the information content of the pulse.

(b) In PDM, the full bandwidth is not used for each transmitted pulse but only when the sampled information signal is such that it produces a narrow pulse.

(c) In PPM, the full bandwidth is used for the transmission of each pulse. It is inherently less susceptible to noise.

(d) The minimum bandwidth necessary to establish the pulse width will affect the bandwidth requirements of these PDM and PPM systems.

(e) In PAM and PDM, sample values equal to zero are usually represented by nonzero amplitude or duration. This is done to prevent missing pulses and to preserve a constant pulse rate.

6.4.4 Characteristics of Pulse-Modulation Systems

Pulse-modulated waves have appreciable dc and low-frequency content, especially near the first few harmonics of $f_s = \frac{1}{T_s}$. It means that direct transmission is very difficult. For pulse-modulated waves, a bandwidth of at least $\frac{1}{2\tau}$ is necessary where τ is the nominal pulse duration. If not, overlapping will take place.

The demodulation of a pulse-modulated wave is done by reconstructing the signal, the sample values are extracted from the modulated wave, converted into weighted implulses and lowpass filtered.

For efficient communication, most pulse systems are carrier modulated in which the pulses are converted to RF pulses. Analytically, it can be written as $X_c(t) = X_p(t)\, A_c \cos \omega_c t$ where $f_c >> f_2$.

Figures RH 6.2(a) and (b) show the complete block diagram of analog pulse-modulation system. The transmitter consists of a sampler, pulse modulator and carrier modulator. The receiver includes carrier demodulation followed by a converter which changes the pulse-modulated wave to a train of weighted impulse $X_s(t)$ to the realizable equivalent. The original signal is recovered by passing the signal to a lowpass filter.

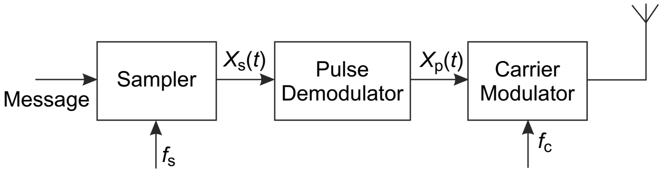


Fig. 6.2(a) Transmitter part of analog pulse modulation

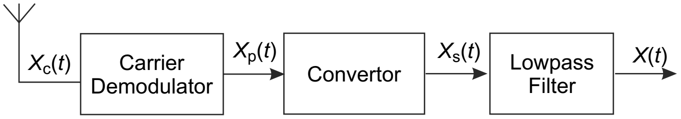


Fig. 6.2(b) Receiver part of analog pulse modulation

According to transmission bandwidth, it is better to have the durations as small as compared to the time between pulses, that is $\tau \le T_s \le \dfrac{1}{2\,\omega}$. As the carrier modulation doubles the baseband width, the transmission bandwidth is $B_T \ge \dfrac{1}{\tau} \ge 2\,\omega$.

6.5 PULSE AMPLITUDE MODULATION (PAM)

6.5.1 Definition

In pulse amplitude modulation, a pulse is generated with amplitude corresponding to that of the modulating waveform. For a PAM process, the signal is sampled at the sampling rate and the carrier pulse is directly amplitude modulated at the sampling frequency. The information is transferred to the pulse at the sampling frequency.

If the sampling frequency is too low then distortion will result in the output. On the other hand, higher sampling rate would provide much greater fidelity, but the time between pulses will limit the number of other signals that can be sampled. In the pulse-amplitude-modulated system, the duration of the pulse, the position of the pulse do not change, because both do not contain any information.

6.5.2 Mathematical Analysis

The pulse train can be represented as follows.

$$e_i(t) = \frac{\tau A}{T} + \frac{2\,\tau A}{T}\sum_{n=1}^{\infty}\frac{\sin n\omega_s \tau/2}{n\omega_s \tau/2}.\cos n\omega_s t$$

$$= A\beta + 2\,A\beta \sum_{n=1}^{\infty}\frac{\sin n\omega_s \tau/2}{n\omega_s \tau/2}.\cos n\omega_s t \qquad (6.1)$$

where A is the amplitude of the pulse,

τ is the pulse width, and

T is the time interval between successive pulses.

If $A = 1$ then

$$e_{\text{PAM}}(t) = (E_{\text{C}} + E_{\text{m}} \sin \omega_{\text{m}} t)\left[\beta + 2\beta \sum_{n=1}^{\infty}\frac{\sin n\omega_s \tau/2}{n\omega_s \tau/2}.\cos n\omega_s t\right]$$

$$= (E_{\text{C}} + E_{\text{m}} \sin \omega_{\text{m}} t)\left[\beta + 2\beta \sum_{n=1}^{\infty}\frac{\sin x}{x}.\cos n\omega_s t\right] \qquad (6.2)$$

where $x = n\omega_s\tau/2$,

E_c is the amplitude of the carrier signal, and

E_m is the amplitude of the message signal.

$$e_{\text{PAM}}(t) = E_C\beta + 2E_C\beta \sum_{n=1}^{\infty} \frac{\sin x}{x}.\cos n\omega_s t.\sin\omega_m t + \beta.E_m \sin\omega_m t + 2\beta.E_m \sin\omega_m t \sum_{n=1}^{\infty} \frac{\sin x}{x}.\cos n\omega_s t$$

$$= E_C\left[\beta + 2\beta \sum_{n=1}^{\infty} \frac{\sin x}{x}.\cos n\omega_s t.\sin\omega_m t + \frac{E_m}{E_C}\beta.\sin\omega_m t + 2\beta.\frac{E_m}{E_C} \sum_{n=1}^{\infty} \frac{\sin x}{x}.\cos n\omega_s t.\sin\omega_m t\right] \quad (6.3)$$

If $E_c = 1$ then

$$e_{\text{PAM}}(t) = \beta + 2\beta \sum_{n=1}^{\infty} \frac{\sin x}{x}.\cos n\omega_s t.\sin\omega_m t + \frac{E_m}{E_C}\beta.\sin\omega_m t + 2\beta.\frac{E_m}{E_c} \sum_{n=1}^{\infty} \frac{\sin x}{x}.\cos n\omega_s t.\sin\omega_m t$$

$$= \beta + \beta.m_a.\sin\omega_m t + 2\beta \sum_{n=1}^{\infty} \frac{\sin x}{x}.\cos n\omega_s t + 2\beta.m_a.\sum_{n=1}^{\infty} \frac{\sin x}{x}.[\sin(n\omega_s + \omega_m)t - \sin(n\omega_s - \omega_m)t] \quad (6.4)$$

Equation (6.4) represents the PAM wave. It can be seen that the wave contains upper and lower side-band frequencies besides the modulating and pulse signals.

6.5.3 Frequency Spectrum of PAM

Figure 6.3 shows the frequency spectrum of PAM.

The unmodulated pulse train has a spectrum of discrete frequencies $f_{c,}$ $2f_{c,}$ $3f_c$, etc. Each of these frequencies will have a pair of side bands USB and LSB as shown in Figure 6.3.

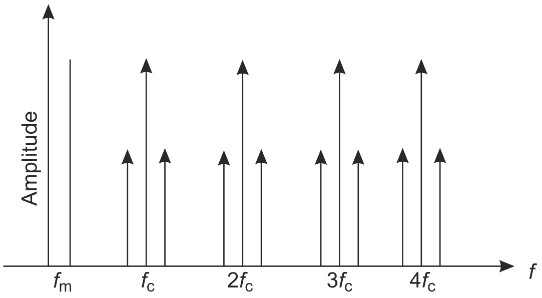


Fig. 6.3 Frequency Spectrum of PAM

6.5.4 Bandwidth of PAM

In PAM signal, the pulse duration τ is considered very small in comparison to the time period T_s between any two samples.

$$\tau << T_{\mathrm{s}} \tag{6.5}$$

Now if the maximum frequency in the modulating signal is f_{m} then the sampling frequency f_{s} must be equal to or higher than the Nyquist rate,

$$f_{\mathrm{s}} \geq 2f_{\mathrm{m}} \tag{6.6}$$

$$\frac{1}{T_{\mathrm{s}}} \geq 2f_{\mathrm{m}} \tag{6.7}$$

$$T_{\mathrm{s}} \leq \frac{1}{2f_{\mathrm{m}}} \tag{6.8}$$

Therefore,

$$\tau << T_{\mathrm{s}} \leq \frac{1}{2f_{\mathrm{m}}} \tag{6.9}$$

If the ON and OFF time of pulse-amplitude-modulated pulse is same then the maximum frequency of the PAM will be

$$f_{\mathrm{max}} = \frac{1}{2\tau} \tag{6.10}$$

Therefore, the bandwidth required for the transmission of a PAM would be equal to the maximum frequency f_{max} given by the above equation.

Thus, the transmission bandwidth will be

$$BW \geq f_{\mathrm{max}} \tag{6.11}$$

But,

$$f_{\mathrm{max}} = \frac{1}{2\tau} \tag{6.12}$$

Therefore,

$$BW \geq \frac{1}{2\tau} \tag{6.13}$$

Since

$$\tau << \frac{1}{2f_{\mathrm{m}}} \tag{6.14}$$

Therefore,

$$BW \geq \frac{1}{2\tau} >> f_{\mathrm{m}} \quad \text{or} \tag{6.15}$$

$$BW >> f_{\mathrm{m}} \tag{6.16}$$

The incoming information sources are sampled every $\frac{1}{2f_{\mathrm{m}}}$ second and these are multiplexed with m signals, so the output will be coming at a rate of $2m.f_{\mathrm{m}}$ per second. These samples are used to AM a carrier and so as the bandwidth is large, to decrease the bandwidth the output is filtered. For PAM, lowpass filter needs a bandwidth of $m \,.\, f_{\text{m}}$ Hz. So, bandwidth of PAM-AM is $2m \,.\, f_{\text{m}}$, which is AM-DSB whereas the bandwidth of AM-SSB is $m \,.\, f_{\text{m}}$.

EXAMPLE 6.1

A pulse-amplitude-modulated transmission of a signal has a maximum frequency of 3 kHz with sampling frequency of 8 kHz and the pulse duration of 0.1T_{s}. Calculate the transmission bandwidth.

Solution

The sampling period is expressed as

$$T_{\text{s}} = \frac{1}{f_{\text{s}}}$$

$$f_{\text{s}} = 8 \text{ kHz}$$

$$\therefore \quad T_{\text{s}} = \frac{1}{f_{\text{s}}} = \frac{1}{8 \times 10^3} = 125\,\mu\text{s}$$

It is given that

$$\tau = 0.1\, T_{\text{s}}$$

$$\tau = 0.1 \times 125\ \mu\text{s} = 12.5\ \mu\text{s}$$

Bandwidth for PAM signal is

$$BW \geq \frac{1}{2\tau}$$

$$\therefore \quad BW \geq \frac{1}{2 \times 12.5 \times 10^{-6}} \geq 40 \text{ kHz}$$

EXAMPLE 6.2

For a PAM transmission of a voice signal with modulating frequency of 5 kHz, find the bandwidth if the sampling frequency is 10 kHz and sampling period is 0.1 $T_{s.}$

Solution

$$T_{\text{s}} = \frac{1}{f_{\text{s}}}$$

$$f_{\text{s}} = 10 \text{ kHz}$$

$$\therefore \quad T_{\text{s}} = \frac{1}{f_{\text{s}}} = \frac{1}{10 \times 10^3} = 100\,\mu\text{s}$$

It is given that

$$\tau = 0.1\, T_{\text{s}}$$

$$\tau = 0.1 \times 100\ \mu\text{s}$$

$$= 10\ \mu\text{s}$$

Bandwidth for PAM signal is

$$BW \geq \frac{1}{2\tau}$$

$$\therefore \quad BW \geq \frac{1}{2 \times 10 \times 10^{-6}} \geq 50\ \text{kHz}$$

6.5.5 Generation of PAM using AND Gate

In PAM, sample values are equal to zero and usually represented by nonzero amplitudes so that fixed dc level is added to the signal. The resulting pulses are always positive. PAM is often not used because the amplitude of pulses is not constant and goes against the basic advantage of pulse systems. Therefore, in practical PAM systems, the pulses are made to perform frequency modulation of the carrier rather than amplitude modulation. This is done as follows.

The signal to be converted to PAM is fed to one input of an AND gate as shown in Figure 6.4. Pulses at the sampling frequency are applied to the other input of AND gates. The output of the gate then consists of pulses at the sampling rate, equal in amplitude to the signal voltage at each instant. The pulses are then passed through a pulse-shaping network, which gives them flat. Frequency modulation is recovered with a standard FM demodulator. They are then fed to an ordinary diode detector, whose outputs are connected to a lowpass filter. The cut-off frequency of the filter is high enough to pass the highest signal frequency, but low enough to remove the sampling frequency ripple, so that an undistorted replica of the original signal is produced.

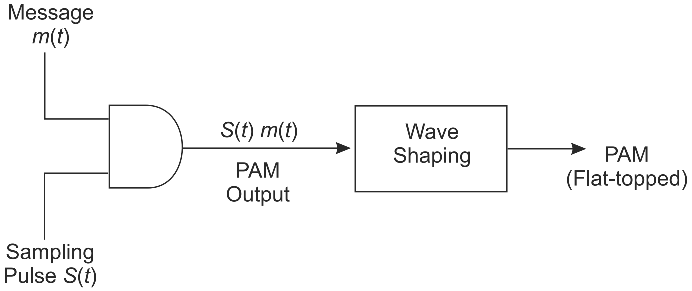


Fig. 6.4 Generation of PAM

6.5.6 Generation of PAM using CE Amplifier

Generation of PAM using a linear amplifier is shown in Figure 6.5. When either one of the signals is present, V_b is not sufficient to turn ON the transistor T, while on the other hand, if both the signals are present, T is switched ON. So the output is equal to message-signal amplitude. This output is present when $S(t)$ is present. For other times, the output is zero.

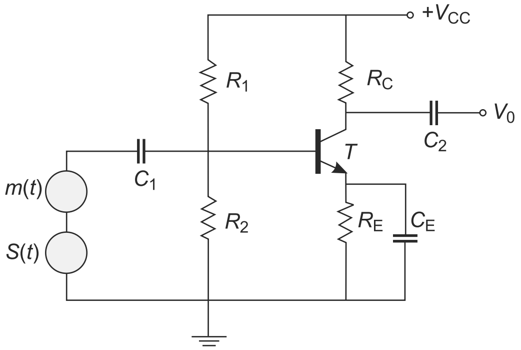


Fig. 6.5 Generation of PAM using CE amplifier

6.5.7 Sampling Considerations in PAM

The minimum sampling rate is equal to twice the highest harmonic in the modulating signal. It is called **Nyquist rate**. The sampling time controls the spacing between harmonics. The fundamental frequency determines the sampling rate. If the sampling time increases then spacing between harmonics decreases. If the width of the pulse is increased, it will interfere with the adjacent channels.

When the sampled signals of all the information signals are combined, it becomes necessary to retain the sharp transition between the pulses. This in turn leads to increase in the system bandwidth. If the bandwidth is limited to twice the highest harmonic of the information signal, i.e. $\frac{1}{\tau}$, then it will produce severe overlapping and much cross-talk. If the sampling time is reduced and the bandwidth is limited to $\frac{2}{\tau}$ then amplitude information is retained.

6.5.8 Demodulation of PAM Signals

Demodulation is the reverse process of modulation in which the modulating signal is recovered back from a modulating signal. For pulse-amplitude-modulated signals, the demodulation is done using a Holding circuit. Figure 6.6 shows the block diagram of a PAM demodulator.

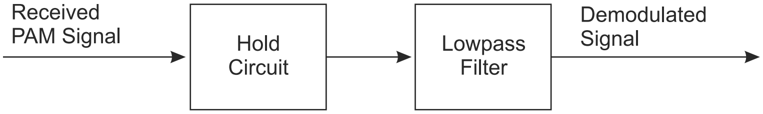
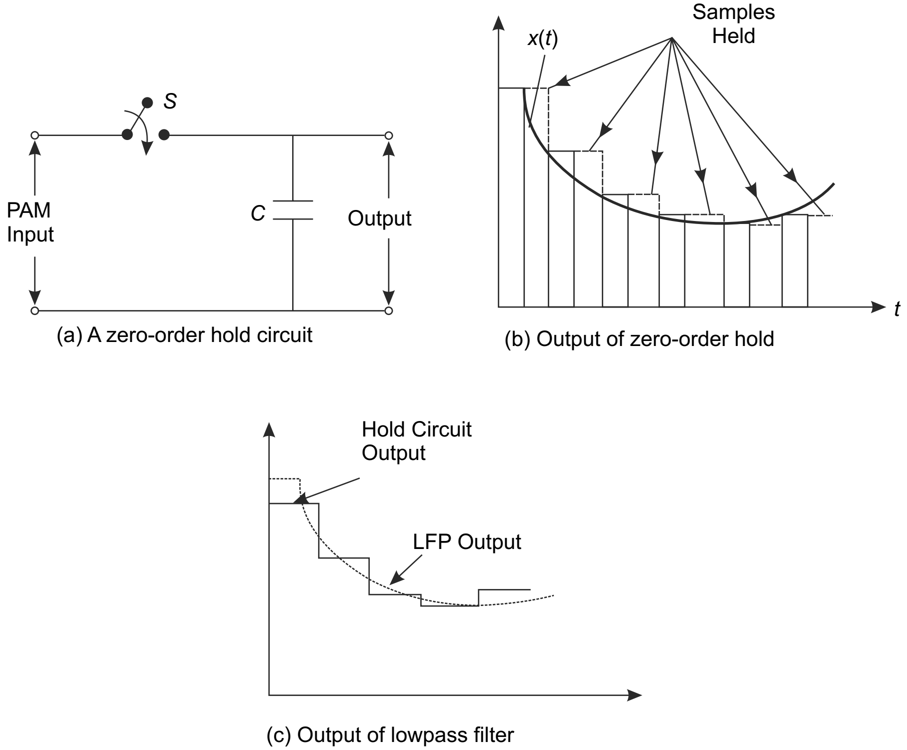


Fig. 6.6 Block diagram of a PAM demodulator

In this method, the received PAM signal is allowed to pass through a holding circuit and a lowpass filter. Figure 6.7 illustrates a zero-order holding circuit.

Here, Switch '*S*' is closed after the arrival of the pulse and it is opened at the end of the pulse. In this way, the capacitor *C* is charged to the pulse-amplitude value and it holds this



(a) A zero-order hold circuit

(b) Output of zero-order hold

(c) Output of lowpass filter

Fig. 6.7 Zero-order hold and its output

value during the interval between the two pulses. Hence, the sampled values are held as shown in Figure 6.7 (b). After this holding circuit, the output is smoothened in a lowpass filter as shown in Figure 6.7(c). The zero-order hold circuit considers only the previous sample to decide the value between the two pulses.

6.5.9 Transmission of PAM Signals

If the PAM signals are to be transmitted directly over a pair of wires then no further signal processing is necessary. However, if they are to be transmitted through the space using an antenna, they must first be amplitude- or frequency- or phase-modulated by a high-frequency carrier and only then can they be transmitted. Thus, the overall system will be taken, then known as PAM-AM or PAM-FM or PAM-PM respectively. At the receiving end, AM or FM or PM detection is first employed to get the PAM signal and then the message signal is recovered from it.

6.5.10 Limitations of PAM

There are some drawbacks of pulse amplitude modulation. They are as follows.

1. The bandwidth required for the transmission of a PAM signal is very large in comparison with the maximum frequency present in the modulating signals.
2. Due to the variation in the amplitude of the PAM pulses in accordance with the modulating signal, the interference of noise is maximum which cannot be removed easily.
3. Due to the variation in the amplitude of the PAM pulses, there is also variation in the peak power required by the transmitter with modulating signal.

6.6 PULSE DURATION MODULATION (PDM)

Pulse Duration Modulation (PDM) is also referred to as **Pulse Width Modulation (PWM)** or **Pulse Length Modulation (PLM)** and along with pulse position modulation, it may be termed as a form of pulse time or pulse frequency modulation.

In general, PWM requires a greater average power than PAM system. Also, PWM requires a larger bandwidth than PAM.

6.6.1 Definition

In this modulation, pulse duration is altered in accordance with the amplitude of the modulating signal to accommodate the information signal. It is apparent that the duration of the pulse can be controlled only to a certain degree.

6.6.2 Process of Pulse Duration Modulation

In PDM system, either the trailing edge or leading edge or both are altered in order to have the pulse accommodated to the intelligence of the information signal. The information signal controls the variation of pulse width, which must occur between certain limits determined by the sampling rate and the number of signals sampled, which determine the maximum pulse sample time. Therefore, the highest modulating frequency in the information signal will be helpful to control the bandwidth.

It is obvious that the maximum pulse width of any one sample cannot exceed the time between adjacent samples. Otherwise two pulses will overlap, which will cause loss of signal information. Theoretically, there is no maximum pulse width, except that of the practically impossible zero. Narrower pulses do allow possibility of increased channel capacity. In this case, the width of sampling signal is varied in accordance with the amplitude of the message signal at the instant of sampling.

6.6.3 Graphical Representation of Pulse Duration Modulation

Figure 6.8 shows the graphical representation of Pulse Width Modulation.

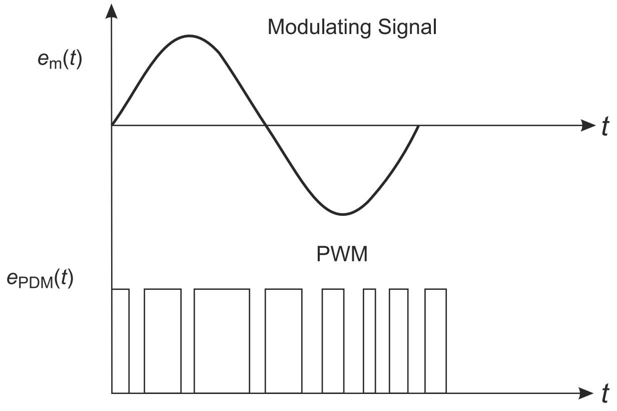


Fig. 6.8 Graphical representation of pulse width modulation

6.6.4 Mathematical Expression of Pulse Duration Modulation

Consider a PDM system with leading and trailing edge of the pulse varied as a function of sampled value. Assume that message has the highest frequency spectral component f_m Hz. If there are n information sources from which messages are to be transmitted then maximum permissible pulse width without overlapping is $\frac{1}{2f_m}$ seconds.

For any given information source $e_m(t)$, pulse width is given by

$$\tau = \frac{T}{2m} + K\frac{T}{2m}e_m(t) \qquad (6.17)$$

$$= \tau_0 + \tau_0 e_m(t) \qquad (6.18)$$

where T is the sampling period for one channel,

m is the number of information sources multiplexed, and

K is an arbitrary constant.

If $+1 \leq e_m(t) \leq -1$ then minimum pulse width is $\frac{T}{2m(1-K)}$

and the maximum pulse width is $\frac{T}{2m(1+K)}$

6.6.5 Representation of Pulse Duration Modulation

Pulse width modulation is the process in which the width of a pulse is varied as a function of the amplitude of the modulating signal.

Let
$$e_i(t) = \frac{\tau A}{T} + \frac{2\tau A}{T}\sum_{n=1}^{\infty}\frac{\sin\frac{n\omega_s\tau}{2}}{\frac{n\omega_s\tau}{2}}\cos n\omega_s t \tag{6.19}$$

$$= \beta + 2\beta\sum_{n=1}^{\infty}\frac{\sin x}{x}\cos n\omega_s t \tag{6.20}$$

where
$$\beta = \frac{\tau A}{T} \tag{6.21}$$

τ is pulse width,

T is sampling interval, and

A is amplitude of the pulse.

According to the definition of PDM, the width is varied with respect to the amplitude of the modulating signal.

$$\therefore \quad \tau = (\tau_0 + \tau_0\, e_m(t)) \tag{6.22}$$

$$e_i(t) = \frac{A}{T}(\tau_0 + \tau_0 e_m(t)) + \frac{\frac{2\tau A}{T}}{\frac{n(2\pi f_s)\tau}{2}}\sum_{n=1}^{\infty}\sin\frac{n\omega_s\tau}{2}\cos n\omega_s t \tag{6.23}$$

$$= \frac{A}{T}(\tau_0 + \tau_0 e_m(t)) + \frac{2}{n\pi}\sum_{n=1}^{\infty}\sin\frac{n\omega_s\tau}{2}\cos n\omega_s t$$

$$= \frac{A}{T}(\tau_0 + \tau_0 e_m(t)) + \frac{2}{n\pi}\sum_{n=1}^{\infty}\sin\left[\frac{n\omega_s}{2}(\tau_0 + \tau_0 e_m(t))\right]\cos n\omega_s t$$

$$= \frac{A}{T}(\tau_0 + \tau_0 e_m(t)) + \sum_{n=1}^{\infty}\frac{2}{n\pi}\sin\left[\frac{n\omega_s\tau_0}{2}(1 + e_m(t))\right]\cos n\omega_s t \tag{6.24}$$

The above equation contains dc terms, the message term and the modulated term. The third term is similar to the frequency-modulated term and hence it may be expanded using Bessel function.

6.6.6 Bandwidth of PDM

PDM needs a sharp rise time and fall time for pulses in order to preserve the message information. Rise time should be very much less than $T_{s.}$

$$t_r << T_s \tag{6.25}$$

and the transmission bandwidth must be
$$BW \geq \frac{1}{2t_r} \tag{6.26}$$

Thus, the transmission bandwidth of PDM is higher than PAM. The power requirement of PDM is less and it may be further reduced by transmitting only edges rather than pulses.

6.6.7 Generation of Pulse Duration Modulation using Sample-and-Hold Circuit

A simple method for the generation of PDM is to sum the message and sawtooth waveform and apply it to a clipping-and-squaring circuit. Figure 6.9(a) and (b) shows the method of generation of PDM and its associated waveforms.

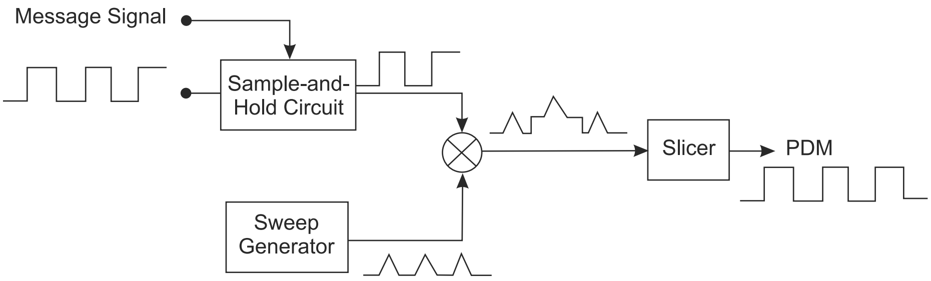


Fig. 6.9(a) Generation of PDM

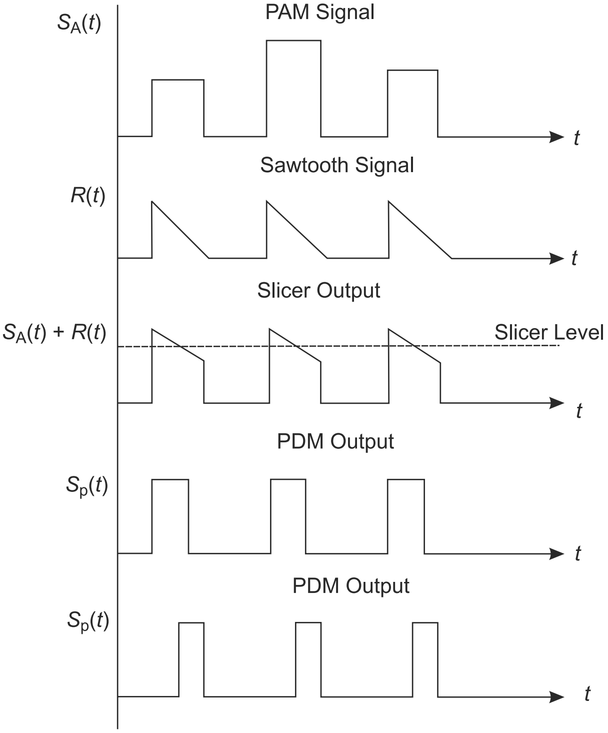


Fig. 6.9(b) PDM from PAM wave

The circuit produces +*A* volts whenever the input exceeds the slicing level, otherwise the output is zero. Then the output is a trailing-edge modulated PDM. For generation of PDM, the sampling and modulation operations have been combined. If the sawtoothed waveform is reversed, a leading-edge modulated waveform will be the result. If a triangular waveform is used instead of sawtooth waveform, it will produce modulation at both edges.

The final duration corresponds to the message samples at the time location of the modulated edge and not the apparent sampling time $t = mT_s$. It means that the sample values are non-uniformly spaced. The difference between uniform and non-uniform sampling is negligible if the pulse duration is small when compared to $T_{s.}$

6.6.8 Generation of Pulse Duration Modulation using Monostable Multivibrator

Figure 6.10 shows the PDM generation using a monostable multivibrator. In this method, trigger pulses are applied at the sampling rate to control the starting time of pulses from a monostable multivibrator and the input signal fed to be sampled for controlling the duration of these pulses.

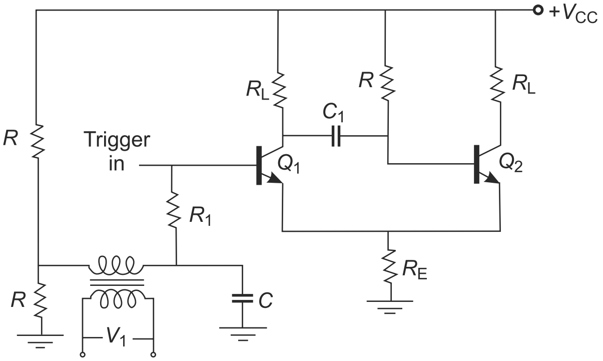


Fig. 6.10 Generation of PDM using a monostable multivibrator

The emitter-coupled monostable multivibrator makes an excellent voltage-to-time converter. Its gate width is dependent on the voltage to which the capacitor is charged. If this voltage is raised in accordance with a signal voltage, a series of rectangular pulses will be obtained with a required variation. This circuit performs the two jobs of sampling and converting samples into PDM.

The stable state for this type of multivibrator is with Q_1 OFF and Q_2 ON. The applied trigger pulse switches Q_1 ON whereupon the voltage at C_1 falls as Q_1 now is beginning to draw collector current. The voltage at B_2 follow suit and Q_2 is switched OFF by regenerative action. As soon as this happens, C_1 begins to charge up to the collector supply potential through *R*.

After a time determined by the supply voltage and the *RC* time constant of the charging network, q_2 turns ON. Q_1 is simultaneously switched OFF by regenerative action and stays

OFF until the arrival of the next trigger pulse. The voltage at the base Q_2 must be (to allow Q_2 to go ON) slightly more positive than the voltage across the common emitter resistor R_E. This voltage depends on the current flowing through which at this time is the collector current of Q_1. The collector current in turn depends on the base bias, which is governed by the instantaneous changes in the applied voltage.

Thus, the applied modulation voltage controls the voltage to make the transistor Q_1 ON. As the rise of V_{B2} is linear, the modulation voltage is seen to control the period of time during which Q_2 is OFF, that is the pulse duration.

6.6.9 PDM Transmission

One method of transmitting more efficiently is to transmit only the pulse signals. Figure 6.11 shows the arrangement for PDM transmission.

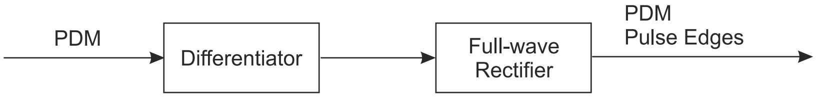


Fig. 6.11 PDM transmission using pulse edges

For proper transmission of PDM in the form of pulse edges, a differentiator and full-wave rectifier are to be used. At the receiving end, the appropriate pulse can be stretched to recover the original pulse.

6.6.10 Demodulation of PDM

For the purpose of PDM demodulation, there are two methods. They are as follows.

1. Conversion of PDM to PAM.
2. Demodulation of PDM is done by means of lowpass filtering provided that the PM side bands do not overlap the message band. This is facilitated by the presence of non-uniform sampling in PDM.

Figure 6.12 shows the PDM to PAM conversion. There is a constant-current generator which assumes a linear voltage rise on the capacitor.

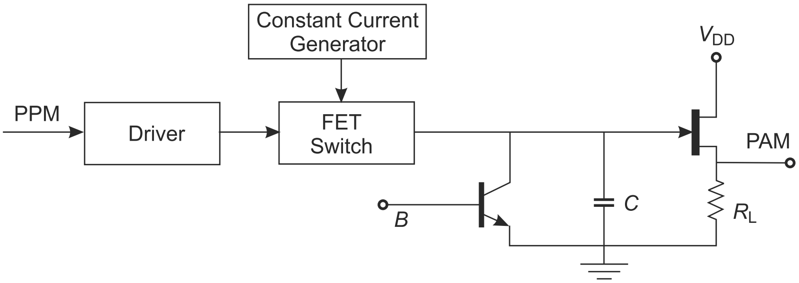


Fig. 6.12 PDM to PAM conversion

The discharging pulses have to be located beyond the widest pulse width. The easiest way to obtain such pulses is to produce narrow pulses by triggering a multivibrator and delaying it with another monostable.

6.7 PULSE POSITION MODULATION (PPM)

6.7.1 Definition

Pulse Position Modulation (PPM) is a process by which the position of a pulse is varied in accordance with information contained in the sample waveform. Since there is no change in pulse width, the bandwidth required for transmission of pulse information remains stationary. Narrow pulse systems require greater bandwidths, but their information capacity is increased.

6.7.2 Generation of PPM from PDM

In PDM generation, the repetition rate of the leading edge is fixed, whereas that of the trailing edge is not a fixed one. The position of the latter depends upon pulse width which, in turn, is determined by the signal amplitude at that instant of time. In other words, the trailing edges of PDM pulses are position modulated. To generate PPM, the leading edges of the PDM pulses have to be removed. The waveform of PPM through PDM is shown in Figure 6.13

The train of pulses obtained now is then differentiated. As a result, another train of pulses will be obtained. This has positive-going narrow pulses corresponding to leading edges and negative-going pulses corresponding to trailing edges. If the position corresponding to the

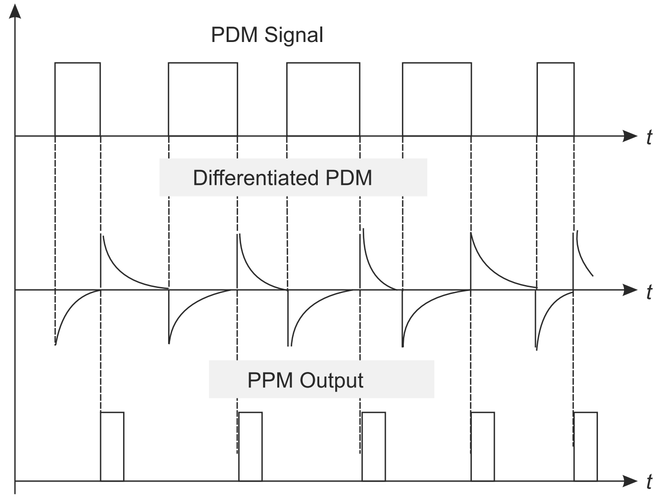


Fig. 6.13 Waveform of PPM through PDM

trailing edge of an unmodulated pulse is counted as zero displacement then other trailing edges will arrive earlier or later. They will, therefore, have a time displacement other than zero; this time displacement is proportional to the instantaneous value of the signal voltage. The differentiated pulse corresponding to the leading edges are removed with a diode clipper or rectifier and the remaining pulses are position modulated.

6.7.3 Generation of PPM from PAM

The block diagram given in Figure 6.14 shows the PPM generation through PAM. The message signal $e_m(t)$ is first converted into PAM signal by means of a sample- and-hold circuit which is used to generate a staircase waveform. The pulse duration of a sample- and-hold circuit is the same as the sampling duration T_s. The staircase waveform is added to a sawtooth wave, yielding the combined signal which is to be applied to a threshold detector.

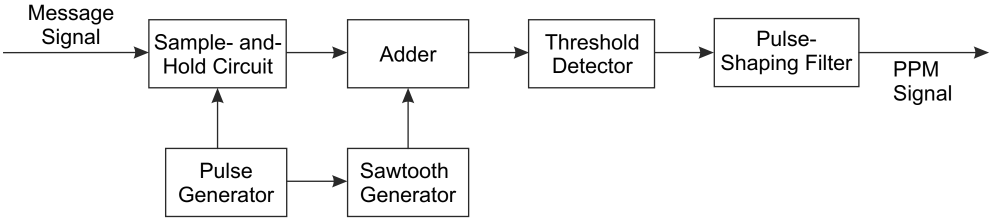


Fig. 6.14 PPM generation through PAM

Threshold detector produces a very narrow pulse each time the adder output passes through a zero crossing in the negative-going direction. Finally, the PPM signal is generated by using this sequence of impulses to excite a filter.

6.7.4 Process of Detection of PPM

In order to detect the PPM output, the following steps are to be followed.

1. Convert the received PPM wave into a PDM wave with the same modulation.
2. Integrate this PDM wave using a device with a finite integration time, thereby computing the area under each pulse of PDM wave.
3. Sample the output of the integrator at a uniform rate to produce a PAM wave whose pulse amplitude is proportional to the signal samples of the original PPM wave.

6.7.5 Comparison with PDM

1. In PDM, due to variation in width of the pulses, the power content is not a constant. This leads to the disadvantage that the transmitter must be powerful enough to handle maximum width pulses.

2. PPM is superior to PDM for message transmission because the wide pulse width of PDM requires more energy than PPM when transmitted.
3. PDM is suitable for communication in the presence of noise. Very high peak narrow pulses can be transmitted and the pulse position can be determined even when the noise level is high.
4. PPM requires large bandwidth for transmitting very narrow pulses. Hence, it is suitable for optical fibre communication.
5. PDM is also useful even though synchronisation between the transmitter and the receiver fails. PPM is not suitable in that situation.

6.7.6 Comparison of Various Pulse-Analog-Modulation Methods

Table 6.1 shows the comparison between various pulse-analog-modulation methods.

Table 6.1 Comparison between various PAM methods

S.No	Pulse Amplitude Modulation (PAM)	Pulse Duration/Width Modulation (PDM/PWM)	Pulse Position Modulation (PPM)
1	Amplitude of the pulse proportional to amplitude of modulating signal	Width of the pulse is proportional to amplitude of modulating signal	The relative position of the pulse is proportional to amplitude of modulating signal
2	Bandwidth of the transmission channel depends on the pulse width	Bandwidth of the transmission channel depends on the rise time of the pulse	Bandwidth of the transmission channel depends on the rising time of the pulse
3	Instantaneous power of the transmitter varies	Instantaneous power of the transmitter varies	Instantaneous power of the transmitter remains constant
4	Noise interference is high	Noise interference is minimum	Noise interference is minimum
5	System is complex to implement	System is simple to implement	System is simple to implement
6	Similar to amplitude modulation	Similar to frequency modulation	Similar to phase modulation

6.8 DIGITAL PULSE MODULATION

In contrast to analog pulse modulation, in digital pulse modulation the amplitude or timing of the transmitted pulse-modulated waveform varies continuously with the amplitude of the modulating signal. In this type, the input signals are encoded which consist of binary digits.

Digital pulse modulation can be further divided into the following types.

1. Pulse Code Modulation (PCM)
2. Delta Modulation (DM)

6.8.1 Pulse Code Modulation

Pulse code modulation is entirely a different type of system than PAM, PDM or PPM. In analog pulse-modulation methods, either the information signal continuously modulated a sinusoidal carrier or the information signal was sampled and the sample periodically modulated a pulse. In all cases, the actual information signal was used to obtain the modulated output directly.

In PCM, the sampled signal is rounded off to the nearest value which is permitted for transmission by the system. The process of rounding off is termed **quantisation** while the possible levels permitted for transmission are called **quantising levels**. If the quantising levels are equally spaced, it is called **uniform quantisation**. If quantising levels are not uniformly spaced, the process is termed **non-uniform quantisation**.

1. Process of Pulse Code Modulation

The signal to be transmitted is first sampled and the samples are given to a quantiser for the rounding-off operation. The quantised pulses are coded into groups using a binary code. In the binary code, only two levels are transmitted, usually 1 and 0, corresponding to carrier ON and OFF. Each pulse group transmitted represents the quantising levels as a binary number. The maximum number of pulses in the pulse group depends upon the total number of quantising levels used in the system. A 5-bit code has 32 quantising levels. In general, an '*n*' bit code has 2^n quantising levels.

However, the actual signals are most likely to have both positive and negative values causing difficulty in coding. This problem is overcome by adding a dc bias to the signal so that the signal will always remain positive. Another problem with speech signals is large amplitude variation which then requires a large number of quantising levels. Amplitude compressor circuits are employed to reduce large peaks in the signals and this reduces the number of quantising levels for a given accuracy and also reduces the channel bandwidth. At the receiver, the expander circuit is to be included to bring the compressed signal back to its original form.

The difference between the analog signal levels and the quantised signal levels is termed as **quantisation noise**. This noise can be reduced by using a large number of quantising levels at the expense of increased cost and bandwidth. A bipolar binary code with voltage levels of +1 V and –1 V leads to some power saving in the system. Here, the positive voltage pulse represents a binary 1 and a negative voltage pulse represents 0.

Figure 6.15 shows the quantising levels and pulse-code modulated trains for a typical biased signal.

The signal is biased by a dc voltage in such a way that it does not become negative at any instant. This biased signal is now sampled at fixed instants $t_0 \rightarrow t_5$ and the signal amplitudes at

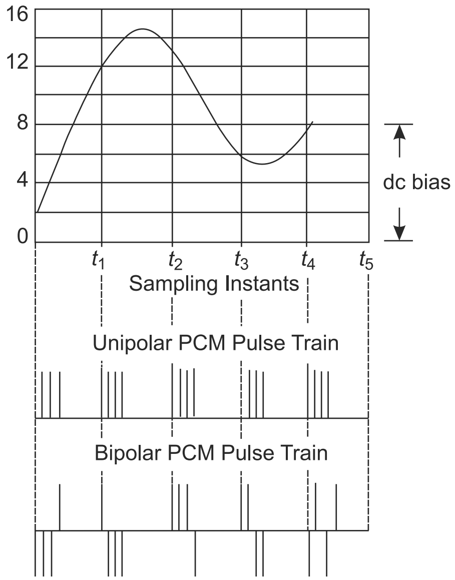


Fig. 6.15 Quantising levels and PCM trains for a typical signal

these instants are converted into binary. Thus, at the instant t_0, sample amplitude equals ONE equalizing level. If a 4-bit binary code is used, this will be represented by 0001. Similarly, at time t_2, the signal amplitude equals 10 units which is represented by 1010 in the binary code. As these pulses are required to be transmitted during the sampling interval allotted for the channel, narrow pulse widths are used for PCM with resultant increase in the bandwidth. If positive and negative pulses are employed for transmission of 1 and 0, the resulting ternary PCM pulse train is produced. Thus, PCM provides a communication system in which the signal is converted into binary. Due to this, the system is commonly referred to as digital communication system.

2. Binary Coding

Assume that an audio signal varies between 0 and 15 volts upon the application of the appropriate bias. It is desired to encode this information into binary digit form. For this, the signal must be sampled at a rate equal to the Nyquist sampling rate which is twice the highest harmonic in the information signal. This will allow sampling of other signals and transmit them in a time-sequential manner. This portion of the sampling and modulation corresponds to PAM.

The pulse-code information must be established before transmission occurs. It is assumed that for this system, the nearest integer voltage will be transmitted. Therefore, a binary group of four pulses is necessary to establish the code group for each message. The possible messages are the amplitude heights, which are 1, 2, 3, …, 15. The messages with their associated code group are shown in Table 6.2

Table 6.2 Pulse codes of amplitude levels

Message	Code	Message	Code
0	00000	9	01001
1	00001	10	01010
2	00010	11	01011
3	00100	12	01100
4	00101	13	01101
5	00110	14	01110
6	00111	15	01111
7	01000	16	10000
8	01001		

Here, eight quantum levels are used corresponding to 0 to 7 V. The encoder translates quantised voltage to its binary equivalent according to the above table.

3. PCM Bandwidth

The bandwidth is determined by the minimum pulse width. This is controlled by the number of pulses needed to establish a message and the rate at which the information is transmitted. Since the information is not in the rise or fall time that establishes the position of the pulse from some reference, the bandwidth requirement is less, provided that the presence of a pulse is detectable. The bandwidths are closer to those of the PAM system.

The rate at which quantised samples occur is $f_s > \omega$ samples/second. It means that there must be vfs coded pulses per second. Assuming there are no spaces in the coded signal, the maximum permitted duration of any pulse will be $\tau = \frac{1}{vfs}$.

Therefore, to ensure pulse resolution, the required bandwidth should at least correspond to $\frac{1}{2\tau}$, that is

$$B_T \geq \frac{1}{2\tau} = \frac{1}{2} vf_s \quad \text{or} \tag{6.27}$$

Since $f_s \geq 2f_m$

$$B_T \geq \frac{1}{2\tau} \geq vf_m \tag{6.28}$$

EXAMPLE 6.3

A TV signal having a bandwidth of 4.5 MHz is transmitted using PCM. Given that the number of quantisation levels is 512, determine (a) code-word length, (b) bandwidth, and (c) signalling rate.

Solution

(a) Determination of code-word length:

$$f_{\mathrm{m}} = 4.5 \text{ MHz}$$

Quantisation levels $q = 512$

$$q = 2^v$$

$$512 = 2^v$$

$$\log_{10} 512 = v \log_{10} 2$$

$$v = \frac{\log_{10} 512}{\log_{10} 2} = 9$$

Thus, the code-word length = 9 bits

(b) Determination of bandwidth:

$$BW \geq vf_{\mathrm{m}}$$

$$\geq 9 \times 4.5 \times 10^6 \quad \geq 40.5 \text{ MHz}$$

(c) Determination of signalling rate:

$$r = vf_s$$

$$f_{\mathrm{s}} \geq 2f_{\mathrm{m}}$$

$$f_{\mathrm{s}} \geq 2 \times 4.5 \text{ MHz}$$

$$f_{\mathrm{s}} \geq 9 \text{ MHz}$$

$$\therefore \quad r = vf_{\mathrm{s}} = 9 \times 9 \times 10^6 = 81 \times 10^6 \text{ bit/s}$$

4. PCM Detection

Detection of the pulse-code modulation requires knowledge of the code and the synchronisation. This information can be transmitted by appropriate pulse-code information.

Although the signal is noise, distortion and bandwidth limitation, the regenerated signal will correspond to the original one. The regenerator compares the signal with threshold decision level or the level midway between the two expected signal levels at each sampling instant. If positive, a positive pulse is generated, if negative a negative pulse is generated.

On a PCM channel, in addition to speech signal information, signalling information such as ON and OFF hook and dialing must also be transmitted. Moreover, synchronising information must be generated so that the receiver can sort out the code words and signalling bits into the proper channels. Alarm codes can also be added to inform the abnormal conditions.

It is obvious that these systems are complex, but the advantages they offer more than compensate for their complexity. The choice of the proper code can greatly aid the demodulation process.

5. PCM Transmission System

Figure 6.16 shows the block diagram of a PCM transmitting system.

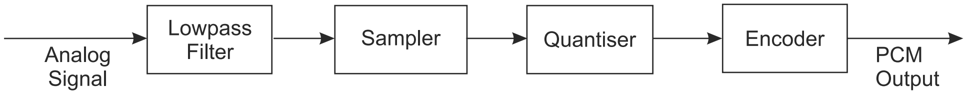


Fig. 6.16 Block diagram of a PCM transmitting system

The transmitting system consists of a lowpass filter with a cut-off frequency half of the sampling frequency. The output of the filter is to a sampler by the quantiser circuit and finally converted as a PCM pulse train by an encoder circuit.

6. PCM Reception System

Figure 6.17 shows the block diagram of a PCM receiving system.

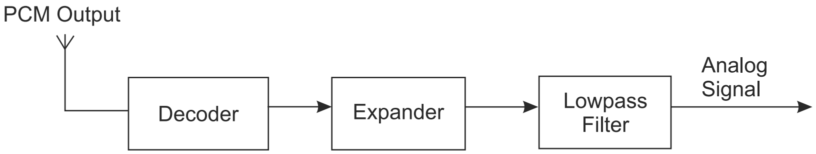


Fig. 6.17 Block diagram of a PCM receiving system

At the receiver side, the received signals from the transmitting section are decoded and converted into equivalent analog signals. Higher frequency components present in the output are attenuated by a lowpass filter. The receiver may include an expander circuit if a compressor circuit has been employed in the transmitter.

7. Differential PCM (DPCM)

The bandwidth of a PCM system depends upon the highest frequency in the analog signal and the number of bits in the PCM signal. for a speech signal of 3.2 kHz, the usual sampling rate is 8 kHz and if the system has 128 quantising levels, the number of bits required is $2^n = 128$ or $n = 7$. The final PCM signal will therefore have a bit rate of 7 × 8 = 56 kilo bits/s. As a result, the system will require a large bandwidth. To reduce this high bit rate in the modulated pulse train, a technique known as **differential pulse code modulation** has been developed.

In DPCM, the difference between two successive signals rather than the samples themselves are quantised and coded into a PCM pulse train. As a result, the bit rate is considerably reduced and the signal-to-noise ratio is also improved. The basic reason for this improvement is that in speech signals, large variations in the amplitude of one sample from the other is most unlikely and if these differences are quantised, it would require a fewer of bits in the transmission system. The diagram of a DPCM modulator is shown in Figure 6.18.

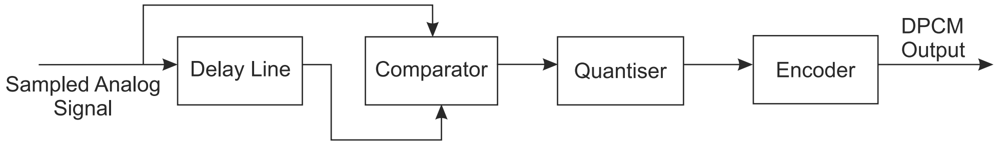


Fig. 6.18 DPCM modulator

The signals are sampled in the usual way and these samples are passed through a delay line to a comparator circuit which also gets the sampled signal input. The comparator output corresponds to the difference between the two successive samples. This difference is then quantised and converted into a pulse-train in the usual way. The system however is highly complicated. The technique has been used in experimental TV transmission.

8. Signal-to-Noise Ratio(S/N) for PCM

In a PCM signal, the signal-to-noise ratio (S/N) is given as

$$\frac{\text{S}}{\text{N}} = \frac{\text{Normalised signal power}}{\text{Normalised noise power}} \tag{6.29}$$

In PCM, the normalised power has been calculated as $P = \frac{\delta^2}{12}$ (6.30)

and the number of bits and quantisation levels are related as $q = 2v$ (6.31)

The input to a linear quantiser has continuous amplitude in the range $-x_{\text{max}}$ to $+x_{\text{max}}$

Therefore the total amplitude range $= x_{\text{max}} - (-x_{\text{max}}) = 2x_{\text{max}}$ (6.32)

Now the step size will be $\delta = \frac{2x_{\text{max}}}{q}$ (6.33)

$$\therefore \qquad \delta = \frac{2x_{\text{max}}}{2^v} \tag{6.34}$$

$$\therefore \quad \frac{\text{S}}{\text{N}} = \frac{P}{\left(\frac{2x_{\text{max}}}{2^v}\right)^2 \cdot \frac{1}{12}} \tag{6.35}$$

$$= \frac{P}{\frac{4x_{\text{max}}^2}{2^{2v}} \cdot \frac{1}{12}}$$

$$= \frac{3P}{x_{\text{max}}^2} . 2^{2v} \text{ bits/sample} \tag{6.36}$$

If the input is assumed to be normalised,

$$x_{\max} = 1 \tag{6.37}$$

Then S/N will be

$$\frac{\mathrm{S}}{\mathrm{N}} = 3 \times 2^{2v} \times P \tag{6.38}$$

Also, if the destination signal power P is normalised,

$$P \leq 1 \tag{6.39}$$

Then S/N will be

$$\frac{\mathrm{S}}{\mathrm{N}} = 3 \times 2^{2v} \tag{6.40}$$

Expressing S/N in decibels,

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)\mathrm{dB} = 10\log_{10}\left(\frac{\mathrm{S}}{\mathrm{N}}\right)\mathrm{dB} \tag{6.41}$$

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)\mathrm{dB} \leq 10\log_{10}\left[3 \times 2^{2v}\right]$$

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)\mathrm{dB} \leq [4.8 + 6v]\mathrm{dB} \tag{6.42}$$

EXAMPLE 6.4

For Example 6.3, find the (S/N) in decibels.

Solution

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right)\mathrm{dB} \leq [4.8 + 6v]\mathrm{dB}$$

Code-word length calculated as $v = 9$

$$\therefore \quad \left(\frac{\mathrm{S}}{\mathrm{N}}\right)\mathrm{dB} \leq [4.8 + 6 \times 9]\mathrm{dB} \quad \leq 58.8\ \mathrm{dB}$$

EXAMPLE 6.5

The bandwidth of an input signal to PCM is restricted to 3 kHz. The input signal varies in amplitude from – 4 V to + 4 V and has the average power of 30 mW. The required S/N is 20 dB. Find the number of bits per sample and transmission bandwidth.

Solution

(a) Determination of number of bits per sample:

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right) = \frac{3P}{x_{\max}^2} \,.\, 2^{2\mathrm{v}}$$

It is given that

$$\left(\frac{\text{S}}{\text{N}}\right) \text{dB} = 10 \log_{10}\left(\frac{\text{S}}{\text{N}}\right) = 20 \text{ dB}$$

$$\therefore \quad \left(\frac{\text{S}}{\text{N}}\right) = 100$$

$$x_{\text{max}} = 4\text{V}$$

$$\text{P} = 30 \text{ mW}$$

$$\therefore \quad 100 = \frac{3 \times 30 \times 10^{-3} \times 2^{2\,\nu}}{4^2}$$

On solving, ν = 7 bits

(b) Determination of transmission bandwidth:

$$BW \geq \nu f_{\text{m}}$$

$$\geq 7 \times 3 \geq 21 \text{ kHz}$$

The signalling rate is two times the transmission bandwidth.

$$\therefore \quad r = 21 \times 2 = 42 \text{ bits/s}$$

EXAMPLE 6.6

An analog voltage signal has a bandwidth of 100 Hz and amplitude range of –10 V to +10 V is transmitted over a PCM system with an accuracy of ±0.1% (full scale). Find the required sampling rate and number of bits in each PCM word.

Solution

Accuracy of PCM = ±0.1% (full scale),

i.e. $$e_{\text{max}} = \pm 0.1\% = \pm 0.001$$

$$e_{\text{max}} = \left|\frac{\delta}{2}\right| = 0.001$$

$$\therefore \quad \delta = 2 \times 0.001 = 0.002$$

It is known that

$$\delta = \frac{2\,x_{\text{max}}}{q}$$

$$\delta = \frac{2\,x_{\text{max}}}{q}$$

$$\therefore \quad 0.002 = \frac{2\times 10}{q}$$

$$q = \frac{20}{0.002} = 10000$$

Therefore, the number of levels =10,000.

$$f_\mathrm{m} = 100 \text{ Hz}$$

But, $$f_\mathrm{s} \geq 2 f_\mathrm{m}$$

$$f_\mathrm{s} \geq 2 \times 100 \geq 200 \text{ Hz}$$

EXAMPLE 6.7

For Example 6.6, find (a) the number of bits for each sample, (b) bit rate, and (c) transmission bandwidth.

Solution

(a) Calculation of the number of bits for each sample:

$$q = 2^\nu$$

The number of levels calculated as 10,000

$$\therefore \quad 10000 = 2^\nu$$

$$\log_{10} 10000 = \nu \log_{10} 2$$

$$\nu = \frac{\log_{10} 10{,}000}{\log_{10} 2}$$

$$= 13.288 \cong 14 \text{ bits}$$

(b) Calculation of bit rate:

The bit rate is expressed as

$$r \geq \nu f_\mathrm{s}$$

$$r \geq 14 \times 200 \quad \geq 2800 \text{ bits/s}$$

(c) Calculation of transmission bandwidth:

$$BW \geq \frac{1}{2} r$$

$$\geq \frac{1}{2} \times 2800 \quad \geq 1400 \text{ Hz}$$

EXAMPLE 6.8

For a PCM system followed by a 7-bit binary encoder, with a bit rate of 50 × 10^6 bits/s, find the maximum message signal bandwidth for satisfactory operation and S/N ratio in dB.

Solution

Calculation of maximum message signal bandwidth:

$$f_\mathrm{s} \geq 2f_\mathrm{m}$$

The number of bits is given as $\nu = 7$

The bit rate is expressed as

$$r \geq \nu f_\mathrm{s}$$

$$r \geq 7 \times 2f_\mathrm{m}$$

$$50 \times 10^6 \geq 14f_\mathrm{m}$$

$$\therefore \qquad f_\mathrm{m} \geq 3.57 \text{ MHz}$$

Calculation of S/N ratio in dB:

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right) \mathrm{dB} = [4.8 + 6\nu]\,\mathrm{dB}$$

$$= [4.8 + 6 \times 7]\,\mathrm{dB} = 43.8 \text{ dB}$$

9. Companding in PCM

The quantisation error in pulse-code-modulation depends on the step size δ. When the steps are uniform in size, the small-signal amplitude signals would have poorer signals to quantisation noise ratio than the large amplitude signals. Since it is necessary to use a fixed number of quantisation levels, the only way to have a uniform S/N ratio is to adjust the step size in such a manner that the ratio remains *a*. This means that the step size must be small for small-signal amplitude signals and large for large-amplitude signals.

The effect of an adaptive step size may also be achieved in a more feasible way by distorting the signal before the quantisation process. So an inverse distortion has to be introduced at the receiver side to make the overall system distortionless.

Hence, the signal is amplified at low signal levels and attenuated at high signal levels. After this process, uniform quantisation is used. This is equivalent to more step size at low signal levels and small step size at high signal levels. At the receiver, a reverse process is done. That is, signal is attenuated at low level signals and amplified at high signal levels to get the original signal. Thus, the compression of signal at transmitter and expansion at the receiver is called **companding**. Figure 6.19 shows companding in PCM.

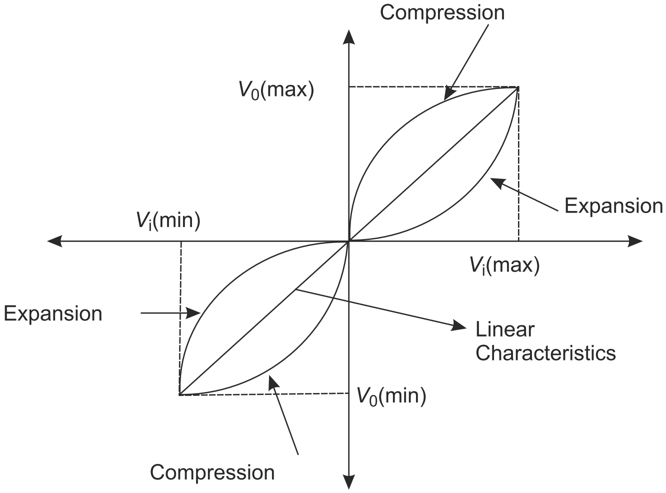


Fig. 6.19 Companding in PCM

There are two types of companding. They are

1. μ-law compander
2. *A*-law compander

The above two companders differ slightly in their compression and expansion curves. These two methods are compatible, but conversion circuits have been developed to convert μ-law to *A*-law and vice versa. The formulas for the conversion from one to one are given below.

μ-*law*:
$$V_{\text{out}} = \frac{V_{\text{m}} \ln\left(1+\mu.\dfrac{V_{\text{in}}}{V_{\text{m}}}\right)}{\ln(1+\mu)} \tag{6.43}$$

A-*law*:
$$V_{\text{out}} = \frac{1+\ln\left(\dfrac{AV_{\text{in}}}{V_{\text{m}}}\right)}{1+\ln A} \tag{6.44}$$

where V_{out} is the output voltage,

V_{m} is the maximum possible input voltage, and

V_{in} is the instantaneous value of the input voltage.

Usually, the value of μ is chosen as 255 and *A* is 87.6.

EXAMPLE 6.9

The input voltage of a PCM compander wit have minimum voltage range of 1 V and a μ of 255 is 0.25. Calculate the output voltage and gain of a μ-law compander.

Solution

For a μ-law compander, the output voltage can be expressed as

$$V_{\text{out}} = \frac{V_{\text{m}} \ln\left(1 + \mu . \dfrac{V_{\text{in}}}{V_{\text{m}}}\right)}{\ln(1+\mu)}$$

$$V_{\text{out}} = \frac{1.0 \cdot \ln\left(1 + 255 \cdot \left(\dfrac{0.25}{1}\right)\right)}{\ln(1+255)}$$

$$= \frac{\ln 64.75}{\ln 256} = 0.75 \text{ V}$$

$$\text{Gain} = \frac{V_{\text{out}}}{V_{\text{in}}}$$

$$= \frac{0.75}{0.25} = 3$$

EXAMPLE 6.10

The input voltage of a PCM compander with a minimum voltage range of 1 V and A of 87.6 is 0.25. Calculate the output voltage and gain of an A-law compander.

Solution

For an A-law compander, the output voltage can be expressed as

$$V_{\text{out}} = \frac{1 + \ln\left(\dfrac{AV_{\text{in}}}{V_{\text{m}}}\right)}{1 + \ln A}$$

$$V_{\text{out}} = \frac{1.0 \cdot \ln\left(1 + 87.6 \cdot \left(\dfrac{0.25}{1}\right)\right)}{\ln(1+87.6)}$$

$$= \frac{\ln 64.75}{\ln 86.6}$$

$$= \frac{3.13}{4.48} = 0.69 \text{ V}$$

$$\text{Gain} = \frac{V_{\text{out}}}{V_{\text{in}}}$$

$$= \frac{0.69}{0.25} = 2.8$$

6.8.2 Delta Modulation

Delta modulation is a method of information transmission with the help of pulses, but it uses a single-digit code that transmits information about the slope of the signal amplitude rather than the actual amplitude, as in PCM. In contrast to pulse-code modulation, delta modulation is not often used. However, it is being used in satellite communications, television and subscriber telephone loops.

1. Principle

In delta modulation, the modulator transmits binary output pulses whose polarity depends upon the difference between the modulating signal and the feedback signal corresponding to the signals previously sent. Figure 6.20 shows a simple delta-modulation system.

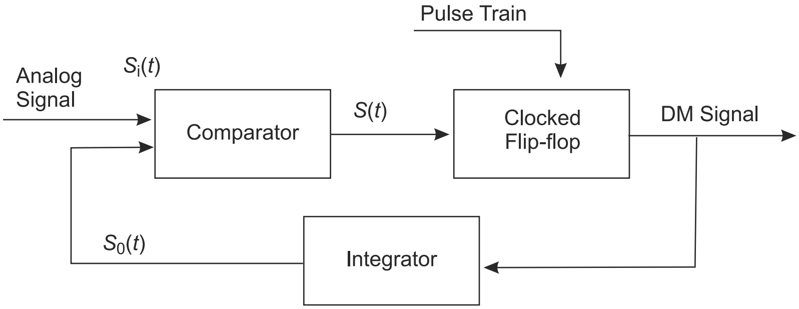


Fig. 6.20 Delta-modulation system

From the above figure, the analog input signal $S_i(t)$ and the integrated output signal $S_0(t)$ are compared in the comparator. The output of the comparator is proportional to the difference between the two inputs.

$$S(t) = S_i(t) - S_0(t)$$

From the comparison between the integrated feedback signal $S_0(t)$ and the input signal $S_i(t)$, it is found that when $S_i(t) > S_0(t)$, a positive pulse forms the output $S(t)$. When $S_i(t) < S_0(t)$, a negative pulse forms the transmitted signal. The output pulse train also forms the approximation $S_0(t)$ after being integrated. Figure 6.21 shows the waveforms of delta modulation.

As shown in the figure, a series of triangular waveforms results for $S_0(t)$ when there is no signal applied to the delta modulator. This is because integration of a constant gives a ramp. In this condition, the transmitted signal consists of alternate positive and negative pulses. The difference between the original signal $S_i(t)$ and the reconstructed signal $S_0(t)$ results in an error signal, which is often called quantisation noise. This noise can be decreased by either decreasing the magnitude of the step size or by increasing the sampling frequency.

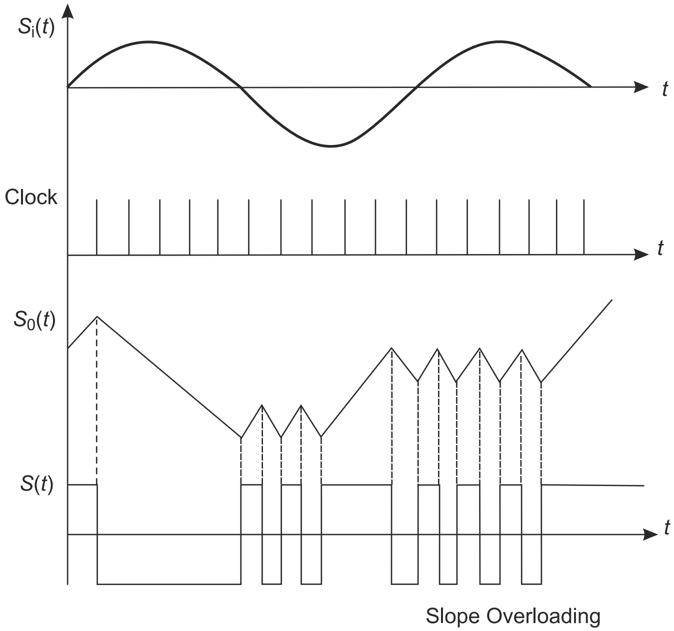


Fig. 6.21 Waveforms of delta modulation

2. Adaptive Delta Modulation

From the above waveforms of delta modulation, when $S_i(t)$ exceeds the feedback signal $S_0(t)$, the transmitted wave train remains positive. When $S_0(t)$ begins to overshoot $S_i(t)$, the encoder output begins transmitting a negative pulse. If the modulating signal changes more rapidly then the encoder can follow the slope overload that occurs. This slope overload causes waveform distortion and is one of the severe limitations of delta modulation. The integrator is unable to closely track large-amplitude and high-frequency signals. The difference between $S_i(t)$ and $S_0(t)$ is called **slope-overload noise**.

Adaptive delta modulation is an improvement over delta modulation and overcomes the problem of slope overload that is often encountered when modulating signals vary at a faster rate. At this faster rate, the output signal will only alternate above and below the modulating signal and hunting results. The output will thus give positive and negative pulses alternately. This condition results in an error in the demodulated signal.

To improve the system performance, a modified circuit is to be used and this circuit includes a variable gain amplifier at the input of the integrator and the gain of this amplifier is controlled by the average rate of pulses being transmitted, no matter whether of positive or negative polarity. Figure 6.22 (a) and (b) show the encoder and decoder of an adaptive delta modulator.

From the above figure, the squaring circuit causes the gain of the feedback amplifier to increase with increase in signal amplitude, regardless of the polarity of the voltage. In

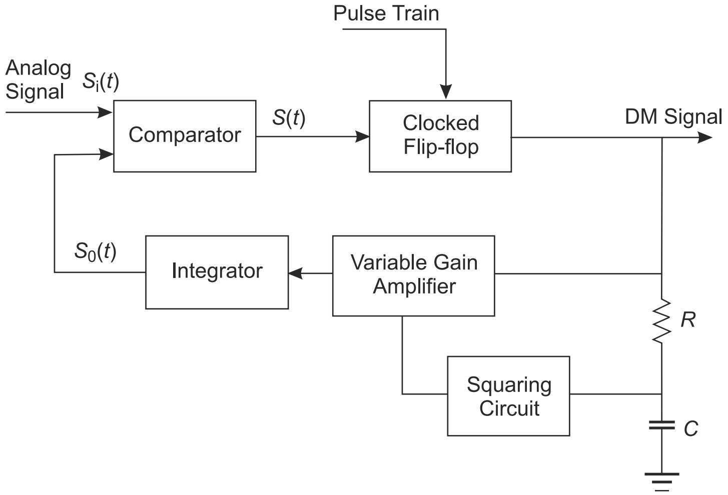


Fig. 6.22(a) Encoder of adaptive delta modulation

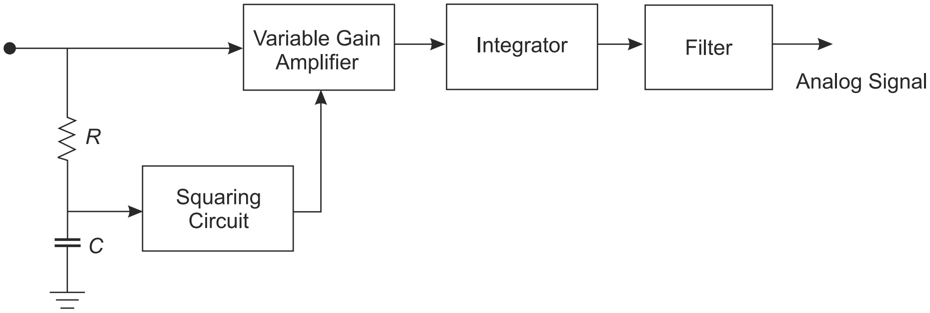


Fig. 6.22(b) Decoder of adaptive delta modulation

addition by squaring the quantisation noise, power is made to vary linear with input signal variations.

3. Advantages and Disadvantages of Delta Modulation

Advantages

(a) It is less complex and, therefore, is less costly than PCM.

(b) It is more tolerant to transmission errors.

(c) It does not require more synchronisation as in case of PCM.

Disadvantages

(a) It is sensitive to slope-overload error.

(b) It is unsuitable for time sharing as an encoder/decoder among multiple channels.

4. Comparison between Pulse Code Modulation (PCM), Delta Modulation (DM) and Adaptive Delta Modulation (ADM)

Table 6.3 illustrates the differences between various digital pulse modulation methods such as Pulse Code Modulation (PCM), Delta Modulation (DM) and Adaptive Delta Modulation (ADM).

Table 6.3 Differences between DPM methods

S. No	Parameter of Comparison	Pulse Code Modulation (PCM)	Delta Modulation (DM)	Adaptive Delta Modulation (ADM)
1	Number of bits	It can use 4, 8 or 16 bits per sample	It uses one bit for one sample	Only one bit is used to encode one sample
2	Levels and step size	The number of levels depends on number of bits. Level size is fixed.	Step size is kept fixed and cannot be varied	Step size varies according to the signal variation
3	Quantisation error and distortion	Quantisation error depends on the number of levels	Slope-overload distortion is present	Quantisation noise is present but no other errors
4	Bandwidth	Highest bandwidth is needed since no. of bits are high	Lowest bandwidth is enough	Lowest bandwidth is enough
5	Feedback	There is no feedback in transmitter or receiver	Feedback exists in the transmitter	Feedback exists in the transmitter
6	Complexity	Complex system to implement	Simple to implement	Simple to implement

Summary

Pulse modulation is divided into two types.

- Analog pulse modulation
- Digital pulse modulation

In an analog pulse modulation, if a message is adequately described by its sample values, it can be transmitted by analog pulse modulation by which the sample values directly modulate a periodic pulse train with one pulse for each sample. The three types of analog pulse modulation systems are

- Pulse Amplitude Modulation (PAM)

- Pulse Duration Modulation (PDM)
- Pulse position Modulation (PPM)

If the amplitude of a pulse is altered in accordance to that of amplitude of modulating signal to make it accommodate the information signal, this modulation process is called Pulse Amplitude Modulation (PAM). In Pulse Duration Modulation (PDM), pulse duration is altered in accordance with the amplitude of the modulating signal to accommodate the information signal. It is apparent that the duration of the pulse can be controlled only to a certain degree. In Pulse Position Modulation (PPM), pulse position is altered in accordance with the amplitude of the modulating signal to accommodate the information signal.

In contrast to analog pulse modulation, in digital pulse modulation the amplitude or timing of the transmitted pulse-modulated waveform varies continuously with the amplitude of the modulating signal. In this type, the input signals are encoded which consists of binary digits.

Digital pulse modulation can be further divided into the following types.

- Pulse Code Modulation (PCM)
- Delta Modulation (DM)

In PCM, the sampled signal is rounded off to the nearest value which is permitted for transmission by the system. The process of rounding off is termed quantisation. Delta modulation is a method of information transmission with the help of pulses, but it uses a single-digit code that transmits information about the slope of the signal amplitude rather than the actual amplitude, as in PCM. In contrast to pulse-code modulation, delta modulation is not often used.

REVIEW QUESTIONS

PART-A

1. What is the purpose of pulse modulation?
2. What are the types of pulse modulation?
3. List the types of analog pulse modulation.
4. Define pulse amplitude modulation.
5. Define pulse duration modulation.
6. Define pulse position modulation.
7. Draw the frequency spectrum of PAM.
8. What is the bandwidth of PAM?
9. What are the sampling considerations of PAM?
10. Give the graphical representation of PDM.

11. What do you mean by digital pulse modulation?
12. List the types of digital pulse modulation.
13. Define pulse code modulation.
14. Define delta modulation.
15. Give the block diagram of a PCM transmitter.
16. Give the block diagram of a PCM receiver.
17. Why do you prefer differential PCM?
18. What do you mean by companding?
19. Define slope overloading.
20. What is an adaptive delta modulation?
21. State the advantages and disadvantages of delta modulation.

PART-B

1. What is pulse amplitude modulation? Give its mathematical analysis and the procedure for generating PAM.
2. Explain the procedure to convert pulse width modulation into pulse position modulation.
3. Explain with a suitable diagram the generation of PPM signal and explain how these signals are demodulated.
4. What is pulse code modulation? Explain the process of PCM and working of PCM transmitter and receiver.
5. Derive the relations for signalling rate and transmission bandwidth in PCM system.
6. Explain the process of adaptive delta modulation and state its advantages and disadvantages.
7. Compare pulse code modulation with delta modulation and adaptive delta modulation methods in detail.

7

DIGITAL MODULATION

Objectives

- To know the purpose of multiplexing in a communication system
- To discuss the concept and different types of multiplexing
- To provide the details about time-division multiplexing and frequency-division multiplexing and their comparison in detail
- To discuss the process of digital modulation and its different types such as Amplitude Shift Keying (ASK), Frequency Shift Keying (FSK) and Phase Shift Keying (PSK) in detail
- To provide the details about the purpose of modems and their applications
- To understand the concepts behind the process of quantization in detail

7.1 INTRODUCTION

Multiplexing involves grouping of several channels in such a way as to transmit them simultaneously on the same physical transmission medium without mixing. At the receiving end, demultiplexing is performed to separate the channels. In the telephone network, each channel provides a bandwidth of 300–3400 Hz for speech signals.

7.2 NEED FOR MULTIPLEXING

Generally, in a communication network, two communicating stations do not utilise the full capacity of a data link. Moreover, when many nodes compete to access the network, it is rare to utilise the data link essentially. A medium can be shared by more than one channel of signals, when the bandwidth of a medium is greater than individual signals to be transmitted through the channel. Multiplexing is a process which makes most effective use of the available channel capacity and the channel capacity can be shared among a number of communicating stations.

7.3 CONCEPT OF MULTIPLEXING

Multiplexing is defined as a process in which multiple analog message signals or digital data streams are combined into one signal over a shared medium. For example, in telecommunications, several phone calls may be transferred using one wire.

The multiplexed signal is transmitted over a communication channel, which may be a physical transmission medium. The multiplexing divides the capacity of the low-level communication channel into several higher-level logical channels, one for each message signal to be transferred. A reverse process called **demultiplexing** can extract the original channels on the receiver side. Figure 7.1 illustrates the functioning of the multiplexing process.

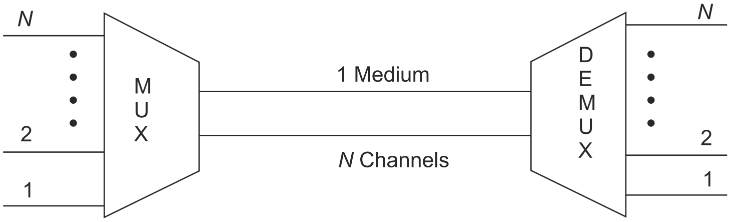


Fig. 7.1 Functioning of multiplexing process

The multiplexer is connected to the de-multiplexer by a single data link. The multiplexer combines data from these *N* input lines and transmits them through the high-capacity data link, which is being de-multiplexed at the other end and is delivered to the appropriate output lines. Thus, multiplexing allows simultaneous transmission of multiple signals across a single data link.

There are two major types of multiplexing. They are as follows.

1. Time-Division Multiplexing (TDM)
2. Frequency-Division Multiplexing (FDM)

7.4 TIME-DIVISION MULTIPLEXING (TDM)

Time-division multiplexing is a type of digital multiplexing in which two or more signals are transferred apparently simultaneously as subchannels in one communication channel, but are physically taking turns on the channel. The time domain is divided into several recurrent time slots of fixed length, one for each subchannel. A sample byte or data block of subchannel 1 is transmitted during time slot 1, subchannel 2 during time slot 2, etc. One TDM frame consists of one time slot per subchannel. After the last subchannel, the cycle starts all over again with a new frame, starting with the second sample, byte or data block from subchannel 1, etc.

As an example, a 24-channel TDM system is assumed for transmitting at a rate of 8000 samples per second. This means that each of the 24 signals is sampled 8000 times every second. Thus, the time period of samples corresponding to different channels equals 1/8000 or 125 microseconds. The magnitude of each sample is represented by seven binary bits and contains an additional bit for synchronization. Thus, a 125-microsecond period available sample from every signal is made to occupy 5 microseconds and the total period comes out to be 24×5=120 microseconds. The remaining 5-microsecond period available is utilised for signalling and synchronisation. The pulse width occupied by each bit in the 8-bit sample equals 5/8 microseconds.

At the receiving end, all the pulses are received and detected by a common detector and then fed to channel gates numbering 1 to 24. These gates are opened for a specific period during which the signal corresponding to that channel is passed to the output. Timing pulses for different gates are provided by a gating generator which is usually a monostable multivibrator and is accurately timed by the synchronisation pulses.

There are two types of TDM circuits.

1. Slow-speed TDM
2. High-speed TDM

7.4.1 Slow-speed TDM

Slow-speed TDM is commonly used for transmission of radio telemetry data and utilises rotating mechanical switches to sample the signals from different channels. The output from all the channels is brought to the rotary switch at the transmitter and the rotor is made to rotate at a slow and constant speed. The rotor remains in contact with each channel for a specified time and during this period, the output of that channel is passed to the output. The receiver also contains a similar rotary switch, the motion of which is synchronised with the corresponding switch in the transmitter section. Figure 7.2 illustrates the principle of slow-speed TDM.

The transmitter as well as receiver has similar rotary mechanical contacts, each having a capability of 16 channels, but for simplicity only 8 lines are shown in the figure. The synchronised motion of the switch wipers is obtained with the help of synchronising timing pulses and this makes the two wipers move together. That is, when the transmitting wiper moves from position 1 to 2, the receiving wiper also makes a similar movement. Signals from all the 16 channels are available at the transmitter rotary switch and samples of each channel signal are taken by the rotor which rotates at a constant and uniform speed. The sampled signals travel through the transmission media and reach the receiver wiper which passes these signals further to the corresponding receiver channels.

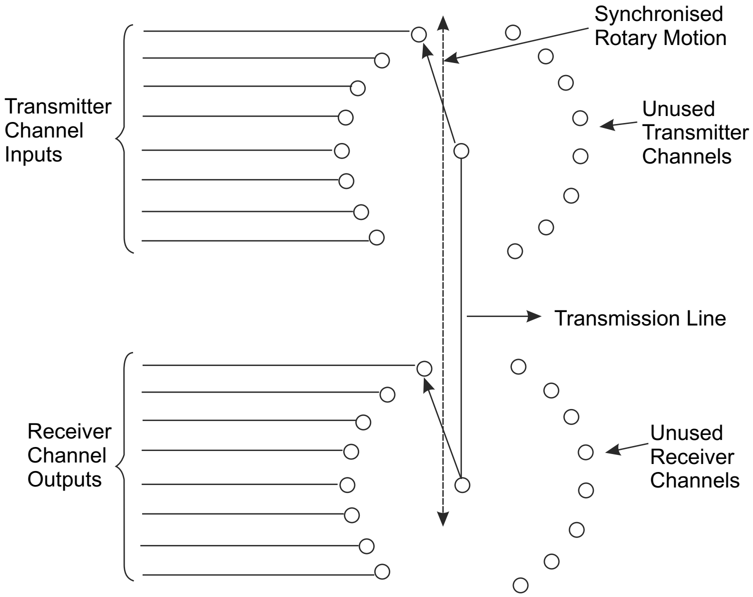


Fig. 7.2 Slow-speed TDM

Theoretically, it is possible to have a large number of channels multiplexed in this manner but this requires high-frequency switching and as a result, bandwidth required is large. Next, the circuits grow more and more complex. As such, frequency-division multiplexing is preferred over time-division multiplexing. In early days, TDM was used for multiplexing telegraph signals.

7.4.2 High-speed TDM

There are high-speed TDM circuits which employ electronic switching and delay lines for their operation. Signals from each channel are sampled by channel-sampling circuits which are triggered simultaneously. The sampled output from each channel is then fed to an adder circuit. The second channel being delayed by 5 microseconds, the third channel by 10 microseconds, and so on.

7.4.3 PAM/TDM System

In Pulse Amplitude Modulation (PAM), the pulse is present for a short duration and for most of the time between two pulses, no signal is present. This free space between the pulses can be occupied by pulses from other channels. This is known as Time-Division Multiplexing (TDM). Thus, TDM makes maximum utilisation of the transmission channel without any gap. So this process is also termed PAM/TDM. Figure 7.3 shows the block diagram of PAM/TDM.

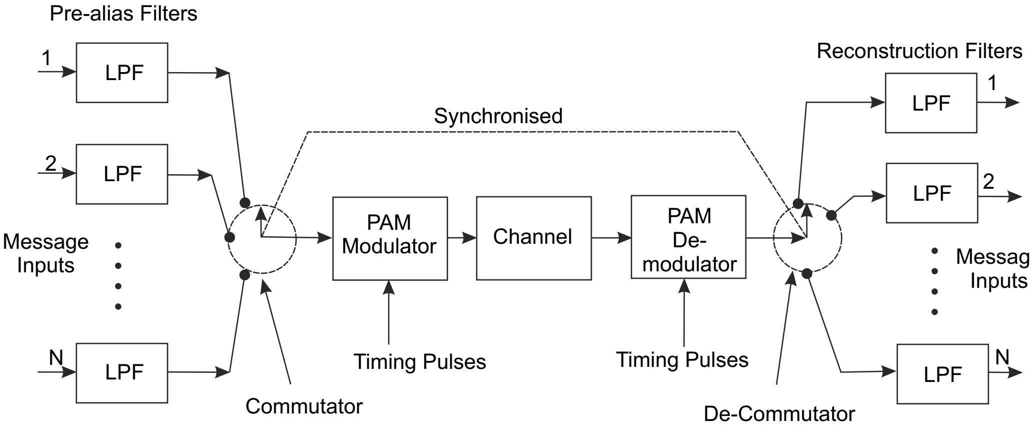


Fig. 7.3 Block diagram of PAM/TDM

Figure 7.3 shows the time-division multiplexing of N PAM channels. Each channel to be transmitted is passed through the lowpass filter. The outputs of the lowpass filters are connected to the rotating sampling switch or commutator. It takes a sample from each channel per rotation and it rotates at the rate of f_s. Thus, the sampling frequency becomes f_s. The single signal composed due to multiplexing of input channels is given to the transmission channel. At the receiver end, the de-commutator separates the time-multiplexed input channels. These channel signals are then passed through lowpass construction filters.

If the highest signal frequency present in all the channels is f_m then the sampling frequency must be

$$f_s \geq 2f_m \tag{7.1}$$

Therefore, the interval between successive samples from any one input will be

$$T_s = \frac{1}{f_s} \tag{7.2}$$

Now it will be

$$T_s \leq \frac{1}{2f_m} \tag{7.3}$$

Let there be N input channels and in each interval, there will be one sample from each of the N samples. This means that there are N samples in each interval. Spacing between two samples is $\frac{T_s}{N}$.

The number of pulses per second or pulse frequency is calculated as

$$\text{Number of pulses per second} = \frac{1}{\text{Spacing between two samples}}$$

$$= \frac{1}{\left(\frac{T_s}{N}\right)} = \frac{N}{T_s} \tag{7.4}$$

Hence, number of pulses per second $= \frac{N}{T_s} = \frac{N}{\left(\frac{1}{f_s}\right)} = Nf_s$ (7.5)

This number of pulses per second is also known as **signalling rate** of TDM signal and is denoted by r.

Signalling rate $r = Nf_s$ (7.6)

Since $f_s \geq 2f_m$,

Signalling rate in PAM/TDM $r \geq 2Nf_m$ (7.7)

In the TDM process, the signal should modulate some carrier. Before modulation, the pulsed signal in TDM is converted to baseband signal $x_b(t)$ which modulates the carrier. The baseband signal is obtained by passing pulsed TDM signal through a lowpass filter. The bandwidth of this lowpass filter is given by half of the signalling rate.

$$B = \frac{1}{2} r = \frac{1}{2} Nf_s \tag{7.8}$$

Therefore, the bandwidth of a TDM channel will be equal to bandwidth of lowpass filter.

Thus, $BW = \frac{1}{2} Nf_s$ (7.9)

$$\therefore \; BW = \frac{1}{2} N. \, 2f_m = Nf_m \tag{7.10}$$

Thus, minimum transmission bandwidth of TDM channel $= Nf_m$

This equation states that if there are a total of N channels in TDM which are bandlimited to f_m then minimum bandwidth of the transmission channel will be Nf_m.

7.4.4 Implementation of TDM

There are two ways of implementing TDM. They are as follows.

1. Synchronous TDM
2. Asynchronous TDM

1. Synchronous TDM

Synchronous TDM is called *synchronous* in which each time slot is pre-assigned to a fixed source. The time slots are transmitted irrespective of whether the sources have any data to send or not. Hence, for the sake of simplicity of implementation, channel capacity is wasted.

Although fixed assignment is used in TDM, devices can handle sources of different data rates. This is done by assigning fewer slots per cycle to the slower input devices than the faster devices. This time slice is allocated even if a device has nothing to transmit. Therefore, the use of synchronous TDM does not guarantee maximum line usage and efficiency.

Figure 7.4 illustrates the process of synchronous TDM.

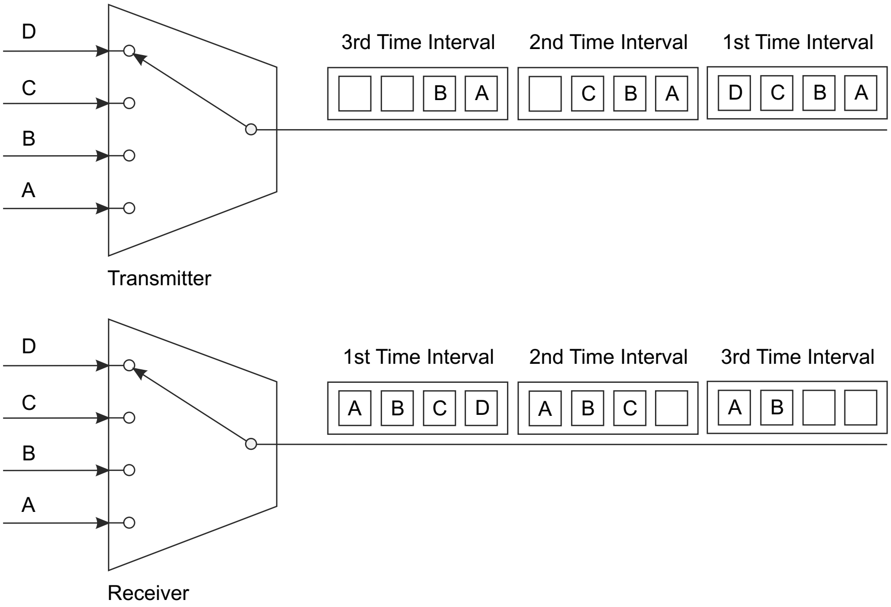


Fig. 7.4 Synchronous TDM

2. Asynchronous TDM

In synchronous TDM, if a particular terminal has no data to transmit at a particular instant of time, an empty time slot will be transmitted. So, many of the time slots in the frame are wasted. This kind of drawback is eliminated by an efficient alternative called asynchronous TDM. It is also termed **intelligent TDM** or **statistical TDM**.

Asynchronous TDM dynamically allocates the time slots on demand to separate input channels, thus saving the channel capacity. In comparison with synchronous TDM, statistical multiplexers also have many I/O lines with a buffer associated to each of them. During the input, the multiplexer scans the input buffers, collecting data until the frame is filled and send the frame. At the receiving end, the de-multiplexer receives the frame and distributes the data to the appropriate buffers.

The difference between synchronous TDM and asynchronous TDM is illustrated by using Figure 7.5.

In case of synchronous TDM, many slots are kept unutilised and the slots are fully utilised leading to smaller time for transmission and better utilisation of bandwidth of the medium. In

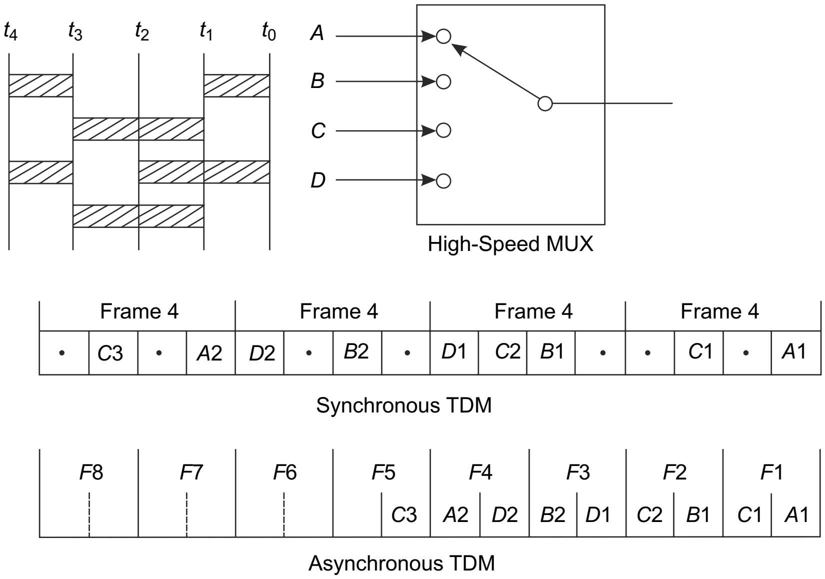


Fig. 7.5 Differences between synchronous TDM and asynchronous TDM

case of statistical TDM, the data in each slot must have an address part, which identifies the source of data. Since data arrive from and are distributed to I/O lines unpredictably, address information is required to assure proper delivery.

Asynchronous TDM is a more flexible method of TDM and it may require more processing by the multiplexer and take longer. However, the time saved by efficient and effective bandwidth utilisation makes it worthwhile.

3. Higher Order Multiplexing

In higher-order multiplex systems, there is a multiplex group of 30 or 24 channels used as a building block for larger numbers of channels. At each level in the hierarchy, several bit streams, known as **tributaries**, are combined by a multiplexer/de-multiplexer, or simply MUX. The output from a multiplexer may serve as a tributary to a multiplexer at the next higher level in the hierarchy, or it may be sent directly over a transmission line. Figure 7.6 illustrates the concept of digital interleaving.

If the inputs to a multiplexer are synchronous, i.e. they have the same bit rate and are in phase, taking a bit or a group of bits from each in turn can interleave them. A switch that samples each input under the control of the multiplex clock can do this. There are two main methods of interleaving digital signals:

1. Bit interleaving
2. Word interleaving or byte interleaving

In bit interleaving, one bit is taken from each tributary in turn. If there are N input signals, each with a rate of f_1 bits/second then the combined rate would be Nf_1 bits/second and each

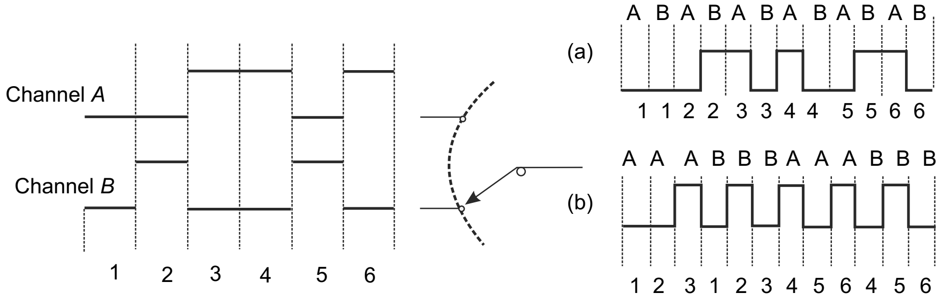


Fig. 7.6 Interleaving digital signals: (a) Bit interleaving (b) Word interleaving

element of the combined signal would have duration equal to 1/N of an input digit. In word interleaving, groups of bits are taken from each tributary in turn and this involves the use of storage at each input to hold the bits waiting to be sampled.

7.5 FREQUENCY-DIVISION MULTIPLEXING (FDM)

Frequency-division multiplexing is a form of signal multiplexing where multiple baseband signals are modulated on different frequency channels and added together to create a composite signal destined for different users.

Frequency-division multiplexing consists in simultaneous transmission of signals belonging to different channels by shifting them in the domain of frequency. In this manner, it is possible to transmit as large a number of telephone channels as 10,800 over a single transmission line. This is needed in big cities where there is a heavy requirement for transmission of data such as telephone, telegraph, etc. from one place to another. Figure 7.7 shows the basic circuit for a frequency-division multiplexing link.

The two-wire channels are converted into four-wire systems and the ‘go’ wires are connected to the multiplexing unit which translates different channels into higher designated frequencies. The allocated frequency then comprises various channels equally spaced in the

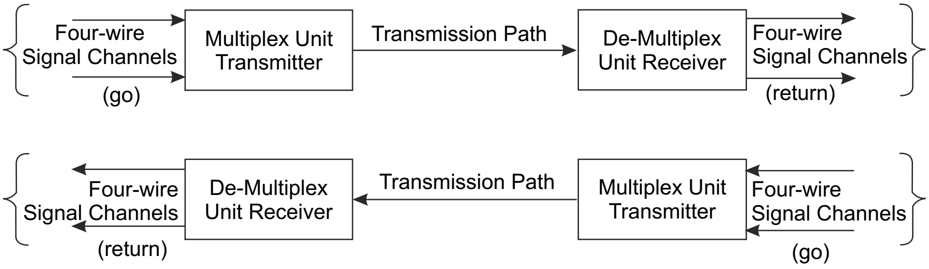


Fig. 7.7 Block diagram of an FDM system

frequency domain which are transmitted through the transmission path. At the receiving end, the demodulating unit covers these channels back to their original frequencies and passes them to respective subscribers.

The heart of the multiplexing and demultiplexing unit is the balanced modulator, to which the carrier is fed, and a voice channel having frequencies in the range of 300 Hz to 3400 Hz. Out of the two components, namely sum and difference frequencies, one is filtered out and the second is taken as output. The system will have as many mixers as the number of channels. Different channels are then combined to form a group.

7.5.1 Simple Illustration of FDM

In FDM, each message of frequency f_m is translated into different frequency spectra using a base carrier. Then they are all combined using an adder circuit and are used to modulate a common carrier using amplitude modulation. At the receiver, a broadband receiver receives this signal and passes it on to a baseband receiver which receives the signals corresponding to the baseband frequency.

If the signals to be transmitted simultaneously are $em_1\,(t)$, $em_2\,(t)$, $em_3\,(t)$ and $em_4\,(t)$, and baseband frequencies are f_1, f_2, f_3 and f_4 then the modulated baseband carriers at the adder input are

$$
\begin{aligned}
e_1\,(t) &= E_1\,[1 + m_a\, em_1(t)] \cos \omega_1 t \\
e_2\,(t) &= E_2\,[1 + m_a\, em_2\,(t)] \cos \omega_2 t \\
e_3\,(t) &= E_3\,[1 + m_a\, em_3(t)] \cos \omega_3 t \\
e_4\,(t) &= E_4\,[1 + m_a\, em_4(t)] \cos \omega_4 t
\end{aligned}
\tag{7.11}
$$

These signals are combined together at the adder and form a common modulating signal for the AM transmitter.

$$e_m = e_1\,(t) + e_2\,(t) + e_3\,(t) + e_4\,(t) \tag{7.12}$$

The transmitter output is

$$e = E_c\,\{1 + m_a\,(e_1\,(t) + e_2\,(t) + e_3\,(t) + e_4\,(t))\}\cos \omega_c t \tag{7.13}$$

Thus, the transmitted wave contains all the original signals and has

$$\text{Bandwidth} = 2\,(fm_1 + fm_2 + fm_3 + fm_4) \tag{7.14}$$

where fm_1, fm_2, fm_3, fm_4 are the highest modulation frequencies in the signal. At the receiver, these signals are separated and passed to respective channels.

To transmit a number of signals over the same channel, the signals must be kept apart, so that they do not interfere with each other. Thus, each signal can be separated at the receiving end. This is achieved by using **guard bands** between different message signals in the frequency spectrum as shown in Figure 7.8.

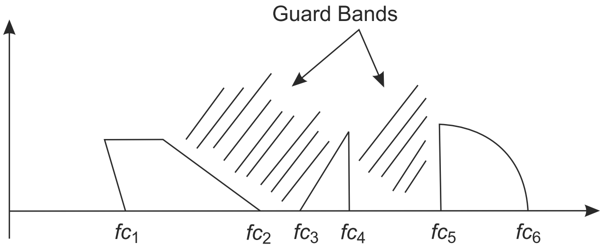


Fig. 7.8 Spectrum of a multiplexed signal

7.5.2 Operation of FDM

Figure 7.9 shows the operation of an FDM system. Each signal input is passed through a lowpass filter, which is designed to remove high-frequency components that do not contribute much towards signal representation. But they are capable of distributing other message signals that share the common channel. The filtered signals are applied to modulators which shift the frequency ranges of the signals so as to occupy mutually exclusive frequency intervals. The bandpass filters following the modulators are used to restrict the band of each modulated wave to its prescribed range. The resulting bandpass filter outputs are next combined in parallel to from the input to the common channel.

At the receiving end, a group of bandpass filters, with their inputs connected in parallel, are used to separate the message signals on a frequency spectrum. Finally, the original message signals are recovered using individual demodulators. The FDM shown above can only operate in one direction. To provide two-way transmission, it is necessary to duplicate the multiplexing facilities with the components connected in reverse order.

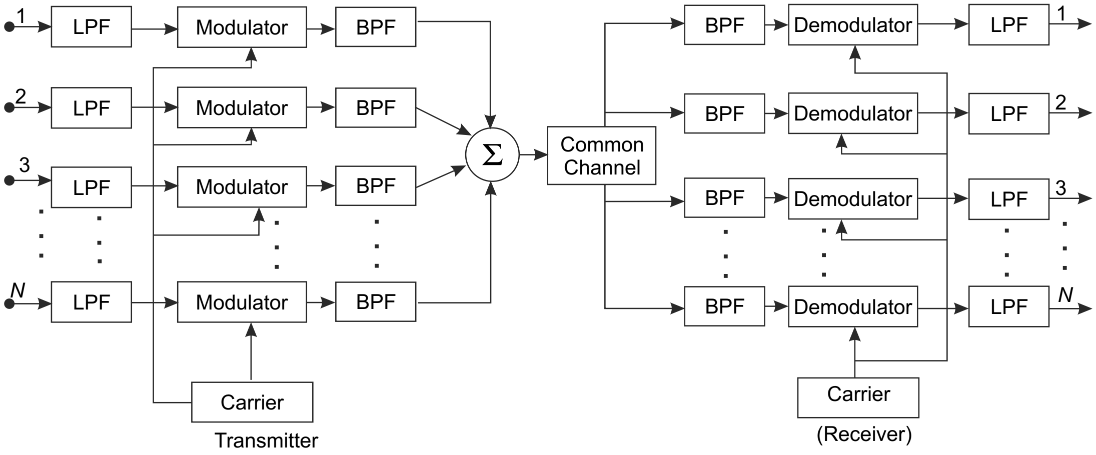


Fig. 7.9 Frequency division multiplexing

7.5.3 Frequency Translation

In the processing of signals in communication, it is necessary to translate the modulated wave upward or downward in frequency, so that it occupies a new frequency band. This is accomplished by multiplication of the signal by a locally generated side band and subsequent filtering.

For example, a DSB SC wave is represented as follows.

$$S(t) = m(t).\cos 2\pi f_c t \tag{7.15}$$

The modulating wave *m(t)* is limited to the frequency band $-\omega \leq \mathrm{f} \leq \omega$

The spectrum of $S(t)$ occupies the bands $(f_c - \omega) \leq f \leq (f_c + \omega) \leq f \leq (-f_c + \omega)$ as shown in Figure 7.10.

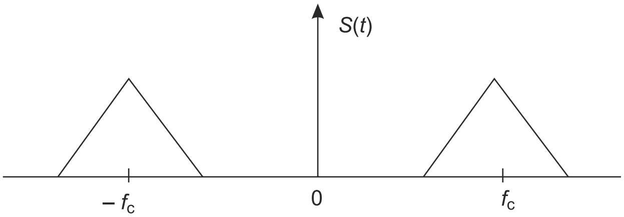


Fig. 7.10 Frequency spectrum of modulated wave

Suppose it is required to translate this modulated wave downward to its frequency, so that its carrier frequency is changed from f_c to a new value f_0, where $f_0<f_c$. To accomplish this requirement, it is necessary to multiply sinusoidal wave of frequency f_1 supplied by a local oscillator to obtain the output $V_1(t)$.

$$\begin{aligned} V_1(t) &= \mathrm{S}(t)\cos 2\pi f_1(t) \\ &= m(t).\cos\,(2\pi f_c t) \times \cos\,(2\pi f_1(t)) \\ &= \frac{1}{2}\,m(t).\cos 2\pi(f_c - f_t)\,t + \frac{1}{2}\,m(t).\cos 2\pi(f_c - f_t)\,t \end{aligned} \tag{7.16}$$

The multiplier output $V_1(t)$ consists of two DSB SC waves, one with a carrier frequency of $f_c - f_1$ and the other with a carrier frequency of $f_c + f_1$. The spectrum of $V_1(t)$ is shown in Figure 7.11.

Let the frequency $f_c - f_1 = f_0$

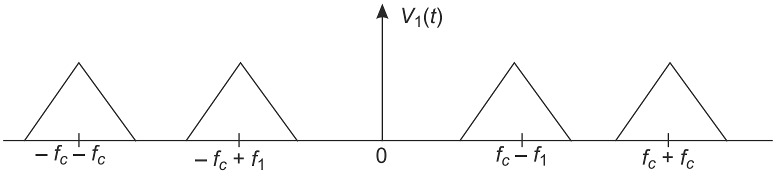


Fig. 7.11 Spectrum of multiplier output $V_1(t)$

From the spectrum of multiplier output, the modulated wave with the desired carrier frequency f_c may be extracted by passing the multiplier output $V_1(t)$ through a bandpass filter of frequency f_0 and bandwidth 2ω, provided that

$$(f_c + f_1 - \omega) > (f_c - f_1 + \omega) \text{ or } f_1 > \omega \tag{7.17}$$

The filter output, represented as $V_2(t)$, is expressed as

$$V_2(t) = \frac{1}{2} m(t).\cos 2\pi (f_c - f_1)t \tag{7.18}$$

$$= \frac{1}{2} m(t).\cos(2\pi f_0 t)$$

This output is the desired modulated wave, translated downwards and is shown in Figure 7.12.

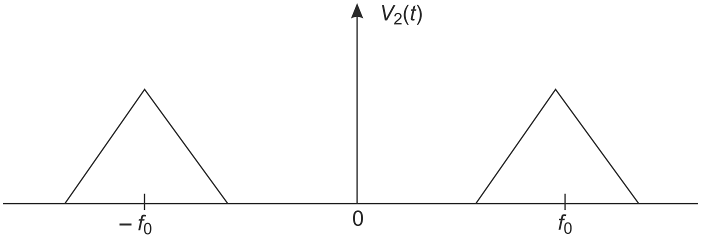


Fig. 7.12 Spectrum of downward translated wave

7.5.4 Mixers

The block diagram of a mixer used for FDM is shown in Figure 7.13. It is used for frequency translation of a modulated wave. The operation is called **mixing** or **super-heterodyning**.

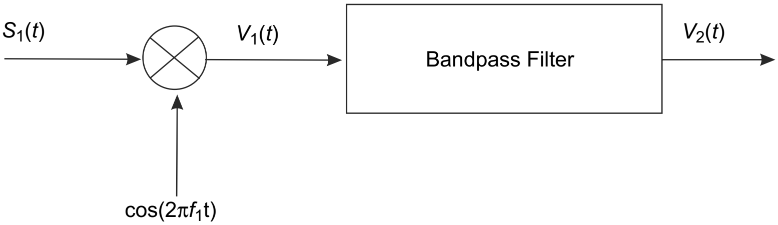


Fig. 7.13 Block diagram of a mixer

The multiplier is usually constructed by using a switching device similar to modulators. Mixing is a linear operation and it preserves the relation of the side bands of the incoming modulated wave.

7.6 COMPARISON BETWEEN TDM AND FDM METHODS

1. TDM channels are assigned to distinct slots but combined together in the time domain whereas in FDM, its channels are assigned to distinct slots but combined together in the frequency domain.
2. TDM involves simple instrumentation whereas FDM requires an analog subcarrier modulation, demodulator and filter for every message channel.
3. TDM synchronisation is slightly more demanding than that of suppressed carrier FDM.
4. TDM is invulnerable to the usual causes of cross-talk. But in FDM, imperfect bandpass filtering gives rise to cross-talk.
5. The use of multiplexers allows a TDM system to accommodate different signals whose bandwidth on pulse rates may differ by more than an order of magnitude.

7.7 DIGITAL MODULATION

The digital modulation provides more information capacity, compatibility with digital data services, higher data security, better quality communications, and quicker system availability. The following constraints are faced by the developers of the communication field.

1. Available bandwidth
2. Permissible power
3. Inherent noise level of the system

Digital modulation schemes have greater capacity to convey large amounts of information than analog modulation schemes. In digital modulation, an analog carrier signal is modulated by a digital bit stream. Digital modulation methods can be considered as digital-to-analog conversion and the corresponding demodulation is termed analog-to-digital conversion.

Digital modulation is widely employed in several applications. Some of the applications are listed below.

1. Modem in personal computer
2. Digital Subscriber Lines (DSL)
3. Digital microwave
4. Satellite communication system
5. Cellular telephone Personal Communication Systems (PCS)

7.7.1 Types of Digital Modulation Methods

There are four types of digital modulation techniques:

1. Amplitude Shift Keying (ASK)

In the case of amplitude shift keying, a finite number of amplitudes are used. If the amplitude of the carrier is varied proportional to the modulating signal, ASK is used.

2. Quadrature Amplitude Modulation (QAM)

In the case of quadrature amplitude modulation, a finite number of at least two phases and at least two amplitudes are used. In QAM, an in-phase signal (the *I* signal, for example a cosine waveform) and a quadrature phase signal (the *Q* signal, for example a sine wave) are amplitude-modulated with a finite number of amplitudes, and summed. It can be seen as a two-channel system, each channel using ASK. The resulting signal is equivalent to a combination of PSK and ASK.

3. Frequency Shift Keying (FSK)

In the case of frequency shift keying, a finite number of frequencies are used. If the frequency of the carrier is varied proportional to the modulating signal, FSK is used.

4. Phase Shift Keying (PSK)

In the case of phase shift keying, a finite number of phases are used. If the phase of the carrier is varied proportional to the modulating signal, PSK is used.

7.7.2 Principles of Digital Modulation

Using the principle of quadrature amplitude modulation method, all other methods are often modulated and demodulated. *I* and *Q* signals can be combined into a complex valued signal $I + jQ$. The resulting signal is a complex-valued representation of the real-valued modulating signal or simply RF signal.

The procedural steps to be followed by the modulator at the transmitter side to transmit the data are given as follows.

1. Incoming data are grouped and then these bits are encoded and from the result, one for each symbol will be transmitted.
2. The code words are mapped to amplitudes of *I* and *Q* signals or frequency or phase values.
3. Filtering methods or any pulse-shaping methods are applied to limit the bandwidth and form the spectrum of the equivalent lowpass signal.
4. Digital-to-analog conversion (DAC) of *I* and *Q* signals is to be performed.
5. A high-frequency sine-wave carrier waveform is then generated which is a cosine quadrature component. Then the modulation is carried out by means of multiplying the

sine and cosine waveform with *I* and *Q* signals which results in the equivalent lowpass signal. It is then frequency shifted into a modulated passband signal or RF signal.

6. Amplification and analog bandpass filtering is carried out to avoid harmonic distortion.

Similarly, the procedural steps to be followed by the demodulator at the receiver side to receive the data are given as follows :

1. Bandpass filtering
2. To compensate attenuation, Automatic Gain Control (AGC) is to be carried out.
3. Frequency shifting of the RF signals to the equivalent *I* and *Q* signals or to an intermediate frequency (IF) signal is to be performed by multiplying the RF signal with a local oscillator sine-wave and cosine-wave frequency.
4. Sampling and Analog-to-Digital Conversion (ADC) is to be carried out.
5. The amplitudes of *I* and *Q* signals, or the frequency or phase of the IF signal, are to be detected.
6. The amplitudes, frequencies or phases to the nearest allowed symbol values are to be quantised.
7. The quantised amplitudes, frequencies or phases to code words are to be mapped.
8. Parallel-to-serial conversion of the code words into a bit stream is to be made.

Then, the resultant bit stream is to be passed on for further processing such as removal of any error-correcting codes.

7.7.3 I/Q Formats

In digital communications, modulation is often expressed in terms of *I* and *Q*. This is a rectangular representation of the polar diagram. On a polar diagram, *I* axis lies on the zero-degree phase reference, and *Q* axis is rotated by 90 degrees. The signal vector's projection onto *I* axis is its "*I*" component and the projection onto the *Q* axis is its "*Q*" component. Figure 7.14 shows the *I/Q* formats.

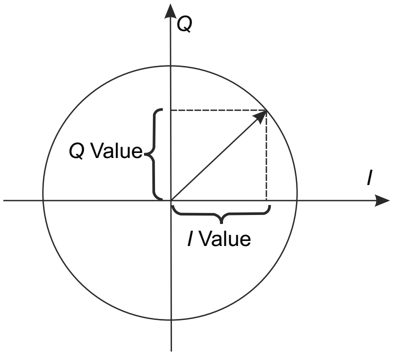


Fig. 7.14 I/Q formats

7.7.4 Error-Rate Calculation

In order to determine the error rate mathematically, the following terms are to be considered.

E_b = Energy per bit

E_s = Energy per symbol = kE_b with k bits per symbol

T_b = Bit duration

T_s = Symbol duration

$\frac{N_0}{2}$ = Noise power spectral density

P_b = Probability of bit error

P_s = Probability of symbol error

$Q(x)$ will give the probability that a single sample taken from a random process with zero-mean and unit-variance Gaussian probability density function will be greater or equal to x. It is a scaled form of the Gaussian error function which is expressed as follows.

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-t^2/2}\, dt \tag{7.19}$$

$$= \frac{1}{2} erfc\left(\frac{x}{\sqrt{2}}\right), \quad x \geq 0 \tag{7.20}$$

7.7.5 Constellation Diagram

A constellation diagram is a representation of a signal modulated by a digital modulation scheme such as quadrature amplitude modulation or phase-shift keying. It displays the signal as a two-dimensional scatter diagram in the complex plane at symbol sampling instants. It represents the possible symbols that may be selected by a given modulation scheme as points in the complex plane. Measured constellation diagrams can be used to recognise the type of interference and distortion in a signal. So, a diagram of the ideal positions in a modulation scheme is referred to as constellation diagram.

7.8 AMPLITUDE SHIFT KEYING (ASK)

In Amplitude Shift Keying (ASK), the amplitude A of the carrier signal $A\cos(\omega_c t)$ is switched between the two levels, which correspond to the level of the input binary signal. The two levels of the binary signal can be 0 volt (Bit 0) and 1 volt (Bit 1). Bit 1 is transmitted by a carrier of particular amplitude. To transmit 0, the amplitude is changed by keeping the frequency as constant. On-Off Keying (OOK) is a special from of ASK where one of the amplitudes is zero as shown in Figure 7.15 and binary ASK signal is shown in Figure 7.16.

$$x(t) = \begin{cases} A\cos(2\pi f_c t), & \text{binary 1} \\ 0, & \text{binary 0} \end{cases} \tag{7.21}$$

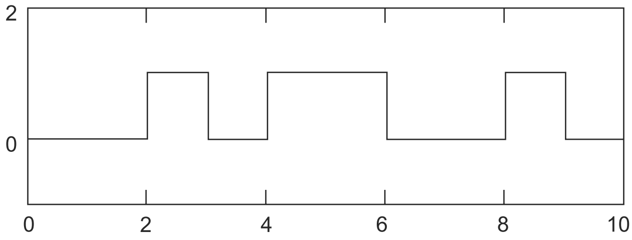


Fig. 7.15 Baseband information sequence—0010110010

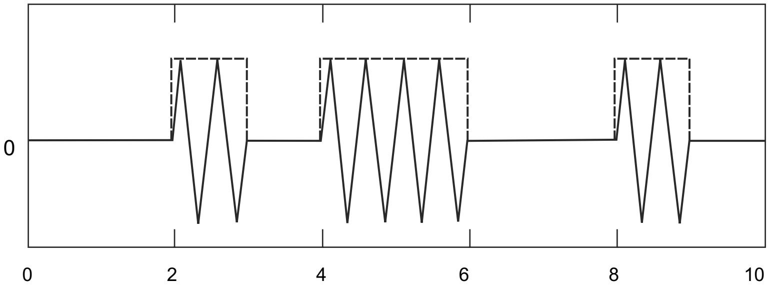


Fig. 7.16 Binary ASK signal

In amplitude shift keying, bandwidth is expressed as

$$B = \frac{f_b}{N} \tag{7.22}$$

where f_b is the Nyquist frequency, and

N is the number of bits encoded into the signalling element.

Baud is expressed as

$$\text{Baud} = \frac{1}{t_s} \tag{7.23}$$

where t_s is the time of the signalling element.

7.8.1 Generation of ASK

There are two methods of generating ASK signals.

1. First Method

An ASK signal can be generated by simply applying the incoming binary data of the baseband signal and the periodic signal such as sinusoidal carrier to the two inputs of a product modulator. The resulting output will be the ASK waveform. Figure 7.17 shows ASK generation using a product modulator.

Let a baseband signal $x_b(t)$ be multiplied by any periodic signal $S(t)$so that the result will be

$$x(t) = x_b(t).\ S(t) \tag{7.24}$$

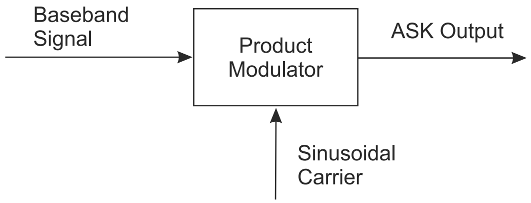


Fig. 7.17 ASK generation using product modulator

The product $x(t)$ contains a series of AM waves with carrier frequencies that are harmonic multiples of the fundamental frequency f_c. A bandpass filter is used to extract any of the harmonics, thus generating the ASK signal.

2. Second Method

The second form of an ASK modulator utilises a square-law device which may be a diode. Here, the baseband signal is added to the carrier oscillations and squaring the sum gives the cross-product, which is the desired modulation term. That is,

$$(x_b(t) + \cos\omega_c t)^2 = x_b(t)^2 + \cos^2 \omega_c t + 2x_b(t)\cos\omega_c t \tag{7.25}$$

3. ASK Generation Using 555 Timer

Using the 555 timer in astable mode, an ASK signal can be generated. Figure 7.18 shows the circuit diagram of ASK generation using 555 timer. The *RC* network (R_A, R_B and C) will determine the carrier frequency of ASK.

$$T = \frac{1}{f} = 0.69C(R_A + R_B) \tag{7.26}$$

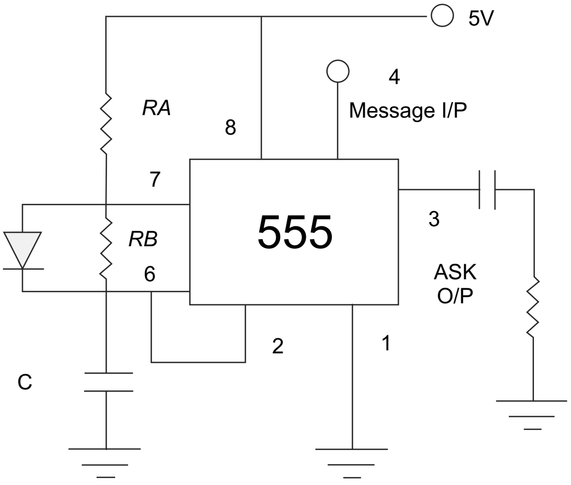


Fig. 7.18 ASK generation using 555 timer

The principle is very simple. Pin No. 4 of the 555 timer is the RESET pin. That means if this PIN is high, the IC will be activated. Otherwise if this pin is grounded, output will be absent. Thus, applying the message information in 4th pin, we can get ASK signal.

7.8.2 ASK Detection

In the amplitude-shift-keying process, some carrier cycles are transmitted to send '1' and no signal is transmitted for binary '0'. Thus,

For binary '1' $\Rightarrow x_1(t) = \sqrt{2P}\cos(2\pi\omega_c t)$ (7.27)

For binary '0' $\Rightarrow x_2(t) = 0$ (no signal) (7.28)

Here, P is the normaliszed power of the signal.

$$P = \frac{A^2}{2} \tag{7.29}$$

ASK detection can be of two types. They are as follows.

1. Coherent detection
2. Noncoherent detection

Coherent demodulators maintain precise timing or simply phase of the incoming carrier. **Non-coherent demodulators** do not maintain this phase and essentially perform a nonlinear operation on the modulating signal to retrieve the baseband amplitude.

1. Coherent Detection

The synchronous demodulator is an example of coherent detection. It simply retranslates the frequencies of the incoming waveform down to the baseband. This is done by multiplying or heterodyning the incoming ASK waveform with a local oscillator matched to the carrier. The lowpass filter will remove the $\cos(2\omega_c t)$ component. The output of the filter having a response in ω_c exactly matches that of the transmitter carrier oscillator. Figure 7.19 shows coherent detection.

From Figure 7.19, an ASK signal is applied to the correlator consisting of a multiplier and integrator. The locally generated coherent carrier is applied to the multiplier. The output of the multiplier is integrated over one bit period. The decision device takes the decision at the end of every bit period. It compares the output of the integrator with the threshold. A decision

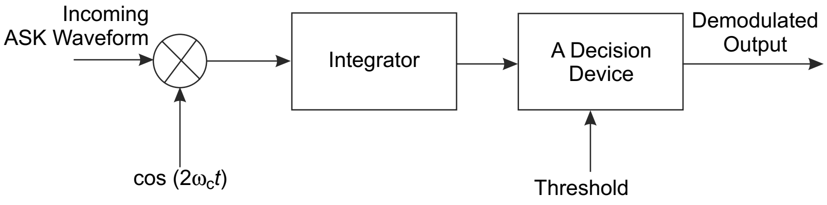


Fig. 7.19 Coherent detection

is taken in favour of '1' when threshold is exceeded, and the decision is taken as '0' if the threshold is not exceeded.

2. Noncoherent Detection

The square-law demodulator is an example of a noncoherent detection. Here, a square-law device is used whose output is passed through a lowpass filter. The output of the filter is then fed to a nonlinear device which is a decision device to take its square root so that the baseband amplitude is retrieved. Figure 7.20 shows noncoherent detection.

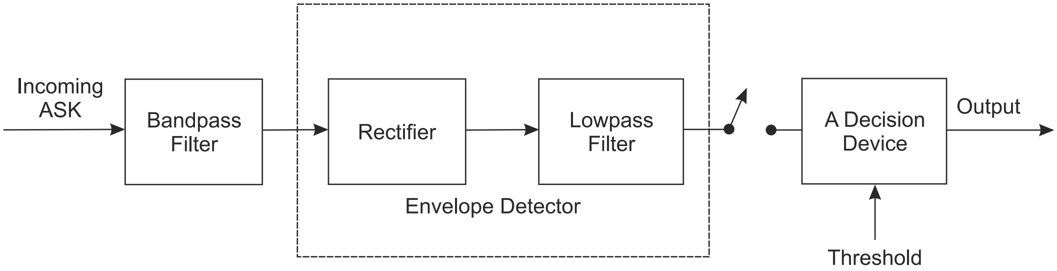


Fig. 7.20 Noncoherent detection

3. ASK Detection Using Comparator

Practically speaking, noncoherent detection is more preferred than coherent detection because generation of the same carrier signal in the receiver side requires complicated circuitry and expensive. An envelope detector is sufficient to detect an ASK signal. An **envelope detector** is a combination of a diode and a parallel *RC* network. The signal is rectified in a diode and the *RC* network is designed in such a way that it keeps the peak amplitude voltage for a small amount of time for proper detection. Comparators are then used for taking decisions for bits 1 or 0.

Comparators are operated in differential mode. One of the input terminals is kept at reference voltage and the signal is applied at the other terminal. There are two type of comparators: positive and negative comparators. If signal is applied to a non-inverting terminal then it is a **positive comparator**. A positive comparator gives high when signal level is greater than reference voltage. If the signal is applied to an inverting terminal then it is a **negative comparator**. A negative comparator gives high when the signal level is less than the reference voltage.

In Figure 7.21, a simple envelope detector is followed by three-stage magnitude comparator and a level translator.

After the envelope detection, a signal is fed to three-stage magnitude comparator. A three-stage comparator is used for reliable signal detection and noise rejection. At the last stage, a level translator is used to get output voltage.

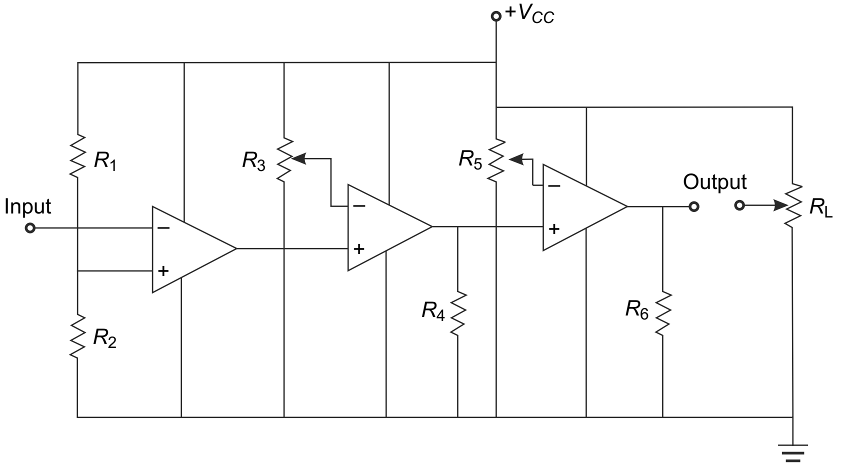


Fig. 7.21 ASK Detection using comparator

EXAMPLE 7.1

Determine the baud rate and minimum bandwidth necessary to pass a 20 kbps binary-signal amplitude shift keying.

Solution

$$\text{Bandwidth} = \frac{f_{\text{b}}}{\text{N}}$$

For ASK, $N = 1$

$$\text{Bandwidth} = \frac{20{,}000}{1} = 20{,}000$$

$$\text{Baud} = \frac{20{,}000}{1} = 20{,}000$$

7.8.3 Advantages and Drawbacks of ASK

Advantages

1. ASK method is used for up to 1200 bps on voice grade lines
2. Speed of this method is very high over optical fiber
3. Simple method
4. Easy to implement

Drawbacks

1. Susceptible to noise
2. Inefficient technique

7.9 FREQUENCY-SHIFT KEYING (FSK)

Frequency-Shift Keying (FSK) is a frequency-modulation scheme in which digital information is transmitted through discrete frequency changes of a carrier wave. Frequency shift keying is a data signal converted into a specific frequency or tone in order to transmit it over wires, cables, optical fibres or wireless media to a destination point.

Frequency modulation and phase modulation are closely related. A static frequency shift of +1 Hz means that the phase is constantly advancing at the rate of 360°/second (2π rad/s), relative to the phase of the unshifted signal. In FSK, the frequency of the carrier is changed as a function of the modulating signal (data) being transmitted. Amplitude remains unchanged.

The general expression for FSK is

$$v_{\text{fsk}}(t) = V_c \cos\{2\pi[f_c + v_m(t)\,\Delta f]t\} \tag{7.30}$$

where $v_{\text{fsk}}(t)$ is binary FSK waveform,

V_c is peak analog carrier amplitude,

f_c is analog carrier centre frequency,

Δf is the shift in the analog carrier frequency, and

$v_m(t)$ is input modulating signal.

From Equation (7.30), it can be seen that the peak shift in the carrier frequency (Δf) is proportional to the amplitude of the binary input signal and the direction of the shift is determined by the polarity. The modulating signal is a normalised binary waveform where a logic 1= +1 V and a logic 0 = –1 V. Thus, for a logic 1 input, $v_m(t) = +1$. Therefore, the equation can be written as

$$v_{\text{fsk}}(t) = V_c \cos\{2\pi[f_c + \Delta f]t\} \tag{7.31}$$

For a logic 0 input, $v_m(t) = -1$. Therefore, the equation can be written as

$$v_{\text{fsk}}(t) = V_c \cos\{2\pi[f_c - \Delta f]t\} \tag{7.32}$$

Figure 7.22 illustrates frequency-shift keying.

where $f_0 = A\cos(\omega_c - \Delta\omega)t$ and $f_1 = A\cos(\omega_c + \Delta\omega)t$

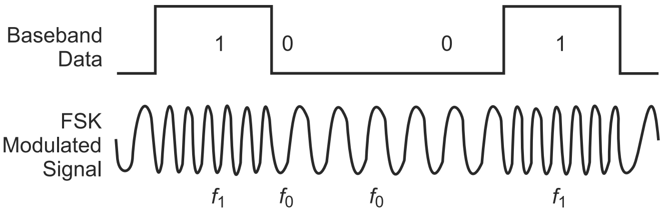


Fig. 7.22 Frequency shift keying

7.9.1 FSK Generation

There are two methods of FSK generation. They are as follows.

1. Separate oscillator method
2. Single oscillator method

Figure 7.23 shows the block diagram of the first method.

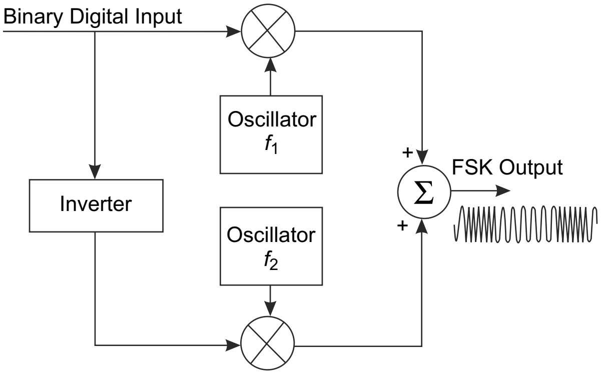


Fig. 7.23 Separate oscillator method

The two carriers may be generated from separate oscillators independent of one another. Those two signals are combined and the combined signal can, therefore, have discontinuities in amplitude and phase, which are undesirable.

On the other hand, the modulation can be achieved by frequency modulating a common carrier which prevents the discontinuities from occurring.

Figure 7.24 shows a single oscillator method of FSK generation.

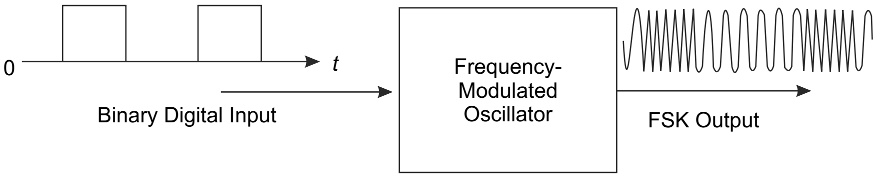


Fig. 7.24 Single oscillator method

The mean carrier frequency is denoted by f_c, and a binary '1' results in $f_1 = f_c + \Delta f$ and a binary '0' results in $f_0 = f_c + \Delta f$, where $2\Delta f$ is the difference between the two signalling frequencies. The modulated signal is given by

$$\text{Binary '0': } v_0\,(t) = A \cos 2\pi f_0\, t \tag{7.33}$$

$$\text{Binary '1': } v_1\,(t) = A \cos 2\pi f_1\, t \tag{7.34}$$

where the fixed phase angle has been set equal to zero for each signal. Where a single oscillator is frequency modulated by the digital signal, the method is called **Continuous Phase Frequency Shift Keying (CPFSK)**.

7.9.2 Frequency-Shift Keying (FSK) Detection

In frequency-shift keying, the modulating signals shift the output frequency between predetermined levels. There are two types of frequency-shift keying. They are as follows.

1. Coherent FSK
2. Noncoherent FSK

In coherent frequency-shift keying or binary FSK (BPSK), there is no phase discontinuity in the output signal. The BFSK method uses a couple of discrete frequencies to transmit binary information comprising '1' and '0'. With this scheme, the '1' is called the **mark frequency** and the '0' is called the **space frequency**. In noncoherent FSK, the instantaneous frequency is shifted between two discrete values named mark and space frequency, respectively.

1. Coherent Detection of FSK

The coherent detection of FSK is shown in Figure 7.25. From the figure, the outputs are combined to form a polar binary signal which is then passed to a matched filter. The outputs from the filters are combined to form a polar waveform which is then next passed as input to the pulse-regenerating circuit.

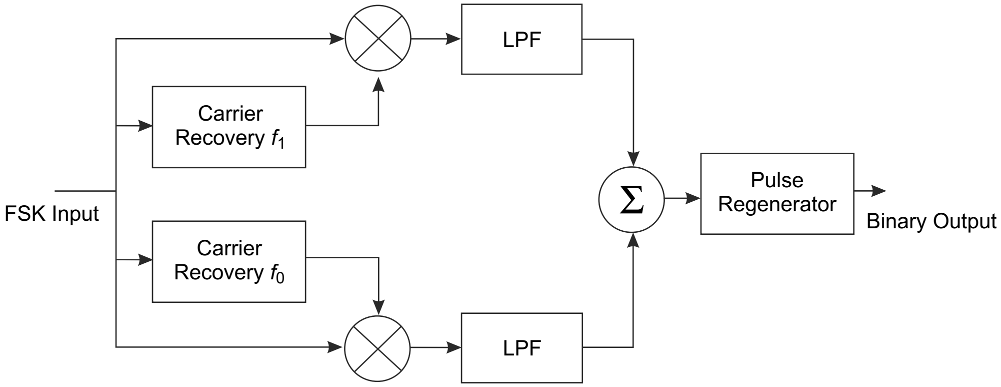


Fig. 7.25 Coherent detection of FSK

2. Non-coherent Detection of FSK

Non-coherent detection of FSK signals need two separate paths with two bandpass filters which are tuned to the individual frequencies, as shown in Figure 7.26. Each filter is followed by an envelope detector. The outputs are combined to form a polar waveform which is then passed as input to the pulse regeneration circuit at zero voltage threshold.

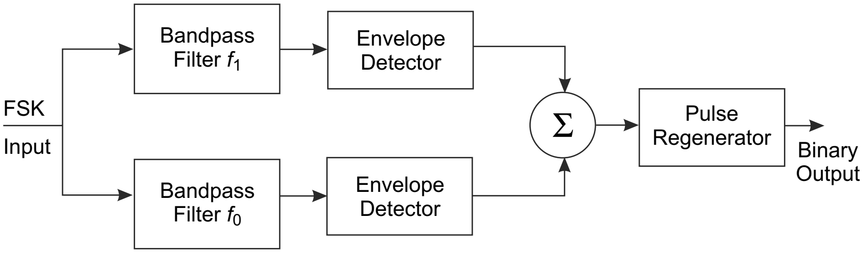


Fig. 7.26 Non-coherent detection of FSK

EXAMPLE 7.2

Determine the peak frequency deviation for a binary FSK signal with a mark frequency of 50 kHz, a space frequency of 54 kHz and input rate of 2 kbps.

Solution

The peak frequency deviation $\Delta f = \dfrac{|f_m - f_s|}{2}$

$$= \frac{|50-54|}{2} = 2 \text{ kHz}$$

EXAMPLE 7.3

For the above problem, find the minimum bandwidth and baud rate.

Solution

Minimum bandwidth is determined from

$$B = 2(\Delta f + f_b)$$

$$= 2(2000 + 2000)$$

$$= 8000 \text{ Hz} = 8 \text{ kHz}$$

$$\text{Baud rate} = \frac{f_b}{N}$$

$$= \frac{2000}{1} = 2000$$

3. Advantages of FSK

1. FSK method is less susceptible to noise.
2. It can be used up to 1200 bps on voice grade lines.
3. It can be used for high-frequency radio (3 to 30 MHz).
4. It can also be used for even higher frequencies on LANs using a coaxial cable.

7.9.3 Minimum-Shift Keying (MSK)

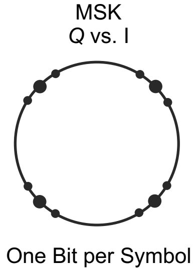


Fig. 7.27 Minimum-shift keying

Since a frequency shift produces an advancing or retarding phase, frequency shifts can be detected by sampling phase at each symbol period. Phase shifts of $(2N + 1)\ \pi/2$ radians are easily detected with an *I/Q* demodulator. At even-numbered symbols, the polarity of I channel conveys the transmitted data, while at odd-numbered symbols the polarity of the *Q* channel conveys the data. Figure 7.27 shows minimum-shift keying.

This orthogonality between *I* and *Q* simplifies detection algorithms and, hence, reduces power consumption in a mobile receiver. The minimum frequency shift which yields orthogonality of *I* and *Q* is that which results in a phase shift of ± π/2 radians per symbol (90° per symbol). FSK with this deviation is called MSK (Minimum-Shift Keying). The deviation must be accurate in order to generate repeatable 90° phase shifts. MSK is used in the GSM cellular standard. A phase shift of + 90° degrees represents a data bit equal to "1," while – 90° represents a "0." The peak-to-peak frequency shift of an MSK signal is equal to one half of the bit rate.

FSK and MSK produce constant-envelope carrier signals, which have no amplitude variations. This is a desirable characteristic for improving the power efficiency of transmitters. Amplitude variations can exercise nonlinearities in an amplifier's amplitude-transfer function, generating spectral re-growth, a component of adjacent channel power. Therefore, more efficient amplifiers can be used with constant-envelope signals, reducing power consumption.

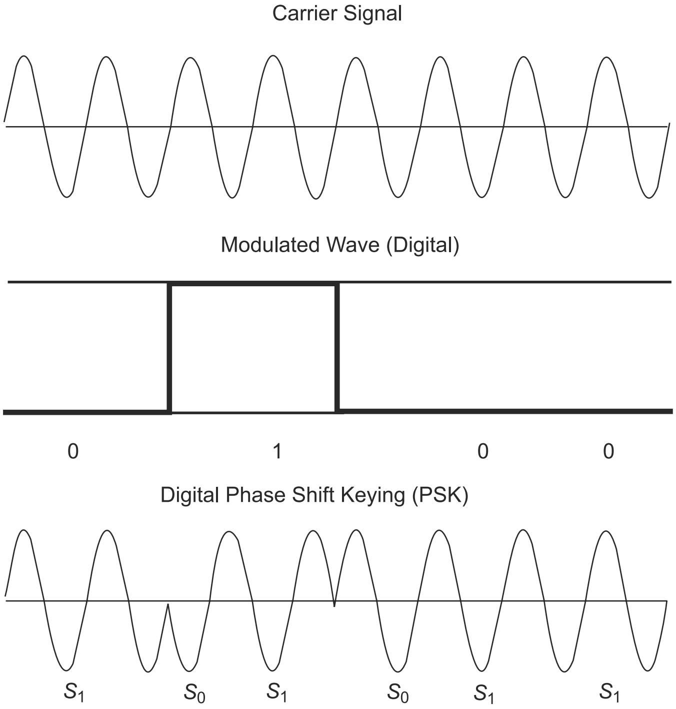
MSK has a narrower spectrum than wider deviation forms of FSK. The width of the spectrum is also influenced by the waveforms causing the frequency shift. If those waveforms have fast transitions or a high slew rate then the spectrum of the transmitter will be broad. To get a narrow spectrum, the waveforms are filtered with a Gaussian filter. In addition, the Gaussian filter has no time-domain overshoot, which would broaden the spectrum by increasing the peak deviation. MSK with a Gaussian filter is termed GMSK or Gaussian MSK.

7.10 PHASE-SHIFT KEYING (PSK)

Phase-Shift Keying (PSK) is a digital modulation scheme that conveys data by changing, or modulating, the phase of a reference signal such as the carrier wave. Figure 7.28 shows phase-shift keying.

The PSK method uses a base reference carrier and always sets the phase relative to that carrier. There are two problems with this approach.

1. If the phase of the signal can be disturbed, it will be very difficult to read at the receiver.
2. If there is a series of bits that translates to the same phase shift over and over again, it creates synchronisation problems



where $s_0 = -A \cos \omega_c t$ and $s_1 = A \cos \omega_c t$

Fig. 7.28 Phase-shift keying

7.10.1 Representation of Phase-Shift Keying

A convenient way to represent PSK schemes is on a constellation diagram. This shows the points in the argand plane where the real and imaginary axes are termed the in-phase and quadrature axes respectively due to their 90° separation. Such a representation on perpendicular axes lends itself to straightforward implementation. The amplitude of each point along the in-phase axis is used to modulate a cosine (or sine) wave and the amplitude along the quadrature axis to modulate a sine (or cosine) wave.

In PSK, the constellation points chosen are usually positioned with uniform angular spacing around a circle. This gives maximum phase separation between adjacent points. The points are positioned on a circle so that they can be transmitted with the same energy.

1. Coherent Phase-Shift Keying (CPSK)

In order to represent the digital data, all the types of digital modulation methods use a finite number of distinct signals. Phase-shift keying uses a finite number of phases and each is assigned a unique pattern of binary bits. Usually, each phase encodes an equal number of bits and each pattern of bits forms the symbol represented by the particular phase. The demodulator, which is designed specifically for the symbol-set used by the modulator, determines the phase of the received signal and maps it back to the symbol it represents and thus the original data is recovered. This requires the receiver to be able to compare the phase of the received signal to a reference signal and such a system is termed coherent phase-shift keying.

2. Differential Phase-Shift Keying (DPSK)

Instead of using the bit patterns to set the phase of the wave, it can be used to change the phase by a specified amount. The demodulator then determines the particular changes in the phase of the received signal rather than the phase itself. Since this scheme depends on the difference between successive phases, it is termed Differential Phase-Shift Keying (DPSK).

In DPSK, there is no need for the demodulator to have a copy of the reference signal to determine the exact phase of the received signal. So this method is much simpler to implement than ordinary PSK. It is also termed **noncoherent phase-shift keying**.

7.10.2 Binary Phase-Shift Keying (BPSK)

Binary phase-shift keying is otherwise termed as Phase Reversal Keying (PRK) or 2-PSK. Since this method uses two phases which are separated by 180°, it can be termed 2-PSK. The binary signal is used to switch the phase between 0° and 180°.

Figure 7.29 shows the constellation diagram of BPSK.

This modulation is the most robust of all the phase-shift keying methods since it takes the highest level of noise or distortion which leads wrong demodulation at the receiver side. This method can modulate at 1 bit/symbol and so it is unsuitable for high data-rate applications when bandwidth is limited. In the presence of an arbitrary phase-shift introduced by the communication channel, the demodulator is unable to identify the particular constellation point. Due to this reason, the data is often differentially encoded prior to modulation.

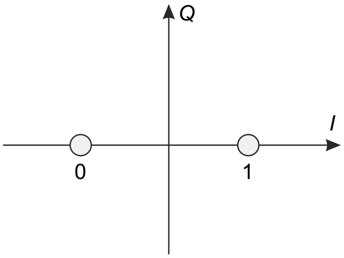


Fig. 7.29 Constellation diagram of BPSK

1. Implementation of BPSK

The binary data used in BPSK are ‘0’ and ‘1’ which are often conveyed with the following signals:

$$S_0\,(t) = V_c \cos(2\pi f_c t + \phi + 180°) \text{ for binary ‘0’} \qquad (7.35)$$

$$S_1\,(t) = V_c \cos(2\pi f_c t + \phi) \text{ for binary ‘1’} \qquad (7.36)$$

where f_c is the frequency of the carrier wave.

2. Error Rate Calculation of BPSK

The Bit Error Rate (BER) of BPSK can be calculated as follows.

$$P_b = \frac{1}{2} erfc \sqrt{\frac{E_b}{N_0}} \tag{7.37}$$

where $\frac{E_b}{N_0}$ is the received signal-to-noise ratio.

3. BPSK Generation

The circuit for the generation of BPSK is a balanced modulator as shown in Figure 7.30.

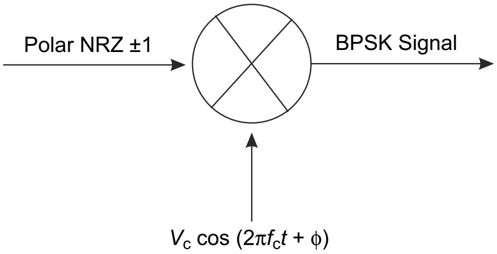


Fig. 7.30 BPSK generation

The modulating signal is polar NRZ and when this value is +1, the modulated output is $V_c \cos(2\pi f_c t + \phi)$ and when this signal is –1, the modulated output is $V_c \cos(2\pi f_c t + \phi + 180°)$.

4. Coherent Detection of BPSK

In order to recover the original NRZ signals, coherent detection circuit is used as per Figure 7.31.

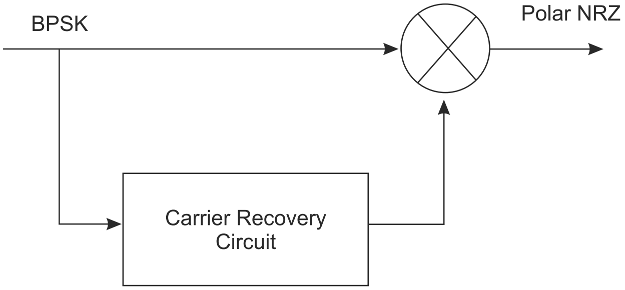


Fig. 7.31 Coherent detection of BPSK

Since the envelope does not contain the modulating information, coherent detection must be necessarily used. Carrier recovery circuit is used from which the signal is multiplied with the BPSK signal and the resultant product will be the polar NRZ signal.

EXAMPLE 7.4

What are the phase states of the carrier when the bit stream 1 0 1 1 1 0 0 1 0 0 is applied to a 4-PSK modulator?

Solution

1	Modulator input	1 0	1 1	1 0	0 1	0 0
	Phase states of the transmitted carrier	7π/4	5π/4	7π/4	3π/4	π/4
2	Relative phase with respect to the recovered carrier	3π/4	π/4	3π/4	7π/4	5π/4
	Output of the demodulator	0 1	0 0	0 1	1 0	1 1

EXAMPLE 7.5

Write the phase states of the differential BPSK carrier for input data stream 100110101. The starting phase of the carrier can be taken as 0.

Solution

A		1	0	0	1	1	0	1	0	1
$\Delta\phi$		π	0	0	π	π	0	π	0	π
ϕ	0	π	π	π	0	π	π	0	0	π

EXAMPLE 7.6

The following bit stream is applied to the differential 4-PSK modulator. Write the carrier phase states taking the initial carrier phase as reference.

Solution

Bit stream	1	0	1	1	1	1	0	0	0	1
$\Delta\phi$		3π/2		π		π		0		π/2
ϕ	0	3π/2		π/2		3π/2		3π/2		0

EXAMPLE 7.7

For an 8-PSK system, operating with information bit rate of 24 kbps, determine the baud rate.

Solution

$$\text{Baud rate} = \frac{f_\text{b}}{N}$$

$$= \frac{24000}{3} = 8000$$

EXAMPLE 7.8

For Example 7.7 determine the bandwidth efficiency.

Solution

$$\text{Bandwidth efficiency} = \frac{\text{Transmission bit rate}}{\text{Minimum bandwidth}}$$

$$= \frac{24000}{8000} = 3 \text{ bits/ second/cycle of bandwidth}$$

7.10.3 Quadrature Phase-Shift Keying (QPSK)

Quadrature phase-shift keying (QPSK) is otherwise termed **quadriphase-PSK** or **4-PSK.** Since this method uses four distinct levels of phase shift, it is widely used in multilevel modulation. With four phases, QPSK can encode two bits per symbol to minimise the BER. This bit error rate is twice the rate of BPSK.

Since there are four points on the constellation diagram, equispaced around a circle, it is referred to as 4-PSK. Figure 7.32 shows the constellation diagram for QPSK in which each adjacent symbol only differs by one bit.

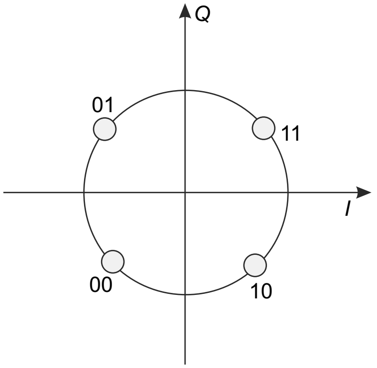


Fig. 7.32 Constellation diagram of QPSK

1. Implementation of QPSK

It is easy to implement QPSK than that of BPSK. It is considered as an implementation of higher-order PSK.

$$s_{\mathrm{i}}(t) = V_{\mathrm{c}} \cos\left(2\pi f_{\mathrm{c}} t + (2i-1)\frac{\pi}{4}\right), \quad i = 1,2,3,4 \tag{7.38}$$

This yields the four phases π/4, 3π/4, 5π/4 and 7π/4 as needed.

2. Error Rate Calculation of QPSK

The probability of bit error for QPSK is the same as for BPSK.

$$P_{\mathrm{b}} = \frac{1}{2}\sqrt{\frac{E_{\mathrm{b}}}{N_0}} \tag{7.39}$$

However, in order to achieve the same bit-error probability as BPSK, QPSK uses twice the power (since two bits are transmitted simultaneously).

The symbol error rate is given by

$$P_{\mathrm{s}} = 1 - (1 - P_{\mathrm{b}})^2$$

$$= \frac{1}{4}\left(\frac{E_b}{N_0}\right) - \left(\sqrt{\frac{E_b}{N_0}}\right) \quad (7.40)$$

If the Signal-to-Noise Ratio (SNR) is high, the probability of symbol error may be approximated as follows.

$$P_s \approx \left(\sqrt{\frac{E_b}{N_0}}\right) \quad (7.41)$$

where $\frac{E_b}{N_0}$ is the received signal-to-noise ratio.

3. QPSK Generation

Figure 7.33 shows the block diagram of QPSK generation and Figure 7.34 shows different generation states of QPSK.

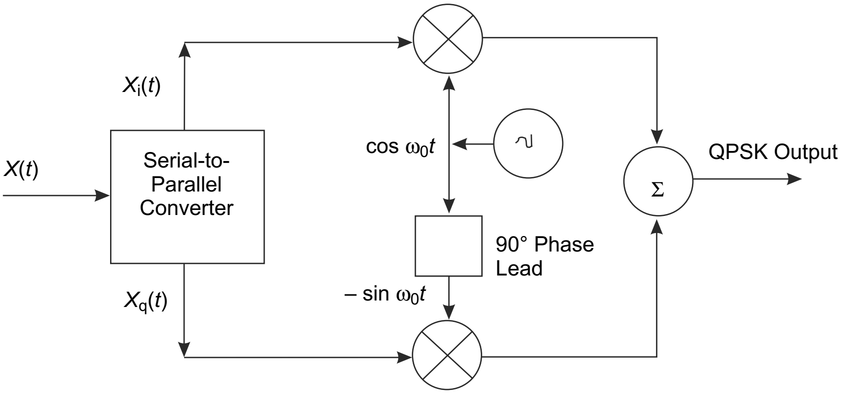


Fig 7.33 QPSK generation

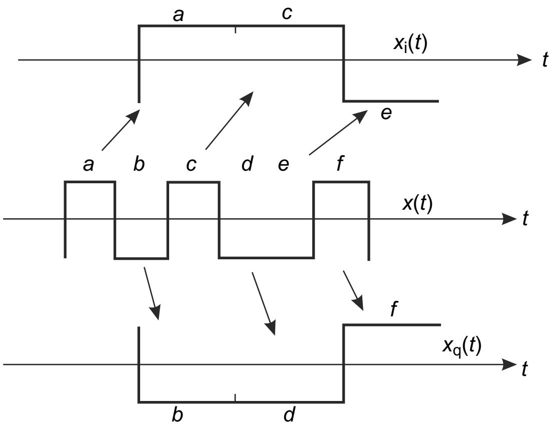


Fig. 7.34 Generation states of QPSK

In QPSK generation circuit, a serial-to-parallel converter is used to convert the binary signal into two separate binary signals in which the period is doubled. These two binary signals are labeled as $x_i(t)$ for in-phase and $x_q(t)$ for quadrature-phase components respectively.

The in-phase component modulates a carrier to produce a BPSK signal while the quadrature component modulates a carrier component shifted by 90° which is also to produce a BPSK signal. The two BPSK signals are added to produce the QPSK signal. Thus, the QPSK signal is equivalent to two BPSK signals, but with the carriers 90° out of phase with one another. The following table shows the different generation states of QPSK.

$x_i(t)$	$x_q(t)$	QPSK generation states
1	1	$\cos \omega_0(t) - \sin \omega_0(t) = \sqrt{2}\cos \omega_0 (t + 45°)$
1	–1	$\cos \omega_0(t) + \sin \omega_0(t) = \sqrt{2}\cos \omega_0 (t - 45°)$
–1	1	$-\cos \omega_0(t) - \sin \omega_0(t) = \sqrt{2}\cos \omega_0 (t + 135°)$
–1	–1	$-\cos \omega_0(t) + \sin \omega_0(t) = \sqrt{2}\cos \omega_0 (t - 135°)$

The two BPSK signals do not interfere with one another because of the phase difference between the carriers. Thus, QPSK signalling requires one-half the bandwidth of BPSK signalling for the same input bit rate in both the cases.

4. QPSK Detection

The detection procedure for QPSK signal is same as that of BPSK detection. The only difference between them is that the recovered carriers in the QPSK method of detection must also have the 90° phase difference. Figure 7.35 shows the QPSK detection circuit. The demodulated output from the circuit is followed by a matched filter detector.

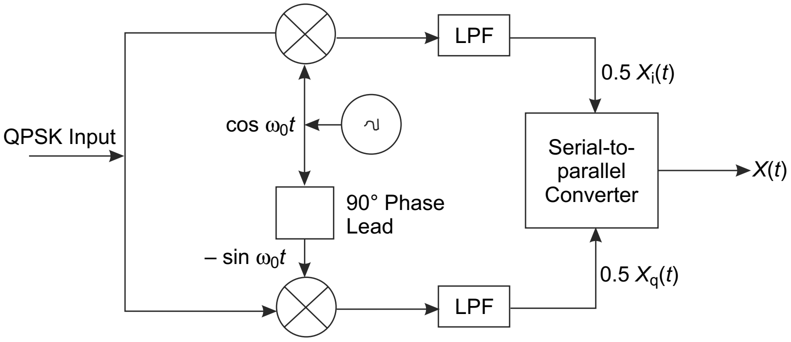


Fig. 7.35 QPSK detection circuit

5. QPSK Signal in Time Domain

The QPSK modulated signal for a random binary data stream is shown in Figure 7.28, from which the odd-numbered bits have been assigned to the in-phase component and the even-numbered bits to the quadrature component. The sum of the two components is also shown in Figure 7.36. PSK changes the phase on each component at the start of each bit period, indicated by the jumps in phase.

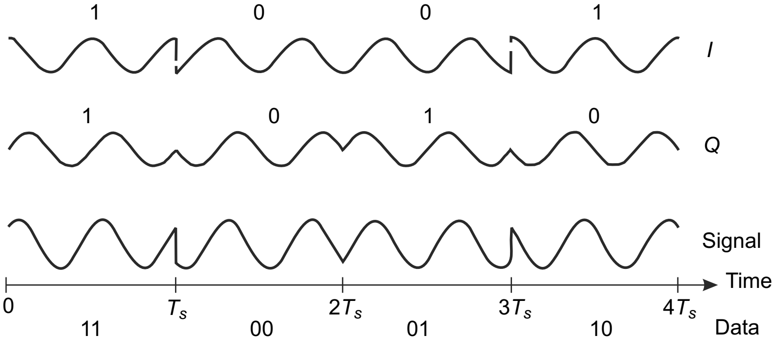


Fig. 7.36 Timing diagram for QPSK

The binary data that is conveyed by the above timing diagram waveform is 1 1 0 0 0 1 1 0 from which the odd bits contribute to the in-phase component such as $\underline{1}$ 1 $\underline{0}$ 0 $\underline{0}$ 1 $\underline{1}$ 0 and the even bits contribute to the quadrature-phase component such as 1 $\underline{1}$ 0 $\underline{0}$ 0 $\underline{1}$ 1 $\underline{0}$.

6. Offset QPSK (OQPSK)

In Offset Quadrature Phase-Shift Keying (OQPSK), which is a variant of phase-shift keying modulation, there are four different values of the phases used to transmit the information. This method is also known as **Staggered Quadrature Phase-Shift Keying (SQPSK)**.

In order to construct a QPSK signal, four values of the phase or simply two bits are taken at a time. This allows the phase of the signal to jump by as much as 180° at a time. When the signal is lowpass filtered, these phase shifts result in large amplitude fluctuations and there will be an undesirable quality in communication systems. By offsetting the timing of the odd and even bits by one bit period, or half a symbol period, the in-phase and quadrature components will never change at the same time.

Figure 7.37 shows the constellation diagram of OQPSK and from which it can be seen that this will limit the phase shift to no more than 90° at a time. This yields much lower amplitude fluctuations than non-offset QPSK.

The difference in the behaviour of the phase between ordinary between QPSK and OQPSK is shown in Figure 7.38. From the above figure, it can be seen that in the first plot, the phase can change by 180° at once, while in OQPSK, the changes are never greater than 90°.

For the offset QPSK, the timing diagram of the modulated signal is shown for a random binary data stream in Figure 7.39.

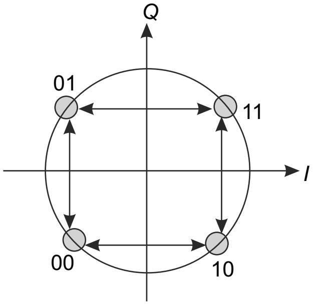


Fig. 7.37 Constellation diagram of OQPSK

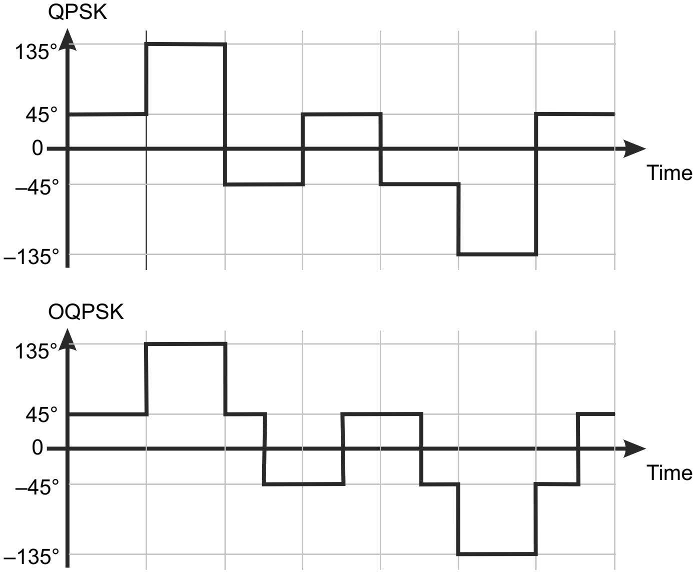


Fig. 7.38 Difference between QPSK and OQPSK

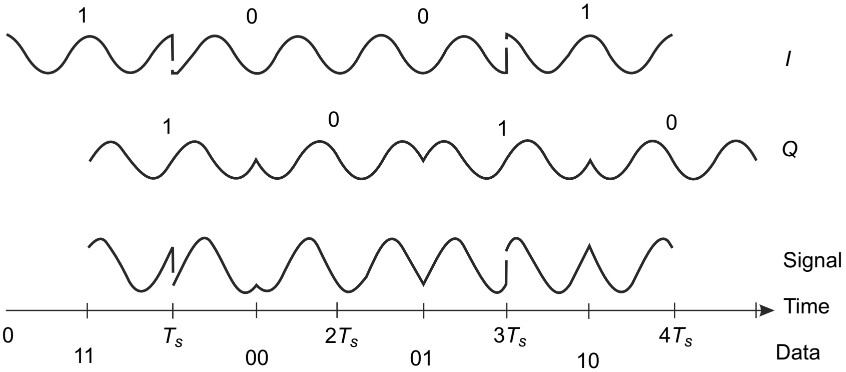


Fig. 7.39 Timing diagram of OQPSK

From the above timing diagram of OQPSK, it is to be noted that the half-symbol-period is offset between the two component waves. The sudden phase shifts occur about twice as often as for QPSK and the magnitude of jumps is smaller in OQPSK when compared to QPSK.

7. π/4 QPSK

In π/4 QPSK method, which is also one variant of QPSK, there are two separate constellations which are rotated by 45° with respect to each other. Figure 7.40 shows the constellation diagram of π/4 QPSK.

Generally, from any one of the constellations, either the even or odd symbols are to be selected and the other symbols select points from the other constellation. This method also reduces the phase shifts from a maximum of 180°, but only to a maximum of 135° and so the amplitude fluctuations of π/4 QPSK are between OQPSK and non-offset QPSK.

The timing diagram of π/4 QPSK modulated signal is shown in Figure 7.41 for a random binary data stream. The construction is the same as for ordinary QPSK. From the above timing

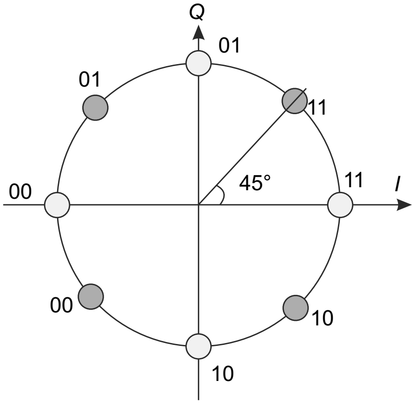


Fig. 7.40 Constellation diagram of π/4 QPSK

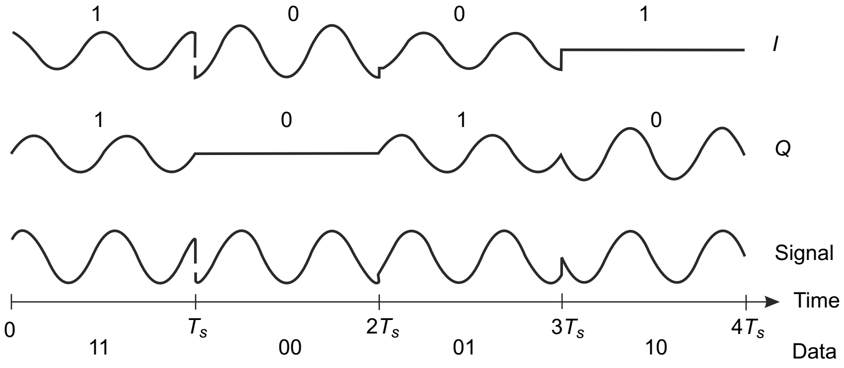


Fig. 7.41 Timing diagram of π/4 QPSK modulated signal

diagram, it is noted that successive symbols are taken from the two constellations. Then, the first symbol (1 1) is taken from the 'white' constellation and the second symbol (0 0) is taken from the 'grey' constellation. Note that magnitudes of the two component waves change as they switch between constellations, but the total signal's magnitude remains constant. The phase shifts are between those of the two previous timing diagrams.

8. Advantage of QPSK

1. As bits are grouped and transmitted in pairs, the bandwidth needed is half compared to binary PSK.
2. In QPSK method, there is a possibility to transmit in the same frequency band double the information, while the number of errors and the signal to-noise-ratio relation are the same.
3. QPSK has more number of samples when compared with the BPSK method and the error rate will also be low.

7.10.4 Differential Phase-Shift Keying (DPSK)

Differential Phase Shift Keying (DPSK) is a phase-modulation process that conveys data by changing the phase of the carrier wave. In previously discussed methods such as BPSK and QPSK, there is an ambiguity of phase if the constellation is rotated by some effect in the communication channel through which the signal passes. This problem can be eliminated by using the data to change rather than set the phase.

In the differentially encoded BPSK (DBPSK) method, a binary '1' may be transmitted by adding 180° to the current phase and a binary '0' may be transmitted by adding 0° to the current phase, whereas in an another variant of DPSK called **Symmetric Differential Phase Shift keying (SDPSK)**, encoding would be +90° for a binary '1' and –90° for a binary '0'.

In differentially encoded QPSK (DQPSK), the phase shifts are 0°, 90°, 180°, –90° corresponding to data '00', '01', '11' and '10'. This kind of encoding may be demodulated in the same way as for non-differential PSK but the phase ambiguities can be ignored. Thus, each received symbol is demodulated to one of the points in the constellation and then a comparator computes the difference in phase between this received signal and the preceding one. Then data encoding will take place by the difference.

The timing diagram of DPSK is shown in Figure 7.42 in which the modulated signal is shown for both DBPSK and DQPSK methods.

In the figure, it is to be noted that the signal starts with zero phase and so there is a phase shift in both signals at $t = 0$.

1. DPSK Generation

A binary input signal is represented as x_k and its differentially encoded signal is represented as e_k whereas e_k sequence phase modulates a carrier $\cos\omega_c t$. The output from the modulator is a

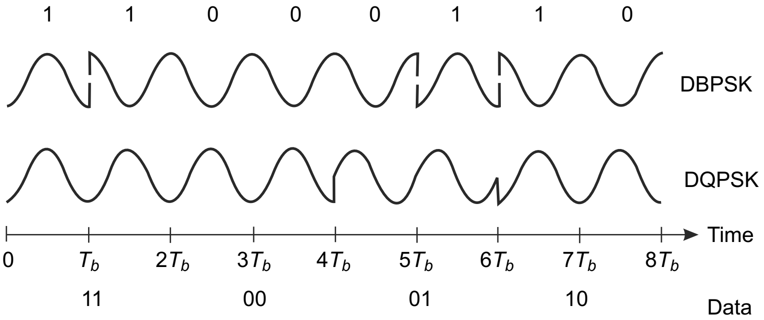


Fig. 7.42 Timing diagram of DPSK

BPSK signal. But since it is modulated by a differentially encoded signal, it will be referred to as Differentially Phase Shift Keying (DPSK). Figure 7.43 shows the block diagram of DPSK generation method.

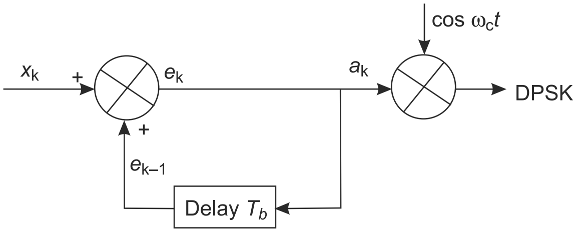


Fig. 7.43 DPSK generation

Let a_k = +1 V for a binary '1' and a_k = –1 V for a binary '0'. The DPSK signal can be represented by

$$DPSK = a_k \cos\omega_c t. \tag{7.42}$$

It is to be noted that a binary '1' corresponds to no phase shift and a binary '0' corresponds to 180° phase shift in the carrier signal.

2. DPSK Detection

In order to detect the DPSK signal, a decoder which is used at the receiving end consists of a multiplier in which the DPSK input signal is to be multiplied with 1-bit delayed version of the same DPSK signal. The block diagram of DPSK detection is shown in Figure 7.44.

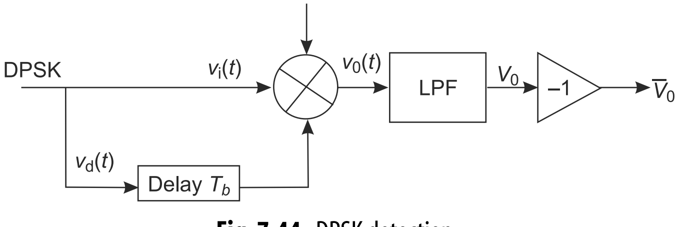


Fig. 7.44 DPSK detection

The output from the multiplier is

$$v_0\,(t) = kv_{\mathrm{i}}\,(t)v_{\mathrm{d}}(t)$$

$$= k(a_{\mathrm{k}} \cos \omega_{\mathrm{c}}t) \times (a_{\mathrm{k-1}} \cos \omega_{\mathrm{c}}\,(t - T_{\mathrm{b}})) \tag{7.43}$$

where k is the multiplier constant.

By making $f_c = \dfrac{N}{T_b}$, where N is an integer, $\cos \omega_{\mathrm{c}}\,(t - T_{\mathrm{b}}) = \cos\omega_{\mathrm{c}}t,$

$$v_0\,(t) = ka_{\mathrm{k}}a_{\mathrm{k-1}} \cos^2 \omega_{\mathrm{c}}t$$

$$= ka_k\, a_{k-1}\left(\frac{1}{2} + \frac{1}{2}\cos 2\,\omega_c\, t\right) \tag{7.44}$$

The lowpass filter at the output of the multiplier removes the second harmonic term of the carrier, leaving only the baseband component. Representing the constants by A, the output of the lowpass filter is

$$v_0 = Aa_{\mathrm{k}}a_{\mathrm{k-1}} \tag{7.45}$$

$$= +\mathrm{A} \qquad \text{for} \quad a_{\mathrm{k}} = a_{\mathrm{k-1}} \tag{7.46}$$

$$= -\mathrm{A} \qquad \text{for} \quad a_{\mathrm{k}} \neq a_{\mathrm{k-1}}$$

The different logic levels corresponding to V_0 are shown in the following table.

d_{k} (logic levels)	1	1	0	1	0	0	1
e_{k} (logic levels)	1	0	0	1	1	1	0
a_{k} (volts)	+1	–1	–1	+1	+1	+1	–1
a_{k} (volts)	+1	–1	–1	+1	+1	+1	–1
$a_{\mathrm{k-1}}$ (volts)	–1*	+1	–1	–1	+1	+1	+1
V_0 (volts)	–1	–1	+1	–1	+1	+1	–1
V_0 (logic levels)	0	0	1	0	1	1	0
$\overline{V}_0$ (logic levels)	1	1	0	1	0	0	1

The levels shown in the above table are the message levels inverted and an inverter is required to restore the message sequence at the output. A disadvantage of DPSK method is that bit errors tend to occur in pairs, because the polarity of a given bit depends on the polarity of the preceding bit.

7.11 MODEMS

The term **modem** is derived from the words, **Modulator** and **DEModulator**. A modem contains a modulator as well as a demodulator. The digital modulation/demodulation schemes are implemented in the modems. Most of the modems are designed for utilising the analog voice-band service offered by the telecommunication network. Therefore, the modulated carrier generated by a modem fits into the 300–3400 Hz bandwidth of the speech channel.

7.11.1 Connection Set-Up using Modems

A typical data connection set-up using modems is shown in Figure 7.45.

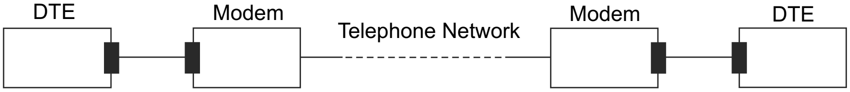


Fig. 7.45 A data circuit implemented using modems

The digital terminal devices which exchange digital signals are called **Data Terminal Equipment (DTE).** Two modems are always required, one at each end. The modem at the transmitting end converts the digital signal from the DTE into an analog signal by modulating a carrier. The modem at the receiving end demodulates the carrier and hands over the demodulated digital signal to the DTE.

The transmission medium between the two modems can be a dedicated leased circuit or a switched telephone circuit. In the latter case, modems are connected to the local telephone exchanges. Whenever data transmission is required, connection between the modems is established through the telephone exchanges. Modems are also required within a building to connect terminals which are located at distances usually more than 15 metres from the host.

7.11.2 Building blocks of a Modem

A modem comprises a transmitter, a receiver and two interfaces. Figure 7.46 shows the building blocks of a modem.

The digital signal to be transmitted is applied to the transmitter. The modulated carrier which is received from the distant end is applied to the receiver. The digital interface connects

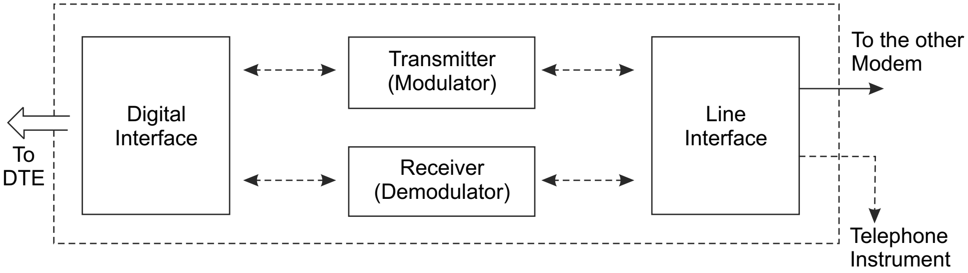


Fig. 7.46 Building blocks of a modem

the modem to the DTE which generates and receives the digital signals. The line interface connects the modem to the transmission channel for transmitting and receiving the modulated signals. Modems connected to telephone exchanges have additional provision for connecting a telephone instrument. The telephone instrument enables establishment of the telephone connection.

The transmitter and receiver in a modem comprise several signal processing circuits which include a modulator in the transmitter and a demodulator in the receiver.

7.11.3 Features of Modems

The following are the features of data modems.

1. Speed

The speed at which the modem can send data is represented in bits per second. Typical modem speeds are 300, 600, 1200, 2400, 4800, 9600, 14.4K, 19.2K, 28.8K bps.

2. Auto Dial/Redial

Smart modems can dial the phone number automatically and redial if a busy signal is received.

3. Auto Answer

Most modems can automatically answer the phone when an incoming call comes in. They have ring detect capability.

4. Self-Testing

New modems have self-testing features. They can test the digital connection to the terminal/ computer and the analog connection to a remote modem. They can also check the modem's internal connections.

5. Voice over Data

Voices over data modems allow a voice conversation to take place while data is being transmitted. This requires both the source and destination modems to have this feature.

6. Synchronous or Asynchronous Transmission

Newer modems allow a choice of synchronous or asynchronous transmission of data. Normally, modem transmission is asynchronous. Individual characters are sent with just start and stop bits.

7.11.4 Types of Modems

Modems can be of several types and they can be classified in a number of ways. The classification is usually based on the following basic modem features.

1. *Directional capability*
   a. Half-duplex modem
   b. Full-duplex modem
2. *Connection to the line*
   a. 2-wire modem
   b. 4-wire modem
3. Transmission mode
   a. Asynchronous modem
   b. Synchronous modem

1. Half-Duplex Modem and Full-Duplex Modem

A half-duplex modem permits transmission in one direction at a time. If a carrier is detected on the line by the modem, it gives an indication of the incoming carrier to the DTE through a control signal of its digital interfaces. So long as the carrier is being received, the modem does not give clearance to the DTE to transmit. Figure 7.47 shows a half-duplex modem.

A full-duplex modem allows simultaneous transmission in both directions. Thus, there are two carriers on the line, one outgoing and the other incoming. Figure 7.48 shows a full-duplex modem.

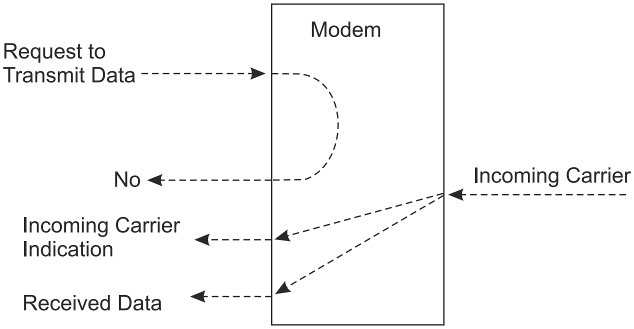


Fig. 7.47 Half-duplex modem

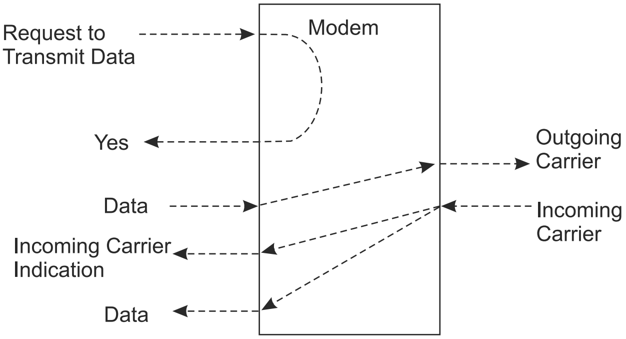


Fig. 7.48 Full-duplex modem

2. 4-Wire Modem and 2-Wire Modem

In a 4-wire connection shown in Figure 7.49, one pair of wires is used for the outgoing carrier and the other is used for the incoming carrier.

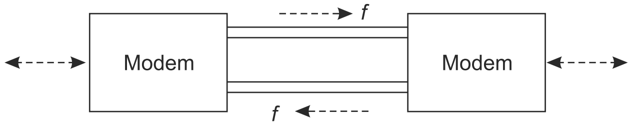


Fig. 7.49 4-wire modem

Full-duplex and half-duplex modes of transmission are possible on a 4-wire connection. As the physical transmission path for each direction is separate, the same carrier frequency can be used for both the directions.

A 2-wire connection, shown in Figure 7.50, is cheaper than a 4-wire connection because only one pair of wire is extended to the subscriber's premises. The data connection established through telephone exchanges is also a 2-wire connection. For the 2-wire connections, modems with a 2-wire line interface are required. Such modems use the same pair of wires for outgoing and incoming carriers.

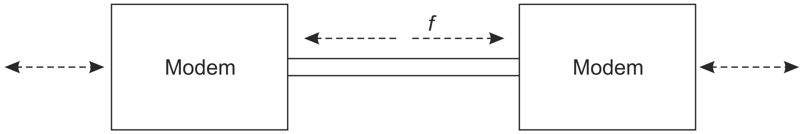


Fig. 7.50 2-wire modem for half-duplex mode

Half-duplex mode of transmission using the same frequency for the incoming and outgoing carriers can be easily implemented. In a half-duplex modem, transmit and receive carrier frequencies can be the same because only one of them is present on the line at a time.

For a full-duplex mode of operation on a 2-wire connection shown in Figure 7.51, it is necessary to have two transmission channels, one for the transmit direction and the other for the receive direction.

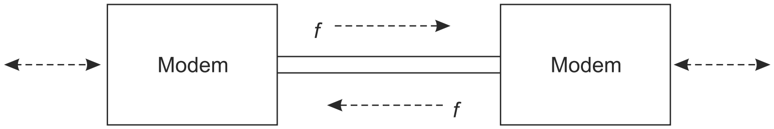


Fig. 7.51 2-wire modem for full-duplex mode

This is achieved by frequency-division multiplexing of two different carrier frequencies. These carriers are placed within the bandwidth of the speech channel. A modem transmits data on one carrier and receives data from the other end on the other carrier. A hybrid is provided in the 2-wire modem to couple the line to its modulator and demodulator which is shown in Figure 7.52.

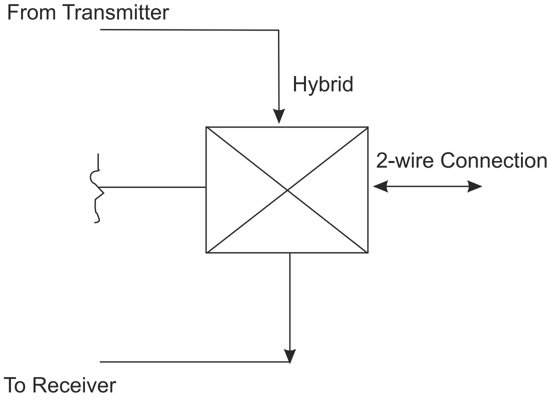


Fig. 7.52 Line interconnection in a 2-wire full-duplex modem

It is to be noted that the available bandwidth for each carrier is reduced to half. Therefore, the baud rate is also reduced to half. There is a special technique which allows simultaneous transmission of incoming and outgoing carriers having the same frequency on the 2-wire transmission medium. Full bandwidth of the speech channel is available to both the carriers simultaneously. This technique is called **echo cancellation technique** and is implemented in high-speed 2-wire full-duplex modems.

3. Asynchronous and Synchronous Modems

Modems for asynchronous and synchronous transmission are of different types. An asynchronous modem can only handle data bytes with start and stop bits. There is no separate timing signal or clock between the modem and the DTE. It is shown in Figure 7.53. The internal timing pulses are synchronised repeatedly to the leading edge of the start pulse.

A synchronous modem can handle a continuous stream of data bits but requires a clock signal, which is shown in Figure 7.54.

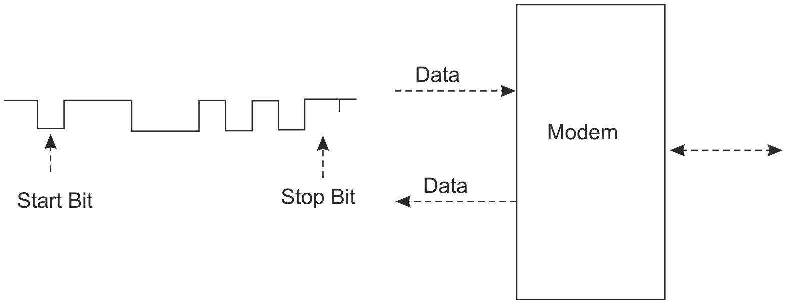


Fig. 7.53 Asynchronous modem

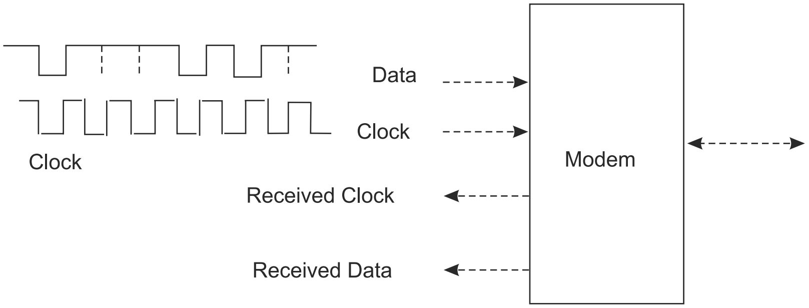


Fig. 7.54 Synchronous modem

The data bits are always synchronised to the clock signal. There are separate clocks for the data bits being transmitted and received. For synchronous transmission of data bits, the DTE can use its internal clock and supply the same to the modem. Otherwise it can take the clock from the modem and send data bits on each occurrence of the clock pulse. At the receiving end, the modem recovers the clock signal from the received data signal and supplies it to the DTE. However, it is necessary that the received data signal contains enough transitions to ensure that the timing extraction circuit remains in synchronisation. High-speed modems are provided with scramblers and descramblers for this purpose.

4. Bell System-103 compatible Modem

The Bell system-103 compatible modem operates in full-duplex mode over two-wire switched telephone lines. With this modem, the data rate is up to 300 bps. There are two types of data channels. One is low-band channel with a bandwidth of (300–1650) Hz and a highband channel of bandwidth of (1650–3000) Hz. The baud rate of Bell 103 modem is about 300 bauds.

5. Bell System-202 T/S Modem

Bell system 202T and Bell system 202S are identical. These methods differ in their mode of operation and line connection. The Bell-202T modem specifies 4-wire full-duplex operation and Bell-202S specifies 2-wire half-duplex operation. They have 1200 baud transceiver with the transmission bit rate of 1200 bps.

6. Low-Speed and High-Speed Modems

Low-speed modems are designed to operate asynchronously. Examples for low-speed modems are Bell-103 compatible modem, Bell-202T/S modem and Bell-212 modem.

High-speed modems are designed to operate in synchronous mode. Examples for high-speed modems are Bell-208 compatible modem, Bell-201T/S modem and Bell-209 modem.

7.11.5 Additional Features of Modems

Modems vary in design and complexity depending on speed, mode of transmission, modulation methods and their application. The driving force for the developments in the modem has been the high cost of the transmission medium.

By more efficient utilisation of the available bandwidth and increasing the effective throughput, the high cost of transmission can be neutralised. Echo cancellers and secondary channel are the two additional features of modems in this direction. For ease of operation, modems are also equipped with test loops.

7.12 QUANISATION PROCESS

Generally, a signal can be defined as a variable which changes subject to any other independent variable. If the independent variable is assumed as time, denoted by *t,* the dependent variable could be any physical measurement variable which changes over time. An example of such a signal is a sinusoidal signal and if it is to be represented in digital domain, the following are to be done.

1. Sampling
2. Quantisation

7.12.1 Sampling

Sampling is a process used to obtain signal values from the continuous signal at regular time-intervals in which the sampling interval is denoted as T_s and its reciprocal, the **sampling frequency** or **sample rate** is denoted as f_s, where $f_s = \frac{1}{T_s}$. The result of this process is just a sequence of numbers.

Having defined our sampling interval T_s, sampling just extracts the signal's value at all integer multiples of T_s such that our discrete time sequence $x(n)$ can be written as follows.

$$x(n) = x(n.T_s) \tag{7.47}$$

After sampling, the input signal is not yet completely digitalised because the values $x(n)$ can still take on any number from a continuous range. Figure 7.55 illustrates the process of sampling a continuous sinusoidal signal.

The resultant waveform shows a sequence of numbers which are called **sampling points**. In this sampling process, it is important to consider the reconstruction of the original sinusoidal signal from the sampled signal.

7.12.2 Interpolation

Consider a continuous signal, $x(t)$, which is defined for all values of t whereas the discrete-time signal is defined for times which are integer multiples of T(s). To reconstruct a continuous

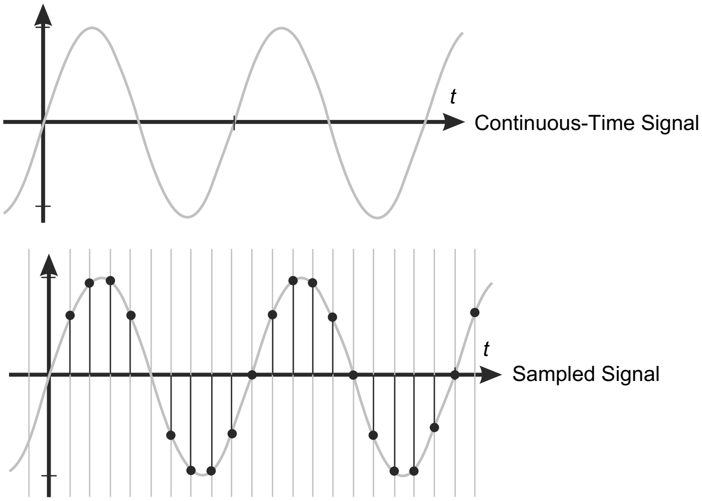


Fig. 7.55 Process of sampling

signal from the samples, interpolation is the process of identifying signal values at arbitrary instants of time. Interpolation creates a continuous-time signal and can be seen as an inverse process to sampling. The continuous signal obtained from interpolation should be equal to the original continuous signal.

The simpler interpolation scheme is a piecewise constant interpolation in which the value of one of the neighbouring samples is taken as identified signal value at any instant of time in between. The resultant reconstructed interpolated function will be in stair-step shape. Another popular interpolation method is linear interpolation which is used to reconstruct a signal value by simply connecting the values at the sampling instants with straight lines. Figure 7.56 shows the reconstructed continuous signals for piecewise constant and linear interpolation.

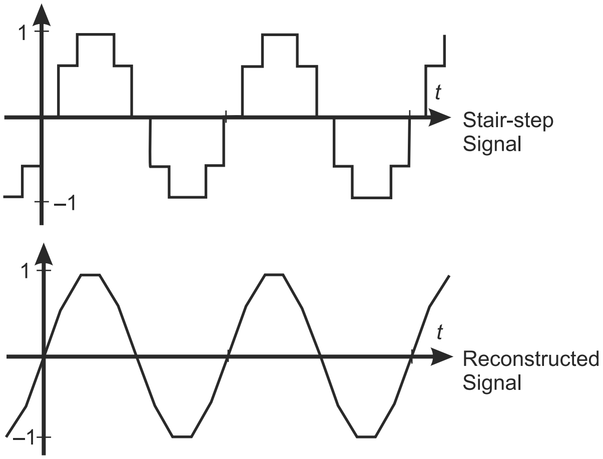


Fig. 7.56 Process of signal reconstruction

7.12.3 The Sampling Theorem

The sampling theorem defines conditions under which a continuous-time signal $x(t)$ can be reconstructed exactly from its samples and also defines the interpolation algorithm which should be used to achieve this exact reconstruction. In other words, the sampling theorem states that the original continuous-time signal can be reconstructed from its samples exactly, when the highest frequency, denoted as f_h, present in the signal is lower than a half of the sampling frequency f_s:

$$f_h < \frac{f_s}{2} \tag{7.48}$$

where $\frac{f_s}{2}$ is also often called **Nyquist frequency**. When this condition is satisfied, the continuous-time signal can be reconstructed exactly. When the above condition is not met, sampling and reconstructing the signal is done anyway but with errors called aliasing.

7.12.4 Aliasing

According to the sampling theorem, the signal shown in Figure 7.55 is sampled properly. As this particular signal is just a single sinusoidal signal, the highest frequency present in this signal is just the frequency of this sinusoidal signal. The sampling frequency is 10 times the sinusoidal frequency, which is expressed as follows.

$$f_h = \frac{f_s}{10} < \frac{f_s}{2} \tag{7.49}$$

To see this situation, around 10 samples are taken from each period of the signal and this is more than enough to exactly reconstruct the underlying continuous-time sinusoidal signal. To make it more precise, the sinusoidal signal is assumed to have a frequency of 1Hz and the sample rate was 10 Hz. Now let's see what happens when the 8 Hz sinusoidal signal is sampled at a sample rate of 10 Hz. In Figure 7.57, the sinusoidal signal at $f = 8$ Hz is aliased into another sinusoid at $f = 2$ Hz, at sampling frequency $f_s = 10$ Hz because both produce the

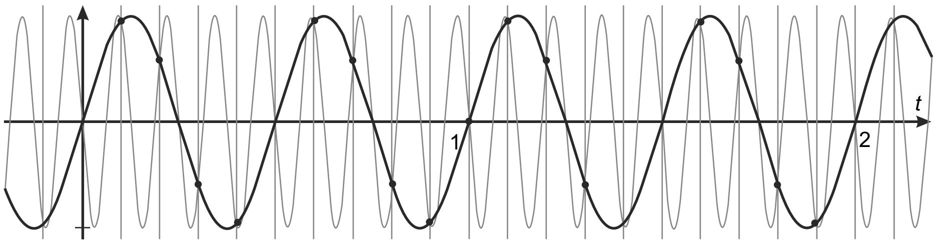


Fig. 7.57 Aliasing

same sequence of samples. The high-frequency sinusoid at 8 Hz produces a certain sequence of samples as usual.

However, there exists a sinusoidal signal of frequency 2 Hz which would give rise to the exact same sequence of samples and it is this lower frequency sinusoid which will be reconstructed by the interpolation algorithm. The general rule can be expressed as follows.

"Whenever a sinusoidal signal of a frequency f above half the sample rate but below the sample rate $f_s > f > \frac{f_s}{2}$ is sampled, its alias will appear at a frequency of $f' = f_s - f$ or $f' = \frac{f_s}{2} - \left(f - \frac{f_s}{2}\right)$ where the second form makes it clear that there will be reflection of the excess of the signal frequency over the Nyquist frequency at the Nyquist frequency to obtain its alias'.

7.12.5 Anti-Aliasing Filter

If the sampling hardware is not fast enough or if the bandwidth of the signal is not known, the anti-aliasing filter is to be used. Figure 7.58 shows the use of an anti-aliasing filter.

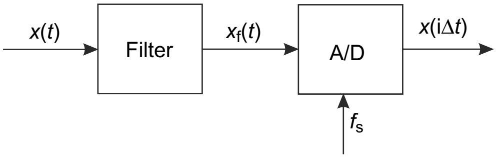


Fig. 7.58 Use of anti-aliasing filter

The anti-aliasing filter is a lowpass filter and the main goal is to eliminate, before sampling, all frequencies in the signal that are, at least, above the Nyquist frequency and therefore avoid aliasing. Note that filtering the original signal causes losing data from the original signal, but it ensures good reconstruction of the filtered signal.

7.12.6 Quantisation

After the sampling process, the resultant waveform is a sequence of numbers which can take any value on a continuous range of values. Because this range is continuous, there are infinitely many possible values for each number. In order to represent each number from such a continuous range, it is required to have an infinite number of digits. Instead, it is necessary to represent the numbers with a finite number of digits. That means after discretising the time variable, the amplitude variable is also to be discretised. This discretisation of the amplitude values is called quantisation.

Assume that our sequence takes values in the range between –1... + 1. Now it is assumed that each number must be represented from this range with just two decimal digits: one before and one after the point. The possible amplitude values are –1.0, –0.9. . . ,–0.1, 0.0, 0.1. . . , 0.9 and 1.0. These are exactly 21 distinct levels for the amplitude and this number of quantisation levels is denoted as N_Q. Each level is a step of 0.1 higher than its predecessor and we will denote this quantisation step size. Now we assign to each number from our continuous range that quantisation level which is closest to our actual amplitude: the range –0.05... + 0.05 maps to quantisation level 0.0, the range 0.05...0.15 maps to 0.1 and so on. That mapping can be viewed as a piecewise constant function acting on our continuous amplitude variable *x*. This is depicted in Figure 7.59.

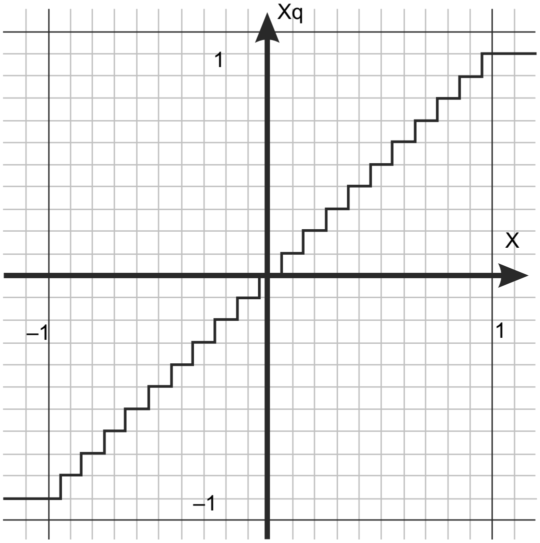


Fig. 7.59 Quantiser characteristics

7.12.7 Quantisation Noise

The difference between the actual analog value and quantised digital value is called quantisation error. This error is either due to rounding or truncation. The error signal is sometimes considered as an additional random signal called quantisation noise. When forcing an arbitrary signal value *x* to its closest quantisation level x_q, this x_q value can be seen as *x* plus some error. This error is denoted as e_q and so:

$$x_q = x + e_q \Leftrightarrow e_q = x_q = x \tag{7.50}$$

The quantisation error is restricted to the range $-\frac{q}{2}....+\frac{q}{2}$. This quantisation error should not be larger than half the quantisation step size.

Summary

Multiplexing is defined as a process in which multiple analog message signals or digital data streams are combined into one signal over a shared medium. The multiplexed signal is transmitted over a communication channel, which may be a physical transmission medium. The multiplexing divides the capacity of the low-level communication channel into several higher-level logical channels, one for each message signal to be transferred. A reverse process called demultiplexing can extract the original channels on the receiver side.

A device that performs the multiplexing at the transmission side is called a multiplexer (MUX), and a device that performs the reverse process at the receiver side is called a demultiplexer (DEMUX).

There are two major types of multiplexing.

- Time-Division Multiplexing (TDM)
- Frequency-Division Multiplexing (FDM)

Time-division multiplexing is a type of digital multiplexing in which two or more signals are transferred apparently simultaneously as subchannels in one communication channel, but are physically taking turns on the channel. Slow-speed TDM and high-speed TDM are the two types of time-division multiplexing.

Frequency-division multiplexing is a form of signal multiplexing where multiple baseband signals are modulated on different frequency channels and added together to create a composite signal destined for different users.

Digital modulation schemes have greater capacity to convey large amounts of information than analog modulation schemes. In digital modulation, an analog carrier signal is modulated by a digital bit stream. Digital modulation is widely employed in several applications.

There are four types of digital-modulation techniques:

- *Amplitude Shift Keying (ASK)*

  In the case of amplitude shift keying, a finite number of amplitudes are used. If the amplitude of the carrier is varied proportional to the modulating signal, ASK is used.
- *Quadrature Amplitude Modulation (QAM)*

  In the case of quadrature amplitude modulation, a finite number of at least two phases and at least two amplitudes are used. In QAM, an in-phase signal and a quadrature phase signal are amplitude-modulated with a finite number of amplitudes, and summed.
- *Frequency-Shift Keying (FSK)*

  In the case of frequency-shift keying, a finite number of frequencies are used. If the frequency of the carrier is varied proportional to the modulating signal, FSK is used.
- *Phase-Shift Keying (PSK)*

  In the case of phase-shift keying, a finite number of phases are used. If the phase of the carrier is varied proportional to the modulating signal, PSK is used.

A constellation diagram is a representation of a signal modulated by a digital modulation scheme such as quadrature amplitude modulation or phase-shift keying. It displays the signal as a two-dimensional scatter diagram in the complex plane at symbol sampling instants. It represents the possible symbols that may be selected by a given modulation scheme as points in the complex plane.

The term modem is derived from the words, modulator and DEModulator. A modem contains a modulator as well as a demodulator. The digital modulation/demodulation schemes are implemented in the modems. A modem comprises a transmitter, a receiver and two interfaces.

Modems can be of several types and they can be classified in a number of ways. The classification is usually based on the following basic modem features.

- *Directional capability*
  Half-duplex modem
  Full-duplex modem
- *Connection to the line*
  2-wire modem
  4-wire modem
- *Transmission mode*
  Asynchronous modem
  Synchronous modem

REVIEW QUESTIONS

PART-A

1. Define multiplexing.
2. What are the types of multiplexing? Compare them.
3. What do you mean by TDM? State its significance.
4. What is FDM? Draw its spectrum.
5. What is digital modulation?
6. State the applications of digital modulation.
7. What are the types of digital modulation?
8. What is meant by *I/Q* format?
9. State the purpose of constellation diagram.
10. Define phase-shift keying.
11. What is DPSK? State its advantage.

12. Draw the constellation diagram of BPSK.
13. What is QPSK? Draw its constellation diagram.
14. Define frequency-shift keying. What are its types?
15. What is amplitude shift keying?
16. What do you mean by MODEM? Mention its significance in digital communication.
17. List the important features of modems.
18. How will you classify modems?
19. Differentiate between half-duplex and full-duplex modems.
20. Differentiate between 2-wire and 4-wire modems.
21. Differentiate between synchronous and asynchronous modems.
22. What is quantisation?
23. Define sampling theorem.
24. What is Nyquist criterion?
25. Define quantisation error.
26. What is aliasing? How will you overcome it?
27. Draw the characteristics of a quantizer.

PART-B

1. What is multiplexing? Explain the principle of operation of its two types in detail.
2. What is digital modulation? Give the procedural steps to be followed by digital modulation.
3. What are the types of digital modulation? Compare their characteristics in detail with respect to their constellation diagram and frequency spectrum.
4. Define phase-shift keying. Explain its various types in brief.
5. What is frequency-shift keying? What are its types? Explain them in brief.
6. Draw the block diagram of MODEM and explain its building blocks.
7. How will you classify modems? Explain the types in detail with neat sketches.
8. Explain the quantisation process with neat sketches.
9. How will you reconstruct the original waveform from the samples waveform? Explain it.
10. Draw the characteristics of a quantiser and explain the process of calculation of quantisation error.

8

INFORMATION THEORY AND CODING

Objectives

- To know the purpose of information theory in a communication system
- To discuss about entropy and conditional entropy
- To provide details about the calculation of information rate
- To discuss channel capacity and its examples
- To provide details about the significance of source coding and its types
- To discuss the procedures of Shannon–Fano coding and Huffman coding in detail
- To know about the causes of intersymbol interference and interpretation of eye pattern
- To provide details about various error-control and error-detection methods

8.1 INTRODUCTION

The purpose of a communication system is to carry information-bearing baseband signals from one place to another place over a communication channel. Information theory deals with mathematical modelling and analysis of a communication system rather than with physical sources and physical channels. Generally, coding offers the most significant application of information theory. Its main purpose is to improve the efficiency of the communication system. Popular coding methods include Shannon–Fano, Huffman coding, etc.

8.2 INFORMATION THEORY

Information theory deals with mathematical modelling and analysis of a communication system rather than with physical sources and physical channels. An information source is a mathematical model for a physical entity that produces a succession of symbols called 'outputs' in a random manner. The symbols produced may be real numbers such as voltage measurements from a transducer, binary numbers as in computer data, two-dimensional intensity fields as in a sequence of images, continuous or discontinuous waveforms, and so on. The space containing all of the possible output symbols is called the alphabet of the source and a source is essentially an assignment of a probability measure to events consisting of sets of sequences of symbols from the alphabet.

8.2.1 Measure of Information

Suppose a discrete information source emits n possible messages such as $m_1,m_2,...m_n$ with probability of occurrences $p_1,p_2,...p_n$ where $p_1 + p_2 +...+ p_n= 1$. The symbols emitted by the source during successive signalling intervals are assumed as statistically independent. A source having the above properties is known as a **discrete memoryless source**.

The information content or the amount of information in the k^{th} message, denoted by I_{k} must be inversely related to p_{k}. The following requirements are to be satisfied by $I(m_{\text{k}})$.

1. $I(m_{\text{k}})$ must approach 0 as p_{k} approaches infinity.
2. The information content $I(m_{\text{k}})$ must be a non-negative quantity since each message contains some information. Rarely, $I(m_{\text{k}})$ can be zero.
3. The information content of a message having higher probability of occurrence is less than the information content of a message having lower probability.

The above three characteristics of the information content of messages are summarised as follows.

$$\begin{aligned} &I\,(m_{\text{k}}) > I(m_j) && \text{for} \quad p_k < p_j \\ &I\,(m_{\text{k}}) \rightarrow\ 0 && \text{for} \quad p_k \rightarrow 1 \\ &I\,(m_{\text{k}}) \geq\ 0 && \text{for} \quad 0 \leq p_k \leq 1 \end{aligned} \tag{8.1}$$

If two individual messages are received, the information content in the combine message is same as the sum of the information contained in the two messages. Mathematically,

$$I = I(m_{\text{k}}\ .\ m_{\text{j}})$$

$$(m_{\text{k}}) \text{ and } (m_{\text{j}}) = I(m_{\text{k}}) + I(m_{\text{j}}) \tag{8.2}$$

Therefore, the mathematical representation of the information content of a message must satisfy the conditions given in Equation (8.1) and Equation (8.2). If a continuous function of p_{k} satisfies the above requirements then it is a logarithmic function. For this logarithmic function, the measure of information is expressed as

$$I\,(m_k) = \log_b\left(\frac{1}{p_k}\right) \tag{8.3}$$

The quantity $I(m_k)$ is known as the self-information of message m_k and this form of information measure is due to Shannon.

Taking 2 as the logarithmic base, the measure of information can be described as

$$I\,(m_k) = \log_2\left(\frac{1}{p_k}\right) \text{ bits} \tag{8.4}$$

The importance of binary digits lies in the fact that any two things can be represented by the two binary digits. If the required logarithmic base is other than 2, it can also be converted to natural logarithm using the following formula.

$$\log_2 \nu = \frac{\ln \nu}{\ln 2} \tag{8.5}$$

or

$$\log_2 \nu = \frac{\log_{10} \nu}{\log_{10} 2} \tag{8.6}$$

EXAMPLE 8.1

An information source produces one of three possible messages during each interval having probabilities $p_1 = \frac{1}{2}, p_2 = \frac{1}{4}, p_3 = \frac{1}{8}$. *Determine the information content of each of these messages.*

Solution

The measure of information can be expressed as

$$I\,(m_k) = \log_2\left(\frac{1}{p_k}\right) \text{bits}$$

$$I\,(m_1) = \log_2\left(\frac{1}{2}\right)$$

$$= \log_2 2 = 1 \text{ bit}$$

$$I\,(m_2) = \log_2\left(\frac{1}{4}\right)$$

$$= \log_2 2^2 = 2 \text{ bits}$$

$$I\,(m_3) = \log_2\left(\frac{1}{8}\right)$$

$$= \log_2 2^3 = 3 \text{ bits}$$

EXAMPLE 8.2

Calculate the amount of information if it is given that $p_1 = \frac{1}{8}$.

Solution

The measure of information can be expressed as

$$I\,(m_k) = \log_2 \left(\frac{1}{p_k}\right) \text{bits}$$

$$= \frac{\log_{10}\left(\frac{1}{p_k}\right)}{\log_{10} 2}$$

It is given that $p_1 = \frac{1}{8}$.

$$\therefore \qquad I\,(m_k) = \frac{\log_{10} 8}{\log_{10} 2} = 4 \text{ bits}$$

EXAMPLE 8.3

In a binary PCM system, if a binary '0' occurs with probability $p_1 = \frac{1}{4}$ *and a binary '1' occurs with probability* $p_2 = \frac{3}{4}$ *then calculate the amount of information covered by each bit.*

Solution

It is given that $p_1 = \frac{1}{4}$ and $p_2 = \frac{3}{4}$

The amount of information is given as

$$I\,(m_k) = \log_2 \left(\frac{1}{p_k}\right) \text{bits}$$

With $\qquad p_1 = \frac{1}{4},\ I\,(m_1) = \log_2 4$

$$p_1 = \frac{\log_{10} 4}{\log_{10} 2} = 2 \text{ bits}$$

And with

$$p_2 = \frac{3}{4}, \quad I\,(m_2) = \log_2 \left(\frac{4}{3}\right)$$

$$p_2 = \frac{\log_{10}\left(\frac{4}{3}\right)}{\log_{10} 2} = 0.415 \text{ bits}$$

8.3 ENTROPY

Messages produced by an information source consist of sequences of symbols which correspond to the message. In a communication system, the process of sending the message has to deal with each alphabet and symbol. It is necessary to define the information content of the symbols. In order to get the information content of the symbol, it is essential to consider the flow of information in the system due to randomness involved into the selection of the symbols.

The following assumptions are to be made for qualitative representation of average information per symbol.

1. The source is stationary so that the probabilities may remain constant with time.
2. The successive symbols are statistically independent and form the source at an average rate of r symbols per second.

Let $m_1, m_2,...m_n$ are M messages with probability of occurrences p_1, p_2,...p_n. Suppose that a sequence of L messages is transmitted. If L is very large then it is said that

p_1L messages of $m_1$are transmitted,

p_2L messages of m_2 are transmitted,

p_3L messages of m_3 are transmitted,

...

$p_\text{m}L$ messages of m_M are transmitted,

Thus, the information due to message $m_1$will be,

$$I_1 = \log_2\left(\frac{1}{p_1}\right) \tag{8.7}$$

Since there are p_1L messages of m_1, the total information due to all messages of m_1 will be

$$I_{1(\text{total})} = p_1\, L \log_2\left(\frac{1}{p_1}\right) \tag{8.8}$$

Similarly, the total information due to all messages of m_2 will be

$$I_{2(\text{total})} = p_2\, L \log_2\left(\frac{1}{p_2}\right), \text{ and so on.} \tag{8.9}$$

Hence, the total information carried out due to the sequence of L messages will be

$$I_\text{total} = I_{1(\text{total})} + I_{2(\text{total})} + ... + I_{M(\text{total})} \tag{8.10}$$

$$I_\text{total} = p_1\, L \log_2\left(\frac{1}{p_1}\right) + p_2\, L \log_2\left(\frac{1}{p_2}\right) + ... + p_M\, L \log_2\left(\frac{1}{p_M}\right) \tag{8.11}$$

The average information per message will be

$$\text{Average information} = \frac{\text{Total information}}{\text{Number of messages}} \tag{8.12}$$

$$= \frac{I_{\text{total}}}{L} \tag{8.13}$$

Average information is also known as **entropy**. It is represented as *H*.

$$\text{Hence, Entropy } (H) = \frac{I_{\text{total}}}{L} \tag{8.14}$$

Equation (8.14) can be written as

$$\text{Entropy } (H) = p_1\, L \log_2\left(\frac{1}{p_1}\right) + p_2\, L \log_2\left(\frac{1}{p_2}\right) + \ldots + p_M\, L \log_2\left(\frac{1}{p_M}\right)$$

$$= \sum_{k=1}^{M} p_k \log_2\left(\frac{1}{p_k}\right) \tag{8.15}$$

This is the required expression for entropy of an information source.

EXAMPLE 8.4

Find entropy if $p_k = 1$ and when $p_k = 0$.

Solution

Entropy can be expressed as

$$\text{Entropy } (H) = \sum_{k=1}^{M} p_k \log_2\left(\frac{1}{p_k}\right)$$

With $p_k = 1$,

$$H = \sum_{k=1}^{M} 1 \times \log_2\left(\frac{1}{1}\right)$$

$$H = \sum_{k=1}^{M} \frac{\log_{10}(1)}{\log_{10}(2)}$$

$$H = 0 \; (\because \log_{10}(1) = 0)$$

With $p_k = 0$,

If p_k is tending to 0,

$$H = \sum_{k=1}^{M} \lim_{p_k \to 0} p_k \log_2\left(\frac{1}{p_k}\right)$$

When $p_k \to 0$, $H = 0$

Thus, entropy is zero for both the messages.

8.4 CONDITIONAL ENTROPY

Let there be two symbols *x* and *y* with associated probabilities of occurrence as *P*(*k*) and *P*(*j*) respectively. Then the entropy is denoted as *H*(*y*/*x*) which is known as conditional entropy. It is expressed as follows.

$$H(y/x) = -\sum_{k=1}^{m}\sum_{j=1}^{n} p(y=j, x=k)\log_2\, p(y-j \mid x=k) \tag{8.16}$$

Or

$$H(y/x) = -\sum_{k=1}^{m}\sum_{j=1}^{n} p(k, j)\log_2\, p(j \mid k) \tag{8.17}$$

Further, the joint probability can be related to the conditional probability such as $p(k, j)\cdot p(j \mid k)$.

8.5 INFORMATION RATE

The information rate can be represented by *R* and it is expressed as

$$\text{Information rate} = R = rH \tag{8.18}$$

where

R is information rate,

H is entropy or average information, and

r is the rate at which messages are generated.

Information rate *R* is represented in average number of bits of information per second. It is calculated as

$$R = \left(r \text{ in} \frac{\text{Messages}}{\text{Second}}\right) \times \left(H \text{ in} \frac{\text{Information bits}}{\text{Message}}\right) \tag{8.19a}$$

$$R = \text{Information bits per second} \tag{8.19b}$$

EXAMPLE 8.5

A source produces one of four possible messages during each interval having probabilities $p_1 = \frac{1}{2}, p_2 = \frac{1}{4}, p_3 = \frac{1}{8}$ and $p_4 = \frac{1}{16}$. *Determine the information rate if all the messages are equally likely.*

Solution

Since there are three messages and they are equally likely,

$$p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$$

The average information per message or entropy is given as

$$H = p_1 \log_2\left(\frac{1}{p_1}\right) + p_2 \log_2\left(\frac{1}{p_2}\right) + p_3 \log_2\left(\frac{1}{p_3}\right) + p_4 \log_2\left(\frac{1}{p_4}\right)$$

Since $p_1 = p_2 = p_3 = p_4$

$$H = 4p \log_2\left(\frac{1}{p}\right)$$

Again since $p_1 = p_2 = p_3 = p_4 = \dfrac{1}{4}$

$$H = \log_2 4$$

$$H = 2 \text{ bits/message}$$

The information rate is expressed as

$$R = rH$$

Here, $r = 2f_m$ messages/second

$\therefore$ $$R = 2f_m \text{ messages/second} \times 2 \text{ bits/message}$$

$$= 4\, f_m \text{ bits/second}$$

EXAMPLE 8.6

There are five possible sources with their probabilities as $p_1 = \dfrac{1}{2}$, $p_2 = \dfrac{1}{4}$, $p_3 = \dfrac{1}{8}$, $p_4 = \dfrac{1}{16}$ *and* $p_5 = \dfrac{1}{16}$. *Determine the entropy and information rate if there are 16 outcomes per second.*

Solution

$$\text{Entropy } (H) = \sum_{k=1}^{5} p_k \log_2\left(\frac{1}{p_k}\right) \text{bits/symbol}$$

$$H = p_1 \log_2\left(\frac{1}{p_1}\right) + p_2 \log_2\left(\frac{1}{p_2}\right) + p_3 \log_2\left(\frac{1}{p_3}\right) + p_4 \log_2\left(\frac{1}{p_4}\right) + p_5 \log_2\left(\frac{1}{p_5}\right)$$

$$= \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{8}\log_2 8 + \frac{1}{16}\log_2 16 + \frac{1}{16}\log_2 16$$

$$= \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + \frac{1}{16}\times 4$$

$$= \frac{1}{2} + \frac{1}{2} + \frac{3}{8} + \frac{1}{4} + \frac{1}{4} = \frac{4+4+3+2+2}{8}$$

$$= \frac{15}{8} = 1.875 \text{ bits/outcome}$$

The information rate

$$R = rH$$

$$= 16 \times \frac{15}{8}$$

$$R = 30 \text{ bits/second}$$

8.6 CHANNEL CAPACITY

Consider a mathematical analog of a physical signalling system shown in Figure 8.1.

Here, source symbols from some finite alphabet are mapped into some sequence of channel symbols, which then produces the output sequence of the channel. The output sequence is random but has a distribution that depends on the input sequence. From the output sequence, the transmitted message is to be recovered. Each of the possible input sequences induces a probability distribution on the output sequences, and the input sequences at the output can be transmitted and source message can be reconstructed at the output side. The maximum rate at which this can be done is called the **capacity of the channel**.

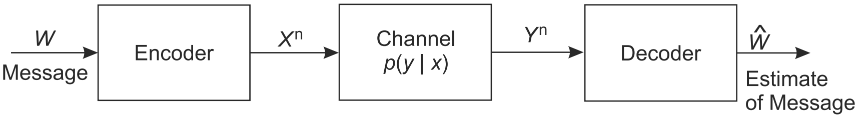


Fig. 8.1 A communication system

Figure 8.2 shows a communication channel with input message (X) and output information (Y), where X is the input from the source S_x, and Y is the output from the source S_Y.

Here, the $I(X : Y)$ can represent the amount of information transmitted and then the channel capacity can be defined as follows.

$$C = \max_{p(x)} I(X:Y) \tag{8.20}$$

$$= \max_{p(x)} \sum p(x).\, p(y|x).\log \frac{p(x)p(y|x)}{p(x)\left(\sum_x p(x)p(y|x)\right)} \tag{8.21}$$

X
Communication
Channel
Y

Fig. 8.2 A communication channel

Stated that $p(x) \geq 0,\ \sum_x p(x) = 1$

8.6.1 Properties of Channel Capacity

There are several important properties to be considered in a communication system. They are as follows.

1. $C \geq 0$, because $I(X : Y) \geq 0$.
2. $C \leq \log |x|$ because $C = I(X : Y) \leq \max H(X) = \log |x|$
3. $C \leq \log |y|$ for the same reason.
4. $I(X : Y)$ is a continuous function of $p(x)$.
5. $I(X : Y)$ is a concave function of $p(x)$.

The channel capacity of a communication channel also depends on the bandwidth of the channel (BW) and (S/N) ratio. It can be expressed as

$$C = B\log_2\left(1+\frac{S}{N}\right) \text{ bits/s} \tag{8.22}$$

$$\frac{S}{N} = \frac{P}{N_0 B} \tag{8.23}$$

where P is the average transmitted power, and

$N_0 B$ is the total noise power.

The channel capacity can be increased for a noisy channel by expanding the bandwidth than increasing the transmitted power in the channel. There are three points to be noted. They are as follows.

1. It is to be noted that the channel capacity is infinity for a noiseless channel.

   If there is no noise, $N = 0,\ \therefore \frac{S}{N} = \infty$

   $$C = B \log_2 (1 + \infty) = \infty$$

2. Infinite bandwidth channel has limited capacity.

   $$\because \qquad N = N_0 B$$

   As bandwidth increases, noise can also be increased and $\frac{S}{N}$ will be decreased.

3. With bandwidth approaching infinity, channel capacity approaches an upper limit.

$$C_\infty = \operatorname*{Lim}_{B\to\infty} C = 1.44\frac{S}{N_0}$$

8.7 EXAMPLES OF CHANNEL CAPACITY

There are some important examples considered for the channel capacity of a communication channel. Let them be discussed as follows.

8.7.1 Noiseless Binary Channel (NBC)

Suppose it is necessary to have a channel with its binary input reproduced exactly at the output. This channel is illustrated in Figure 8.3.

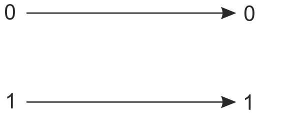


Fig. 8.3 Noiseless binary channel

In this case, any transmitted bit is received without error. Hence, 1 error-free bit can be transmitted per use of the channel, and the capacity is 1 bit. The information capacity can be calculated as

$$C = \max_{p(x)} I(X:Y) = 1 \text{ bit}$$

It can be achieved by using $p(x) = \left(\frac{1}{2}, \frac{1}{2}\right)$.

8.7.2 Noisy Channel with Non-overlapping Outputs (NCNO)

Consider a noisy channel has two possible outputs corresponding to each of the two inputs, and it is illustrated in Figure 8.4.

The channel shown in the above figure appears to be noisy, but really is not noisy. Even though the output of the channel is a random consequence of the input, the input can be determined from the output, and hence every transmitted bit can be recovered without error. The capacity of this channel is also 1 bit per transmission. The information capacity can be calculated as

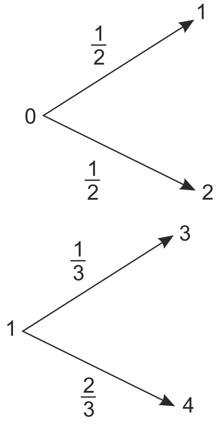


Fig. 8.4 Noisy channel with non-overlapping outputs

$$C = \max_{p(x)} I(X : Y) = 1 \text{ bit}$$

It can be achieved by using $p(x) = \left(\frac{1}{2}, \frac{1}{2}\right)$.

8.7.3 Binary Symmetric Channel (BSC)

Consider the Binary Symmetric Channel (BSC), which is shown in Figure 8.5.

This is a binary channel in which the input symbols are complemented with probability p and it is the simplest model of a channel with errors. When an error occurs, a binary '0' is received as a binary '1' and vice versa. The received bits do not reveal where the errors have occurred which shows that all the received bits are unreliable.

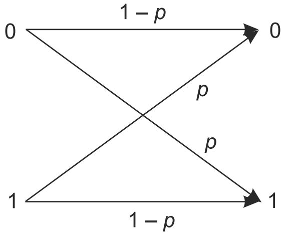


Fig. 8.5 Binary symmetric channel

Let the mutual information be bounded by

$$\begin{aligned} I(X : Y) &= H(Y) - H(Y \mid X = x) \\ &= H(Y) - \sum p(x)\, H(Y \mid X = x) \\ &= H(Y) - \sum p(x)\, H(p) \\ &= H(Y) - H(p) \\ &\leq 1 - H(p) \qquad (8.24) \end{aligned}$$

Hence, the information capacity of a binary symmetric channel with parameter p is

$$C = 1 - H(p) \text{ bits} \qquad (8.25)$$

8.7.4 Binary Erasure Channel (BEC)

A Binary Erasure Channel (BEC) is simply an analog of the binary symmetric channel in which some bits are lost rather than corrupted. In the binary erasure channel, a fraction α of the bits are erased. The receiver knows which bits have been erased. The binary erasure channel has two inputs and three outputs as shown in Figure 8.6.

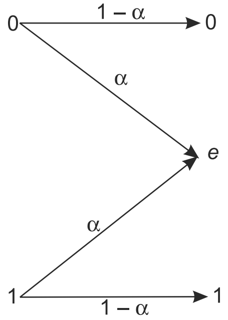


Fig. 8.6 Binary erasure channel

The calculation of the capacity of the binary erasure channel is given as follows.

$$\begin{aligned} C &= \max_{p(x)} I(X : Y) \\ &= \max_{p(x)} (H(Y) - H(Y \mid X = x)) \end{aligned}$$

$$= \max_{p(x)} H(Y) - H(\alpha) \tag{8.26}$$

Letting E be the event $\{Y = e\}$, and then using the expansion

$$H(Y) = H(Y, E) = H(E) + H(Y \mid E), \tag{8.27}$$

and letting $P(X = 1) = \pi$,

$$H(Y) = H((1 - \pi)\ 1 - \alpha, \alpha, \pi\ (1 - \alpha)) = H(\alpha) + (1 - \alpha)\ H(\pi). \tag{8.28}$$

Hence,

$$\begin{aligned} C &= \max_{p(x)} H(Y) - H(\alpha) \\ &= \max_{\pi}\ (1 - \alpha)\ H(\pi) + (H(\alpha) - H(\alpha) \\ &= \max_{\pi}\ (1 - \alpha)\ H(\pi) \\ &= 1 - \alpha \end{aligned} \tag{8.29}$$

where capacity is achieved by $\pi = \dfrac{1}{2}$

The above expression for the capacity has some meaning. Since a proportion α of the bits are lost in the channel, we can recover (at most) a proportion $1 - \alpha$ of the bits. Hence, the capacity is almost $1 - \alpha$. It is not immediately obvious that it is possible to achieve this rate.

In many practical channels, the sender receives some feedback from the receiver. If feedback is available for the binary erasure channel, it is very clear to do the next step. If a bit is lost, it will be retransmitted until it gets through. Since the bits get through with probability $1 - \alpha$, the effective rate of transmission is $1- \alpha$. In this way, it is easy to achieve a capacity of $1 - \alpha$ with feedback.

8.8 SOURCE CODING

Generally, coding offers the most significant application of the information theory. Its main purpose is to improve the efficiency of the communication system. It is a procedure for mapping a given set of messages $\{m_1,m_2,...m_\text{n}\}$into a new set of encoded messages $\{C_1,C_2,... C_\text{n}\}$ in such a manner that the transformation is in the form of one to one. This means that for each message, there is only one encoded message. This is called source coding. In source coding, the symbols produced by the information source are given to the source encoder. These symbols cannot be transmitted directly. These are first converted into digital form by the source encoder. Every binary '1' and '0' is known as a bit. The group of bits is called a **code word**. The source encoder assigns code words to the symbols. Thus, for each distinct symbol, there is a unique code word. It can be of 4, 8, 16 or 32 bits length. Also, as the number of bits is increased in each code word, the symbols which can be represented are increased.

If the number of bits is 8, it will have $2^8 = 256$ distinct code words. So, 8 bits can be used to represent 256 symbols. Similarly, 16 bits can represent $2^{16} = 65536$ symbols, and so on. In the above examples, the bits in every code word are throughout the same. This means 8 bits in first case and 16 bits in the next case respectively. This type of coding is called **fixed-length coding**. If the bits are not constant throughout such as in the Morse code, it is called **varying length coding**.

Hence, an efficient source encoder should satisfy the following requirements.

1. The code words produced by the encoder are in binary form.
2. The source code is uniquely decodable so that the original source sequence can be reconstructed perfectly from the encoded binary sequence.

8.8.1 Source-Coding Efficiency

Consider that a discrete memoryless information source producing M equally likely symbols has a source information rate given as

$$\text{Information Rate} = R = rH \tag{8.30}$$

where

R is information rate,

H is entropy or average information, and

r is the rate at which messages are generated.

If all symbols convey the same amount of information then

$$H = r \log_2 M \tag{8.31}$$

In such cases, an efficient transmission may take the form of M-ary signalling with a signalling rate equal to the symbol rate r.

On the other hand, if the symbols have different probabilities then

$$R = rH < r \log_2 M \tag{8.32}$$

In this case, efficient transmission requires an encoding process which takes into account the variable amount of information per symbol.

A binary source encoder is shown in Figure 8.7. It converses the incoming symbols to code words consisting of binary digits produced at some fixed rate, say r_b.

At the output side, the encoder appears to be a binary source with entropy $\Omega(p)$ and information rate is given as

$$r_b\,\Omega(p) \le r_b \log_2 2$$

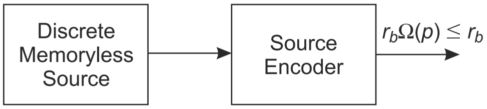


Fig. 8.7 Source encoding

$$r_b\Omega(p) \le r_b \cdot 1$$

$$r_b\Omega(p) \le r_b \tag{8.33}$$

So, coding does not produce any additional information and also does not destroy the original information. However, this is true when the code is uniquely separable.

$$\therefore \qquad R = rH = r_b\Omega(p) \le r_b \tag{8.34}$$

This means that

$$\frac{r_b}{r} \ge H \tag{8.35}$$

where $\frac{r_b}{r}$ is the average code length denoted as $\overline{N}$.

Hence, $$\overline{N} = \frac{r_b}{r} \tag{8.36}$$

where $\overline{N}$ can be seen as the average number of binary digits per source symbols used in the source-encoding process. Let $N_{\min}$ denote the minimum possible value of $\overline{N}$. Then the coding efficiency of the source encoder is expressed as follows.

$$\eta = \frac{N_{\min}}{N} \tag{8.37}$$

With $\overline{N} = N_{\min}$, $\eta << 1$. The source encoder is said to be efficient when η approaches unity.

Mathematically, $$\overline{N} = \sum_{k=1}^{M} p_k N_k \tag{8.38}$$

Here, N_k is the length of the code word for the k^{th} symbol, and

p_k is the probability.

The above expression is known as **Shannon's source-coding theorem**.

8.8.2 Source-Coding Theorem

According to the source-coding theorem, the minimum value of $\overline{N}$ is bounded as

$$H \le \overline{N} < H + \varepsilon \tag{8.39}$$

Here, ε is a positive integer that can be made as small as possible. For an optimum source coding,

$$\overline{N} = H \tag{8.40}$$

The following ratio can be used as the measure of the efficiency of the sub-optimum codes.

$$\frac{R}{r_b} = \frac{H}{\overline{N}} \le 1 \tag{8.41}$$

For necessary and sufficient condition for a binary code to be uniquely separable, the word length N_i must be such that

$$K = \sum_{i=1}^{M} 2^{-N_i} \le 1 \tag{8.42}$$

The above condition is known as **Kraft inequality**.

The simple encoding process involves generation of fixed-length code in which all code words have the same length given by

$$N_i = \overline{N} \tag{8.43}$$

In this case, the value of K is obtained as

$$K = M^{2-\overline{N}} \leq 1 \tag{8.44}$$

This means that separability in the case of fixed coding needs that

$$\overline{N} \leq \log_2 M \tag{8.45}$$

The resulting efficiency can be calculated as

$$\frac{H}{N} \leq \frac{H}{\log_2 M} \tag{8.46}$$

When $H < \log_2 M$, it is essential to reduce the average code length $\overline{N}$ to obtain higher efficiency. The average code length $\overline{N}$ may be reduced by using a variable-length coding.

8.9 SHANNON–FANO CODING

Shannon–Fano coding is a technique for constructing a prefix code based on a set of symbols and their probabilities. In this coding, the symbols are arranged in order from the most probable to the least probable and then divided into two sets whose total probabilities are as close as possible to being equal. All symbols then have the first digits of their codes assigned and symbols in the first set receive a binary '0' and symbols in the second set receive a binary '1'. The same process is repeated on those sets to determine successive digits of their codes for all the sets with more than one member remaining. When a set has been reduced to one symbol, it means the symbol's code is complete and further it will not form the prefix of any other symbol's code.

In this coding method, when the two smaller sets produced by a partitioning are in fact of equal probability, the one bit of information used to distinguish them is used most efficiently. Sometimes, Shannon–Fano does not produce optimal prefix codes. The set of probabilities $\{0.35, 0.17, 0.17, 0.16, 0.15\}$ is an example of one that will be assigned non-optimal codes by Shannon–Fano coding.

In order to develop a Shannon–Fano tree, the following algorithm is to be followed.

1. Develop a corresponding list of probabilities or frequency counts for a given list of symbols, so that each symbol's relative frequency of occurrence is known.
2. Arrange the lists of symbols according to frequency, with the most frequently occurring symbols at the left and the least common at the right.

3. Divide the list into two parts, with the total frequency counts of the left half being as close to the total of the right as possible.
4. The left half of the list is assigned the binary digit 0, and the right half is assigned the digit 1. This means that the codes for the symbols in the first half will all start with 0 and the codes in the second half will all start with 1.
5. Recursively, apply the steps 3 and 4 to each of the two halves, subdividing groups and adding bits to the codes until each symbol has become a corresponding code leaf on the tree.

The following example shows the construction of the Shannon code for a small alphabet. There are five symbols which can be coded having the following frequencies listed in the following table.

Symbol	*A*	*B*	*C*	*D*	*E*
Count	15	7	6	6	5
Probabilities	0.3846	0.1794	0.1538	0.1538	0.1282

The procedural steps of the Shannon–Fano coding are schematically shown in Figure 8.8. All the symbols (from *A* to *E*) are sorted by their frequency of occurrence, from left to right. It is shown in Figure 8.8(a). By putting the dividing line between symbols *B* and *C* results in a total of 22 in the left group and a total of 17 in the right group. This minimises the difference in totals between the two groups.

With this division, *A* and *B* will each have a code that starts with a binary '0' bit, and the *C*, *D*, and *E* codes will all start with a binary '1', as shown in Figure 8.8(b). Subsequently, the left half of the tree gets a new division between *A* and *B*, which puts *A* on a leaf with code 00 and *B* on a leaf with code 01.

After four division procedures, a tree of codes will result. In the final tree, the three symbols with the highest frequencies have all been assigned 2-bit codes, and two symbols with lower counts have 3-bit codes as shown in the table below.

Symbol	A	B	C	D	E
Code	00	01	10	110	111

Results in 2 bits for *A*, *B* and *C* and per 3 bits for *D* and *E*,

$$\therefore \quad \text{average bit number} = \frac{2\text{ bit}\cdot(15+7+6)+3\text{ bit}\cdot(6+5)}{39\text{ symbols}}$$

$$= 2.28 \text{ bits/symbol}$$

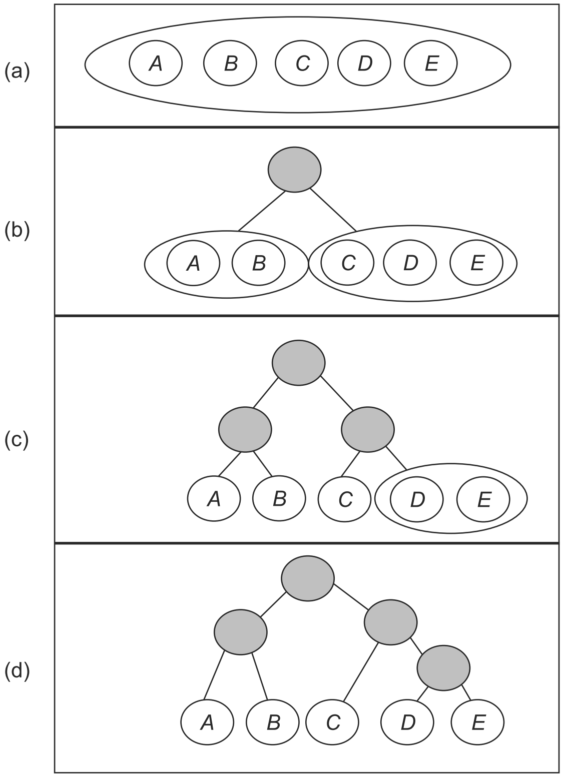


Fig. 8.8 Shannon–Fano coding

EXAMPLE 8.7

A discrete memoryless source has five symbols A, B, C, D and E with their probabilities P_A = 0.4, P_B = 0.19 P_C = 0.16 P_D = 0.15 and P_E = 0.1. Construct a Shannon–Fano code and also calculate the efficiency of the code.

Solution

Symbols	*A*	*B*	*C*	*D*	*E*
Probabilities	0.4	0.19	0.16	0.15	0.1

The development of Shannon–Fano tree is as follows.

After the Shannon–Fano tree, the final encoding values are given in the following table.

Symbols	Probabilities	Step-1	Step-2	Step-3	Code
A	0.4	0	0		00
B	0.19	0	1		01

C	0.16	1	0		10
D	0.15	1	1	0	110
E	0.1	1	1	1	111

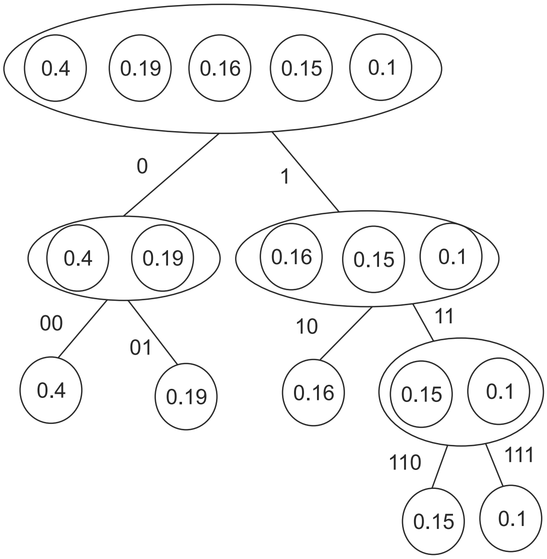


Fig. 8.9

$$H = \sum_{k=1}^{5} p_k \log_2 \left(\frac{1}{p_k}\right) \text{bits/symbol}$$

$$= 0.4 \log_2 0.4 + 0.19 \log_2 0.19 + 0.16 \log_2 0.16 + 0.15 \log_2 0.15 + 0.1 \log_2 0.1 = 2.15$$

$$\overline{N} = \sum_{k=1}^{5} p_k \; n_k$$

$$= 0.4(2) + 0.19(2) + 0.16(2) + 0.15(3) + 0.1(3) = 2.25.$$

$$\text{Efficiency } \eta = \frac{H}{\overline{N}}$$

$$= \frac{2.15}{2.25} = 95.6\%$$

EXAMPLE 8.8

Apply the Shannon–Fano coding procedure for the following message ensemble and also find the efficiency of the coding.

Symbols	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Probabilities	0.4	0.2	0.12	0.08	0.08	0.08	0.04

Solution

Symbols	Probabilities	Step-1	Step-2	Step-3	Step-3	Code	Length
A	0.4	0	0			00	2
B	0.2	0	1			01	2
C	0.12	1	0	0		100	3
D	0.08	1	0	1		101	3
E	0.08	1	1	0		110	3
F	0.08	1	1	1	0	1110	4
G	0.04	1	1	1	1	1111	4

$$H = \sum_{k=1}^{7} p_k \ \log_2 \left(\frac{1}{p_k}\right) \text{bits/symbol}$$

$$= 0.4 \log_2 0.4 + 0.2 \log_2 0.2 + 0.12 \log_2 0.12 + 3 \times (0.08 \log_2 0.08) + 0.041 \log_2 0.04$$

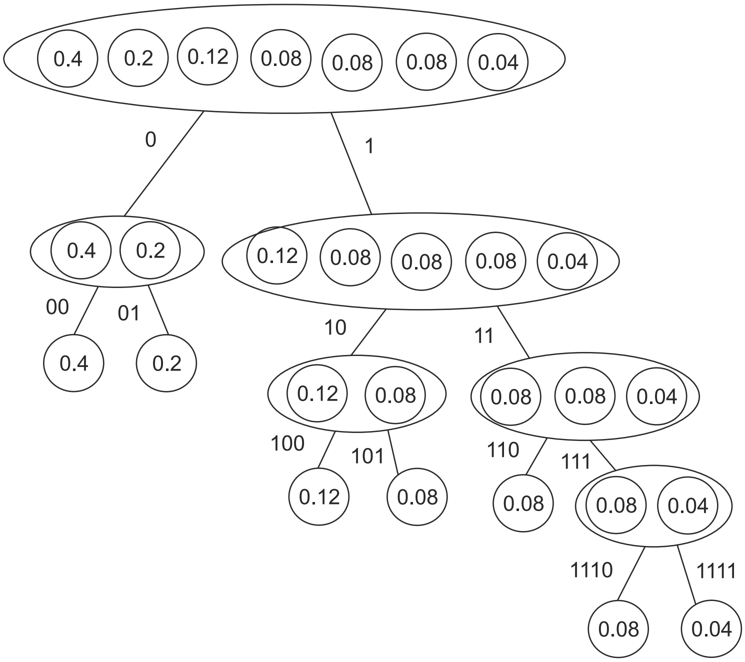


Fig. 8.10

$$= 2.42 \text{ bits/message}$$

$$\overline{N} = \sum_{k=1}^{7} p_k \; n_k$$

$$= 0.4\,(2) + 0.2\,(2) + 0.12\,(3) + 0.08\,(3) + 0.08\,(3) + 0.08\,(4) + 0.04\,(4)$$

$$= 2.52 \text{ letters/message}$$

$$\text{Efficiency } \eta = \frac{H}{\bar{N}}$$

$$= \frac{2.42}{2.52} = 96.03\%$$

8.10 HUFFMAN CODING

When the two smaller sets produced by a partitioning are of equal probability and the one bit of information used to distinguish them is used most efficiently, Shannon–Fano coding will produce fairly efficient variable-length encodings. But it does not always produce optimal prefix codes. For example, the set of probabilities {0.35, 0.17, 0.17, 0.16, 0.15} is one that will be assigned non-optimal codes by Shannon–Fano coding.

Due to this reason, Shannon–Fano is not always preferred. In that situation, Huffman coding is preferably used which is computationally simple and it produces prefix codes that always achieve the lowest expected code-word length under the constraints that each symbol is represented by a code formed of an integral number of bits. This is a constraint not often needed, since the codes will be packed end-to-end in long sequences. If groups of codes are to be considered at a time, Huffman coding is only optimal if the probabilities of the symbols are independent.

In comparison with the Shannon–Fano algorithm which does not always generate an optimal code, Huffman coding is a different algorithm that always produces an optimal tree for any given probabilities. While the Shannon–Fano tree is created from the root to the leaves, the Huffman algorithm works from leaves to the root in the opposite direction.

In order to develop a Huffman-coding tree, the following algorithm is to be followed.

1. Create a leaf node for each symbol and add it to the frequency of occurrence.
2. While there is more than one node in the queue:
    i. Remove the two nodes of lowest probability or frequency from the queue.
    ii. Prepend 0 and 1 respectively to any code already assigned to these nodes.
    iii. Create a new internal node with these two nodes as children and with probability equal to the sum of the two nodes' probabilities.
    iv. Add the new node to the queue.
3. The remaining node is the root node and the tree is complete.

To have a better comparison, the same example is taken as in Shannon–Fano algorithm. There are five symbols which can be coded having the following frequencies listed in the following table.

Symbol	*A*	*B*	*C*	*D*	*E*
Count	15	7	6	6	5
Probabilities	0.3846	0.1794	0.1538	0.1538	0.1282

The procedural steps of the Huffman coding are schematically shown in Figure 8.11.

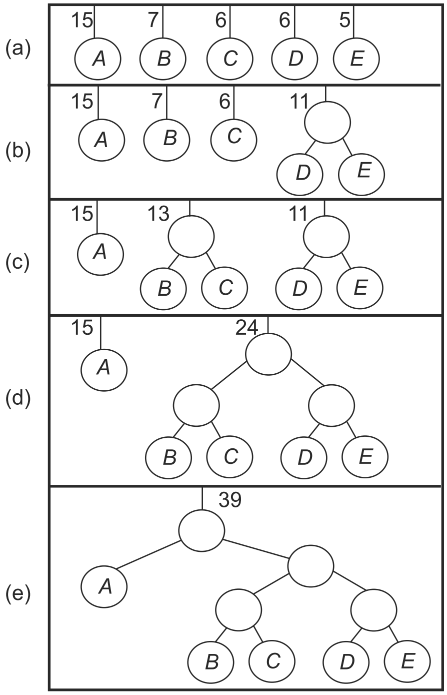


Fig. 8.11 Huffman coding

In this case, *D* and *E* have the lowest frequencies and so are allocated 0 and 1 respectively and grouped together with a combined probability of 0.2820. The lowest pair now is *B* and *C*, so they are allocated 0 and 1 and grouped together with a combined probability of 0.3332. This leaves *BC* and *DE* now with the lowest probabilities, so 0 and 1 are prepended to their codes and they are combined. This then leaves just *A* and *BCDE*, which have 0 and 1 prepended respectively and are then combined. This leaves us with a single node and our algorithm is complete.

The code lengths for the different characters this time are 1 bit for *A* and 3 bits for all other characters.

Symbol	*A*	*B*	*C*	*D*	*E*
Code	0	100	101	110	111

Results in 1 bit for *A* and per 3 bits for *B*, *C*, *D* and *E* an average bit number of

$$\therefore \quad \text{Average bit number} = \frac{1\text{ bit}\cdot 15 + 3\text{ bit}\cdot(7+6+6+5)}{39\text{ symbols}} = 2.23\text{ bits/symbols}$$

EXAMPLE 8.9

Apply Huffman-coding procedure for the data given in the following table.

Character	Frequencies of Occurrences (*n*)
e	3320
h	1458
l	1067
o	1749
p	547
t	2474
w	266
Total	10881

Solution

As per the procedural steps, the Huffman tree is developed which is given below and the resulting code is as follows.

Character	Binary Code
e	00
h	011
l	110
o	010
p	1110
t	10
w	1111

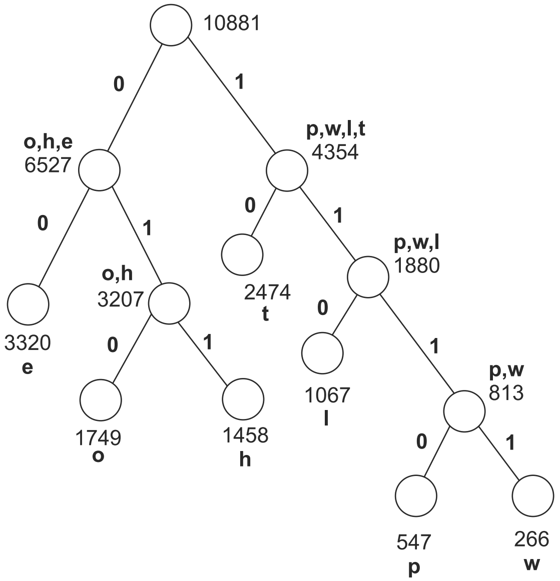


Fig. 8.12

8.11 INTERSYMBOL INTERFERENCE (ISI)

Intersymbol Interference (ISI) arises due to the dispersive nature of a communication channel in a digital baseband transmission. It is a form of distortion of a signal in which one symbol interferes with subsequent symbols. This is an unwanted phenomenon as the previous symbols have similar effect as noise, thus making the communication less reliable.

Intersymbol interference is usually caused by multipath propagation or the inherent nonlinear frequency response of a channel causing successive symbols to blur together. The system with the presence of ISI introduces errors in the decision device at the receiver output. Therefore, while designing the transmitting and receiving filters, it is essential to minimise the effects of intersymbol interference and thereby deliver the digital data to its destination with the smallest error rate possible.

ISI is mainly due to the dispersive nature of a communication channel in a digital baseband transmission. In order to transmit the baseband signal, discrete pulse amplitude modulation is used. Figure 8.13 shows the block diagram of the baseband binary PAM system. The input to the system is binary sequence $\{b_k\}$of symbols '1' and '0'. The duration of each symbol is T_b.

In Figure 8.13, the pulse-amplitude modulator converts this input sequence into polar form, i.e.

$$a_k = 1 \quad \text{if } b_k = 1 \tag{8.47}$$

$$a_k = -1 \quad \text{if } b_k = 0 \tag{8.48}$$

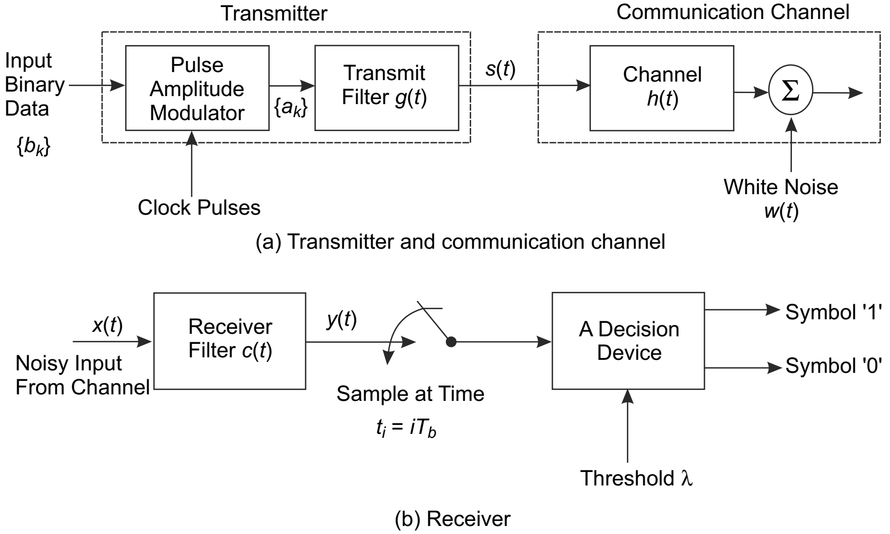


Fig. 8.13 A baseband binary data-transmission system

The polar sequence $\{a_k\}$ of short pulse is made to apply to a transmit filter having an impulse response $g(t)$. The output of the transmit filter is $s(t)$ and it is transmitted over a communication channel of impulse response $h(t)$. Then the channel adds a random noise to the signal being transmitted. Thus, the signal $x(t)$ at the end of the communication channel is a noisy signal. This signal is then given to the input of the receiver filter $y(t)$, sampled synchronously with the transmitter. This means that $y(t)$ is sampled at the instants at which the pulse was transmitted. The decision device then compares the sampled signal $y(t_i)$ with some threshold λ. Depending on the value of $y(t_i)$ with respect to threshold λ, the decision is taken.

The sequence $\{a_k\}$ at the transmitter is passed through three blocks such as transmit filter, channel and receive filter. At the output of the receive filter, the signal $y(t)$ is obtained. The combined impulse response of the above three blocks is obtained by double convolution as

$$\mu p(t) = g(t) \otimes h(t) \otimes c(t) \tag{8.49}$$

where μ is the scaling factor which represents amplitude changes during the signal transmission. $p(t)$ is the combined impulse response.

The output of the receive filter may be obtained by multiplication of input sequence and the combined impulse response of the three blocks obtained by Equation (8.49), i.e.

$$y(t) = \mu \sum_{k=-\infty}^{\infty} a_k\, p(t - kT_b) + n(t) \tag{8.50}$$

where $n(t)$ is the noise at the output of the receive filter generated due to additive noise and $p(t - kT_b)$ is the combined impulse response delayed by kT_b duration for k^{th} symbol in the sequence a_k.

When $y(t)$ is sampled at time $t_i = iT$, Equation (8.50) becomes,

$$y(t_i) = \mu \sum_{k=-\infty}^{\infty} a_k p(t_i - kT_b) + n(t_i) \tag{8.51}$$

Since $t_i = iT$,

$$y(t_i) = \mu \sum_{k=-\infty}^{\infty} a_k p(iT_b - kT_b) + n(t_i) \tag{8.52}$$

The sample time t_i is synchronised with the transmitter clock. This means that the instant at which pulse a_k is transmitted is same as the time at which $y(t)$ is sampled. There is some delay during transmission. By assuming zero delay, the pulse is received as soon as it is transmitted. Equation (8.52) is expressed as

$$y(t_i) = a_i\,(0) + \sum_{\substack{k=-\infty \\ k\neq i}}^{\infty} a_k p(i-k)T_b + n(t_i) \tag{8.53}$$

From the above equation, the first term $\mu a_i(0)$ represents the contribution of i^{th} bit and the second term represents the residual effect of all other bits transmitted before and after the sampling instant t_i. This means that in the output $y(t_i)$ for i^{th} bit a_i, the outputs due to other bits $\{a_k p(i-k)T_{\text{b}}\}$ is also present. Thus, the presence of the outputs due to other bits interferes with the output of the required bit $\mu\ a_i(0)$. This effect is called as InterSymbol Interference (ISI). In the absence of noise $n(t_i)$ and ISI, the output $y(t_i)$ from Equation (8.53) will be

$$y(t_i) = \mu\ a_i p(0) \tag{8.54}$$

If the impulse response $p(t)$ is normalized, then

$p(0) = 1$ and let $\mu = 1$, then the equation will be

$$y(t_i) = a_i \tag{8.55}$$

This reveals that the output would be completely error-free in the absence of noise and ISI. Generally, noise and ISI combinely introduce errors in the output. Hence, it is necessary to make efforts to minimise the ISI by increasing the signal to noise ratio.

8.11.1 Causes of Intersymbol Interference

There are two main causes of intersymbol interference. They are

1. Multipath propagation
2. Bandlimited channels

1. Multipath Propagation

Multipath propagation is one of the causes of intersymbol interference in which a wireless signal from a transmitter reaches the receiver via many different paths. The causes of this include reflection, refraction and atmospheric effects. All of these paths are different lengths which results in the different versions of the signal arriving at different times. This delay means that a part or all of a given symbol will be spread into the subsequent symbols, thereby interfering with the correct detection of those symbols. In addition, various paths of the signal often distort the amplitude and/or phase of the signal, thereby causing further interference with the received signal.

2. Bandlimited Channels

The transmission of a signal through a bandlimited channel is also a cause of intersymbol interference. Passing a signal through such a channel results in the removal of frequency components above this cut-off frequency and the amplitude of the frequency components below the cut-off frequency may also be attenuated by the channel.

This filtering of the transmitted signal affects the shape of the pulse that arrives at the receiver. The effects of filtering a rectangular pulse not only changes the shape of the pulse within the first symbol period, but it is also spread out over the subsequent symbol periods. When a message is transmitted through such a channel, the spread pulse of each individual symbol will interfere with the following symbols.

8.12 EYE PATTERN

When the sequence is transmitted over a baseband binary data-transmission system, the signal obtained at the output will be a continuous-time signal. Ideally this signal must go high and low depending on the symbol which was transmitted. However, because of the nature of the transmission channel, the signal becomes continuous with increasing and decreasing amplitudes. Figure 8.14 (a) shows the binary sequence which is transmitted and Figure 8.14 (b) shows that the signal $y(t)$ is obtained at the output with various sampling instants.

Hence, based on the signal obtained over the period between two sampling instants, a decision is taken by the decision device. If the signal $y(t)$ is cut in each interval and placed over one another, the eye pattern is shown in Figure 8.14 (c). The name *eye pattern* is given because it looks like an eye.

In telecommunication, an eye pattern can be obtained on a CRO display in which digital data signal from a receiver is repetitively sampled and applied to the vertical input, while the data rate is used to trigger the horizontal sweep. When there is a large number of bits of the sequence, the eye pattern is shown in Figure 8.14 (d).

There are several points related to the eye pattern. They are listed as follows.

1. The width of an eye opening describes the interval over which the received wave can be sampled without error from intersymbol interference. It is preferable to sample the instant at which the eye is open widest.
2. The sensitivity of the system to timing error is determined by the rate of closure of the eye as the sampling time is varied.
3. The height of the eye opening at the specified sampling time is called the margin over the noise.

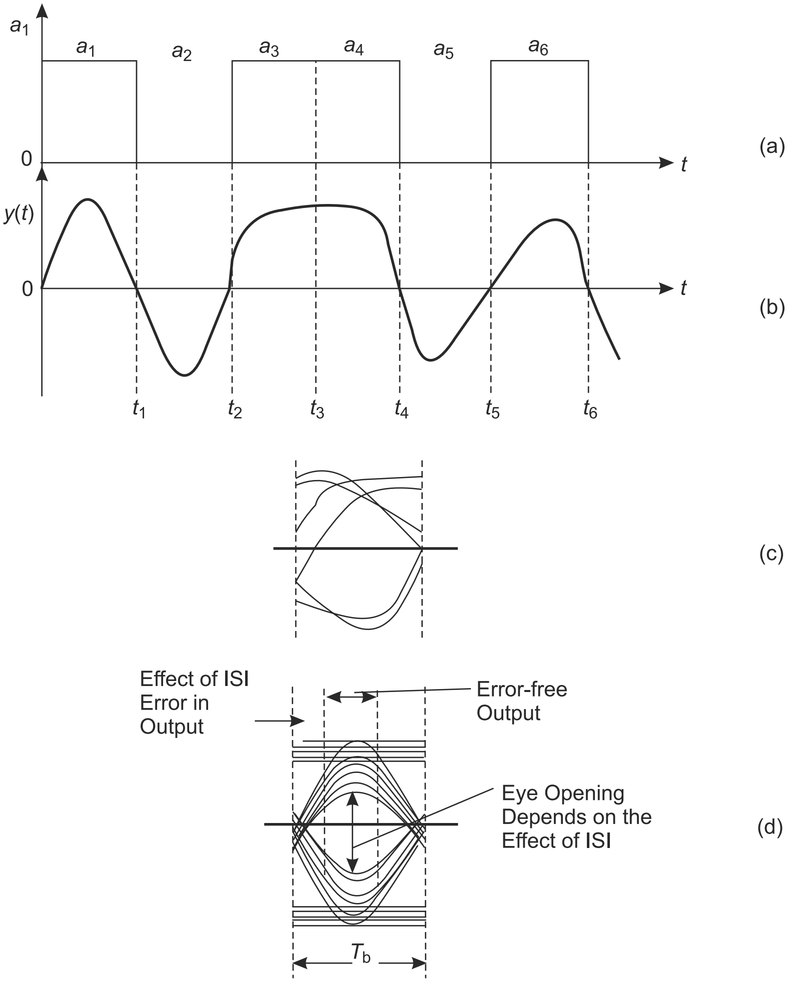


Fig. 8.14 Eye diagram (a) Transmission of binary sequence, (b) Received signal by baseband transmission, (c) Eye pattern of signal in (b) and (d) Eye pattern for large number of bits in waveform

As the effect of intersymbol interference increases, the eye opening reduces. If the eye is closed completely then it is not possible to avoid errors in the output.

8.12.1 Interpretation of Eye Pattern

The eye diagram is created by taking the time-domain signal and overlapping the traces for a certain number of symbols. If a signal is sampled at a rate of 10 samples per second and two symbols are to be taken then the signal is cut at every 20 samples and overlapping takes place. This overlapped signals show a lot of useful information and this is called the eye diagram. Figure 8.15 shows the interpretation of an eye pattern.

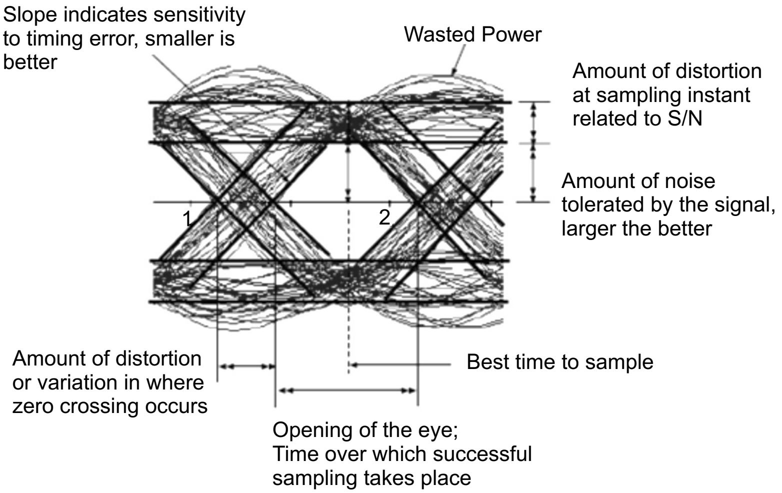

Slope indicates sensitivity to timing error, smaller is better
Wasted Power
Amount of distortion at sampling instant related to S/N
Amount of noise tolerated by the signal, larger the better
1
2
Amount of distortion or variation in where zero crossing occurs
Best time to sample
Opening of the eye; Time over which successful sampling takes place

Fig. 8.15 Interpretation of eye pattern

The open part of the signal represents the time at which the signal can be sampled successfully. If the opening of the eye is larger, the sampling will be better. A smaller opening will lead to larger errors if not sampled at the best sampling time which occurs at the centre of the eye.

The horizontal band represents the amount of signal variation at the time it is sampled. This variation is directly related to *S/N* of the signal. A small band means that there is a large *S/N*. The slope of the eye determines how sensitive the signal is to timing errors. A small slope allows the eye to be opened more and hence there is less sensitivity to timing errors. The width of the crossover represents the amount of jitter present in the signal. Small is better.

8.13 ERROR CONTROL

For the transmission of electrical signals as electrical signals, it suffers from many drawbacks which ultimately result in introduction of errors in the bit stream. Digital systems are very sensitive to errors and may function wrongly if the error rate is above a certain level. Therefore, there is a need of built-in error-control mechanisms in digital systems. Error control in data communication is based on the detection of errors in a message and its retransmission.

8.13.1 Transmission Errors

Errors are introduced in the data bits during their transmission. These errors can be classified as follows.

1. Content errors
2. Flow-integrity errors

Content errors are errors in the content of a message, e.g. a ‘1’ may be received as a ‘0’. Such errors creep in due to impairment of the electrical signal in the transmission medium. **Flow-integrity errors** refer to missing blocks of data. For example, a data block may be lost in the network due to its having been delivered to a wrong destination.

During voice transmission, the listener can tolerate a good deal of signal distortion and make sense of the received signal but digital systems are very sensitive to errors. Therefore, several measures are considered to counteract the effect of errors in a data-communication system. These measures include the following.

1. Introduction of additional check bits in the data bits to detect content errors
2. Correction of the errors
3. Establishment of procedures of data exchange which enable detection of missing blocks of data
4. Recovery of the corrupted messages

8.13.2 Coding for Error Detection and Correction

For error detection and correction, it is necessary to add some check bits to a block of data bits. Those check bits are also called **redundant bits** because they do not carry any user information. Check bits are chosen that the resulting bit sequence has a unique characteristic which enables error detection.

Coding is the process of adding the check bits to the block of data bits. The block of data bits to which check bits are added is called a **data word**. The bigger block containing check bits is called the **code word**. Hamming distance between two code words is the number of disagreements between them. By considering Figure 8.16, the distance between the two words given below is 3.

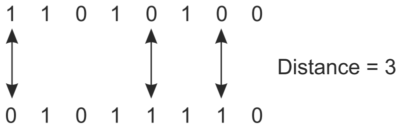


Fig. 8.16 Hamming distance

The weight of a code word is the number of ‘1’s in the code word. For example, the code 11001100 has a weight of 4. A code set consists of all valid code words. All the valid code words have a built-in characteristic of the code set.

8.13.3 Error Detection

When a code word is transmitted, one or more of its bits may be reversed due to signal impairment. The receiver can detect these errors if the received code word is not one of the valid code words of the code set.

When errors occur, the distance between the transmitted and received code words becomes equal to the number of erroneous bits, shown in Figure 8.17.

Transmitted Code Word	Received Code Word	Number of Errors	Distance
11001100	11001110	1	1
10010010	00011010	2	2
10101010	10100100	3	3

Fig. 8.17 Hamming distance between transmitted and received code words

In other words, the valid code words must be separated by a distance of more than 1. Otherwise, even a single bit error will generate another valid code word and the error will not be detected. The number of errors which can be detected depends on the distance between any two valid code words. For example, if the valid code words are separated by a distance of 4, up to 3 errors in a code word can be detected. By adding a certain number of check bits and properly choosing the algorithm for generating them, it is ensured that there is some minimum distance between any two valid code words of a code set.

8.13.4 Error Correction

After the detection of an error, there are two approaches to correct the errors.

1. Reverse error correction
2. Forward error correction

In reverse error-correction method, the receiver request for retransmission of the code word whenever it detects an error. In forward error-correction method, the code set is so designed that it is possible for the receiver to detect and correct the errors as well. The receiver locates the errors by analysing the received code word and reverses the erroneous bits.

Another way of forward error correction is to search for the most likely correct code word. When an error is detected, the distances of all the valid code words from the received invalid code word are measured. The nearest valid code word is the most likely correct version of the received word. It is illustrated in Figure 8.18.

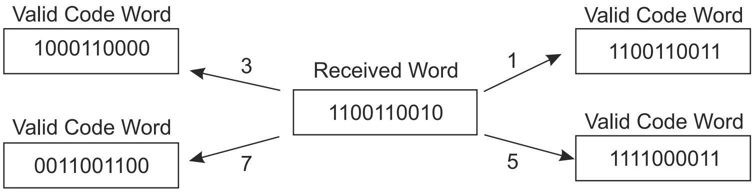


Fig. 8.18 Error correction by least Hamming distance

If the minimum distance between valid code words is D, up to $\frac{D}{2} - 1$ errors can be corrected. More than $\frac{D}{2}$ – 1 errors will cause the received code word to be nearer to the wrong valid code word.

8.13.5 Bit Error Rate (BER)

In analog transmission, signal quality is specified in terms of signal-to-noise ratio, which is usually expressed in dB. In digital transmission, the quality of received digital signal is expressed in terms of Bit Error Rate (BER) which is the number of errors in a fixed number of transmitted bits. A typical error rate on a high-quality leased telephone line is as low as 1 error in 10^6 bits.

Similar to bit error rate, Character Error Rate (CER) and Frame Error Rate (FER) can also be used. CER is the average number of characters received with at least one error in a large sample of transmitted characters. For low values of BER, CER and FER can be calculated from BER as below.

$$\text{CER} = b \times \text{BER} \tag{8.56}$$

$$\text{FER} = F \times \text{BER} \tag{8.57}$$

where b is the number of bits per character and f is the number of bits per frame.

EXAMPLE 8.10

If the average BER is 1 in 10^5, what is the probability of having

(a) Single bit error ?

(b) Single bit correct ?

(c) At least one error in an eight-bit byte?

Solution

(a) Probability of having single bit error $=1/10^5 = 0.00001$

(b) Probability of having single bit correct $= 1 - 0.00001 = 0.99999$

(c) Probability of having at least one error in an eight-bit byte $= 1 - (0.99999)^8$

$= 0.00008$

8.13.6 Residual Error Rate (RER)

Whatever be the methods of error control, errors cannot be eliminated. There is always some residual error which goes undetected. Residual error rate is the error rate in the data bits after error control has been performed.

8.14 METHODS OF ERROR DETECTION

Error detection is the process of monitoring data transmission and determining when errors have occurred. It is not possible to correct the errors and also to identify the particular bits in error by the error-detection techniques. They are used to indicate only when an error has occurred.

The purpose of error detection is not to prevent errors from occurring but to prevent undetected errors from occurring. The most common methods of error detection are

1. Parity checking

2. Check-sum error detection
3. Redundancy checking which includes
   (a) Vertical Redundancy Checking (VRC)
   (b) Longitudinal Redundancy Checking (LRC)
   (c) Cyclic Redundancy Checking (CRC)

8.14.1 Parity Checking

In parity-checking methods, an additional bit called a 'parity' bit is added to each data word. The additional bit is chosen that the weight of a code word so formed is either even or odd, which is illustrated in Figure 8.19.

Even Parity

P	Data Word
0	1001011
1	0010110

Odd Parity

P	Data Word
0	1001011
1	0010110

Fig. 8.19 Even and odd parity bits

When a single error or an odd number of errors occurs during transmission, there is a change in the parity of the code word. It is shown in Figure 8.20. Parity of the code word is checked at the receiving end and violation of the parity rule indicates errors somewhere in the code word.

Transmitted Code	10010110	Even Parity
Received Code (Single Error)	00010110	Odd Parity (Error is detected)
Received Code (Double Error)	00011110	Even Parity (Error is not detected)

Fig. 8.20 Error detection by change in parity

It is to be noted that double or any even number of errors will go undetected because the resulting parity of the code word will not change. Thus, a simple parity-checking method has its limitations. It is not suitable for multiple errors. To keep the possibility of occurrence of multiple errors low, the size of the data word is usually restricted to a single byte.

Parity checking does not reveal the location of the erroneous bit. Also, the received code word with an error is always at equal distance from two valid code words. Therefore, errors cannot be corrected by the parity-checking method.

8.14.2 Check-Sum Error Detection

In this method, a check-sum is transmitted along with every block of data bytes. The characters within a message are combined together to produce an error-checking character or the check-sum. The check-sum is appended to the end of the message. The receiver replicates the combing operation and determines its own check-sum. The receiver's check-sum is compared to the check-sum appended to the message, and if they are same, it is assumed that there are no transmission errors. If the two check-sums are different, a transmission error has definitely occurred.

8.14.3 Redundancy Checking

Duplicating each data unit for the purpose of detecting errors is a form of error detection called redundancy. It is more efficient to add bits to data units that check for transmission errors. Adding bits for the single purpose of error detection is called redundancy checking.

1. Vertical Redundancy Checking (VRC)

It is the simplest error-detection scheme and is generally referred to as character parity or simply parity. With character parity, each character has its own error-detection bit, called the **parity bit**. Since it is not actually part of the character, it is considered a redundant bit. An n-character message would have n redundant parity bits. Therefore, the number of error-detection bits is directly proportional to the length of the message.

With character parity, a single parity is added to each character to force the total number of logic 1s in the character, including the parity bit, to be either an odd number (odd parity) or an even number (even parity).

For example, the ASCII code for the letter C is 43H or P1000011 binary, where the P bit is the parity bit. There are three logic 1s in the code, not counting the parity bit. If odd parity is used, the P bit is made logic 0, keeping the total number of logic 1s at three, which is an odd number. If even parity is used, the P bit is made logic 1, making the total number of logic 1s four, which is an even number.

Its drawback is that when an even number of bits are received in error, the parity checker will not detect them because when the logic condition of an even number of bits is changed, the parity of the character remains the same. Consequently, over a long time, parity will detect only 50% of the transmission errors.

2. Longitudinal Redundancy Checking (LRC)

It is a redundancy error-detection scheme that uses parity to determine if a transmission error has occurred within a message and it is also called **message parity**. With LRC, each bit position has a parity bit. In other words, b_0 from each character is XORed with b_0 from all other characters in the message. Similarly, b_1, b_2, and so on, are XORed with their respective bits from all the characters in the message. Essentially, LRC is the result of XORing the character codes that make up the message, whereas VRC is the XORing of the bits within a single character. With LRC, even parity is generally used whereas with VRC, odd parity is used. Figure 8.21 illustrates the process of LRC.

The LRC bits are computed in the transmitter while the data are being sent and then appended to the end of the message as a redundant character. In the receiver, the LRC is recomputed from the data and the recomputed LRC is compared to the LRC appended to the message. If the two LRC characters are the same, most likely no transmission errors have occurred. If they are different, one or more transmission errors have occurred.

With longitudinal redundancy checking, all messages have the same number of error-detection characters. This characteristic alone makes LRC a better choice for systems that

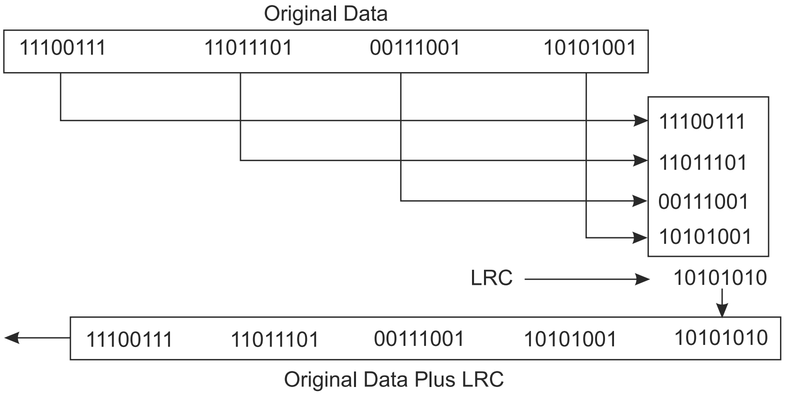


Fig. 8.21 Longitudinal redundancy checking

typically send long messages. Possibly, LRC detects between 95% and 98% of all transmission errors. It will not detect transmission errors when an even number of characters have an error in the same bit position.

3. Cyclic Redundancy Checking (CRC)

This is the most reliable method for error detection. With the help of the cyclic redundancy checking method, almost 99.99% of all transmission errors are detected. These codes are very powerful and are almost universally employed. This code provides a better measure of protection at the lower level of redundancy and can be easily implemented. It is also treated as a systematic code.

A CRC code word of length N with m-bit data word is referred to as (N, m) cyclic code and contains $(N - m)$ check bits, which is also said to the length of Block Check Character (BCC).

$$\text{BCC} = N - m \tag{8.58}$$

For a CRC-16 CODE, the BCC is the remainder of a binary-division process from which a data message polynomial $G(x)$ is divided by a unique generator polynomial function $P(x)$. The quotient of the division process is discarded and the remainder is truncated to 16 bits and which is appended to the message as a **Block Check Sequence (BCS).**

In CRC code, the division is not accomplished with standard arithmetic division. Instead, Modulo-2 division is used, where the remainder is derived from an XOR operation. In the receiver, the data stream, including the CRC code, is divided by the same generating function $P(x)$. If there are no transmission errors, the remainder will be zero. In the receiver, message and CRC character pass through a block check register. After the entire message has passed through the register, its contents should be zero if the receive message contains no errors. Figure 8.22 illustrates the process of CRC.

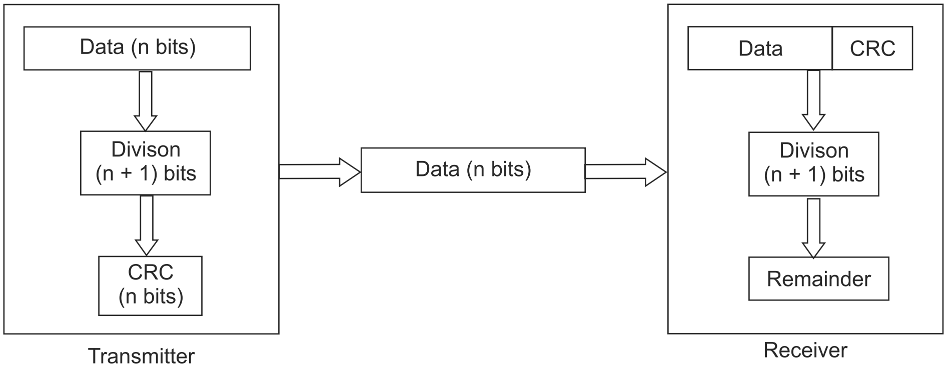


Fig. 8.22 CRC process

EXAMPLE 8.11

Determine the BCS for the following data and CRC generating polynomials:

$$G(x) = x^7 + x^5 + x^4 + x^2 + x^1 + x^0 = 10110111$$

$$\text{CRC } P(x) = x^5 + x^4 + x^1 + x^0 = 110011$$

Solution

G(x) is multiplied by the number of bits in the CRC code, which is 5.

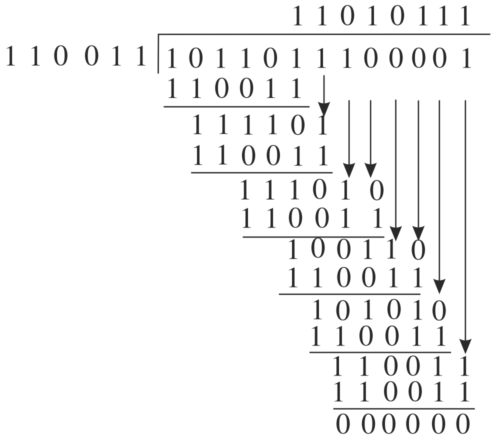
$$x^5 (x^7 + x^5 + x^4 + x^2 + x^1 + x^0) = x^{12} + x^{10} + x^9 + x^7 + x^6 + x^5 = 1011011100000$$

```
                          1 1 0 1 0 1 1 1
1 1 0  0 1 1 | 1 0 1 1 0 1 1 1 0 0 0 0 0
               1 1 0 0 1 1
                 1 1 1 1 0 1
                 1 1 0 0 1 1
                     1 1 1 0 1 0
                     1 1 0 0 1 1
                         1 0 0 1 0 0
                         1 1 0 0 1 1
                             1 0 1 1 1 0
                             1 1 0 0 1 1
                               1 1 1 0 1 0
                               1 1 0 0 1 1
                                   1 0 0 1 = CRC
```

The CRC is appended to the data to give the data stream as

$$\overbrace{101101110}^{G(x)} \quad \overbrace{1001}^{\text{CRC}}$$

At the receiver side, the data is again divided by $P(x)$.



The remainder is zero in the above division process which means that there were no transmission errors.

Not all the types of errors can be detected by a CRC code. The probability of error detection and the types of errors which can be detected depends on the choice of the divisor. If the number of check bits in the CRC code is n, the probabilities of error detection for various types of errors are as given below.

1. Single errors	100%
2. Two-bit errors	100%
3. Odd number of bits in error	100%
4. Error bursts of length < $(n + 1)$	100%
5. Error bursts of length = $(n + 1)$	$1 - (1/2)^{n-1}$
6. Error bursts of length > $(n + 1)$	$1 - (1/2)^{n}$

EXAMPLE 8.12

Determine the BCS for the following data and CRC generating polynomials:

$$G(x) = x^{10} + x^9 + x^7 + x^5 + x^3 + x^2 + x^1 + x^0 = \text{and CRC } P(x) = x^5 + x^4 + x^1 + x^0$$

Solution

$$G(x) = x^{10} + x^9 + x^7 + x^5 + x^3 + x^2 + x^1 + x^0 = 11010101111$$

$$\text{CRC } P(x) = x^5 + x^4 + x^1 + x^0 = 110011$$

$G(x)$ is multiplied by the number of bits in the CRC code, which is 5.

$$x^5 (x^{10} + x^9 + x^7 + x^5 + x^3 + x^2 + x^1 + x^0) = x^{15} + x^{14} + x^{12} + x^{10} + x^8 + x^7 + x^6 + x^5$$

$$= 1101010111100000$$

Dividing the polynomial by CRC $P(x)$.

```
                 10010000010
        110011 | 1101010111100000
                 110011
                   0110011
                    110011
                    000000110000
                          110011
                          0000110
```

The CRC is appended to the data to give the data stream as

$$\overbrace{11010101111}^{G(x)}\quad\overbrace{00110}^{\text{CRC}}$$

At the receiver side, the data is again divided by $P(x)$.

```
                 10010000010
        110011 | 1101010111100110
                 110011
                   0110011
                    110011
                    000000110011
                          110011
                          0000000
```

The remainder is zero in the above division process which means that there were no transmission errors.

EXAMPLE 8.13

Determine the CRC code for the following data and CRC generating polynomials:

$$G(x) = x^8 + x^7 + x^5 + x^3 + x^1 \text{ and CRC } P(x) = x^5 + x^3 + x^0$$

Solution

$$G(x) = x^8 + x^7 + x^5 + x^3 + x^1 = 110101010$$

$$\text{CRC } P(x) = x^5 + x^3 + x^0 = 10101$$

```
               111000111
        10101 | 1101010100000
                10101
                 11111
                 10101
                  10100
                  10101
                      11000
                      10101
                       11010
                       10101
                        11110
                        10101
                         1011   Remainder
```

The CRC code word is received by appending the remainder with the polynomial data.

```
       1 1 0 1 0 1 0 1 0 0 0 0 0
                         1 0 1 1
       -------------------------
   + – 1 1 0 1 0 1 0 1 0 1 0 1 1   CRC Code
```

EXAMPLE 8.14

For the CRC code word of Example 8.12, check if there are errors.

Solution

CRC code word = 1 1 0 1 0 1 0 1 0 1 0 1 1

Dividing the code word by CRC $P(x) = x^5 + x^3 + x^0 = 1\,0\,1\,0\,1$

```
                 111000111
          0101 | 1101010101011
                 10101
                  11111
                  10101
                   10100
                   10101
                    00011010
                        10101
                        011111
                          10101
                           10101
                           10101
                           00000   Remainder
```

The remainder is zero in the above division process which means that there were no transmission errors.

EXAMPLE 8.15

If the CRC code word of Example 8.12 be received as 1100100101011, check if there are errors.

Solution

CRC code word = 1 1 0 0 1 0 0 1 0 1 0 1 1

Dividing the code word by CRC $P(x) = x^5 + x^3 + x^0 = 1\ 0\ 1\ 0\ 1$

```
              1 1 1 0 0 0 1 1 1
          ___________________________
1 0 1 0 1 | 1 1 0 0 1 0 0 1 0 1 0 1 1
            1 0 1 0 1
            _________
              1 1 0 0 0
              1 0 1 0 1
              _________
                1 1 0 1 0
                1 0 1 0 1
                _________
                  1 1 1 1 1
                  1 0 1 0 1
                  _________
                    1 0 1 0 0
                    1 0 1 0 1
                    _________
                            1 1 0 1 1
                            1 0 1 0 1
                            _________
                            0 1 1 1 0     Remainder
```

The nonzero remainder indicates that there are errors in the received code word.

8.15 METHODS OF ERROR CORRECTION

Error correction in data communication can be of two major types.

1. Forward error correction
2. Reverse error correction

8.15.1 Forward Error Correction

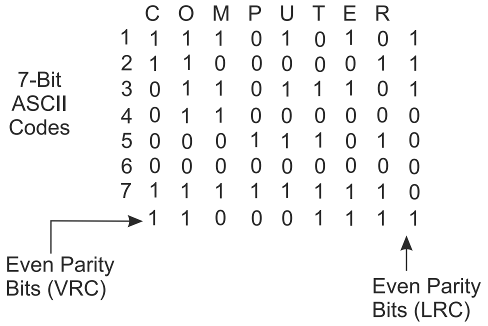
Location of errors and correction require a bigger overhead in terms of number of check bits in the code word. Under this category, there are three error-correction codes.

1. Block parity
2. Hamming code
3. Convolution code

1. Block Parity

The main purpose of parity checking is to detect and correct single errors. The data block is arranged in a rectangular matrix form as shown in Figure 8.19 and from which two sets of parity bits are generated, namely

1. Longitudinal Redundancy Check (LRC)
2. Vertical Redundancy Check (VRC)



	C	O	M	P	U	T	E	R	
1	1	1	1	0	1	0	1	0	1
2	1	1	0	0	0	0	0	1	1
3	0	1	1	0	1	1	1	0	1
4	0	1	1	0	0	0	0	0	0
5	0	0	0	1	1	1	0	1	0
6	0	0	0	0	0	0	0	0	0
7	1	1	1	1	1	1	1	1	0
	1	1	0	0	0	1	1	1	1

Bit transmission sequence

11000011 11110011 10110010 00001010 10101010 00101011 10100011 01001011 11100001

Fig. 8.23 Vertical and longitudinal parity-check bits

VRC is the parity bit associated with the character code and LRC is generated over the rows of the bits. LRC is appended to the end of a data block. The bit 8 of the LRC represents the VRC of the other 7 bits of the LRC. In Figure 8.23, even parity is used for the LRC and the VRC.

Even a single error in any bit results in failure of longitudinal redundancy check in one of the rows and vertical redundancy check in one of the columns. The bit which is common to the row and column is the bit in error. Multiple errors in rows and columns can be detected but cannot be corrected as the bits which are in error cannot be located.

EXAMPLE 8.16

The following bit stream is encoded using VRC, LRC and even parity. Correct the error if any.

11000011 11110011 10110010 00001010 10111010 00101011
10100011 01001011 11100001

Solution

$$\begin{array}{cccccccccl}
1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & \\
0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & \\
0 & 1 & 1 & 0 & \textcircled{1} & 0 & 0 & 0 & 0 & \longleftarrow \text{Wrong Parity} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & \\
1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & \\
 & & & & \uparrow \text{Wrong Parity} & & & & &
\end{array}$$

The fourth bit of the fifth byte is in error. It should be ‘0’.

EXAMPLE 8.17

The following bit stream is encoded using VRC, LRC and odd parity. Correct the error if any.

11000001 11110001 10110110 00001110 10111011 00101100
10100001 01100001

Solution

$$\begin{array}{ccccccccl}
1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & \\
1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & \\
1 & 0 & \textcircled{1} & 1 & 1 & 0 & 1 & 1 & \longleftarrow \\
0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & \text{Wrong Parity} \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & \\
 & & & & \uparrow \text{Wrong Parity} & & & &
\end{array}$$

The fifth bit of the third byte is in error. It should be ‘0’.

EXAMPLE 8.18

Generate the code word for ASCII character ‘K’ = 1001011. Assume even parity.

Solution

	1	2	3	4	5	6	7	8	9	10	1 1	
	P_1	P_2	1	P_4	0	0	1	P_8	0	1	1	
First parity bit	P_1		1		0		1		0		1	$P_1 = 1$
Second parity bit		P_2	1			0	1			1	1	$P_2 = 0$
Third parity bit				P_4	0	0	1					$P_4 = 1$
Fourth parity bit								P_8	0	1	1	$P_8 = 0$
Code Word	1	0	1	1	0	0	1	0	0	1	1	

EXAMPLE 8.19

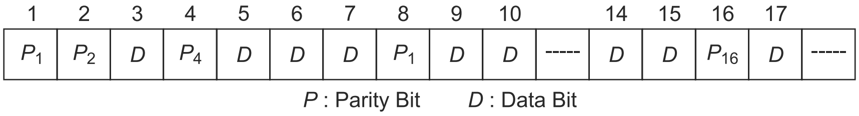
Write the ASCII code of the word 'HELLO' using even parity.

Solution

Bit positions	8	7	6	5	4	3	2	1
H	0	1	0	0	1	0	0	0
E	1	1	0	0	0	1	0	1
L	1	1	0	0	1	1	0	0
L	1	1	0	0	1	1	0	0
O	1	1	0	0	1	1	1	1

2. Hamming Code

It is the single-error correcting code given by Hamming. In this code, there are multiple parity bits in a code word. Bit positions 1,2,4,8.., etc. of the code are reserved for the parity bits. The other bit positions are for the data bits. It is illustrated in Figure 8.24.



1	2	3	4	5	6	7	8	9	10		14	15	16	17	
P_1	P_2	D	P_4	D	D	D	P_1	D	D	-----	D	D	P_{16}	D	-----

P : Parity Bit D : Data Bit

Fig. 8.24 Location of parity bits in the Hamming code

The number of parity bits required for correcting single-bit errors depends on the length of the code word. A code word of length n contains m parity bits, where m is the smallest integer satisfying the following condition,

$$2^m \geq n + 1 \tag{8.59}$$

The MSB of the data word is on the right-hand side and its opposition is third in Figure 8.24. LSB is transmitted first in the usual manner.

Each data bit is checked by a number of parity bits. Data-bit position expressed as a sum of the powers of 2 determines parity-bit positions which check the data bit. For example, a data bit in the position 6 is checked by parity bits P_4 and P_2 ($6 = 2^2 + 2^1$). Similarly, the data bit in the position 11 is checked by parity bits P_8, P_2 and P_1 ($11 = 2^3 + 2^1 + 2^0$). The following table shows the possible parity bit positions which check the various data positions.

Data-bit positions	Parity-bit positions					
	P_1	P_2	P_4	P_8		
3	×	×				
5	×		×			
6		×	×			
7	×	×	×			
Bit positions					Parity	Check
9	×			×		
10		×		×		
11	×	×		×		
12			×	×		

Each parity bit is determined by the data bits it checks. Even or odd parity can be used. For example, if even parity is used, P_2 is such that the number of '1's in 2nd, 3rd, 6th, 7th, 10th and 11th positions is even. The logic behind this way of generating the parity bits is that when a code word suffers an error, all the parity bits which check the erroneous bit will indicate violation of the parity condition and the parity-bit positions will indicate the position of the erroneous bit. For example, if the 11th bit is in error, parity bits P_8, P_2 and P_1 will indicate error and $8 + 2 + 1 = 11$ will immediately point to the 11th bit.

EXAMPLE 8.20

Detect and correct the single error in the received Hamming code word 10110010111. Assume even parity.

Solution

	1	2	3	4	5	6	7	8	9	10	11		
	P_1	P_2	D	P_4	D	D	D	P_8	D	D	D		
Code Word	1	0	1	1	0	0	1	0	1	1	1		
First check (P_1, 3, 5, 7, 9, 11)	1		1		0		1		1		1	Odd	Fail

Second check	0	1			0	1			1	1	Even	Pass	
(P_2, 3, 6, 7, 10, 11)													
Third check			1	0	0	1					Even	Pass	
(P_4, 5, 6, 7)													
Fourth check							0	1	1	1	Odd	fail	8
(P_8, 9, 10, 11)													9

3. Convolutional Code

In block codes, the check bits are computed for a block of data. But convolutional codes are generated over a span of data bits. That means that, a convolutional code of constraint length 3 is generated bit by bit always using the 'last 3 data bits'.

Figure 8.25 shows a simple convolutional encoder which consisting a shift register having three stages and XOR gates which generate two output bits for each input bit. It is called a **half-rate convolutional encoder**.

The state-transition diagram of the convolutional encoder is shown in Figure 8.26.

Each circle in the diagram represents a state of the encoder, which is the content of the two leftmost stages of the shift register. There are four possible states 00, 01, 10, 11. The arrows represent the state transitions for the input bit which can be 0 or 1. The label on each arrow

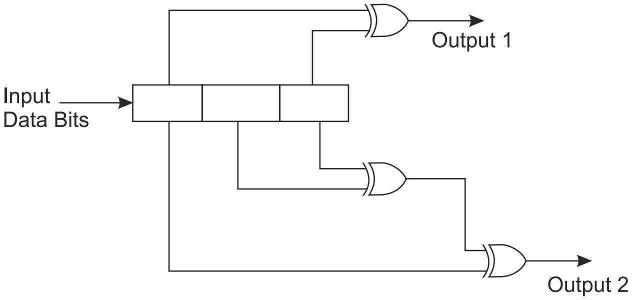


Fig. 8.25 Half-rate convolutional encoder

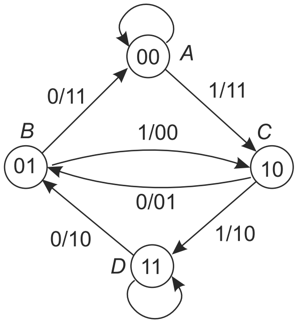


Fig. 8.26 State-transition diagram of convolutional encoder

shows the input data bit by which the transition is caused and the corresponding output bits. As an example, suppose the initial state of the encoder is 00 and the input data sequence is 1011. The corresponding output sequence of the encoder will then be 11010010.

The decoder for the convolutional code is based on the maximum likelihood principle. Knowing the encoder behaviour and the received sequence of bits, the maximum likely transmitted sequence is found out by analysing all the possible paths. The path which results in the output sequence, the nearest to the received sequence, is chosen and the corresponding input bits are the decoded data bits.

8.15.2 Reverse Error Correction

With respect to the number of check bits, the reverse error-correction method is more economical than the forward error-correction method. So, error-detection methods are implemented preferably with error-correction mechanism which requires the receiver to request the transmitter for retransmission of the code word received with errors. There are three types of error correction.

1. Stop and wait
2. Go-back-N
3. Selective retransmission

1. Stop and Wait

In this method, the transmitting end transmits one block of data at a time and then waits for acknowledgement from the receiver. If the receiver detects any error in the data block, it transmits a request for retransmission in the form of negative acknowledgement. If there is no error, the receiver sends a positive acknowledgment in which case the transmitting end transmits the next block of data. Figure 8.27 shows the method of error correction by the stop-and-wait method.

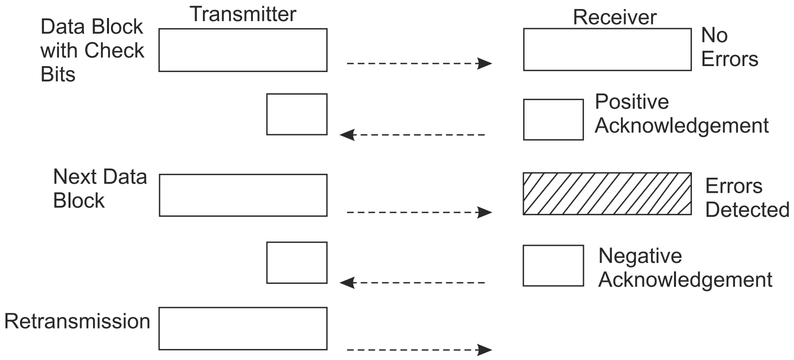


Fig. 8.27 Reverse error correction by stop-and-wait method

2. Go-Back-N

With this method, all the data blocks are numbered and the transmitting end keeps transmitting the data blocks with check bits. Whenever the receiver detects an error in a block, it sends a retransmission request indicating the sequence number of the data block received with errors. The sending end then starts retransmission of all the data blocks from the requested data block onwards. Figure 8.28 illustrates the Go-Back-N method of error correction.

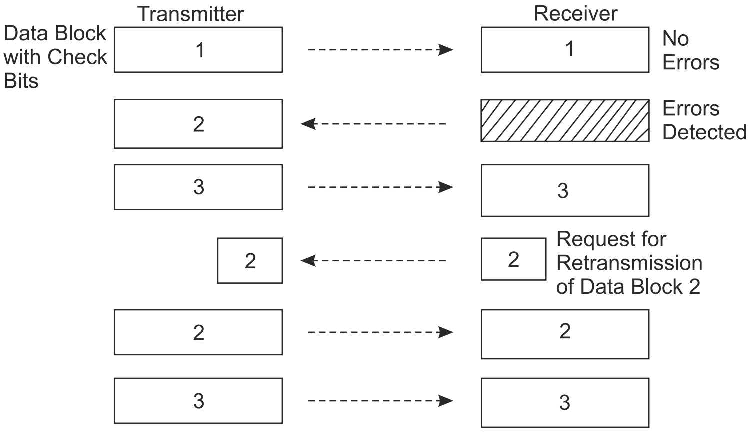

Transmitter
Receiver
Data Block with Check Bits
1
1
No Errors
2
Errors Detected
3
3
2
2
Request for Retransmission of Data Block 2
2
2
3
3

Fig. 8.28 Go-Back-N method of error correction

3. Selective Retransmission

If the receiver is equipped with the capability to resequence the data blocks, it requests for selective retransmission of the data block containing errors. On the receipt of a request, the sending end retransmits the data block but skips the following data blocks already transmitted and continues with the next data block. Figure 8.29 illustrates the error correction by selective retransmission.

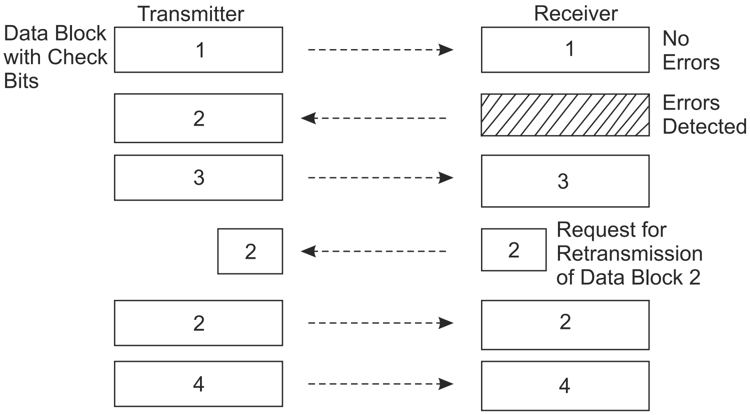

Transmitter
Receiver
Data Block with Check Bits
1
1
No Errors
2
Errors Detected
3
3
2
2
Request for Retransmission of Data Block 2
2
2
4
4

Fig. 8.29 Error correction by selective retransmission

Summary

Information theory deals with mathematical modelling and analysis of a communication system rather than with physical sources and physical channels. An information source is a mathematical model for a physical entity that produces a succession of symbols called 'outputs' in a random manner. The symbols produced may be real numbers such as voltage measurements from a transducer, binary numbers as in computer data, two-dimensional intensity fields as in a sequence of images, continuous or discontinuous waveforms, and so on.

In any communication system, each of the possible input sequences induces a probability distribution on the output sequences and the input sequences at the output can be transmitted and the source message can be reconstructed at the output side. The maximum rate at which this can be done is called the capacity of the channel.

Source coding offers the most significant application of the information theory. Its main purpose is to improve the efficiency of the communication system. It is a procedure for mapping a given set of messages $\{m_1, m_2,...m_n\}$ into a new set of encoded messages $\{C_1, C_2,...C_n\}$ in such a manner that the transformation is in the form of one to one.

Shannon–Fano coding is a technique for constructing a prefix code based on a set of symbols and their probabilities. In comparison with the Shannon–Fano algorithm which does not always generate an optimal code, Huffman coding is a different algorithm that always produces an optimal tree for any given probabilities. Huffman coding produces prefix codes that always achieve the lowest expected code-word length under the constraints that each symbol is represented by a code formed of an integral number of bits.

Intersymbol Interference (ISI) arises due to the dispersive nature of a communication channel in a digital baseband transmission. It is a form of distortion of a signal in which one symbol interferes with subsequent symbols. This is an unwanted phenomenon as the previous symbols have similar effect as noise, thus making the communication less reliable. There are two main causes of Intersymbol interference. They are

- Multipath propagation and
- Bandlimited channels

As the effect of intersymbol interference increases, the eye opening reduces. If the eye is closed completely then it is not possible to avoid errors in the output.

Error control in data communication is based on the detection of errors in a message and its retransmission. For error detection and correction, it is necessary to add some check bits to a block of data bits. Those check bits are also called redundant bits because they do not carry any user information.

The most common methods of error detection are

- Parity checking
- Check-sum error detection
- Redundancy checking which includes

- Vertical Redundancy Checking (VRC)
- Longitudinal Redundancy Checking (LRC)
- Cyclic Redundancy Checking (CRC)

In parity-checking methods, an additional bit called a 'parity' bit is added to each data word. The additional bit is so chosen that the weight of a code word so formed is either even or odd. In the check-sum error-detection method, a check-sum is transmitted along with every block of data bytes. Duplicating each data unit for the purpose of detecting errors is a form of error detection called redundancy.

Vertical Redundancy Checking (VRC) is the simplest error-detection scheme and is generally referred to as character parity or simply parity. With character parity, each character has its own error-detection bit called the parity bit. Longitudinal Redundancy Checking (LRC) is a redundancy error-detection scheme that uses parity to determine if a transmission error has occurred within a message and it is also called message parity. Cyclic Redundancy Checking (CRC) is the most reliable method for error detection. With the help of the cyclic redundancy checking method, almost 99.99% of all transmission errors are detected. These codes are very powerful and are almost universally employed.

After the detection of an error, there are two approaches to correct the errors.

- Reverse Error Correction
- Forward Error Correction

In the reverse error-correction method, the receiver requests for retransmission of the code word whenever it detects an error. In the forward error-correction method, the code set is so designed that it is possible for the receiver to detect and correct the errors as well. The receiver locates the errors by analysing the received code word and reverses the erroneous bits. Hamming code and convolutional code are the examples of error correction.

An interface that can be used for serial communication is generally said to be a serial interface and in which only one bit is transmitted at a time. It is a general-purpose interface that can be used for any kind of devices including modems, scanners and printers. Some of the popular serial interfaces include RS-232, RS-422 and RS-485. A parallel interface is used for communication between a computer and an external device such as printer.

REVIEW QUESTIONS

PART-A

1. How will you measure the information carried by a message?
2. Define entropy.
3. What is meant by conditional entropy?

4. How will you calculate the information rate?
5. What is channel capacity of a communication channel?
6. State the properties of channel capacity.
7. What do you mean by BSC?
8. What is the significance of BEC?
9. How will you find the efficiency of a source coding?
10. Shannon–Fano coding is not often preferred. Why?
11. What are the causes of intersymbol interference?
12. What is the significance of an eye pattern?
13. Classify errors during data transmission.
14. What are the error measures to be considered in data communication?
15. List out the approaches to correct the errors.
16. Define Bit Error Rate (BER).
17. Mention the popular methods of error detection.
18. What do you mean by parity checking?
19. What is check-sum error detection?
20. Define redundancy checking.
21. Compare VRC and LRC.
22. What are the advantages of CRC?

PART-B

1. Explain logarithmic measurement of information in detail.
2. What is channel capacity? Explain with binary symmetric channel and binary erasure channel as examples.
3. What is an information rate? How will you measure it?
4. With your own illustration, explain the procedure of Shannon–Fano coding in detail.
5. How is Huffman coding superior to Shannon–Fano coding? Explain the procedure for the development of a Huffman coding tree.
6. Describe the causes of intersymbol interference and interpretation of eye pattern in detail.
7. What are the methods of error detection? Explain them in detail.
8. Explain the following methods of error-detection methods.
   i) Longitudinal redundancy checking
   ii) Vertical redundancy checking

9. How will you detect the error by using Cyclic Redundancy Checking? Explain with an example.
10. Generate a CRC code for the data word 1010001011 using the divisor 11101.
11. If the CRC code is 10100010111100 and the generating polynomial is 11101, check if there is any error in the code word.
12. The received Hamming code word is 11110000101. Even parity is used. Locate and correct the bit in error.

9

DATA COMMUNICATION AND NETWORKS

Objectives

- To discuss different data communication codes in detail
- To provide details about data communication, its hardware and interfaces
- To discuss about data communication networks with their functions, components and topologies
- To provide details about the ISDN and LAN

9.1 INTRODUCTION

Data communication refers to exchange of digital information, the information that is stored in digital form, between two digital devices. The fundamental concepts of data communication include data communication code, error detection and error correction and character synchronisation. The hardware for data communication includes various pieces of computer and networking equipment such as line control units, serial interfaces and data communication modems.

9.2 DATA COMMUNICATION CODES

The main purpose of data communication codes is to represent characters and symbols such as letters and digits and punctuation marks. For this reason, data communication codes are also referred to as character codes, character sets or character languages.

9.2.1 Data Representation

A binary digit or bit has only two states, ‘0’ and ‘1’ and can represent only two symbols. But the simplest form of communication between computers requires a much larger set of symbols like

- 52 capital and small-case letters
- 10 numbers from 0 to 9
- Punctuation marks and other special symbols
- Terminal control characters.

Therefore, a group of bits is used as a code to represent a symbol. The code is usually 5 to 8 bits long. 5-bit code can have 2^5 =32 combinations and which can represent 32 symbols. Similarly, an 8-bit code can represent 2^8 = 256 symbols. A code set is the set of these codes representing the symbols.

1. ASCII Code

ASCII refers to American Standard Code for Information Interchange. ASCII is defined by American National Standards Institute (ANSI) in ANSI X3.4. It is the standard character set

Table 9.1 ASCII Code Set

Bit numbers		7 6 5	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
4321			0	1	2	3	4	5	6	7
0000		0	NUL	DLE	SPACE	0	@	P		p
0001		1	SOH	DC1	!	1	A	Q	a	q
0010		2	STX	DC2	“	2	B	R	b	r
0011		3	ETX	DC3	#	3	C	S	c	s
0100		4	EOT	DC4	$	4	D	T	d	t
0101		5	ENQ	NAK	%	5	E	U	e	u
0110		6	ACK	SYN	&	6	F	V	f	v
0111		7	BEL	ETB	‘	7	G	W	g	w
1000		8	BS	CAN	(	8	H	X	h	x
1001		9	HT	EM	)	9	I	Y	i	y
1010		A	LF	SUB	*	:	J	Z	j	z
1011		B	VT	ESC	+	;	K	[	k	{
1100		C	FF	FS	,	<	L	\	l	|
1101		D	CR	GS	-	+	M	]	m	}
1110		E	SO	RS	.	>	N	^	n	~
1111		F	SI	US	/	?	O		o	DEL

for source coding the alphanumeric character set that humans understand but computers are not able to understand, which understands only 1s and 0s. It is a 7-bit code and all the possible 128 codes have defined meanings. The code set consists of the following symbols.

- 96 graphic symbols (columns 2 to 7), comprising 94 printable characters, SPACE and DEL, etc., characters.
- 32 control symbols (columns 1 and 1)

The complete ASCII code set is shown in Table 9.1.

The binary representation of a particular character can be easily determined from its hexadecimal coordinates. For example, the coordinates of ‘K’ are (4, B). Therefore, its binary code is 100 1011.

The control symbols are codes reserved for special functions. Table 9.2 lists the control symbols. Some important functions and the corresponding control symbols are

- Functions relating to basic operation of the terminal device like a printer

  CR (Carriage Return)

  LF (Line Feed)

Table 9.2 Control symbols

ACK	Acknowledgement	FF	Form Feed
BEL	Bell	FS	File Separator
BS	Backspace	GS	Group Separator
CAN	Cancel	HT	Horizontal Tabulation
CR	Carriage Return	LF	Line Feed
DC1	Device Control 1	NAK	Negative Acknowledgement
DC2	Device Control 2	NUL	Null
DC3	Device Control 3	RS	Record Separator
DC4	Device Control 4	SI	Shift In
DEL	Delete	SO	Shift Out
DLE	Data Line Escape	SOH	Start of Heading
EM	End of Medium	STX	Start of Text
ENQ	Enquiry	SUB	Substitute Character
EOT	End of transmission	SYN	Synchronous Idle
ESC	Escape	US	Unit Separator
ETB	End of Transmission Block	VT	Vertical Tabulation
ETX	End of Text		

- Functions relating to error control

  ACK (Acknowledgement)

  NAK (Negative Acknowledgement)

- Functions relating to blocking of data characters
  STX (Start of Text)
  ETX (End of Text)

DC1, DC2, DC3 and DC4 are user definable. DC1 and DC3 are generally used as X-ON and X-OFF for switching the transmitter.

With the 7-bit fixed length ASCII code, the least significant bit (LSB) is designated b_o and the most significant bit (MSB) is designated b_7 as shown here.

b_7 b_6 b_5 b_4 b_3 b_2 b_1 b_0

MSB $\longrightarrow$ LSB
Direction of Propagation

Table 9.3 Character set of EBCDIC code

Bit	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
num-	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
bers	2	0 0	0 1	1	1	0	0	1	1	0	0	1	1	0	0	1	1
	3			0	1	0	1	0	1	0	1	0	1	0	1	0	1
4321		0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0000	0	NUL	DLE				SP	&	-								0
0001	1	SOH	SBA					/		a	j			A	J		1
0010	2	STX	EUA		SYN					b	k	s		B	K	S	2
0011	3	ETX	IC							c	l	t		C	L	T	3
0100	4									d	m	u		D	M	U	4
0101	5	HT	NL	ETB						e	n	v		E	N	V	5
0110	6			ESC	EOT					f	o	w		F	O	W	6
0111	7									g	p	x		G	P	X	7
1000	8									h	q	y		H	Q	Y	8
1001	9		EM							i	r	z		I	R	Z	9
1010	A					⊄	!	!									
1011	B						$		#								
1100	C		DUP		RA	<	*	%	@								
1101	D		SF	ENQ	NAK	(	)	–									
1110	E		FM			+	;	>	=								
1111	F		ITB		SUB	|	¬	?	“								

Among the 7 bits, b_7 is not part of the ASCII code but is generally utilised during transmission as an error-detection bit called the parity bit.

2. EBCDIC Code

EBCDIC refers to Extended Binary Coded Decimal Interchange Code, which was developed by International Business Machines Corporation (IBM). It is an 8-bit fixed length code with 256 possible combinations. There is no parity bit for error checking in the basic code set. The graphic symbol subset is approximately the same as ASCII. There are several differences in the control characters. This code is not the same for all devices. In EBCDIC, the bit numbering starts from the most significant bit (MSB), but in ASCII, it starts from the least significant bit (LSB). Table 9.3 shows the character set of EBCDIC code.

3. Baudot Teletype Code

Baudot code was the first fixed-length character code, developed specially for machines rather than people. It is also called **telex code** or **ITA-2** (International Telegraph Alphabet Number 2). It is a 5-bit code and is used in electromechanical teletype machines. 32 codes are possible using 5 bits but in this code, there are 58 symbols. The same code is used for two symbols using letter shift/figure shift keys which change the meaning of a code. In telegraph terms, the binary '1' is called **Mark** and binary '0' is called **Space**.

4. Bar Code

Bar codes are those omnipresent black-white striped stickers that seem to appear on virtually every consumer item worldwide. A bar code is a series of vertical black bars separated by vertical white bars, which are called spaces. The widths of the bars and spaces along with their reflective abilities represent binary 1s and 0s and combinations of bits identify specific items.

It also contains information regarding the cost of the product, inventory management and control, security access, shipping and receiving, production counting, document and order processing, automatic billing and many other applications. A typical bar code is shown in Figure 9.1.

Fig. 9.1 A typical bar code

There are several standard bar-code formats. The format selected depends on what types of data are being stored, how the data are being stored, system performance and which format is most popular with business and industry.

Bar codes are generally classified as being discrete, continuous or two-dimensional. A discrete code has spaces or gaps between characters and so each character within the bar code is independent of every other character. A continuous bar code does not include spaces between characters. A 2D bar code stores data in two dimensions and it has a larger storage capacity than one-dimensional bar code (1 kilobyte or more).

5. BCDIC Code

BCDIC refers to Binary Coded Decimal Interchange Code. It is also a fixed-length code which is of 6-bit length code. It consists of 64 symbols.

9.3 DATA TRANSMISSION

There is always need to exchange data, commands and other control information between a computer and its terminals or between two computers. This information is in the form of bits.

Data transmission refers to movement of the bits over some physical medium connecting two or more digital devices. There are two options of transmitting the bits, namely, parallel transmission or serial transmission.

9.3.1 Parallel Transmission

In parallel transmission, all the bits of a byte are transmitted simultaneously on separate wires as shown in Figure 9.2 and hence multiple circuits interconnecting the two devices are required. It is practical only if the two devices are close to each other like a computer and its associated printer.

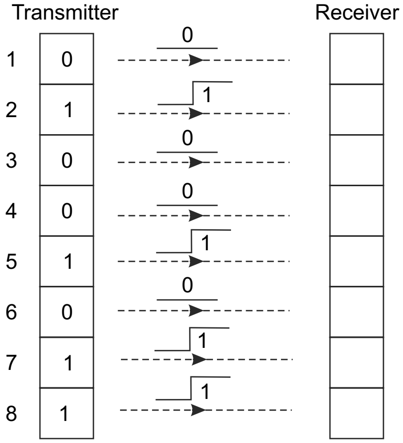


Fig. 9.2 Parallel transmission

9.3.2 Serial Transmission

In serial transmission, bits are transmitted serially one after the other. It is represented in Figure 9.3.

The Least Significant Bit (LSB) is usually transmitted first. By comparing with parallel transmission, serial transmission requires only one circuit interconnecting the two devices. Therefore, serial transmission is suitable for transmission over long distances.

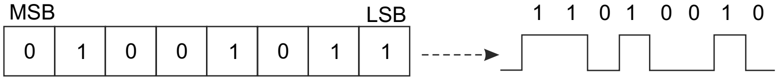


Fig. 9.3 Serial transmission

Bits are transmitted as electrical signals over interconnecting wires. The two binary states ‘1’ and ‘0’ are represented by two voltage levels. If one of these states is assigned a 0-volt level, the transmission is termed unipolar and if binary ‘1’ is the chosen state by a positive voltage +*V* volts and a binary ‘0’by a negative voltage –*V* volts, the transmission is termed bipolar. Bipolar transmission is preferred because the signal does not have any *dc* component. The transmission media usually do not allow the *dc* signals to pass through.

EXAMPLE 9.1

A block of 256 sequential 10-bit data words is transmitted serially in 0.015 s. Calculate (a) time duration of 1 word, (b) time duration of 1 bit, and (c) speed of transmission in bits/s.

Solution

(a) Calculation of time duration of 1 word:

$$t(\text{word}) = \frac{0.015}{256} = 58.6\ \mu\text{s}$$

(b) Calculation of time duration of 1 bit:

$$t(\text{bit}) = \frac{58.6\ \mu\text{s}}{10\ \text{bits}} = 5.86\ \mu\text{s}$$

(c) Calculation of time duration of 1 bit and speed of transmission in bits/s.

$$\text{Bits/second} = \frac{1}{5.86 \times 10^{-6}} = 170.67\ \text{kbps}$$

EXAMPLE 9.2

During serial transmission, a group of 512 sequential 12-bit data words is transmitted in 0.016 s. Find the speed of transmission in bps.

Solution

$$t(\text{word}) = \frac{0.016}{512} = 31.25\ \mu\text{s}$$

$$t(\text{bit}) = \frac{31.25\ \mu\text{s}}{12\ \text{bits}} = 2.60\ \mu\text{s}$$

Speed of transmission in bits/s is

$$\text{Bits/s} = \frac{1}{2.60 \times 10^{-6}} = 384\ \text{kbps}$$

9.4 BIT RATE

Bit rate is simply the number of bits which can be transmitted in a second. If t_p is the duration of a bit, the bit rate R will be $1/\ t_p$. It must be noted that bit duration is not necessarily the pulse duration.

9.5 RECEPTION OF DATA BITS

The signal received at the other end of the transmitting medium is never identical to the transmitted signal as the transmission medium distorts the signal to some extent. As a result, the receiver has to put in considerable effort to identify the bits. The receiver must know the time instant at which it should look for a bit. Therefore, the receiver must have synchronised clock pulses which mark the location of the bits. The received signal is sampled using the clock pulses and depending on the polarity of a sample, the corresponding bit is identified.

It is essential that the received signal is sampled at the right instants as otherwise it could be misinterpreted. For that purpose, the clock frequency should be exactly the same as the transmission bit rate. Even a small difference will build up as timing error and eventually result in sampling at wrong instants. Figure 9.4 shows two situations when the clock frequency is slightly faster or slightly slower than the bit rate. When clock frequency is faster, a bit may be sampled twice and may be missed when it is slower.

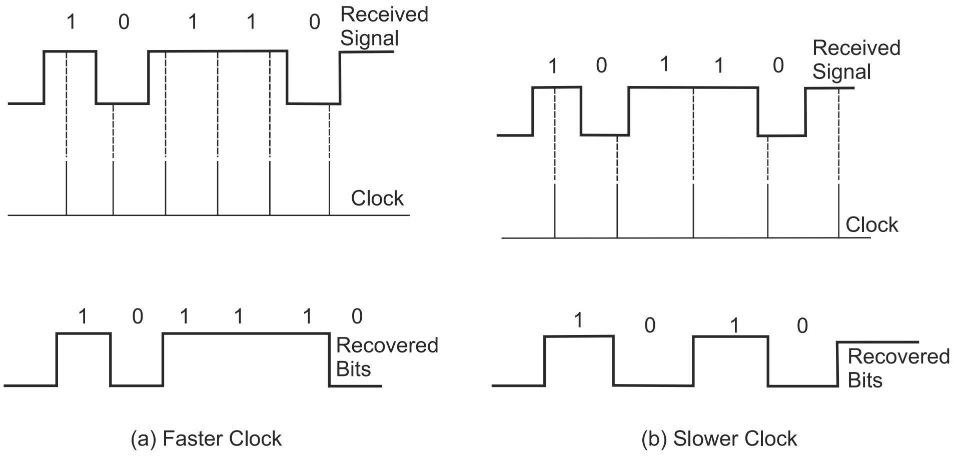


Fig. 9.4 Timing errors

9.6 MODES OF DATA TRANSMISSION

There are two methods of timing control for reception of bits. The transmission modes corresponding to these two timing methods are as follows.

1. Asynchronous transmission
2. Synchronous transmission
3. Isosynchronous transmission

9.6.1 Asynchronous Transmission

Asynchronous transmission refers to the case when the sending end commences transmission of bytes at any instant of time. Only one byte is send at a time and there is no time relation between consecutive bytes. That means, after sending a byte, the next byte can be sent after arbitrary delay. In the idle state, when no byte is being transmitted, the polarity of the electrical signal corresponds to '1'. It is shown in Figure 9.5.

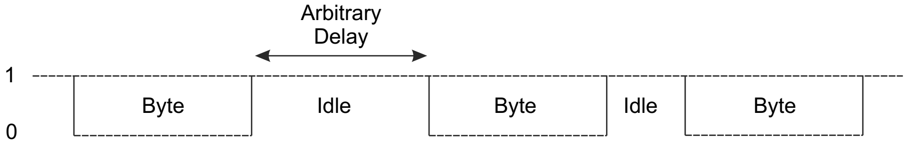


Fig. 9.5 Asynchronous transmission

Due to the arbitrary delay between consecutive bytes, the time occurrences of the clock pulses at the receiving end need to be synchronised repeatedly for each byte. This is achieved by providing two extra bits, a start bit at the beginning and a stop bit at the end of a byte.

1. Start Bit

The start bit is always '0' and is prefixed to each byte. At the onset of transmission of a byte, it ensures that the electrical signal changes from the idle state '1' to '0' and remains at '0' for one bit duration. The leading edge of the start bit is used as a time reference for generating the clock pulses at the required sampling instants, which is illustrated in Figure 9.6. Thus, each onset of a byte results in resynchronisation of the receiver clock.

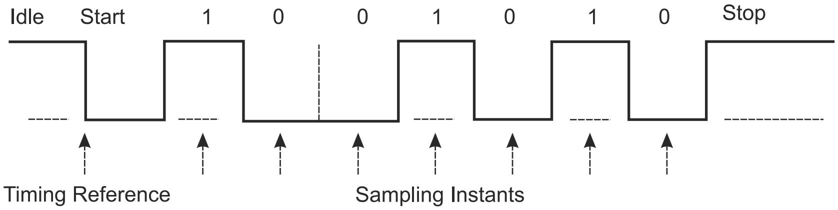


Fig. 9.6 Start and stop bits

2. Stop Bit

To ensure that the transition from ‘1’ to ‘0’ is always present at the beginning of a byte, it is necessary that polarity of the electrical signal should correspond to ‘1’ before occurrence of the start bit. So the idle state is kept as ‘1’. But there may be two bytes, one immediately following the other and if the last bit of the first byte is ‘0’, the transition from ‘1’ to ‘0’ will not occur. Therefore, a stop bit is also suffixed to each byte. It is always ‘1’ and its duration is usually 1, 1.5 or 2 bits.

9.6.2 Synchronous Transmission

A synchronous transmission is carried out under the control of a timing source. In synchronous transmission, bits are always synchronised to a reference clock irrespective of the bytes they belong to. There are no start or stop bits. Bytes are transmitted as a block in a continuous stream of bits. Even the interblock idle time is filled with idle characters. The process of synchronous transmission is shown in Figure 9.7.

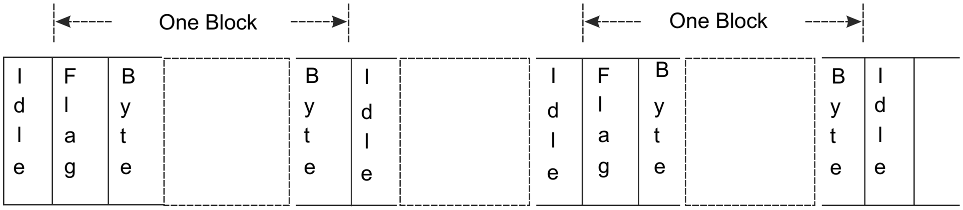


Fig. 9.7 Synchronous transmission

1. Bit Recovery

Continuous transmission of bit enables the receiver to extract the clock from the incoming electrical signal, which is illustrated in Figure 9.8. As this clock is inherently synchronised to the bits, the job of the receiver becomes simpler.

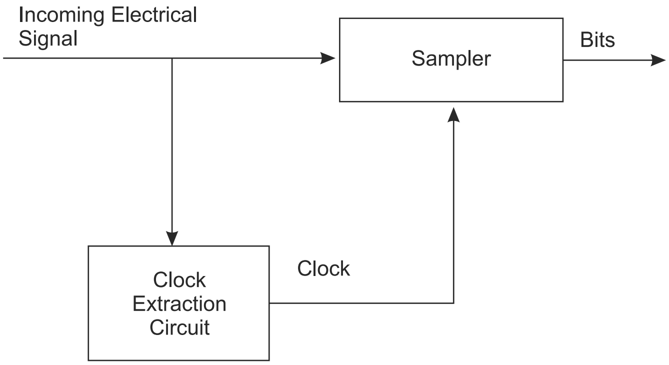


Fig. 9.8 Bit recovery in synchronous transmission

In synchronous transmission, the bytes lose their identity and their boundaries need to be identified. Therefore, a unique sequence of fixed number of bits, called flag, is prefixed to each clock. The flag identifies the start of a block. The receiver first detects the flag and then identifies the boundaries of different bytes using a counter.

9.6.3 Isosynchronous Transmission

The above asynchronous and synchronous transmission methods involve detailed error-checking mechanisms. But the advantage of isochronous communication is a fast, steady, uninterrupted data stream. Isochronous clocking information is derived from or included in the data stream. Communication can be disrupted if the transmitter does not maintain a constant transfer rate or if the receiver is insufficient to store data at the rate it is arriving and then hold it until it can be processed.

To maintain data-transfer speed, error checking is sometimes ignored and there is no hardware mechanism by which to request retransmission of corrupted data. Isochronous communication is best suited for applications where a steady data stream is more important than accuracy.

9.7 TRANSMISSION CHANNEL

A transmission channel transports the electrical signals from the transmitter to the receiver. It is characterised by two basic parameters, namely bandwidth and Signal-to-Noise Ratio (SNR). These parameters determine the ultimate information-carrying capacity of a channel. If B is the bandwidth of a transmission channel which carries a signal having L levels, the maximum data rate R is given by

$$R = 2B \log_2 \mathrm{L} \tag{9.1}$$

The number of levels L can be more than two. By considering the signal-to-noise ratio, the maximum data rate is given by

$$R = B \log_2 \left(1 + \frac{S}{N}\right) \tag{9.2}$$

This equation puts a limit on the number of levels L.

EXAMPLE 9.3

If bandwidth of the transmission channel is 3000 Hz and SNR is 1000 dB, calculate the maximum data rate.

Solution

Maximum data rate can be expressed as

$$R = B\log_2\left(1+\frac{S}{N}\right)$$

$$R = 3000 \times \log_2 (1 + 1000) = 30{,}000 \text{ bits/s}$$

For the data rate of 30,000 bits per second, the number of levels L can be computed as

From $R = 2\text{B} \log_2 L$

$$30{,}000 = 2 \times 3000 \log_2 L$$

$\therefore L = 32$

9.7.1 Bauds

When bits are transmitted as an electrical signal having two levels, the bit rate and the modulation rate of the electrical signal are the same. Modulation rate is the rate at which the electrical signal changes its levels. It is expressed in bauds. Note that there is one-to-one correspondence between bits and electrical levels.

9.7.2 Modem

The electrical state can also be defined in terms of other attributes of an electrical signal such as amplitude, frequency or phase. The basic electrical signal is a sine wave in this case. The binary signal modulates one of these signal attributes. The sine wave carries the information and is termed 'carrier'. The device which performs modulation is called a modulator and the device which recovers the information signal from the modulated carrier is called a demodulator.

In data transmission, the devices used to perform both modulation as well as demodulation function are called 'modems'. They are required when data is to be transmitted over long distances. In a modem, the input signal modulates a carrier which is transmitted to the distance end. At the distant end, another modem demodulates the received carrier to get the digital signal. Thus, a pair of modems is always required.

9.8 DIRECTIONAL CAPABILITIES OF DATA EXCHANGE

There are three possibilities of data exchange.

1. Transfer in both directions at the same time
2. Transfer in either direction, but only in one direction at a time
3. Transfer in one direction only

Terminology used for specifying the directional capabilities is different for data transmission and for data communication. It is shown in Table 9.4.

Table 9.4 Terminology for directional capabilities

Directional Capability	Transmission	Communication
One direction only	Simplex	One way
One direction at a time	Half-duplex	Two-way alternate
Both directions at the same time	Full duplex	Two-way simultaneously

9.9 DATA COMMUNICATION HARDWARE

9.9.1 Data Terminal Equipment (DTE)

Data Terminal Equipment (DTE) is a digital device that functions either as a source or destination of binary digital data. This device does not communicate directly with one another in order to establish and control the communication between the end points of data communication. DTE contains both hardware and software. DTE can be of any type like a terminal, computer, printer, fax machine or any other device that generates or receives digital data. DTE includes the concept of terminals which are used to input sections, output sections and to display the information, clients, hosts and servers.

9.9.2 Data Communication Equipment (DCE)

Data communication Equipment (DCE) is also called data-circuit terminating equipment. It includes any functional unit that transmits or receives information in analog or digital form.

Data communication equipment takes the data generated by the DTEs, converts them to analog or digital form and then sends to the telecommunication links. Data communication equipment includes modems.

The data communication equipment at the receiver takes the signal and converts it to a form usable by the data terminal equipment and delivers the signal. During the data communication, both the sending and receiving data communication equipment must use the same modulating method. For this purpose, data communication equipment includes channel service units, digital service units and data modems.

9.10 SERIAL INTERFACES

The Personal Computer (PC) is a bus-oriented system whereby subsystems or peripherals are interconnected through the bus architectures. The user should be able to print out data or copy to another system as well as collect the data from various instruments.

To design a PC with such features, there is a need of a common understanding of equipment specifications among manufacturers. This is known as standards. In the field of electronics, these standards are generally defined by professional organisations such as IEEE (Institute of Electrical and Electronics Engineers) and Electronics Industry Association (EIA). The need for expandability and modularity gave rise to various bus standards.

An interface that can be used for serial communication is generally said to be serial interface and in which only one bit is transmitted at a time. It is a general-purpose interface that can be used for any kind of devices including modems, scanners and printers.

The Electronics Industry Association (EIA) has produced standards for RS-232, RS-485, RS-422 and RS-423 that deal with data communications. EIA standards were previously marked with the prefix 'RS' to indicate Recommended Standard. However, the standards are now generally indicated as 'EIA' standards to identify the standards organisation.

Electronic data communications between elements will generally fall into two broad categories, namely

1. Single-ended
2. Differential

The single-ended data communications bus allows for data transmission from one transmitter to one receiver data. RS-232 and RS-423 are single-ended data communication buses.

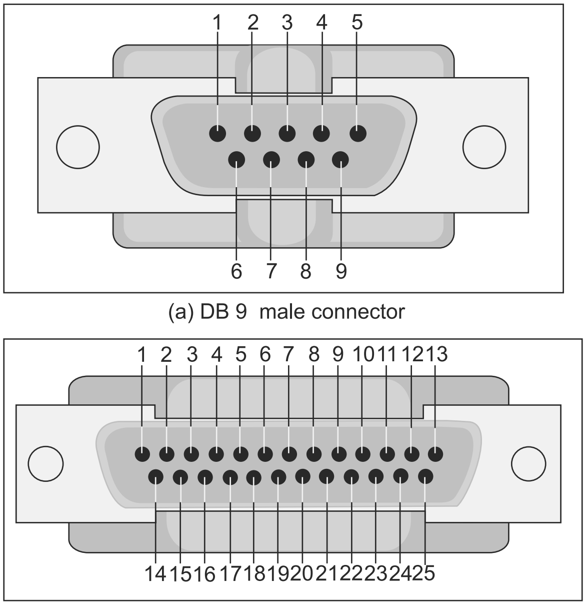
When communicating at high data rates, or over long distances in real-world environments, single-ended methods are often inadequate. Differential data transmission offers superior performance in most applications. Differential signals can help nullify the effects of ground shifts and induced noise signals that can appear as common mode voltages on a network. RS-422 and RS-485 are the differential data communication buses. It was constructed by multipoint network, so that it can communicate data rates up to 100 BPS and distance up to 4000 ft.

9.10.1 RS-232 Interface

RS-232, expanded as Recommended Standard-232, is a standard interface approved by the EIA for the purpose of connecting serial devices. RS-232 is a long-established standard that describes the physical interface and protocol used for relatively low-speed serial data communication between computers and related devices. It is used in computers as an interface to have a communication to talk and exchange data with modems and other serial devices.

The RS-232C standard, a subset of RS-232, specifies a maximum baud rate of 20,000 bits per second and short distances up to 50 feet. There are two sizes frequently used in serial communication. They are D-Type 25 pin connector and the D-Type 9 pin connector, both of which are male on the back of the PC. Figure 9.9 shows the two sizes of RS-232C.

In the DB-25 connector, a minimum number of pins is used and most of the pins are not needed for normal PC communications. Nowadays, PCs are equipped with male D type connectors having only 9 pins. Using a 25- pin B-25 or 9-pin DB-9 connector, its normal cable limitation of 50 feet can be extended to several hundred feet with high-quality cable. There is a standardised pinout for RS-232 on a DB9 and DB25 connectors, as shown in Table 9.5.



(a) DB 9 male connector

(b) DB 25 male connector

Fig. 9.9 Two sizes of RS-232 interface

Table 9.5 Pin configurations of DB-9 and DB 25 male connector

Pin Number		Signal	Signal Description
DB-9	DB-25		
1	8	DCD	Data carrier detect
2	3	RxD	Receive Data
3	2	TxD	Transmit Data
4	20	DTR	Data terminal ready
5	7	GND	Signal ground
6	6	DSR	Data set ready
7	4	RTS	Ready to send
8	5	CTS	Clear to send
9	22	RI	Ring Indicator

The following are the specifications of the pins of RS-232C male connectors.

- **DCD** Data Carrier Detect (DCD) indicates that carrier for the transmit data is ON.
- **RXD** This pin carries data from the serial device to the computer.
- **TxD** This pin carries data from the computer to the serial device.
- **DTR** DTR is used by the computer to signal that it is ready to communicate with the serial device like modem.

- **DSR** Data Set Ready (DSR) is an indication from the Dataset that it is ON.
- **RTS** This pin is used to request clearance to send data to a modem.
- **CTS** This pin is used by the serial device to acknowledge the computer's RTS signal. In most situations, RTS and CTS are constantly on throughout the communication session.
- **CD (Carrier Detect)** Carrier detect is used by a modem to signal that it has a made a connection with another modem, or has detected a carrier tone. In other words, this is used by the modem to signal that a carrier signal has been received from a remote modem.
- **RI (Ring Indicator)** A modem toggles the state of this line when an incoming call rings your phone. It is used by an auto-answer modem to signal the receipt of a telephone ring signal

  The Carrier Detect (CD) and the Ring Indicator (RI) lines are only available in connection to a modem. Because most modems transmit status information to a PC when either a carrier signal is detected or when the line is ringing, these two lines are rarely used.
- **Clock signals (TC, RC, and XTC)** The clock signals are only used for synchronous communications. The modem or DSU extracts the clock from the data stream and provides a steady clock signal to the DTE.

RS-232 has some limitations as an electrical interface. They are listed below.

1. The interface presupposes a common ground between the DTE and DCE. This is a reasonable assumption where a short cable connects a DTE and DCE in the same room, but with longer lines and connections between devices that may be on different electrical buses.
2. A signal on a single line is impossible to reduce the noise. By screening the entire cable, it is possible to reduce the influence of noise, but internally generated noise remains a problem. As the baud rate and line length increase, the effect of capacitance between the cables introduces serious crosstalk until a point is reached where the data itself is unreadable.
3. Crosstalk can be reduced by using a low-capacitance cable. Control of slew rate in the signal also decreases the crosstalk. But the original specifications for RS-232 had no specification for maximum slew rate.
4. Voltage levels with respect to ground represent the RS-232 signals. There is a wire for each signal, together with the ground signal. This interface is useful for point-to-point communication at slow speeds. Due to the way the signals are connected, a common ground is required.
5. RS-232 was designed for communication of local devices and it supports one transmitter and one receiver.

9.10.2 RS-422 Interface

When communicating at high data rates, or over long distances in real-world environments, single-ended methods are often inadequate. To meet the above requirements, EIA has released a new serial interface, RS-422.

This standard was designed for high-speed communication and for greater distances such as 4000 feet. Its baud rate is higher than RS-232 such as 100 Kbits/second. In its simplest form, a pair of converters from RS-232 to RS-422 can be used to form an RS-232 extension cord.

RS-422 devices cannot be used to construct a truly multipoint network. A true multipoint network consists of multiple drivers and receivers connected on a single bus, where any node can transmit or receive data.

9.10.3 RS-485 Interface

RS-485 is a specially designed standard by Electronics Industry Association (EIA) for multipoint communications. It supports several types of connectors, including DB-9 and DB-37. RS-485 is similar to RS-422 but can support more nodes per line and it meets the requirements for a truly multipoint communications network. This standard specifies up to 32 drivers and 32 receivers on a single bus.

RS-485 specifies bi-directional, half-duplex data transmission, and is the only EIA standard that allows multiple receivers and drivers in "bus" configurations whereas RS-422 specifies a single, unidirectional driver with multiple receivers.

9.10.4 Comparison between RS-232, RS-422 and RS-485

Table 9.6 shows the comparison between the specifications of RS-232, RS-422 and RS-485.

Table 9.6 Specifications of RS-232, RS-422 and RS-485

Specifications	RS-232 Interface	RS-422 Interface	RS-485 Interface
Mode of operation	Single-ended	Differential	Differential
Allowed no. of Tx and Rx	1 Tx, 1 Rx	1 Tx, 10 Rx	32 Tx, 32 Rx
Maximum cable length	50 feet	4000 feet	4000 feet
Maximum data rate	20 kbps	100 kbps/10 mbps	100 kbps/10 mbps
Minimum driver output range	±5 V to ±15 V	±2 V	±1.5 V
Maximum driver output range	±25 V	±6 V	±6 V
Tx load impedance (Ohms)	3 k to 7 k	100	54
Rx input sensitivity	±3 V	±200 mV	±200 mV
Rx input voltage range	±15 V	±7 V	–7 V to +12 V
Maximum Rx input resistance (Ohms)	3k to 7k	min 4k	min > = 12k

9.11 PARALLEL INTERFACE

A parallel interface is used for having communication between a computer and an external device such as printer. This parallel port uses a 25-pin connector, DB-25 to connect printers, computers and other devices which need high bandwidth.

Parallel ports, offered by IBM Centronics, are mainly used to connect the following computer peripherals.

1. Printers
2. Scanners
3. CD burners
4. External hard drives
5. Zip removable drives
6. Network adapters
7. Tape backup drives

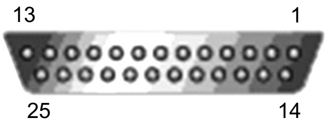


Fig. 9.10 A parallel port by Centronics

Figure 9.10 shows a special cable developed by Centronics to connect the printer to the computer.

When a computer sends data to a printer or other device using a parallel port, it sends 8 bits of data at a time. These 8 bits are transmitted parallel to each other whereas in serial port, the eight bits are being transmitted serially. A standard parallel port is capable of sending 50 to 100 kilobytes of data per second.

Table 9.7 shows given pin configuration of DB-25 pin connector of the parallel port.

Table 9.7 Pin configuration of DB-25 pin connector

Pin	Signal
1	Strobe
2	Data0
3	Data1
4	Data2
5	Data3
6	Data4
7	Data5
8	Data6
9	Data7
10	Acknowledge
11	Busy
12	Paper end
13	Select
14	Auto feed
15	Error
16	Init
17	Select in
18	GND
19	GND
20	GND
21	GND
22	GND
23	GND
24	GND
25	GND

The following are the specifications of the pins of DB-25 parallel port connector.

- Pin 1 carries the strobe signal. It maintains a level of between 2.8 and 5 volts, but drops below 0.5 volts whenever the computer sends a byte of data. This drop in voltage tells the printer that data is being sent.
- Pins 2 through 9 are used to carry data. To indicate that a bit has a value of 1, a charge of 5 volts is sent through the correct pin. No charge on a pin indicates a value of 0. This is a simple but highly effective way to transmit digital information over an analog cable in real time. Pin 10 sends the acknowledge signal from the printer to the computer. Like Pin 1, it maintains a charge and drops the voltage below 0.5 volts to let the computer know that the data was received.
- If the printer is busy, it will charge Pin 11. Then, it will drop the voltage below 0.5 volts to let the computer know it is ready to receive more data.
- The printer lets the computer know if it is out of paper by sending a charge on Pin 12.
- As long as the computer is receiving a charge on Pin 13, it knows that the device is online.
- The computer sends an autofeed signal to the printer through Pin 14 using a 5-volt charge.
- If the printer has any problems, it drops the voltage to less than 0.5 volts on Pin 15 to let the computer know that there is an error.
- Whenever a new print job is ready, the computer drops the charge on Pin 16 to initialise the printer.
- Pin 17 is used by the computer to remotely take the printer offline. This is accomplished by sending a charge to the printer and maintaining it as long as you want the printer offline.
- Pins 18-25 are grounds and are used as a reference signal for the low (below 0.5 volts) charge.

9.12 DATA COMMUNICATION NETWORKS

Data communication network is defined as any group of computer terminals connected together and the process of sharing resources between computer terminals over a data communication network is called **networking**. In other words, networking is two or more computer terminals linked together by means of a common transmission medium for the purpose of sharing the information or data.

The number of links L required between N nodes is calculated as follows.

$$L = \frac{N(N-1)}{2} \tag{9.3}$$

The most important considerations of a data communication network are performance, transmission rate, reliability and security.

EXAMPLE 9.4

If the number of terminals in a data communication network is 6, calculate the number of required links.

Solution

Number of links $L = \frac{N(N-1)}{2}$

Here, $N = 6$

$$L = \frac{6(6-1)}{2} = 15$$

EXAMPLE 9.5

If the number of links required is 21, how many nodes can be connected together in a data-communication network?

Solution

Number of links $L = \frac{N(N-1)}{2}$

It is given that number of links $L = 21$

$$21 = \frac{N(N-1)}{2}$$

$$42 = N(N-1)$$

$$\therefore N = 7$$

9.13 APPLICATIONS OF NETWORKING

There are various applications running on modern computers applicable from company to company. Depending on the type of application, it is necessary to design the data communication network. The specific application affects how well a network will perform. Each network has a finite capacity. Therefore, network designers and engineers must be aware of the type and frequency of the information traffic on the network. Applications of networking are listed as follows.

1. *Standard Office Applications*: E-mail, file transfer and printing.
2. *High-end Office Applications*: Video imaging, computer-aided drafting, computer-aided design and software development.
3. *Manufacturing Automation*: Process and numerical control.
4. *Mainframe Connectivity*: Personal computers, workstations and terminal support.
5. *Multimedia Applications*: Video conferencing.

There are many factors involved when designing a computer network including the following.

1. Network goals as defined by the organisation.
2. Network security
3. Network uptime requirements
4. Network response time requirements
5. Network and resource costs

9.14 COMPONENTS OF A COMPUTER NETWORK

Communication functions are implemented and controlled by using many hardware (physical) and software (logical) components in a computer network.

The following are the physical components.

1. Computer hardware
2. Front-end processor
3. Terminals
4. Modems, concentrators, multiplexers
5. Transmission media
6. Data switching equipment, etc.

The following are the logical components.

1. Operating system
2. File-management system
3. Communication software
4. Application software, etc.

All these components of a computer network function in a coordinated fashion to realise the functional requirements of meaningful communication between the end systems. Design and implementation of such a system is one of the most complex tasks that humans have ever tried.

Figure 9.11 illustrates the basic components of computer networks. No two computers are the same.

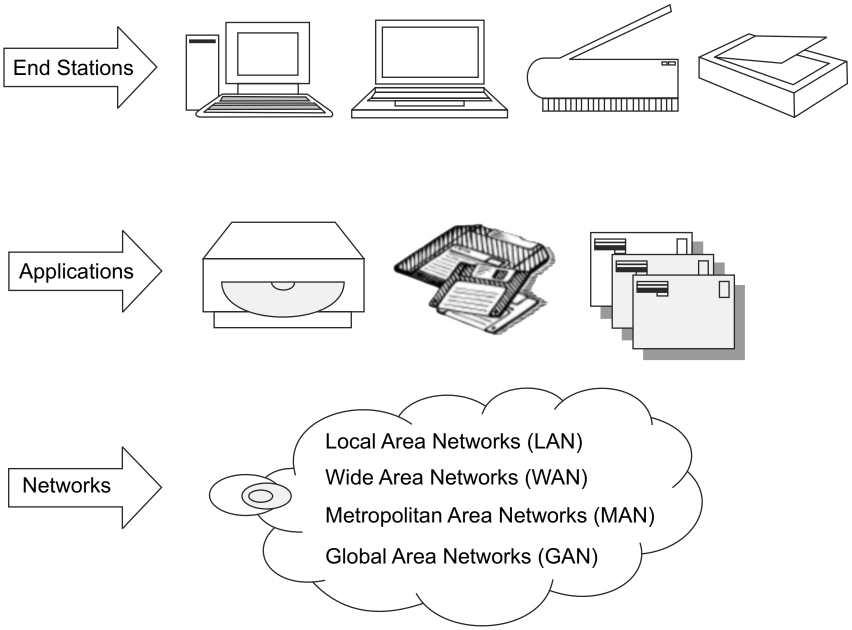

End Stations
Applications
Networks
Local Area Networks (LAN)
Wide Area Networks (WAN)
Metropolitan Area Networks (MAN)
Global Area Networks (GAN)

Fig. 9.11 Basic network components

All the computer networks include some combination of end stations, applications and network that will support the data traffic between the end stations. Computer networks all share common devices, functions and features including servers, clients, transmission media, shared data, shared printers and other peripherals, hardware and software resources, Network Interface Card(NIC), Local Operating System(LOS) and Network Operating System(NOS).

9.15 NETWORK FUNCTIONS

Some important functions and features of various network devices are listed as follows.

9.15.1 Servers

Servers are computers that hold shared files, programs and the network operating system. They provide access to network resources to all the users of the network. They are of different kinds like file servers, print servers, mail servers, communication servers, database servers, fax servers and web servers. One server can provide several functions. Figure 9.12 shows the file-server operation.

A client requests a file from the file server. The file server sends a copy of the file to the requesting user. It allows users to access and manipulate disk resources stored on the other computers. File servers have the following characteristics.

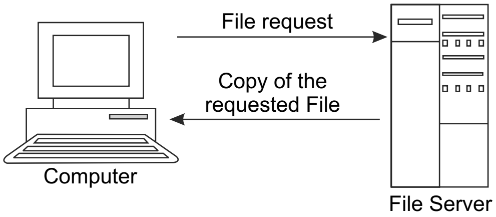


Fig. 9.12 File-server operation

1. File servers are loaded with files, accounts and a record of the access rights of users or groups of users on the networks.
2. The server provides a shareable virtual disk to the clients.
3. File mapping schemes are implemented to provide the virtualness of the files (i.e. the files are made to look like they are on the client's computers).
4. Security systems are installed and configured to provide the server with the required security and protection for the files.
5. Redirector or shell-software programs located on the user's computers transparently activate the client's software on the file server.

9.15.2 Clients

Clients are computers that access and use the network and shared network resources. Client computers are basically the customers of the network, as they request and receive services from the servers.

9.15.3 Transmission Media

Transmission media are the facilities used to interconnect computers in a network, such as twisted-pair wire, coaxial cable and optical fiber cable. Transmission media are sometimes called channels, links or lines.

9.15.4 Shared Data

Shared data are data that file servers provide to clients such as data files, printer access programs and email.

9.15.5 Shared Printers and Other Peripherals

Shared printers and peripherals are hardware resources provided to the users of the network by servers. Resources provided include data files, printers, software or any other items used by clients on the network.

9.15.6 Network Interface Card (NIC)

Each computer in a network has a special expansion card called Network Interface Card (NIC). The NIC prepares and sends data, receives data and controls data flow between the computer and the network. On the transmit side, the NIC passes frames of data on to the physical layer which transmits the data to the physical link. On the receiver side, the NIC processes bits received from the physical layer and processes the message based on its contents. NICs have the following characteristics.

1. The NIC constructs, transmits, receives and processes data to and from the computer and connected network.
2. Each device connected to a network must have a NIC standard.
3. A NIC is generally installed in a computer as a daughterboard, although some computer manufacturers incorporate the NIC into the motherboard during manufacturing.
4. Each NIC has a unique six-byte media access protocol address, which is typically permanently burned into the NIC when it is manufactured.
5. The NIC must be compatible with the network to operate properly.
6. NICs manufactured by different vendors vary in speed, complexity, manageability and cost.
7. The NIC requires drivers to operate on the network.

9.15.7 Local Operating System (LOS)

A Local Operating System (LOS) allows computers to access files, print to a local printer and have and use one or more disk and CD drives that are located on the computer. Some examples of LOSs are MS-DOS, PC-DOS, Unix, Windows 95, Windows 98, Windows 2000 and Linux. Figure 9.13 shows the relationship between a computer and its LOS.

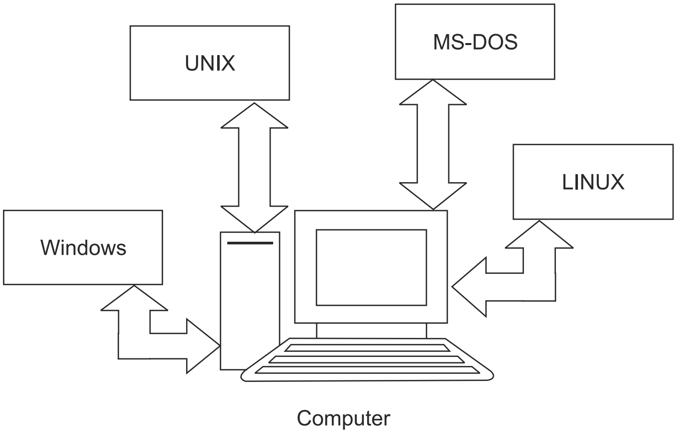


Fig. 9.13 Local operating system

9.15.8 Network Operating System (NOS)

A Network Operating System (NOS) is a program that runs on computers and servers that allows the computers to communicate over a network. The NOS provides services to clients such as log-in features, password authentication, printer access, network administration functions and data file sharing. Some of the popular network operating systems are Unix, Novell Netware, AppleShare, IBM LAN Server, Compaq Open VMS and Microsoft Windows NT server.

The NOS is software that makes communications over a network more manageable. The relationship between clients, servers and the NOS is shown in Figure 9.14.

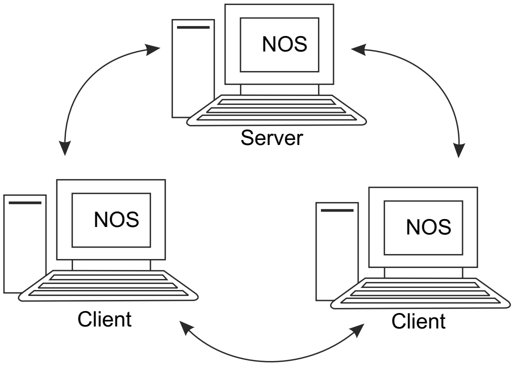


Fig. 9.14 Network operating system

NOSs have the following characteristics.

1. A NOS allows users of a network to interface with the network transparently.
2. A NOS commonly offers the following services: file service, print service, mail service, communications service, database service and directory and security services.
3. The NOS determines whether data are intended for the user's computer or whether the data needs to be redirected out onto the network.
4. The NOS implements client software for the user which allows them to access servers on the network.

9.16 NETWORK MODELS

Computer networks can be represented with two basic network models.

1. Peer-to-peer client/server
2. Dedicated client/server

The client/server method specifies the way in which two computers can communicate with software over a network. Although clients and servers are generally shown as separate units, they are often active in a single computer but not at the same time. With client/server concept, a computer acting as a client initiates a software request from another computer acting as a server. The server computer responds and attempts to satisfy the request from the client. The sever computer might then act as a client and request services from another computer. The client/server is shown in Figure 9.15.

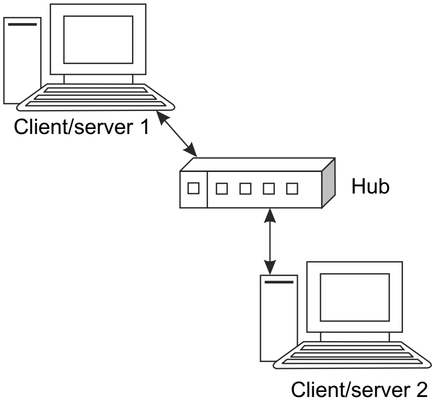


Fig. 9.15 Client/server concept

9.16.1 Peer-to-Peer Client/Server Network

A peer-to-peer client/server network is one in which all computers share their resources such as hard drives, printers, and so on, with all the other computers on the network. Therefore, the peer-to-peer operating system divides its time between servicing the computer on which it is loaded and serving request from other computers. In a peer-to-peer network, there are no dedicated servers among the computers.

Figure 9.16 shows a peer-to-peer client/server network with four clients/servers connected together through a hub.

All computers are equal, hence the name *peer*. Each computer in the network can function as a client and/or a server and no single computer holds the network operating system or shared files. Also, no one computer is assigned network administrative tasks. The users at each computer determine which data on their computer are shared with the other computers

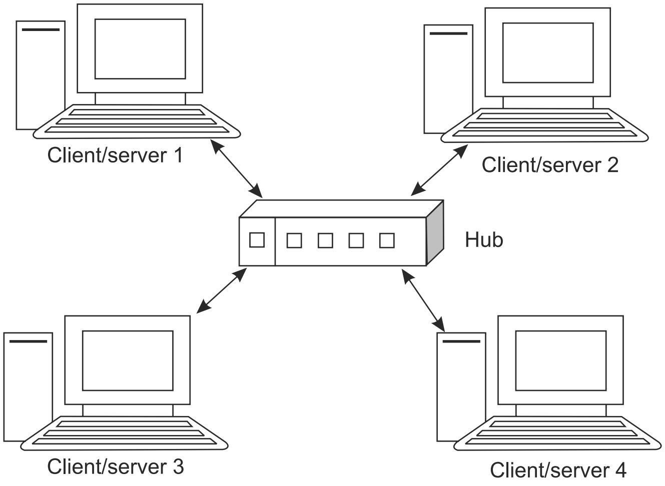


Fig. 9.16 A peer-to-peer client/server network

on the network. Individual users are also responsible for installing and upgrading the software on their computer. Since there is no central controlling computer, a peer-to-peer network is an appropriate choice when there are lesser than 10 users on the network.

Peer-to-peer networks should be small for the following reasons.

1. When operating in the server role, the operating system is not optimised to efficiently handle multiple simultaneous requests.
2. The end user's performance as a client would be regarded.
3. Administrative issues such as security, data backups and data ownership may be compromised in a large peer-to-peer network.

9.16.2 Dedicated Client/Server Network

In a dedicated client/server network, one computer is designated as server and the rest of the computers are clients. As the network grows, additional computers can be designated as servers. Generally, the designated servers function only as servers and are not used as a client or workstation. The servers store all the network's shared files and application programs such as word-processor documents, compilers, database applications, spreadsheets and the network operating system. Client computers can access the servers and have shared files transferred to them over the transmission medium.

Figure 9.17 shows a dedicated client/server based network with three servers and three clients. Each client can access the resources on any one of the servers and also the resources

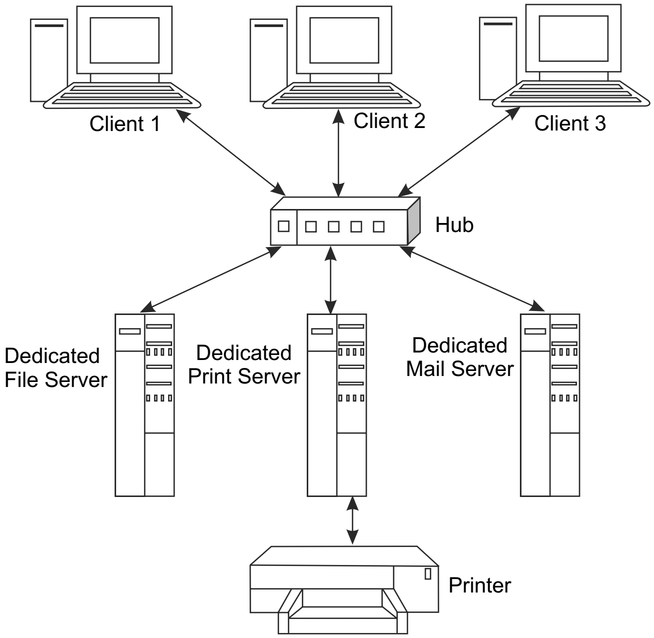


Fig. 9.17 A dedicated client/server based network

on other client computers. The dedicated client/server based network is probably the most commonly used computer networking model. There can be a separate dedicated server for each function or one single general purpose for all services.

In some client/server networks, client computers submit jobs to one of the servers. The server runs the software and completes the job and then sends the results back to the client computer. In this type of network, less information propagates through the network than with the file-server configuration because only data and not application programs are transferred between computers.

The dedicated client/server model is preferable to the peer-to-peer client/server model for general-purpose data networks. The peer-to-peer model client/server model is usually preferable for special purposes such as a small group of users sharing resources.

9.17 NETWORK TOPOLOGIES

The term topology refers to the way a network is laid out, either physically or logically. Two or more devices connect to a link and two or more links form a topology. The topology of a network is the geometric representation of the relationship of all the links and linking devices, usually called **nodes**, to each other. Topology is a major consideration for capacity, cost and reliability when designing a data communications network. There are two basic topologies.

1. Point-to-point line configuration
2. Multipoint configuration

9.17.1 Point-to-Point Line Configuration

A point-to-point line configuration is used in data communication networks that transfer high-speed digital information between only two stations. Point-to-point data circuits involve communications between a mainframe computer and another mainframe computer or some other type of high-capacity digital device. A two-point circuit is shown in Figure 9.18.

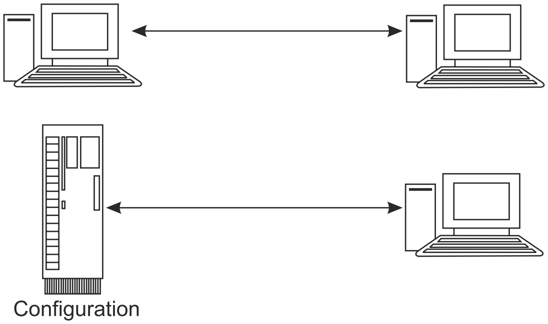


Fig. 9.18 Point-to-point configuration

9.17.2 Multipoint Configuration

A multipoint, also called **multidrop**, configuration is one in which more than two specific devices share a single link. Figure 9.19 shows multipoint configuration.

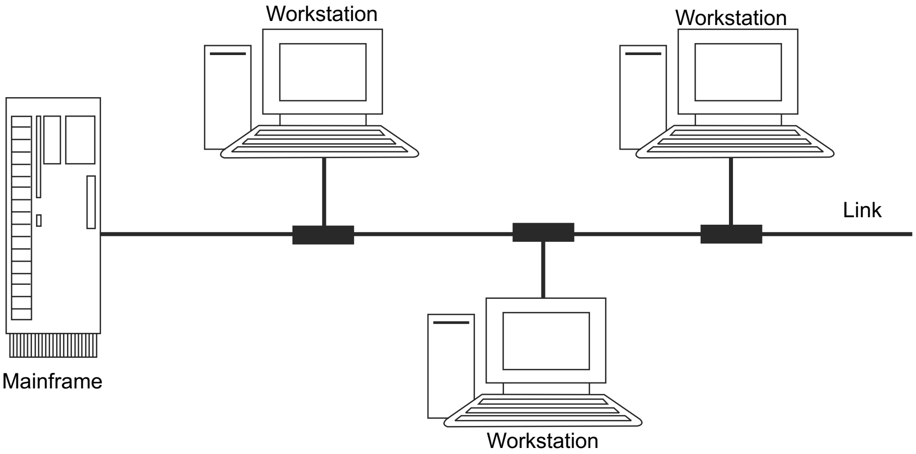


Fig. 9.19 Multipoint configuration

In multipoint configuration, the capacity of the channel is shared, either spatially or temporally. If several devices can use the link simultaneously, it is a spatially shared line configuration. If users must take turns, it is time-shared line configuration.

Examples of multipoint topologies are as follows.

1. Mesh topology
2. Star topology
3. Tree topology
4. Bus topology
5. Ring topology
6. Hybrid type

1. Mesh Topology

In a mesh topology, every device has a dedicated point-to-point link to every other device. The term *dedicated* means that the link carries traffic only between the two devices it connects. A fully connected mesh network,

therefore, has $\frac{n(n-1)}{2}$ physical channels to link *n* devices. To accommodate the many channels, every device on the network must have *n* – 1 input/output ports. Figure 9.20 shows a fully connected mesh topology.

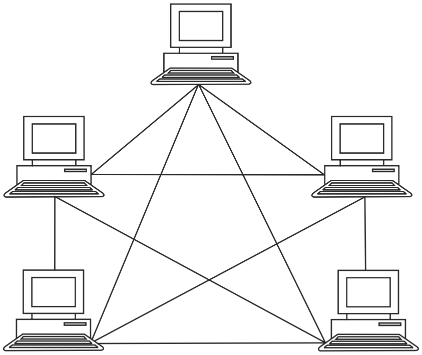


Fig. 9.20 A fully connected mesh topology

A mesh offers several advantages over other network topologies.

1. The use of dedicated links guarantees that each connection can carry its own data load and thus eliminates the traffic problems that can occur when links must be shared by multiple devices.
2. Mesh topology is robust. If one link becomes unusable, it does not incapacitate the entire system.
3. Privacy and security is available in mesh topology. When every message sent travels along a dedicated line, only the intended recipient sees it. Physical boundaries prevent other users from gaining access to messages.
4. Point-to-point links make fault identification and fault isolation easy. Traffic can be routed to avoid links with suspected problems. This facility enables the network manager to find the precise location of the fault and aids in finding its cause and solution.

The following are the disadvantages of the mesh topology.

1. There is a need of amount of cabling and the number of input/output ports required.
2. Since every device must be connected to every other device, installation and reconfiguration is difficult.
3. The sheer bulk of the wiring can be greater than the available space can accommodate.
4. The hardware required to connect each link can be prohibitively expensive.

2. Star Topology

In a star topology, each device has a dedicated point-to-point link only to a central controller, usually called a **hub**. The devices are not directly linked to each other. Unlike a mesh topology, a star topology does not allow direct traffic between devices. The controller acts as an exchange. If one device wants to send data to another, it sends the data to the controller, which then relays the data to the other connected device. Figure 9.21 shows star configuration.

A star topology is less expensive than a mesh topology. In a star, each device needs only one link and one input/output port to connect it to any number of others. This makes it easy to

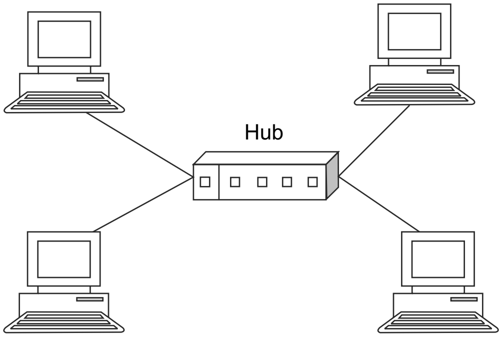


Fig. 9.21 Star configuration

install and reconfigure. Far less cabling needs to be housed and additions, moves and deletions involve only one connection between that device and the hub.

Other advantages include robustness. If one link fails, only that link is affected. Other links remain active. This factor also lends itself to easy fault identification and fault isolation. As long as the hub is working, it can be used to monitor link problems and bypass defective links.

However, although a star requires far less cable than a mesh, each node must be linked to a central hub. For this reason, more cabling is required in a star configuration.

3. Tree Topology

A tree topology is a variation of a star. As in a star, nodes in a tree are linked to a central hub that controls the traffic to the network. The majority of the devices connect to a secondary hub and the turn is connected to the central hub. Figure 9.22 shows the tree configuration.

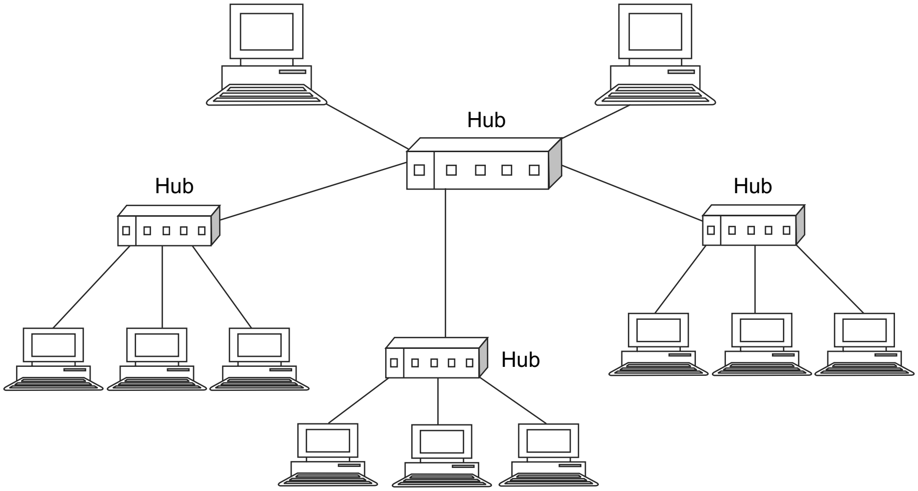


Fig. 9.22 Tree configuration

The central hub in the tree is an active hub. An active hub contains a repeater which is a hardware device that generates the received bit patterns before sending them out. Repeating strengthens transmissions and increases the distance a signal can travel. The secondary hubs may be active or passive hubs. A passive hub provides a simple physical connection between the attached devices.

The advantages and disadvantages of a tree topology are generally the same as those of a star. The addition of secondary hubs, however, brings two further advantages.

(a) It allows more devices to be attached to a single central hub and can, therefore, increase the distance a signal can travel between devices.

(b) It allows the network to isolate and prioritise communications from different computers. For example, the computers attached to one secondary hub can be given priority over computers attached to another secondary hub. In this way, the network designers and operator can guarantee that time-sensitive data will not have to wait for access to the network.

A good example of tree topology can be seen in cable TV technology where the main cable from the main office is divided into main branches and each branch is divided into smaller branches, and so on. The hubs are used when a cable is divided.

4. Bus Topology

A bus topology is a multipoint data communication circuit that makes it relatively simple to control data flow between and among the computers because the configuration allows all stations to receive every transmission over the network. Figure 9.23 shows bus topology.

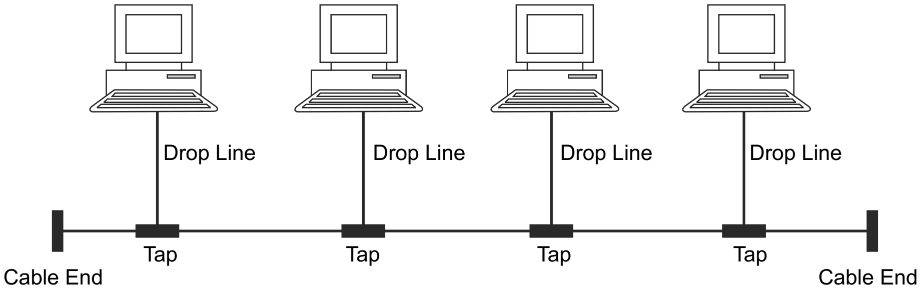


Fig. 9.23 Bus topology

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

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

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

In addition, a fault or break in the bus cable stops all transmission, even between devices on the same side of the problem. The damaged area reflects a signal back in the direction of origin, creating noise in both directions.

5. Ring Topology

In a ring topology, each device has a dedicated point-to-point line configuration only with the two devices on either side of it. A signal is passed along the ring in one direction from device to device, until it reaches its destination. Each device in the ring incorporates a repeater. When a device receives a signal intended for another device, its repeater regenerates the bits and passes them along. Figure 9.24 shows ring topology.

A ring is relatively easy to install and reconfigure. Each device is linked only to its immediate neighbours. To add or delete a device requires moving only two connections. The

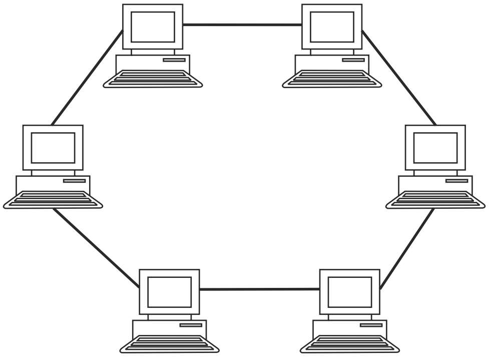


Fig. 9.24 Ring topology

only constraints are media and traffic considerations like maximum ring length and number of devices. In addition, fault isolation is simplified. Generally in a ring, a signal is circulating at all times. If one device does not receive a signal within a specified period, it can issue an alarm. The alarm alerts the network operator to the problem and its location.

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

6. Hybrid Topology

A hybrid topology is simply combining two or more of the traditional topologies to form a larger, more complex topology. Hybrid topologies are sometimes called mixed topologies. For example, one department has a ring. The two can be connected to each other via a central controller in a star topology. An example of a hybrid topology is the star ring topology is shown in Figure 9.25.

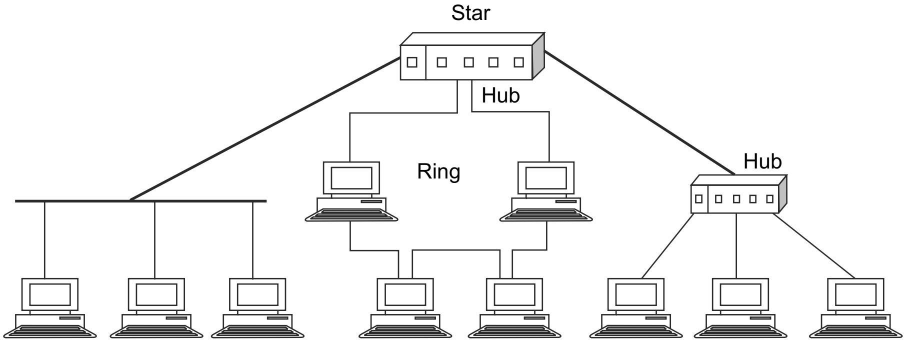


Fig. 9.25 Hybrid topology

9.18 NETWORK CLASSIFICATIONS

Networks are classified based on size which includes geographic area, distance between stations, number of computers, transmission speed, transmission media and the network's physical architecture. The four types of classification of networks are as follows.

1. Local Area Networks (LAN)
2. Metropolitan Area Networks (MAN)
3. Wide Area Networks (WAN)
4. Global Area Networks (GAN)

There are two future computer networks also to be considered.

1. Personal Area Network (PAN) whose idea is to allow people to transfer data through the human body simply by touching each other.
2. Power Line Area Network (PLAN) use existing ac distribution networks to carry data wherever power line go, which is virtually everywhere.

9.18.1 Local Area Networks (LAN)

Local Area Networks (LAN) are typically privately owned data communications networks in which 10 to 40 computer users share data resources with one or more file servers. LANs use a network operating system to provide two-way communications at bit rates typically in the range of 10 Mbps to 100 Mbps and higher between a larger variety of data communications equipment within a relatively small geographical area such as in the same room, building or building complex.

A LAN can be as simple as two personal computers and a printer or could contain dozens of computers, workstations and peripheral devices. Most LANs link equipment that are within a few miles of each other or closer. Because the size of most LANs is limited, the longest transmission time is bounded and known by everyone using the network. Therefore, LANs can utilise configurations that otherwise would not be possible.

LANs were designed for sharing resources between a wide range of digital equipment including personal computers, workstations and printers. The resources shared can be software as well as hardware.

9.18.2 Metropolitan Area Networks (MAN)

A Metropolitan Area Networks (MAN) is a high-speed network similar to a LAN except MANs are designed to encompass larger areas, usually that of an entire city. Most MANs support the transmission of both data and voice and in some cases video. MANs typically operate at speeds of 1.5 Mbps to 10 Mbps and range from 5 miles to a few hundred miles in length. A MAN generally uses only one or two transmission cables and requires no switches. A MAN could be a single network, such as a cable television distribution network or it could be a means of interconnecting two or more LANs into a single, larger network, enabling data resources to be shared LAN to LAN as well as from station to station or computer to computer. Large companies often use MANs to interconnect all their LANs.

A MAN can be owned and operated entirely by a single, private company or it could lease services and facilities on a monthly basis from the local cable or Telephone Company. Some examples of MANs are Fibre Distributed Data Interface (FDDI) and Asynchronous Transfer Mode (ATM).

9.18.3 Wide Area Networks (WAN)

Wide Area Networks (WAN) are the oldest type of data communications network that provide relatively slow speed, long-distance transmission of data, voice and video information over relatively large and widely dispersed geographical areas such as country or entire continent. WANs typically interconnect cities and states. WANs typically operate at bit rates from 1.5 Mbps to 2.4 Gbps and cover a distance of 100 to 1000 miles.

WANs may utilise both public and private communications systems to provide service over an area that is virtually unlimited. However, WANs are generally obtained through service providers and normally come in the form of leased-line or circuit-switching technology. Often WANs interconnect routers in different locations. Examples of WANs are Integrated Services Digital Network (ISDN), T1 and T3 digital carrier systems, frame relay, ATM and using data modems over standard telephone lines.

9.18.4 Global Area Networks (GAN)

Global Area Networks (GAN) provide connects between countries around the entire globe. The Internet is a good example of a GAN, as it is essentially a network comprised of other networks that interconnect virtually every country in the world. GANs operate from 1.5 Mbps to 100 Gbps and cover thousands of miles.

9.19 LAYERED ARCHITECTURE OF A COMPUTER NETWORK

Decomposition of the organisation into offices and each office into hierarchical functional levels and the interaction procedures define the overall organisation architecture. A computer network is also partitioned into end systems interconnected using a subnetwork and the communication process is decomposed into hierarchical functional layers. Figure 9.26 shows the layered architecture of an end system.

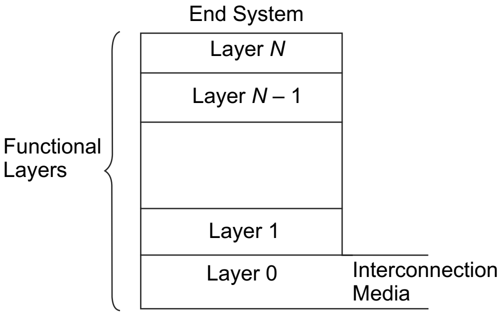


Fig. 9.26 Layered architecture of an end system

Similar to an office, each layer has a distinct identity and a specific set of functions assigned to it. Each layer has an active element, a piece of hardware or software, which carries out the layer functions. It is called **layer entity**.

The general criteria for defining the boundaries of a layer are as follows.

1. Each function is distinctly identified and implemented precisely in one layer.
2. Sequentiality of the functions is ensured by proper design of the hierarchy.
3. Number of layers should be minimum.
4. Boundaries of a layer are defined taking into consideration the existing acceptable implementation.
5. The implementation details of a function are hidden so that any change in the implementation does not affect other layers.

9.19.1 Functionality of the Layered Architecture

The layered architecture emphasises that there is hierarchy of functions. Each layer provides certain services to the next higher layer which uses these services to carry out its assigned functions. Each layer also needs to interact with the peer layer of another end system or the subnetwork to carry out its functions. Since there is no direct path between peer layers, they have interaction using services of the lower layers. Therefore, two types of communication take place in the layered architecture to make it work properly. They are

1. Hierarchical communication
2. Peer-to-peer communication

Figure 9.27 shows hierarchical and peer-to-peer communications in the functional layers.

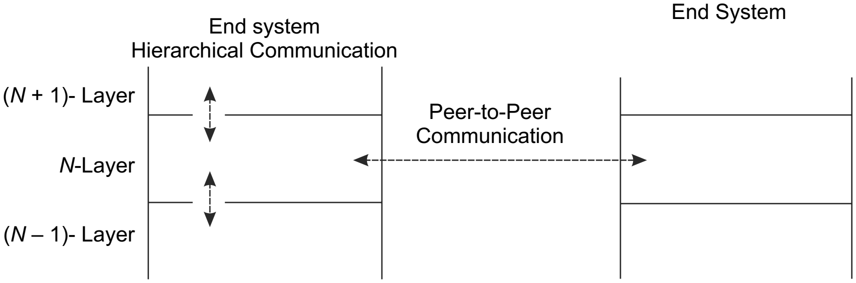


Fig. 9.27 Hierarchical and peer-to-peer communications in the functional layers

1. Hierarchical Communication

Hierarchical communication between adjacent layers of a system is for requesting and receiving services from the lower layer. The rules and procedures for hierarchical communication are specified by the service interface definition which consists of description of the services provided by the lower layer and rules, procedure and parameters for requesting and utilising the services.

The messages exchanged between the adjacent layers during hierarchical communication are called Interface Control Information (ICI). Figure 9.28 shows Interface Control Information for hierarchical communication.

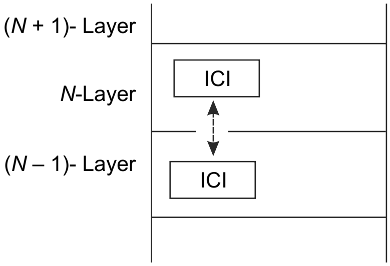


Fig. 9.28 Interface control information for hierarchical communication

2. Peer-to-Peer Communication

Peer-to-peer communication is between the peer layers for carrying out an assigned set of functions. Rules and procedures for peer-to-peer communication are called **protocol**. The messages which are exchanged between the peer layers are called Protocol Control Information (PCI). Figure 9.29 shows protocol control information for peer-to-peer communication.

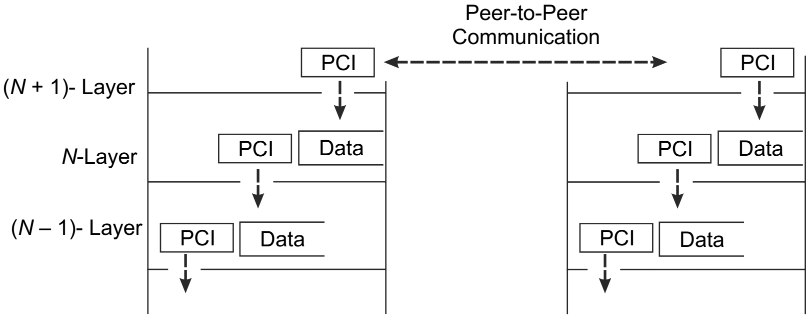


Fig. 9.29 Protocol control information for peer-to-peer communication

Since there is no direct path between the peer layers, protocol control information is exchanged using the services provided by the lower layer. From the above figure, whatever a layer receives from the layer above is treated as 'data'. PCI is added as header to the data and handed over to the layer below. The PCI may consist of one or several fields. The process is repeated at each layer. At the other end, each layer strips off the PCI of its peer layer from the received block of data and hands over the remaining block to the layer above.

9.19.2 Need for Standardisation of Network Architecture

The layered architecture was made into many systems but different vendors defined proprietary protocols and interfaces. The layer partitioning also did not match. As a result, there was total integration incompatibility of architectures developed by different vendors. Standardisation of network architecture can solve these problems and save a lot of effort required for developing interfaces for networking different architectures. Some of the important network architectures are as follows.

1. IBM's System Network Architecture (SNA)
2. Digital's Digital Network Architecture (DNA)
3. Open System Interconnection (OSI) Reference Model developed by ISO (International Organisation for Standardisation)

Among the above, SNA and DNA are vendor-specific layered architectures while the OSI model has been accepted as international standard.

9.20 OPEN SYSTEM INTERCONNECTION (OSI)

Open System Interconnection (OSI) is the name for a set of standards for communicating among computers. The main purpose of OSI standards is to serve as a structural guideline for exchanging information between computers, workstations and networks.

The OSI architecture decomposes the communication process into hierarchical functional layers and identifies the standards necessary for open-system interconnection. It does not specify the standards but provides a common basis for coordination of standards development. The OSI architecture is, therefore, called Reference Model for Open System Interconnection.

9.21 LAYERED ARCHITECTURE OF THE OSI REFERENCE

An ISO standard that covers all aspects of network communications is the Open System Interconnection model. It is a layered framework for the design of network systems that allows for communication across all types of computer systems.

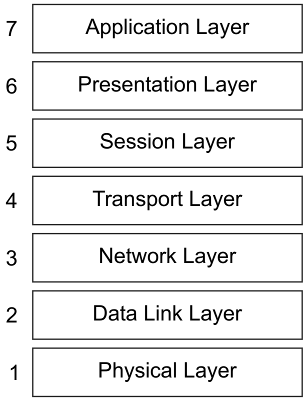


Fig. 9.30 Layered architecture of OSI model

It has seven layers, each of which defines a segment of the process of moving information across a network. The layered architecture is shown in Figure 9.30.

Figure 9.31 shows the layers involved when a message is sent from device A to device B. As the message travels from *A* to *B*, it may pass through many intermediate nodes. These nodes usually involve only the first three layers of the OSI model.

Network functions having related functions are concentrated into distinct layers. Within a single machine, each layer uses the services of the layer just below it. For example, Layer-3 uses the services given by Layer-2 and provides services for Layer-4. Between machines, Layer-*x* on one machine

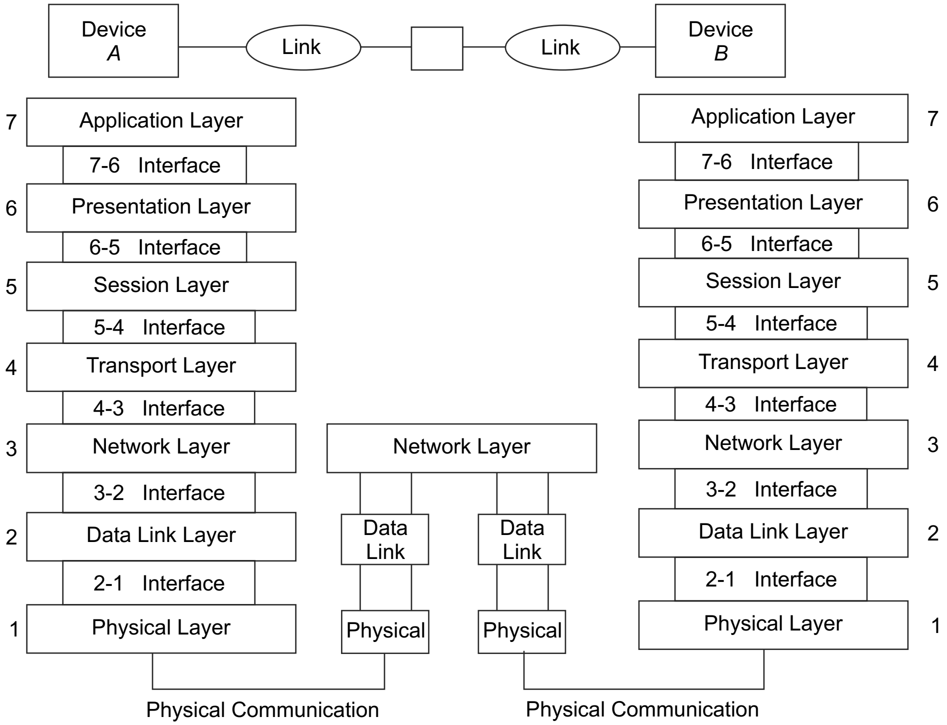


Fig. 9.31 OSI model with seven layers

has communication with Layer-*x* on another machine. The processes on each machine that communicate at a given layer are called peer-to-peer processes.

The passing of the data and network information down through the layers of the sending machine and backup through the layers of the receiving machine is made possible by an interface. Each interface defines what information and services a layer must provide for the layer above it.

Layers 1, 2 and 3 are the network-support layers. They determine the physical aspects of moving data from one device to another. Layers 5, 6 and 7 are called user-support layers. They allow the interoperability between unrelated software systems. Layer-4 ensures end-to-end reliable data transmission while Layer-2 ensures reliable transmission on a single link.

9.21.1 Physical Layer

The physical layer coordinated the functions required to transmit a bit stream over a physical medium. The various functions of the physical layer are as follows.

1. Physical characteristics of interfaces and media.
2. Representation of bits for transmission. Bits must be encoded into signals which may be electrical or optical.

3. The transmission rate or the number of bits sent each second is defined by the physical layer.
4. Synchronisation of bits.
5. Line of configuration which defines the connection of devices to the medium, either point-to-point or the multipoint communication.
6. Physical topology in which how devices are connected to a network, star, mesh, tree, ring or bus topology.
7. Transmission mode which defines the direction of transmission between two devices like simplex, half-duplex or full duplex.

9.21.2 Data-Link Layer

This layer is responsible for node-to-node delivery. The functions of the data-link layer are as follows.

1. *Framing*: It divides the stream of bits received from the network layer into manageable units called frames.
2. *Physical Addressing*: The data-link layer adds a header to the frame to define the physical address of the sender and receiver of the frame.
3. *Flow control* in which data flow is controlled.
4. *Reliable error control* is possible.
5. *Access control* to decide which device has the control over the link.

9.21.3 Network Layer

This layer is responsible for the source-to-destination delivery of a packet across multiple networks. The responsibilities of the network layer are as follows.

1. Logical Addressing If a packet passes the network boundary, an addressing system is needed to help differentiate the source and destination systems. The network layer adds a header to the packet coming from the upper layer that includes the logical addresses of the sender and receiver.

2. Routing When different links are connected to form an internetwork, The connecting devices are called routers which route the packets to their final destination.

9.21.4 Transport Layer

It is responsible for source-to-destination delivery of the entire message. The transport layer ensures that the whole messages arrive intact and in order, overseeing both error control and flow control at the source-to-destination level. The responsibilities of the transport layer are as follows.

1. Service-point Addressing Computers often run several programs at the same time. Hence, source-to-destination delivery means delivery not only from one computer to the next but also from a specific process on the other. The transport layer header, therefore, must include a type of address called a service-point address.

2. Segmentation and Reassembly A message is divided into transmittable segments and each segment has a sequence number. These numbers enable the transport layer to reassemble the message correctly upon arriving at the destination and to identify and replace packets that are lost in transmission.

3. Connection Control It can be either connectionless or connection-oriented. A connectionless transport layer treats each segment as an independent packet and delivers it to the transport layer at the destination. A connection-oriented transport layer makes a connection with the destination transport layer before delivering the packets. After data transfer, the connection is terminated.

4. Flow control and error control are also the responsibilities of transport layer.

9.21.5 Session Layer

The session layer is the network dialog controller. It establishes, maintains and synchronises the interaction between communicating systems. The responsibilities of session layer are as follows.

1. Dialog Control It allows two systems to enter into a dialog communication which is either half-duplex or full duplex.

2. Synchronisation It allows a process to add check points into a stream of data.

9.21.6 Presentation Layer

This layer is concerned with the syntax and semantics of the information exchanged between two systems. The responsibilities of presentation layer are as follows.

1. Translation As different computers use different encoding methods, the presentation layer is responsible for interoperability between these different encoding methods.

2. Encryption For privacy and protecting data, the sender transforms the original information to another form and sends it and it is called encryption. Decryption reverses the original process to transform the message back to its original form.

9.21.7 Application Layer

This layer enables the user to access the network. It provides user interfaces and support for services such as e-mail, remote file access and transfer, shared database management and other types of distributed information systems. Its responsibilities are as follows.

1. Network Virtual Terminal It is a software version of a physical terminal and allows a user to log on to a remote host.

2. File Transfer, Access and Management (FTAM) It allows a user to access files in a remote computer, to retrieve files from a remote computer and to control or manage files.

3. Mail Services It provides the basics for e-mail forwarding and storage.

4. Directory Services It provides distributed database sources and access for global information about various objects and services.

9.22 INTEGRATED SERVICES DIGITAL NETWORK (ISDN)

The Integrated Services Digital Network (ISDN) is a proposed network designed for providing worldwide telecommunication support of voice, data, video and facsimile within the same network. Simply, ISDN is the integrating of a wide range of services into a single multipurpose network. ISDN is a network that proposes to interconnect an unlimited number of independent users through a common communications network.

9.22.1 Principle of ISDN

The main feature of the ISDN concept is to support a wide range of voice (telephone) and nonvoice (digital data) applications in the same network using a limited number of standardised facilities. It supports a wide variety of applications including both switched and nonswitched connections. An ISDN will contain intelligence for the purpose of providing service features, maintenance and network management functions.

9.22.2 Subscriber's Conceptual View of ISDN

Figure 9.32 illustrates how ISDN can be conceptually viewed by a subscriber of the system.

Customers gain access to the ISDN system through a local interface connected to a digital transmission medium called a digital pipe. There are several sizes of pipes available with varying capacities, bit rates, depending on customer need. For example, a residential customer may require only enough capacity to accommodate a telephone and a personal computer. However, an office complex may require a pipe with sufficient capacity to handle a large number of digital telephones interconnected through an on-premise Private Branch Exchange (PBX) or a large number of computers on a LAN.

A single residential telephone is at the low end of the ISDN demand curve, followed by a multidrop arrangement serving a telephone, a personal computer and a home alarm system. Industrial complexes would be at the high end, as they require sufficient capacity to handle hundred of telephones and several LANs. Although a pipe has a fixed capacity, the traffic on the pipe can be comprised of data from a variety of sources with varying signal types and bit rates that have been multiplexed into a single high-capacity pipe. Therefore, a customer can gain access to both circuit and packet-switched services through the same pipe. Due to

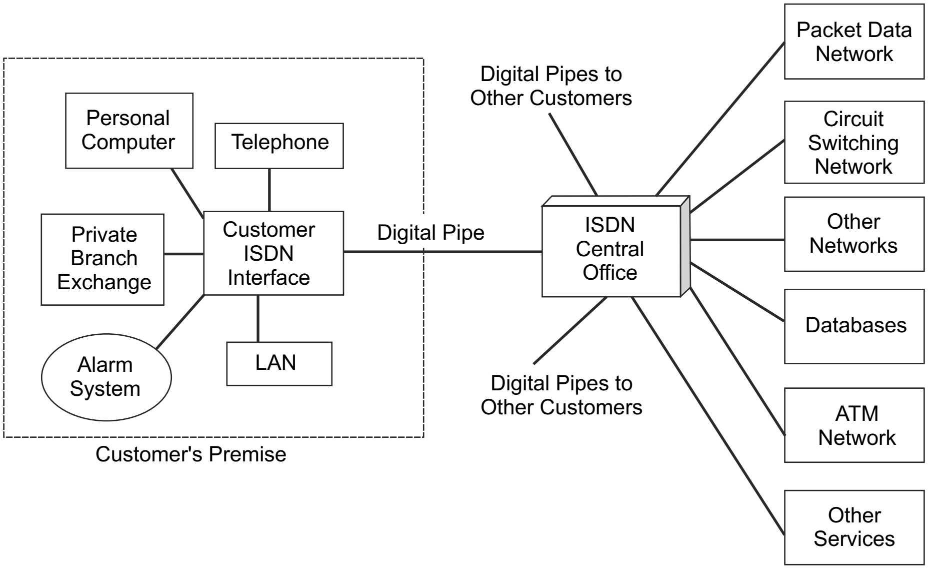


Fig. 9.32 Subscriber's conceptual view of ISDN

the complexity of ISDN, it requires a complex control system to facilitate multiplexing and demultiplexing data to provide the required services.

9.22.3 ISDN Objectives

The basic objectives of ISDN are the following.

1. System Standardisation Ensure universal access to the network.

2. Achieving Transparency Allow customers to use a variety of protocols and applications

3. Separating Functions ISDN should not provide services that preclude competiveness.

4. Variety of Configurations Provide leased and switched services.

5. Addressing Cost-related Tariffs ISDN service should be directly related to cost and independent of the nature of the data.

6. Migration Provide a smooth transition while evolving.

7. Multiplexed Support Provide service to low-capacity personal subscribers as well as to large companies.

9.22.4 ISDN Architecture

The block diagram of the architecture of ISDN functions is shown in Figure 9.33. The ISDN network is designed to support an entirely new physical connector for the customer, a digital subscriber loop and a variety of transmission services.

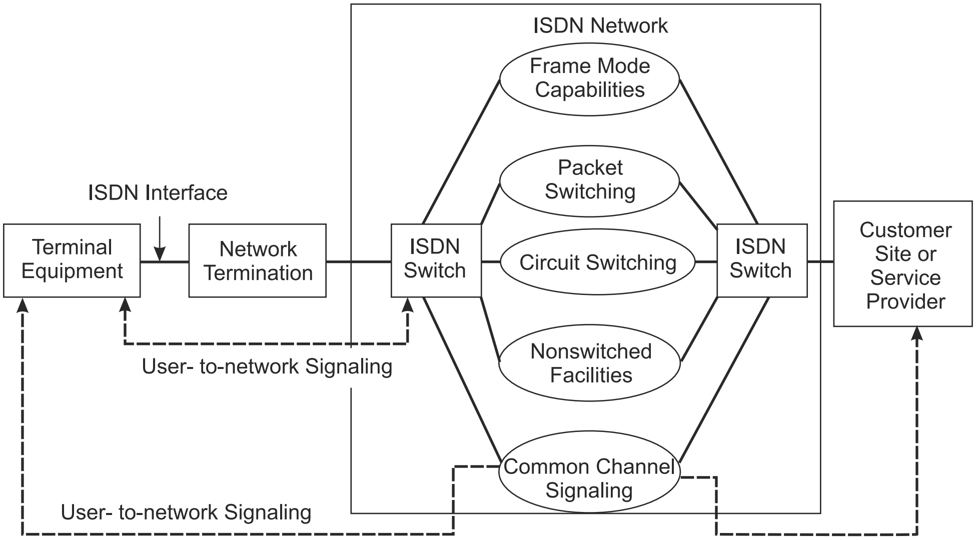


Fig. 9.33 ISDN architecture

A common physical is defined to provide a standard interface connection. A single interface will be used for telephones, computer terminals and video equipment. Therefore, various protocols are provided that allow the exchange of control information between the customer's device and the ISDN network. There are three basic types of ISDN channels.

1. *B channel*: 64 kbps
2. *D channel*: 16 kbps or 64 kbps
3. *H channel*: 384 kbps (H_0), 1536 kbps (H_{11}) or 1920 kbps(H_{12})

ISDN standards specify that residential users of the network be provided a basic access consisting of three full duplex, time division multiplexed digital channels, two operating at 64 kbps (designated the *B* channels) and one at 16 kbps (designated the *D* channel). The *D* channel is used for carrying signalling information and for exchanging network control information. One *B* channel is used for digitally encoded voice and the other for applications such as data transmission. The *H* channels are used to provide higher bit rates for special services such as fast facsimile, video, high-speed data and high-quality audio.

There is another service called the primary service, primary access or Primary Rate Interface (PRI) that will provide multiple 64 kbps channels that can be used by higher-volume subscribers of the network.

The subscriber's loop, as with the twisted pair cable used with a common telephone, provides the physical signal path from the subscriber's equipment to the ISDN central office. The subscriber loop must be capable of supporting full-duplex digital transmission for both basic and primary data rates. In future, metallic cables may be replaced by optic cables.

9.22.5 Broadband ISDN

It is defined as a service that provides transmission channels capable of supporting transmission rates greater than the primary data rate, e.g. video transmission. The working principle is based on Asynchronous Transfer Mode (ATM) that uses optic fibre as transmission medium. The data rates are nearly 11 Mbps, 155 Mbps or 600 Mbps in this service.

The services provided by BISDN are interactive and distribution services. **Interactive services** include those in which there is a two-way exchange of information. Distribution services are those in which data transfer is primarily from service provider to subscribers. Also, conversational services will provide a means for bidirectional end-to-end data transmission in real time between two terminals. Figure 9.34 shows how access to the BISDN network is accomplished.

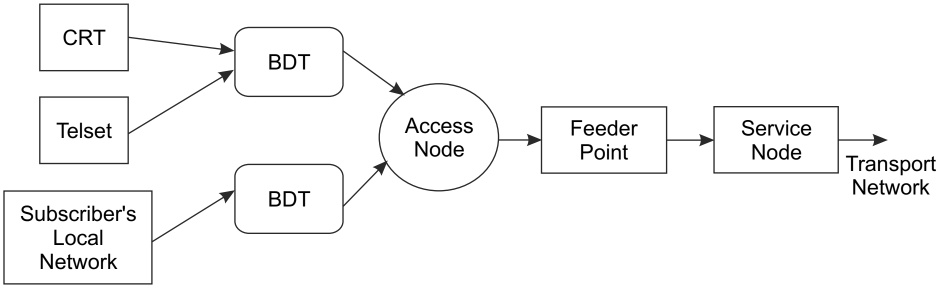


Fig. 9.34 BISDN access

Each peripheral device is interfaced to the access node of BISDN network through a Broadband Distant Terminal (BDT). BDT is responsible for the electrical to optical conversion, multiplexing of peripherals and maintenance of the subscriber's local system. Access nodes concentrate several BDTs into high-speed optical fibre lines directed through a feeder point into a service node. Most of the control functions for system access are managed by the service node such as call processing, administrative functions, and switching and maintenance functions.

The functional modules are interconnected in a star configuration and include switching, administrative, gateway and maintenance modules. The interconnection of the function modules is shown in Figure 9.35.

The central control hub acts as the end-user interface for control signalling and data traffic maintenance. Subscriber terminals near the central office may bypass the access nodes entirely and be directly connected to the BISDN network through a service node. BISDN networks that use optical fibre cables can utilise much wider bandwidths and have higher transmission rates and offer more channel-handling capacity than ISDN systems.

The broadband channel rates are the following.

1. H_{21}: 32.768 Mbps

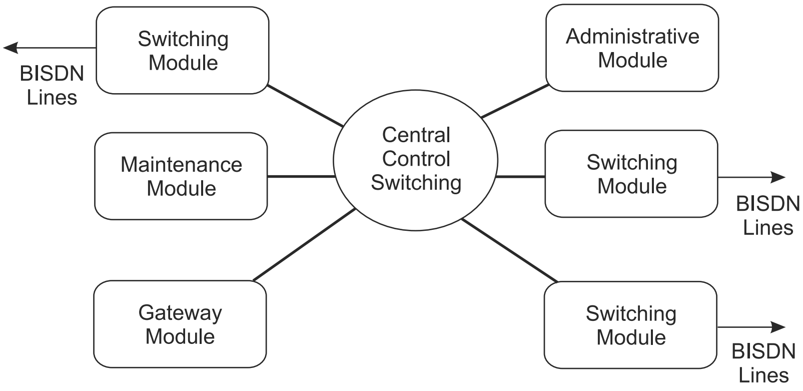


Fig. 9.35 BISDN functional module interconnections

2. H_{22}: 43 Mbps to 45 Mbps
3. H_4: 132 Mbps to 138.24 Mbps

H_{21} and H_{22} are used for full-motion video transmission for video conferencing, video telephone and video messaging. H_4 is used for bulk data transfer of text, facsimile and enhanced video information.

9.23 LOCAL AREA NETWORKS (LAN)

9.23.1 Need for LAN

Local Area Networks (LAN) is a generic term for a network facility spread over a relatively small geographical radius. The LAN concept began with the development of distributed processing. The need of LAN is to interconnect the computers so that the data, software and hardware resources within the premises of an organisation could be shared.

9.23.2 LAN Attributes

A LAN consists of a number of computers, graphic stations and user-terminal stations interconnected through a cabling system. It has the following attributes.

1. Geographic coverage of local area networks is limited to area less than 5 km.
2. The data rate exceed 1Mbps
3. The physical interconnecting medium is privately owned.
4. The physical interconnecting medium is usually shared by the stations connected to the LAN.

9.23.3 Key Elements of LAN

The key elements of a LAN are

1. Topology
2. Transmission medium
3. Layout
4. Medium access control

These elements decide not only the cost and capacity of LAN, but also the type of data to be transmitted, the speed and efficiency of the communications etc.

9.23.4 LAN Topologies

The physical topology of a local area network refers to the way in which the stations are physically interconnected. In the past, LANs were categorised on the basis of physical topology because it also determined the way in which the LANs operated. But today, LANs have the same topology but operating in different ways.

Physical topology of a local area network should have the following desirable features.

1. The topology should be flexible to accommodate changes in physical locations of the stations, increase in the number of stations and increase in the LAN geographical coverage.
2. The cost of physical media and installation should be minimum.
3. The network should not have any single point of complete failure.

Bus topology, ring topology and star topology are common. There can be some other topologies as well such as distributed star, tree, etc. These are extensions of the basic topologies, i.e. bus, ring and star.

The factors that influence the choice of topology are reliability, expandability and performance. There are four.

9.23.5 LAN Transmission Medium

There are four alternative media that can be used for a LAN.

1. Twisted pair
2. Baseband coaxial cable
3. Broadband coaxial cable
4. Optical fibre

The factors that influence the choice of transmission medium include

1. Capacity to support the expected network traffic
2. Reliability to meet requirements for availability
3. Types of data supported
4. Environmental scope

9.23.6 LAN Protocol Architecture

The architecture of a LAN is explained with the layering of protocols that organise the basic function of LAN. It is generally compared with the OSI models, with its lower layers. A diagrammatic scheme of IEEE 8.2 is shown in Figure 9.36.

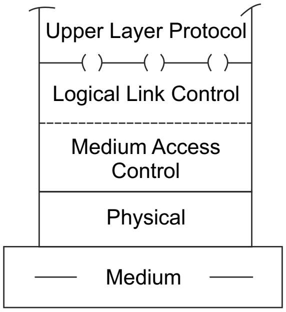


Fig. 9.36 IEEE 802 reference model

The lower layer of the IEEE 802 reference model for LAN corresponds to the physical layer and includes functions like

1. Encoding/decoding of signals
2. Synchronisation
3. Bit transmission and reception
4. Choice of transmission medium
5. Topology selection

The second layer in 802 model has been divided into two sections.

1. Logical Link Control (LLC)
2. Medium Access Control (MAC)

The LLC performs the following functions.

1. On transaction, assemble data into a frame with address and error-detection fields.
2. On reception, disassemble frame and perform address recognition and error detection.
3. Have access to LAN transmission medium.
4. Provide an interface to higher layers.
5. Flow and error control.

LLC specifies the mechanisms for addressing stations across the medium and for controlling the exchange of data between two users. There are services provided as alternatives for attached devices using LLC. They are as follows.

1. Unacknowledged connections service
2. Connection mode service
3. Acknowledged connectionless service

The frame format of LLC protocol is shown in Figure 9.37.

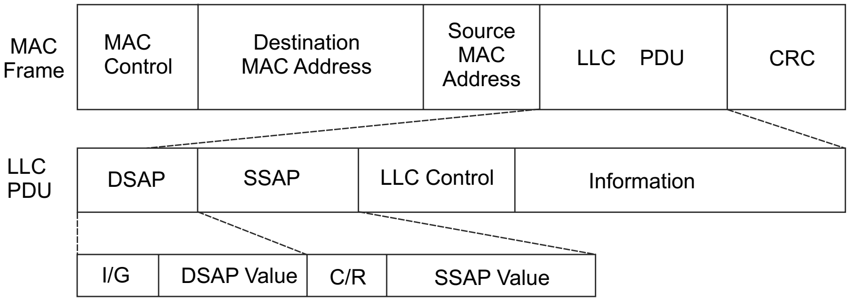


Fig. 9.37 Frame format of LLC protocol

where
I/G—Individual/Group
DSAP—Destination Service Accept
C/R—Connections/Response
SSAP—Source Service Accept

The main function of the MAC protocol is to control the access to the transmission medium for efficient use of the transmission capacity. The MAC layer is also responsible for detection of errors and discarding any frames that are I error.

9.23.7 Media Access Control in LAN

There are several methods of media access control in the local area networks. Each of these methods is applicable to a specific LAN topology. These methods can be classified into two categories.

1. Centrally controlled access
2. Distributed access control

In the first category, access to the media is controlled by a central controller. Polling, demand assigned frequency division or time-division multiple access are some such methods. But distributed access control methods are more common in the local area networks. Centrally controlled access methods suffer from a basic limitation in that they have a single point of network failure.

The distributed control methods have an edge over the centrally controlled methods in that there is no single point of network failure. These methods are available for both bus and ring topologies. For the bus topology, the following are the two methods.

1. Token passing, IEEE 802.4
2. Carrier Sense Multiple Access/ Collision Detection (CSMA/CD), IEEE 802.3.

1. TOKEN BUS—IEEE 802.4

Physically, the token bus is a linear or tree-shaped cable on which the stations are connected in the form of a ring. The highest numbered station sends the first frame. Then it passes the permission to its immediate neighbour by sending a special control frame. This special control frame is called a **token**. Figure 9.38 illustrates the process of token passing.

In this method, the station that holds the token is allowed to transmit data frames on the bus. Once the transmission time of the station is over, it passes the token to the next station in logical sequence. In one cycle, every station gets an opportunity to transmit.

The frame format for IEEE 802.4 is shown in Figure 9.39.

where
Preamble: Used for bit transmission
SD: Frame start delimiter for start of frame
FC: Frame Control to identify data and control then frame

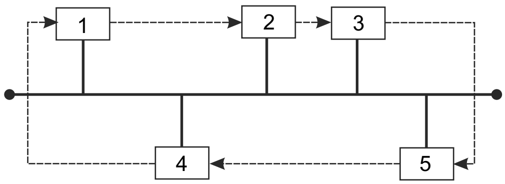


Fig. 9.38 Token Passing—IEEE 802.4

Preamble	SD	FC	DA	SA	Data	FCS	ED
7 bytes	1 byte	1 byte	6 bytes	6 bytes		4 bytes	1 byte

Fig. 9.39 Frame format for IEEE 802.4

DA: Destination Address

SA: Source Address

FCS: Frame Check Sequence

ED: End Delimiter denotes end of frame

The token bus LANs operate at data rates of 1, 5 or 10 Mbps and there are two types of transmission systems.

1. *Carrier Band*: Single-channel bidirectional transmission and it uses phase coherent FSK modulation.
2. *Broadband*: Multiple-channel unidirectional transmission which uses AM and PM.

2. CSMA/CD—IEEE 802.3

Carrier Sensing Multiple Access/Collision Detection (CSMA/CD) is the widely used MAC protocol. When two stations try to access the channel or bus at the same time, a collision may occur. In CSMA, the station continues to transmit the frame till the end of the frame even if collision occurs. This wastes the channel time. In CSMA/CD, the station abandons the transmission of frame as soon as a collision is detected. After a random delay, the stations try again.

When a collision is detected, the transmitting station sends a jam signal to alert the other stations to the collision. For the technique to work properly, the stations should not attempt to transmit again immediately after a collision has occurred. Usually, the stations are given a random back-off delay for retry. If a collision repeats, back-off delay is progressively increased. So the network adapts itself to the traffic. The frame transmission time shall be at least equal to twice of the end-to-end propagation time. By careful design, it is possible to achieve efficiencies of more than 90% using CSMA/CD.

The frame format of CSMA/CD is shown in Figure 9.40.

Preamble	SD	DA	SA	L	Data	FCS
7 bytes	1 byte	1 byte	6 bytes	2 bytes		4 bytes

Fig. 9.40 Frame format of CSMA/CD

- **Preamble** The preamble is an even-bytes long pattern to establish bit synchronization.
- **Start Frame Delimiter (SD)** It is a one byte-long unique bit pattern which marks the start of the frame.
- **Destination Address (DA)** The destination address field is 2 or 6 bytes long.
- **Length (L)** It is a 2-byte field which indicates the number of bytes in the data field.
- **Data Field** It can have 46 to 1500 bytes if the address field option is 6 octets. If data bytes are less than 46, the PAD field makes up the difference. This ensures minimum size of the frame.
- **Frame Check Sequence (FCS)** The frame check sequence is 4 octets long and contains the CRC code. It checks on DA, SA, length, data and PAD fields.

EXAMPLE 9.6

Find the speed of transmission of a 1500-byte group of data transmitted on 10 Mbps Ethernet-IEEE 802.3 packet.

Solution

Time for the transmission of 1 bit (t_{bit}) = $\dfrac{1}{10\times10^6}$ = 100 ns

Time for the transmission of 1 byte (t_{byte}) = 8 × 100 = 800 ns

$\therefore$ time for the transmission of 1526 bytes = 1526 × 800 = 1220.800 ns

3. Token Passing IEEE 802.5

For the ring topology, token passing IEEE 802.5, shown in Figure 9.41, is used. In this method, the numbers of stations are connected in a ring by point-to-point links. Each station is connected to the Ring Interface Unit (RIU). Each RIU regenerates the data frames it receives on the next link after a delay of at least one bit.

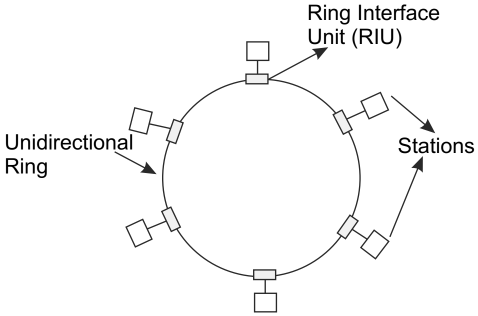


Fig. 9.41 Token ring

A token ring consists of a special bit pattern called the token, which circulates around the ring. When the station wants to transmit data, it captures the token and removes it from the ring before transmitting. There is only one token and only one station can transmit at any given instant. The ring must have sufficient delay to

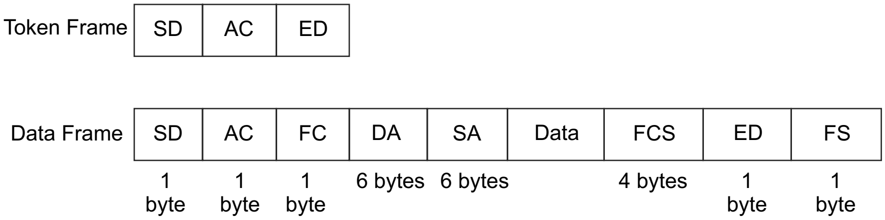


Fig. 9.42 Frame format of token ring IEEE 802.5

contain a complete token circulating across the ring when all the stations are idle. Its frame format is shown in Figure 9.42.

Here the following items are used.

- **Start Delimiter (SD)** It is a one-byte long unique bit pattern which marks the start of the data or token frame.
- **Access Control (AC)** It is a one-byte long field containing priority bits, token bit, monitoring bit and reservation bits.
- **Frame Control (FC)** It is a one-byte long field and indicates the type of the frame, data frame or control frame. It also distinguishes different types of control frames.
- **Date Field** It can have 0 or more bytes. There is no maximum size but the frame transmission time is limited by the token holding timer.
- **Frame Check Sequence** (FCS) The frame check sequence is 4 bytes long and contains CRC code. It checks on DA, SA, FC and data fields.
- **End Delimiter (ED)** It is one byte long and contains a unique bit pattern marking the end of a token or data frame.
- **Frame Status (FS)** This field is one byte long and contains two address-recognised bits, two frame copied bits and reserved bits.

The first three bits of AC field are priority bits. That last three bits are reservation bits. They are used in the priority management.

One of the stations on the ring acts as an active monitoring system. It identifies and rectifies the error condition. Any station can act as a monitor station. The differential Manchester coding is used to transmit the baseband signal.

EXAMPLE 9.7

Find the speed of transmission if a 1500-byte group of data is transmitted on 16 Mbps token ring packet.

Solution

Time for the transmission of 1 bit (t_{bit}) = $\dfrac{1}{16\times10^6}$ = 62.5 ns

Time for the transmission of 1 byte (t_{byte}) = 8 × 62.5 = 500 ns

$\therefore$ time for the transmission of 1526 bytes = 1521 × 500 = 760.500 ns

Summary

Data communication refers to exchange of digital information, the information that is stored in digital form, between two digital devices. The main purpose of data communication codes is to represent characters and symbols such as letters and digits and punctuation marks. There are several ways to represent the data which include binary representation, ASCII code, EBCDIC code, Baudot code and bar code.

Data transmission refers to movement of the bits over some physical medium connecting two or more digital devices. There are two options of transmitting the bits, namely, parallel transmission or serial transmission. In parallel transmission, all the bits of a byte are transmitted simultaneously on separate wires. In serial transmission, bits are transmitted serially one after the other.

There are two methods of timing control for reception of bits. The transmission modes corresponding to these two timing methods are called asynchronous transmission and synchronous transmission. Asynchronous transmission refers to the case when the sending end commences transmission of bytes at any instant of time. In synchronous transmission, bits are always synchronised to a reference clock irrespective of the bytes they belong to. There are no start or stop bits.

In data transmission, the devices used to perform both modulation as well as demodulation function are called 'modems'. They are required when data is to be transmitted over long distances. In a modem, the input signal modulates a carrier which is transmitted to the distance end. At the distant end, another modem demodulates the received carrier to get the digital signal. Thus, a pair of modems is always required.

A data communication network is defined as any group of computer terminals connected together and the process of sharing resources between computer terminals over a data communication network is called networking. Communication functions are implemented and controlled by using many hardware (physical) and software (logical) components in a computer network.

All the computer networks include some combination of end stations, applications and network that will support the data traffic between the end stations. Computer networks all share common devices, functions and features including servers, clients, transmission media, shared data, shared printers and other peripherals, hardware and software resources, Network Interface Card(NIC), Local Operating System(LOS) and Network Operating System(NOS).

Computer networks can be represented with two basic network models.

1. Peer-to-peer client/server
2. Dedicated client/server

The client/server method specifies the way in which two computers can communicate with software over a network. With client/server concept, a computer acting as a client initiates a software request from another computer acting as a server. The server computer might then act as a client and request services from another computer. A peer-to-peer client/server network is one in which all computers share their resources such as hard drives, printers and so on, with all the other computers on the network. Therefore, the peer-to-peer operating system divides its time between servicing the computer on which it is loaded and serving requests from other computers. In a peer-to-peer network, there are no dedicated servers among the computers.

The topology of a network is the geometric representation of the relationship of all the links and linking devices, usually called nodes, to each other. Topology is a major consideration for capacity, cost and reliability when designing a data communications network. There are two basic topologies.

1. Point-to-point line configuration
2. Multipoint configuration

A point-to-point line configuration is used in data communications networks that transfer high-speed digital information between only two stations. A multipoint, also called multidrop configuration is one in which more than two specific devices share a single link.

Examples of multipoint topologies are as follows.

1. Mesh topology
2. Star topology
3. Tree topology
4. Bus topology
5. Ring topology
6. Hybrid type

Networks are classified based on size which includes geographic area, distance between stations, number of computers, transmission speed, transmission media and the network's physical architecture. The four types of classification of networks are as follows.

1. Local Area Networks (LAN)
2. Metropolitan Area Networks (MAN)
3. Wide Area Networks (WAN)
4. Global Area Networks (GAN)

A computer network is also partitioned into end systems interconnected using a subnetwork and the communication process is decomposed into hierarchical functional layers. The layered architecture emphasises that there is hierarchy of functions. Each layer provides certain services to the next higher layer which uses these services to carry out its assigned functions. Each layer also needs to interact with the peer layer of another end system or the subnetwork to carry out its functions. Since there is no direct path between peer layers, they have interaction using services of the lower layers.

Open System Interconnection (OSI) is the name for a set of standards for communicating among computers. The main purpose of OSI standards is to serve as a structural guideline for exchanging information between computers, workstations and networks. The OSI architecture decomposes the communication process into hierarchical functional layers and identifies the standards necessary for open-system interconnection.

OSI has seven layers, each of which defines a segment of the process of moving information across a network. Layers 1, 2 and 3 are the network-support layers. They determine the physical aspects of moving data from one device to another. Layers 5, 6 and 7 are called user-support layers. They allow the interoperability between unrelated software systems. Layer-4 ensures end-to-end reliable data transmission while Layer-2 ensures reliable transmission on a single link.

The Integrated Services Digital Network (ISDN) is a proposed network designed for providing worldwide telecommunication support of voice, data, video and facsimile within the same network. ISDN is a network that proposes to interconnect an unlimited number of independent users through a common communications network. The main feature of the ISDN concept is to support a wide range of voice (telephone) and nonvoice (digital data) applications in the same network using a limited number of standardised facilities.

Local Area Networks (LAN) is a network facility spread over a relatively small geographical radius and it is mainly to interconnect the computers so that the data, software and hardware resources within the premises of an organisation could be shared. It has the following attributes.

- Geographic coverage of local area networks is limited to an area less than 5 km.
- The data rate exceeds 1 Mbps.
- The physical interconnecting medium is privately owned.
- The physical interconnecting medium is usually shared by the stations connected to the LAN.

REVIEW QUESTIONS

PART-A

1. What is data communication?
2. What are the hardware components of data communication?
3. How will you represent the data?

4. What is an ASCII code? What is it used for?
5. What is EBCDIC code? What is its purpose?
6. Which one is the first fixed-length code? What is its use?
7. What is the purpose of bar code?
8. Define data transmission. What are the types of data transmission?
9. What do you mean by serial and parallel transmission?
10. What are the modes of data transmission?
11. What is synchronous transmission?
12. What is asynchronous transmission?
13. Mention the significance of start and stop bits.
14. Define modulation rate. What is its unit?
15. What is meant by MODEM?
16. What is serial interface? Give two examples.
17. What is a parallel interface? What is its use?
18. What is a data communication network?
19. Define networking.
20. What are the factors to be considered for data communication networks?
21. List out the applications of computer networking.
22. What are the physical and logical components of a computer network?
23. What are the functions of a server?
24. Define Network Interface Card (NIC).
25. List a few characteristics of the network interface card.
26. What is local operating system?
27. Define network operating system.
28. What are the characteristics of a network operating system?
29. Peer-to-peer networks should be small. Why?
30. State the importance of network topologies.
31. What do you mean by point-to-point line configuration?
32. What do you mean by multipoint configuration?
33. Name the multipoint topologies.
34. State the concept of mesh topology and mention its advantages.
35. State the advantages and disadvantages of star topology.
36. What are advantages and disadvantages of tree topology?
37. What are the advantages and disadvantages of bus topology?
38. Define hybrid topology.

39. How will you classify computer networks?
40. What is a local area network?
41. What is a metropolitan area network?
42. What is meant by wide area networks?
43. What is meant by global area networks?
44. State the significance of open system interconnection.
45. List out the layers involved in an OSI model.
46. State the functions of the physical layer.
47. What are the functions of the data-link layer?
48. What are the responsibilities of the network layer?
49. What is ISDN? State its principle.
50. What is LAN? What are its attributes?
51. What are the key elements of LAN?
52. State the different transmission media used for LAN.

PART-B

1. What are the various data communication codes? Explain in detail.
2. Explain about the following data communication codes in detail.
   i) ASCII code ii) EBCDIC code
3. What are the different types of data transmission? Explain in detail.
4. What are the different types of modes of data transmission? Explain them in detail.
5. What are the different serial interfaces available? Explain any one in detail.
6. What is meant by parallel interface? Explain in detail about the functionality and uses of parallel interface.
7. What is network? What are the components, functions of a computer network? Explain them.
8. What are the different network topologies? With neat sketches, explain their significance in detail.
9. Explain about the layered architecture of OSI reference model with the responsibilities of individual layers.
10. What is ISDN? Explain the principle, objectives and architecture of ISDN with its neat sketches.
11. With neat sketches, explain about the need, attributes and different topologies of LAN in detail.

10

OPTICAL-FIBRE COMMUNICATION

Objectives

- To know the purpose and different generations and wavelength spectra of optical fibre communication
- To discuss details about the propagation characteristics of optical fibres
- To discuss the structure and types of optical fibres and optical-fibre connectors
- To provide details about various losses to be considered in optical fibres
- To provide the details about different optical sources and optical detectors used in optical-fibre communication and also about various applications

10.1 INTRODUCTION

Optical communication systems may be defined as systems for transferring information from a source to a destination, using light as an information carrier. A general perception is that optical communication has had its origin somewhere in the 1960s and 1970s, with the invention of lasers and low-loss optical fibres. From the historical survey, it is shown that optical communication is one of the earliest known and used long-distance communication techniques by humankind.

Fibre-optic communication is a method of transmission of information from one place to another place by means of light signals through an optical fibre. The light signal forms an electromagnetic carrier signal that is modulated to carry the information.

The process of communicating using fibre optics involves the following basic steps: creating the optical signal involving the use of a transmitter, relaying the signal along the fibre, ensuring that the signal does not become too distorted or weak, receiving the optical signal, and converting it into an electrical signal.

10.2 OPTICAL-FIBRE COMMUNICATION

Figure 10.1 shows the basic components of optical-fibre communication.

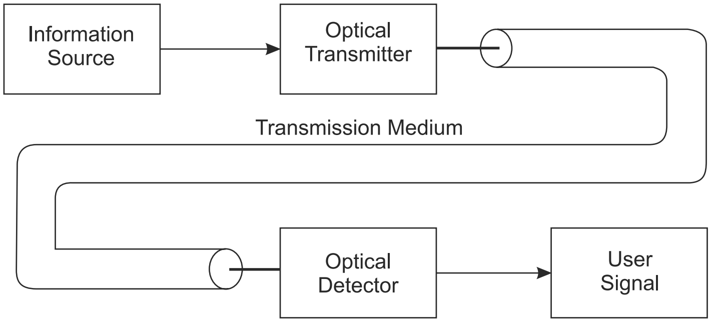


Fig. 10.1 Block diagram of optical-fibre communication

10.2.1 Information Source

The information to be transmitted through an optical fibre may be voice, picture or any kind of data. This information is first to be converted into a series of digital pulses. For this purpose, A/D convertor or encoders are to be used.

10.2.2 Optical Transmitter

The digital signal from the A/D converter/encoder is then modulated by pulse-code modulation and given by electrical to optical converter, which is the optical transmitter. This is usually a Light Emitting Diode (LED) or LASER light. It produces modulated light signals.

10.2.3 Transmission Medium

These light signals are then transmitted through an optical fibre which is the transmission medium. Optical fibres are used for transmission of optical signals in the same manner as coaxial cables for radio-wave communication. With the fibre optical cable, it is possible for very high-density information transmission.

10.2.4 Optical Detector

The modulated light wave, carried by the optical fibre, is picked up by a photodetector. This is usually a photodiode or phototransistor. Their conduction is varied by means of light. This light signal is then demodulated to recover the originally transmitted signal, voice, picture or any form of data.

10.3 GENERATIONS OF OPTICAL-FIBRE COMMUNICATION

Light has been used for line-of-sight communications for thousands of years. Real advances of optic communication began in the 19th century with several improved line-of-sight communication schemes. With the aid of optical fibres, guided transmission of light over fibre became a reality in the early 1970s.

In 1966, optical fibres were proposed for communication purpose by Charles K Kao and George Hockham. In 1970, optical fibres were successfully developed with attenuation low enough for communication purposes. At the same time, semiconductor Lasers were also developed for transmitting light through fibre-optic cables for long distances.

In the year 1975, the first-generation commercial telephone system was operated at a wavelength of 820 nm using Laser diode. The fibre attenuation was in the range of 3 to 8 dB/km. This allowed a 90 Mb/s transmission rate over 8 to 12 km.

The second generation of fibre-optic communication was developed for commercial use in the early 1980s, operated at a wavelength of 1.3 μm without repeaters and used in GaAsP semiconductor lasers. Fibre attenuation was less than 0.5 dB/km. This provided 565 Mb/s transmission rate up to 45 km.

Third-generation fibre-optic systems operated at 1.55 μm and had losses of about 0.2 dB/km. It is anticipated that they will employ single-mode fibre with less than 0.3 dB/km. It will provide a transmission rate of 1.3 Gb/s over a distance of 45 km.

The fourth generation of fibre-optic communication systems, in 1992, used several amplifiers for amplification which reduced the need for repeaters and to increase the channel capacity. Its transmission rate is 2 Gb/s over a fibre link of 1330 km.

In order to satisfy the needs of longer wavelength for operation, fifth-generation of fibre-optic communication systems were developed. From 1990 onwards, fibre-optic communication grows exponentially due to increased use of the Internet and commercialisation of various bandwidth-intensive consumer services.

10.4 WAVELENGTH AND SPECTRA

Normally, light can be characterised in terms of its wavelength, which is analogous to the frequency of a radio signal. The wavelength of the light is expressed in microns or nanometres. The spectrum of visible light ranges from ultraviolet to infra-red. Optical fibre systems operate in three IR windows around 800 nm, 1310 nm and 1550 nm. Figure 10.2 shows the spectrum of light by considering the wavelength in nanometres.

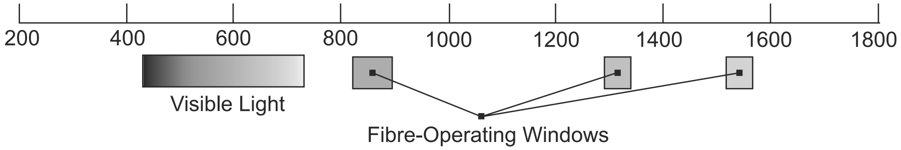


Fig. 10.2 Spectrum of light

The above light spectrum can be divided into three categories of frequency bands. They are as follows.

1. Infrared Band

Generally, an optical-fibre communication system operates in the infrared band. This band has the light frequencies ranging between 770 nm and 10^6 nm and this band is too high to be seen by the human eye. Infrared light is sensed as heat. This type of radiation can affect photographic plate coated with special chemicals.

2. Visible Band

The next band of frequencies in the electromagnetic spectrum is visible light. Since this band is visible to the human eye, it is called visible band. This band has the light frequencies ranging between 390 nm and 770 nm. This band occupies only a small part of the spectrum and it produces the sensation of vision.

When light is composed of a mixture of many different wavelengths, each of which can be seen by the human eye as a colour. It consists of different colours such as violet, blue, green, yellow, orange and red. When an ordinary white light is passed through a prism, it is separated into a rainbow of colours known as visible spectrum.

3. Ultraviolet Band

This ultraviolet band is available just above the visible light on the electromagnetic spectrum. This is the band of frequencies that are too low to be seen by the human eye. This band has the light frequencies ranging between 10 nm and 390 nm. Normally, ultraviolet light can be detected by a photographic plate and this band is very much useful in the medical field as it produces chemical effect on our human body.

There are some common properties to all of above regions. They are listed as follows.

1. Electromagnetic waves travel through free space in straight lines at a speed of 3×10^8 m/s.
2. They are associated with oscillating electric and magnetic fields and are traverse in nature.
3. Electromagnetic waves omitted from a point source in free space obey an inverse square law.

10.5 ADVANTAGES OF OPTICAL-FIBRE COMMUNICATION

The following are the major advantages of optical-fibre communication.

1. Attenuation

Attenuation in a fibre is very much lower than that of a coaxial cable or twisted pair and is constant over a very wide range of transmission and a wide range of distance is possible without repeaters.

2. Smaller Size and Lighter Weight

Optical fibres are considerably thinner than the coaxial cable or bundled twisted pair cable. So they occupy less space.

3. Electromagnetic Isolation

Electromagnetic waves generated from electrical disturbances or electrical noises do not interfere with light signals. As a result, the system is not vulnerable to interference, impulse noise or crosstalk.

4. No Physical Electrical Connection

In an optical-fibre communication, there is no physical electrical connection required between the transmitter and the receiver.

5. Reliability

The fibre is much more reliable, because it can better withstand environmental conditions such as pollution, radiation and the salt produces no corrosion. Moreover, it is less affected by the nuclear radiation. Its life is much longer than that of copper wires.

6. Security

Since fibre-optic communication does not permit any crosstalk, the transmission is more secure and private. No tapping is permitted in the fibre.

7. Bandwidth

Bandwidth of the optical fibre is higher than that of an equivalent wire-transmission line.

8. Dielectrics

As fibres are very good dielectrics, isolation coating is not required in the fibre-optic communication system.

9. Higher Data Rate

In a fibre, data rate is much higher and hence much information can be carried by each fibre than by equivalent copper cables.

10. Lower Cost

The cost per channel is lower than that of an equivalent wire cable system. It is expected that in the near future, the optical-fibre communication system will be more economical than any other type of communication system.

11. No Radiation

Due to non-inductive and non-conductive nature of a fibre, there is no radiation and interference on the other circuits and systems.

12. Greater Repeater Spacing

Fewer repeaters indicate lower cost and fewer sources of error. It has been observed that a fibre transmission system can achieve a data rate of 5 Gbps over a distance of 111 km without repeaters, whereas coaxial and twisted pair systems generally have repeaters every few kilometres.

13. Availability

Optical fibres are made of silica which is abundantly available as a natural resource

10.6 DISADVANTAGES OF OPTICAL-FIBRE COMMUNICATION

Although there are many advantages, there are also some disadvantages due to the use of optical-fibre cables in fibre-optic communication. They are listed as follows.

1. Cost

The optical fibre with its necessary interfacing will lead the high cost of communication system. To use the optical fibre for any application, it must be connected to necessary electronic facilities which require expensive interfaces.

2. Tensile Strength

Optical fibres have a significantly lower tensile strength than coaxial cables. They can be improved by the use of fibres with specialised coating and protective jacket.

3. Need of Remote Electrical Power

It is necessary to provide electrical power to remote interface. For that purpose, additional metallic cables must be included in the cable assembly.

4. Susceptibility to Losses

Optical fibre cables are more susceptible to losses which are introduced by bending the cable. Bending of cables cause irregularities in the cable dimensions and it will result in a loss of signal power.

5. Need of Specialised Measurements

Optical-fibre cables require specialised test and measurement tools to splicing process and cable repairing. To perform the work with optical-fibre cables, the technicians should also be skilled and trained. In addition, the location of faults is difficult to identify because of no electrical continuity in optical-fibre cables.

10.7 PROPAGATION OF LIGHT WAVES IN AN OPTICAL FIBRE

Generally, the electromagnetic energy of light is a form of electromagnetic radiation. Light and similar forms of radiation are made up of moving electric and magnetic forces. A simple example of motion similar to these radiation waves can be made by dropping a pebble into a pool of water. Here, the water is not actually being moved by the outward motion of the wave, but rather by the up-and-down motion of the water. The up-and-down motion is at right angles to the outward motion of the waves. This type of wave motion is called transverse wave motion. The transverse waves spread out in expanding circles until they reach the edge of the pool, in much the same manner as the transverse waves of light spread from the sun.

Light radiates from its source in all directions until it is absorbed or diverted by some substance. The lines drawn from the light source to any point on one of the transverse waves indicate the direction that the wavefronts are moving in. These lines are generally called light rays. Figure 10.3 shows the light rays and wavefronts in front of a light source.

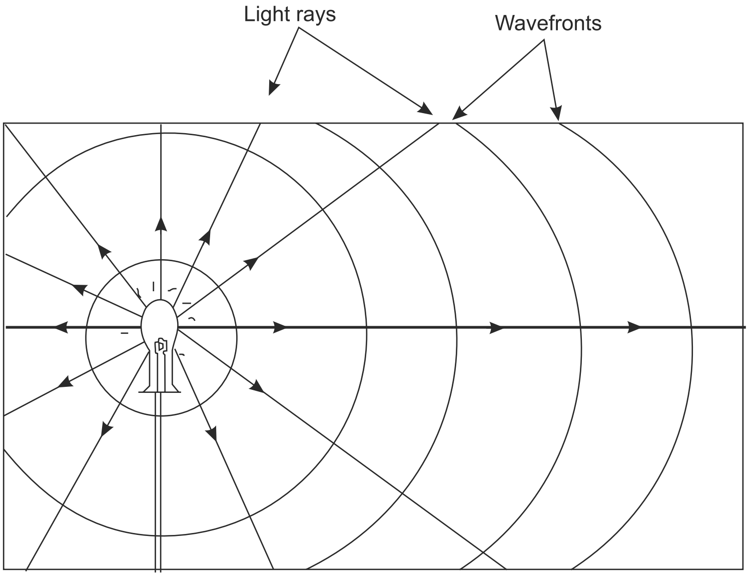


Fig. 10.3 Light rays and wavefronts in front of a light source

10.7.1 Basic Concepts of Physics

Light is simply a range of electromagnetic radiation that can be detected by the human eye. Electromagnetic radiation has a dual nature as both particles and waves. According to Maxwell's theoretical statement, the electromagnetic radiation contains a series of oscillating waves comprised of an electric and a magnetic field at 90° angles. This electromagnetic wave has **amplitude**, which is the brightness of the light, **wavelength**, which is the colour of the light, and an **angle** at which it is vibrating, called **polarisation**. This was the classical interpretation, crystallised in Maxwell's equations, which held sway until Planck, Einstein and others came along with the quantum theory. In terms of the modern quantum theory, electromagnetic radiation consists of particles called photons, which are packets of energy moving at the speed of light. In this particle view of light, the brightness of the light is the number of photons, the colour of the light is the energy contained in each photon, and four numbers (X, Y, Z and T) are the polarisation.

As proposed by Einstein, light is composed of photons which are very small packets of energy. The reason that photons are able to travel at light speeds (C) is due to the fact that they have no mass (M) and, therefore, Einstein's infamous equation $E = MC^2$ cannot be used.

Later, Planck proved that when light is emitted or absorbed, it behaves like an electromagnetic wave and also like a particle called photon which possesses energy proportional to its frequency. This effect is named **Planck's law**.

Planck's law is stated as *'when visible light or high-frequency electromagnetic radiation illuminates a metallic surface, there will be emission of electrons and these emitted electrons produce an electric current*. It is mathematically expressed as

$$E_p = hf \tag{10.1}$$

where E_p is energy of the photon (J),

h is Planck's constant = 6.625×10^{-34} J/s, and

f is the frequency of light emitted (Hz).

Photon energy may also be expressed in terms of wavelength. Now Equation (10.4) can be written as

$$E_p = \frac{hc}{\lambda} \tag{10.2}$$

An atom has several energy levels, among which the ground level has the lowest energy. Any energy level above the ground state is called an **excited state**. If an atom in one energy level decays to a lower energy level, the loss of energy is emitted as a photon of light. The photon energy is equal to the difference between the energy of two levels.

Atoms can be irradiated by a light source whose energy is equal to the difference between ground level and an energy level. This can cause an electron to change from one energy level to another by absorbing light energy. In the process of moving from one energy level to another, the atom absorbs a photon.

The energy absorbed or emitted is equal to the difference between the two energy levels. Mathematically,

$$E_p = E_2 - E_1 \tag{10.3}$$

where E_p is the photon energy in joules.

10.7.2 Refraction

In an optical fibre, the basic mechanism of propagation results from so-called refraction. Refraction results in a change of direction for a light ray. Figure 10.4 shows the refraction between air and water.

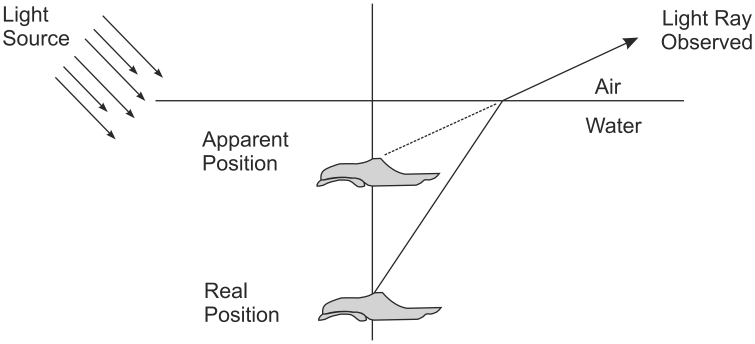


Fig. 10.4 Refraction between air and water

For light wave frequencies, electromagnetic waves travel through air at approximately the same velocity as through a vacuum. Figure 10.5 shows the refraction of light through a prism.

From the above Figure 10.5, it is shown that a light ray is refracted as it passes from a less dense medium (air) into a more dense material (glass). Actually, light is not bending, instead the direction of light transmission is changed at the interface. Refraction occurs at both air/glass (less to dense media) interfaces. The violet wavelengths are refracted the most,

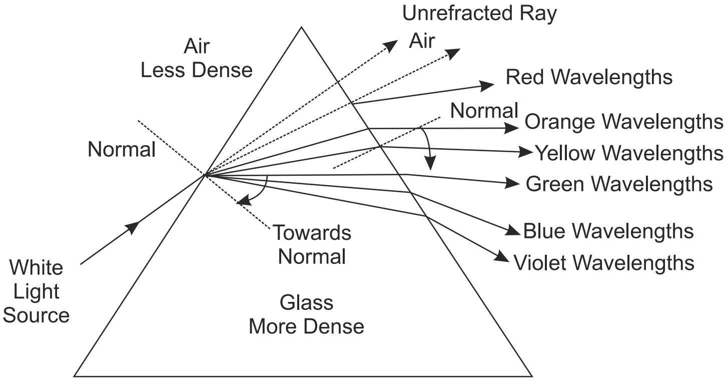


Fig. 10.5 Refraction of light through a prism

whereas the red wavelengths are refracted the least. The spectral separation of white light in this manner is called **prismatic refraction**. This prism splits the sunlight into the various wavelengths, thus developing a visible spectrum of colour.

10.7.3 Refractive Index and Snell's Law

When a ray is incident on the interface between two media of different indices, refraction will take place. The refractive index of a medium is defined as the ratio of velocity of light in free space (vacuum) to the velocity of light in the medium (given material).
Mathematically, refractive index is expressed as

$$n = \frac{c}{\nu} \tag{10.4}$$

where n is refractive index,

c is the speed of light in vacuum, and

ν is the speed of light in a given material.

Several materials have different refractive indices. Table 10.1 shows refractive indices of different materials.

Table 10.1 Various refractive indices for various materials

Material	Refractive Index	Speed of Light
Air	1.00028	299, 706 km/s
Ice	1.310	228, 847 km/s
Water	1.333	224, 900 km/s
Perspex	1.495	200, 528 km/s
Crown Glass	1.52	197, 230 km/s
Flint Glass	1.62	185, 055 km/s
Diamond	2.42	123, 880 km/s
Typical fibre core	1.487	201,607 km/s

Snell's law defines the relationship between refractive indices and the light ray angles. By assuming that $n_1 > n_2$, $\theta_1 > \theta_2$, Snell's law for refraction is expressed as follows.

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \tag{10.5}$$

where n_1 is the refractive index of the more dense material (Medium-1),

n_2 is the refractive index of the less dense material (Medium-2),

θ_1 is the angle of incidence (in degrees), and

θ_2 is the angle of refraction (in degrees).

The angle at which the propagating ray strikes the interface with respect to the normal is termed the **angle of incidence** (θ_1) and the angle formed between the propagating ray and the normal after the ray has entered the next medium is termed the **angle of refraction** (θ_2). Figure 10.6 shows an example of a refractive model in which a light ray is travelling from one medium to another

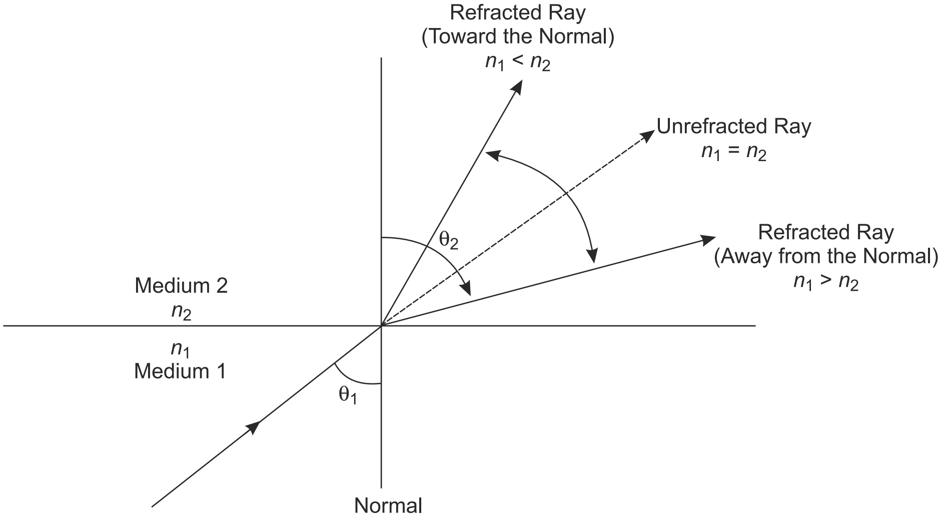


Fig. 10.6 Light refraction

From Figure 10.6, the light ray is refracted as it travels from Medium-1 of more density to Medium-2 of less density. It is noted that the direction of light is changed at the interface and the angle of refraction is greater than the angle of incidence. Consequently, when a light ray enters Medium-2 of less density, it gets refracted away from the normal. In the same way, when a light ray enters a Medium-1 of more density, it gets refracted toward the normal.

10.7.4 Total Internal Reflection

If light wave enters at one end of a fibre in proper conditions, most of the light is propagated down the length of the fibre and comes out from the other end of the fibre. There may be some loss due to a small fraction leakage through the side walls of the fibre. This type of fibre is called **light guide**. The reason of confining the light beam inside the fibre is the total internal reflection of light waves by the inside surface of the fibre.

Optical fibre obeys laws of reflection and refraction of light waves. The light which enters at one end of a fibre at a slight angle to the axis of the fibre follows a zig-zag path due to series of reflections down the length of the fibre.

For the phenomenon of total internal reflection to take place, the following two conditions are to be satisfied.

1. The glass at around the centre of the fibre (core) should have higher refractive index (n_1) than that of the material (cladding) surrounding the fibre (n_2).
2. The light should be incident at an angle of θ which will be greater than the critical angle θ_c.

$$\sin \theta_c = \frac{n_2}{n_1} \qquad (10.6)$$

Reflection, refraction and total internal reflection of light waves are shown in Figure 10.7.

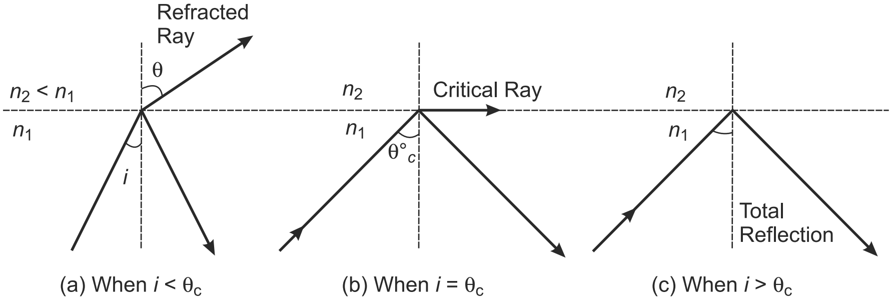


Fig. 10.7 Total internal reflection of light waves

The conditions of reflection, refraction and total internal reflection are as follows.

1. In reflection, the angle of incidence is equal to the angle of reflection.
2. In refraction, $n_1 \sin i = n_2 \sin \theta$. The refracted wave should move towards the normal, if the light wave is incident from the optically lighter medium to an optically denser medium. And the refracted light wave should move away from the normal, if the light wave travels from the optically denser to optically lighter medium.
3. The condition for total internal reflection is vide Equation (10.4).

Total internal reflections can be demonstrated using a semicircular glass block shown in Figure 10.8.

A **ray box** shines a narrow beam of light onto the glass. The semicircular shape ensures that a ray pointing towards the centre of the flat face will hit the curved surface at a right angle and which will prevent refraction at the air/glass boundary of the curved surface. θ_c is the critical angle measured normal to the surface.

If $\theta < \theta_c$, the ray will split. Some of the ray will reflect off the boundary, and some will refract as it passes through. If $\theta > \theta_c$, the entire ray reflects from the boundary. None passes through. This is called total internal reflection.

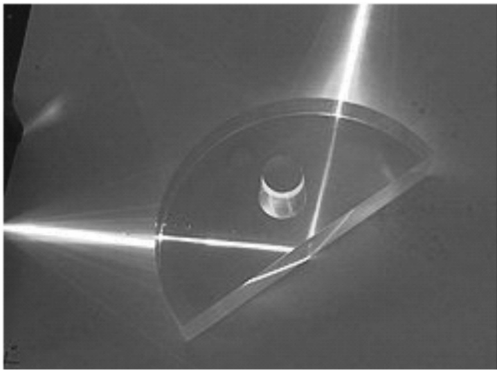


Fig. 10.8 Total internal reflection

10.7.5 Propagation in an Optical Fibre

It is well known that the refractive index of the core is greater than that of cladding. From Figure 10.9, light ray-1 propagates because at *B*, it undergoes total internal reflection and is reflected back into the core and light ray-2 does not undergo total internal reflection and is thus lost in the cladding.

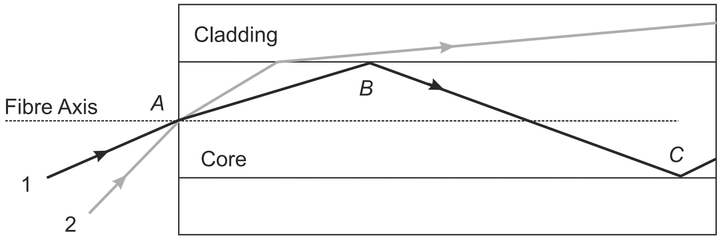


Fig. 10.9 Propagation of light through an optical fibre

The following are the conditions for the propagation.

1. Angle of incidence (θ_i) must be greater than the critical angle (θ_c) for the fibre if total internal reflection is to take place.
2. Only light rays which enter the core with an angle less than the acceptance angle will propagate.

10.7.6 Critical Angle (θ_c)

Critical angle (θ_c) of a medium is defined as the value of the incident angle at which the angle of refraction is 90°. It is the angle of incidence above which total internal reflection occurs. The angle of incidence is measured with respect to the normal at the refractive boundary. The critical angle θ_c is given by

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) \tag{10.7}$$

where n_2 is the refractive index of the cladding and n_1 is the refractive index of the core.

If the incident ray is precisely at the critical angle, the refracted ray is tangent to the boundary at the point of incidence. For example, visible light were travelling through glass with an index of refraction of 1.50 into air with an index of refraction of 1.00. The calculation would give the critical angle for light from acrylic into air, which is

$$\theta_c = \sin^{-1}\left(\frac{1.00}{1.50}\right) = 41.8^\circ \tag{10.8}$$

Light incident on the border with an angle less than 41.8° would be partially transmitted, while light incident on the border at larger angles with respect to normal would be totally internally reflected.

The critical angle for diamond in air is about 24.4°, which means that light is much more likely to be internally reflected within a diamond. If the fraction $\frac{n_2}{n_1}$ is greater than 1 then arcsine is not defined which means that total internal reflection does not occur even at very shallow or grazing incident angles. So the critical angle is only defined when $\frac{n_2}{n_1}$ is less than 1.

10.7.7 Transmission of Light Ray in a Perfect Optical Fibre

Figure 10.10 illustrates the transmission of a light ray in an optical fibre via a series of total internal reflections at the interface of the silica core and the slightly lowered refractive index, silica cladding.

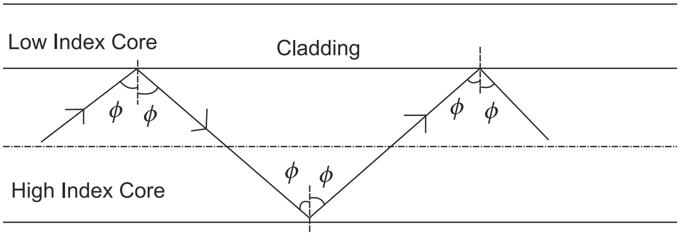


Fig. 10.10 Transmission of a light ray in a perfect optical fibre

The light ray shown in Figure 10.8 is known as a **meridional ray** as it passes through the axis of the fibre core. The ray has an angle of incidence ϕ at the interface which is greater than the critical angle and is reflected at the same angle to the normal.

The light transmission assumes a perfect fibre and that any discontinuities or imperfections at the core–cladding surface would probably result in refraction rather than total internal reflection with the subsequent loss of light ray into the cladding.

10.7.8 Acceptance Angle (Θ_a)

Acceptance angle (θ_a) is defined as the maximum angle to the axis at which the light may enter the fibre in order to be propagated, at which the angle of incidence at the core–cladding boundary is equal to the critical angle of the core medium. It is illustrated in Figure 10.11.

Acceptance angle (θ_a) $= \sin^{-1}\sqrt{n_1^2 - n_2^2}$ (10.9)

The analysis of the acceptance angle is carried out from Figure 10.12.

Assuming that $\theta = \theta_c$.

From the above figure, $\theta_2 = 90° - \theta_c$

By Snell's law,

$$n_0 \sin \theta_1 = n_1 \sin \theta_2 \tag{10.10}$$

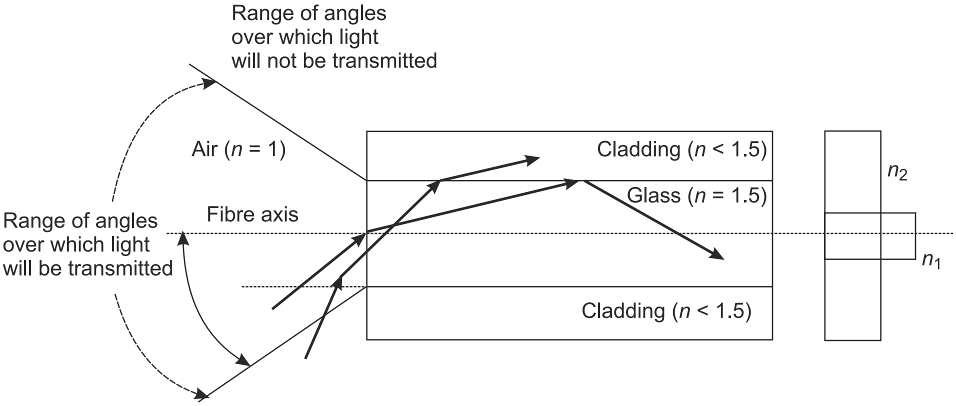


Fig. 10.11 Acceptance angle

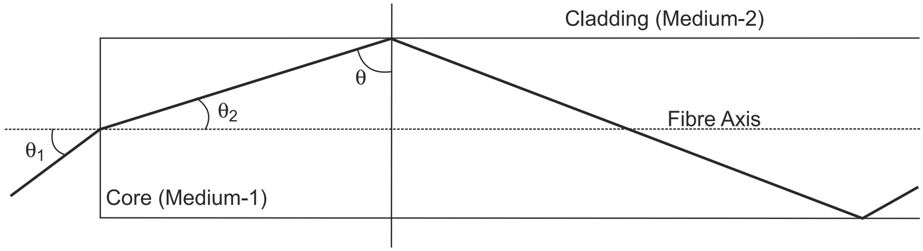


Fig. 10.12 Analysis of the acceptance angle

where n_0 is the refractive index of air =1.

So,
$$\theta_1 = \sin^{-1}(n_1 \sin(90° - \theta_c)) \tag{10.11}$$

$$\theta_1 = \sin^{-1}\left(n_1 \sqrt{\cos^2 \theta_c}\right)$$

$$\theta_1 = \sin^{-1}\left(n_1 \sqrt{1 - \frac{n_2^2}{n_1^2}}\right)$$

Hence,
$$\theta_1 = \sin^{-1}\left(\sqrt{n_1^2 - n_2^2}\right) \tag{10.12}$$

This last value is the maximum value that θ_1 can take on if total internal reflection is to take place and it is, therefore, called the **fibre acceptance angle**.

Figure 10.13 shows the visualisation of acceptance angle in an optical fibre from which the acceptance angle is half angle of a cone, visualised at the fibre input.

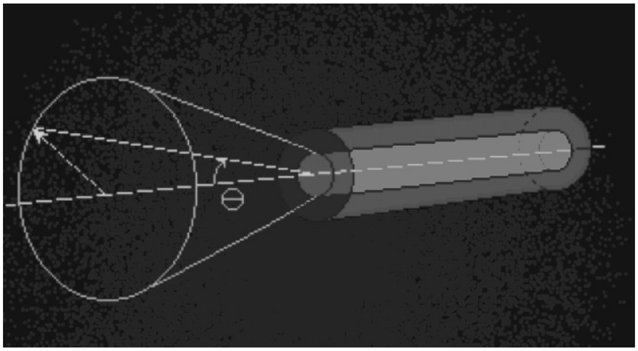


Fig. 10.13 Visualisation of acceptance angle in an optical fibre

10.7.9 Numerical Aperture (NA)

Numerical Aperture (NA) of the fibre is the light-collecting efficiency of the fibre and it is the measure of the amount of light rays that can be accepted by the fibre. This factor gives the relationship between the acceptance angle and the refractive indices of the three media involved, namely the core, cladding and the air. The NA is related to the acceptance angle θ_a, which indicates the size of a cone of light that can be accepted by the fibre.

Figure 10.14 shows a light ray incident on the fibre core at an angle θ_1 to the fibre axis which is less than the acceptance angle for the fibre θ_a. The ray enters the fibre from a medium (air) of refractive index n_0 and the fibre core has refractive index n_1 which is slightly greater than the cladding refractive index n_2.

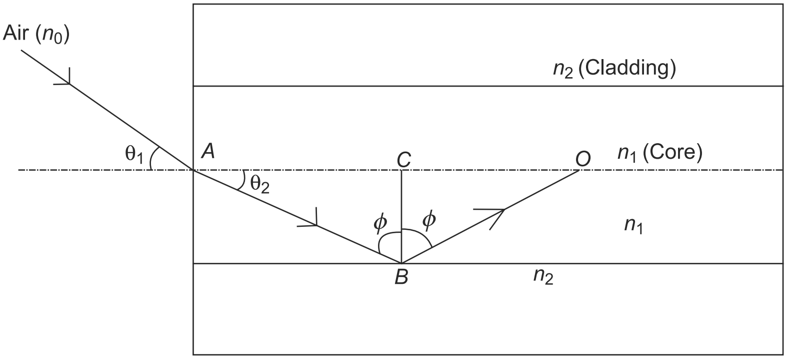


Fig. 10.14 Light ray path for a meridional ray

Assuming the entrance face at the fibre core to be normal to the axis then by considering the refraction at the air-core interface and using Snell's law,

$$n_0 \sin \theta_1 = n_1 \sin \theta_2 \tag{10.13}$$

Considering the right-angled triangle *ABC*,

$$\theta_2 + \pi/2 + \phi = 180°$$

$$\theta_2 = \pi/2 - \phi \tag{10.14}$$

Substitute Equation (10.11) in Equation (10.8),

$$n_0 \sin \theta_1 = n_1 \sin(\pi/2 - \phi)$$

$$n_0 \sin \theta_1 = n_1 \cos \phi \tag{10.15}$$

By using the following known relationship,

$$\sin^2 \phi + \cos^2 \phi = 1 \tag{10.16}$$

Equation (10.13) becomes

$$n_0 \sin \theta_1 = n_1 (1 - \sin^2 \phi_c)^{1/2}$$

By considering the total internal reflection, ϕ becomes equal to the critical angle for core-cladding interface.

i.e., $\phi = \phi_c$

In this case, θ_1 becomes the acceptance angel for the fibre (θ_a).

$$n_0 \sin \theta_a = n_1 (1 - \sin^2 \phi_c)^{1/2} \tag{10.17}$$

$$n_0 \sin \theta_a = n_1 \left(1 - \left(\frac{n_2}{n_1}\right)^2\right)^{1/2} \tag{10.18}$$

$$n_0 \sin \theta_a = n_1 \left(\frac{n_1^2 - n_2^2}{n_1^2}\right)^{1/2}$$

$$n_0 \sin \theta_a = NA \left(n_1^2 - n_2^2\right)^{1/2} \tag{10.19}$$

This is the equation of numerical aperture. When fibre is used in air, $n_0 = 1$ and the numerical aperture equals $\sin \theta_a$.

The numerical aperture may also be given in terms of relative refractive index difference Δ between the core and the cladding which is given by

$$\Delta = \frac{n_1^2 - n_2^2}{2 n_1^2} \tag{10.20}$$

$$NA = n_0 \sin \theta_a = n_1 \sqrt{2\Delta} \tag{10.21}$$

This equation also gives numerical aperture of the fibre.

EXAMPLE 10.1

A silica optical fibre has a core refractive index of 1.50 and a cladding refractive index of 1.46. Determine (a) critical angle at the core-cladding interface, (b) numerical aperture for the fibre, and (c) acceptance angle in air for the fibre.

Solution

(a) Critical angle $\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right)$

$$= \sin^{-1}\left(\frac{1.46}{1.5}\right) = 76.73°$$

(b) Numerical aperture $NA = \left(n_1^2 - n_2^2\right)^{1/2}$

$$= (1.50^2 - 1.46^2)^{1/2} = 0.344$$

(c) Acceptance angle = $\theta_a = \sin^{-1}\sqrt{n_1^2 - n_2^2} = 20.12°$

EXAMPLE 10.2

Compute numerical aperture and the acceptance angle of an optical fibre with n_1 as 1.55 and n_2 as 1.50.

Solution

Numerical Aperture $NA = (n_1^2 - n_2^2)^{1/2}$

$$= (1.55^2 - 1.50^2)^{1/2} = 0.390$$

Acceptance angle = $\theta_a = \sin^{-1}(0.390) = 23°$

EXAMPLE 10.3

Determine the numerical aperture for an optical fibre in air if the relative refractive-index difference for the optical fibre is 0.05 and the refractive index of the core is 1.46.

Solution

Numerical aperture NA = $n_1\sqrt{2\Delta}$

$$= 1.46\sqrt{2 \times 0.05} = 0.46$$

10.7.10 Performance Considerations

The amount of light that can be coupled into the core through the external acceptance angle is directly proportional to the efficiency of the fibre-optic cable. The greater the amount of light that can be coupled into the core, the lower the Bit Error Rate (BER), because more light reaches the receiver. The attenuation a light ray experiences in propagating down the core is inversely proportional to the efficiency of the optical cable because the lower the attenuation in propagating down the core, the lower the BER. This is because more light reaches the receiver. Also, the less chromatic dispersion realised in propagating down the core, the faster the signalling rate and the higher the end-to-end data rate from source to destination. The major factors that affect performance considerations are the size of the fibre, the composition of the fibre and the mode of propagation.

10.7.11 Optical Power Measurement

It is too wide to express the power level in optical communications on a linear scale. Normally, a logarithmic scale known as decibel (dB) is used to express power in optical communications. The wide range of power values makes decibel unit to express the power levels that are associated with an optical system. The gain of an amplifier or attenuation in fibre is expressed in decibels. The decibel does not give a magnitude of power, but it is a ratio of the output power to the input power.

$$\text{Loss or gain} = 10\log\left(\frac{P_{\text{output}}}{P_{\text{input}}}\right) \tag{10.22}$$

The decibel milliwatt (dBm) is the power level related to 1 milliwatt (mW). Transmitter power and receiver dynamic ranges are measured in dBm. A 1 mW signal has a level of 0 dBm. Signals weaker than 1 mW have negative dBm values, whereas signals stronger than 1 mW have positive dBm values.

$$\text{dBm} = 10\log\left(\frac{\text{Power (mW)}}{1\text{mW}}\right) \tag{10.23}$$

10.7.12 Skew Ray Propagation

Skew rays are the light rays propagated through graded-index fibres. They are rays which describe angular helices as they progress along the fibre. They follow a helical path around the axis of the fibre and these rays do not cross the axis of the fibre. These have a larger acceptance angle which is greater than the acceptance angle for meridional rays. The light-gathering ability or numerical aperture is also more for these fibres. Figure 10.15 shows the helical ray propagation.

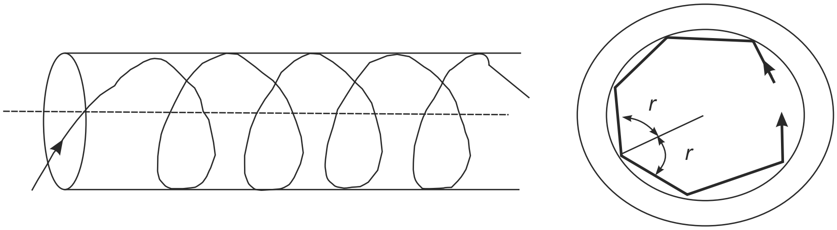


Fig. 10.15 Helical ray propagation and cross section of the fibre

With meridional rays at the fibre output, the angle depends on the input angle. For skew rays this is not so; instead the output angle depends on the number of reflection undergone. Thus, skew rays tend to make the light output from a fibre more uniform.

$$\text{Acceptance angle for skew rays} = \sin^{-1}\left[\frac{\sqrt{(n_1^2 - n_2^2)}}{\cos\gamma}\right] \tag{10.24}$$

γ is the angle of reflection for skew rays within the fibre.

Since cos γ <1, acceptance angle is higher for skew rays.

EXAMPLE 10.4

An optical fibre in air has a numerical aperture of 0.3. Find the acceptance angle for skew rays which change direction by 90° at each reflection.

Solution

$$\text{Acceptance angle for skew rays} = \sin^{-1}\left[\frac{\sqrt{(n_1^2 - n_2^2)}}{\cos\gamma}\right]$$

$$= \sin^{-1}\left[\frac{0.3}{\cos 45^\circ}\right] = 25.12^\circ$$

10.8 OPTICAL FIBRES

Optical fibres are widely used in fibre-optic communication, which permits transmission over longer distances and at wider bandwidths (data rates) than other forms of communications. The optical fibre used in the optical communication has the potential to transmit simultaneously a relatively larger number of telephone signals in the form of light waves than the coaxial cables. In optical fibres, the transmission of 15,000 or more simultaneous telephone messages is possible utilising light as the carrier.

Optical fibres can cover the long distances between local phone systems and can also provide the backbone for many network systems. Some of the optic-fibre users include cable-television services, university campuses, office buildings, industrial plants, and electric utility companies.

Fibres are used instead of metal wires because signals travel along them with less loss of signals, and they are also immune to electromagnetic interference. Fibres are also used for illumination and are wrapped in bundles so they can be used to carry images, thus allowing viewing in tight spaces. Figure 10.16 shows the photographic view of a bundle of optical wires

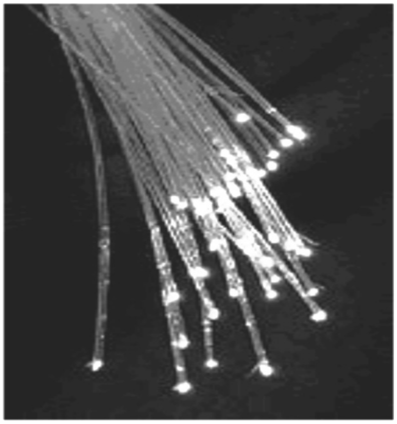


Fig. 10.16 Photographic view of a bundle of optical wires

10.9 STRUCTURE OF AN OPTICAL FIBRE

An optical fibre is physically very thin and a flexible medium. It has a cylindrical shape which consists of three sections.

1. Core
2. Cladding
3. Protective enclosure

Figure 10.17 shows a fibre and its components.

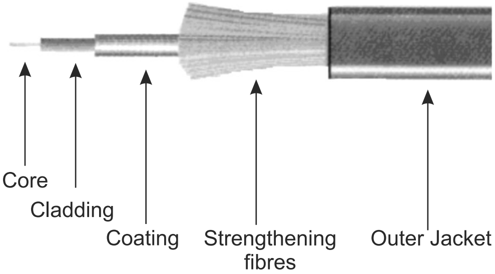


Fig. 10.17 A fibre and its components

10.9.1 The Core

The core is a cylindrical rod of dielectric material. Dielectric material conducts no electricity. Light propagates mainly along the core of the fibre. The core is generally made of glass. The core is described as having smaller radius and an index of refraction n_1. The core is surrounded by a layer of material called the cladding.

10.9.2 The Cladding

The cladding layer is made of a dielectric material with an index of refraction n_2. The index of refraction of the cladding material is less than that of the core material. The cladding is generally made of glass or plastic. Cladding is necessary to provide proper light guidance, i.e., to retain the light wave within the core as well as to provide high mechanical strength and safety to the core from damages.

The functions of the cladding are as follows.

1. Reduces loss of light from the core into the surrounding air
2. Reduces scattering loss at the surface of the core
3. Protects the fibre from absorbing surface contaminants
4. Adds mechanical strength

10.9.3 The Protective Jacket

The outer section of the optical fibre is the protection jacket or enclosure made of plastic or polymer. The main purpose of this jacket is to protect the fibre from the effects of moisture, absorption, crushing and other environmental effects.

10.10 TYPES OF OPTICAL FIBRES

In general, optical fibres are characterised by their structure and by their properties of transmission. Based on fibres used in communication, they are classified into two major types.

1. Step-index fibres
2. Graded-index fibres

Basically, optical fibres are classified based on the number of modes that propagate along the fibre. It is known that the structure of the fibre can permit or restrict modes from propagating in a fibre. The basic structural difference is the core size.

10.10.1 Step-index Fibres

Step-index fibres have a uniform core with one index of refraction, and a uniform cladding with a smaller index of refraction. A step-index fibre is characterised by the core and cladding refractive indices n_1 and n_2 and the core and cladding radii a and b.

The fractional refractive-index change can be calculated by the formulae,

$$\Delta = \frac{n_1 - n_2}{n_1} << 1 \tag{10.25}$$

For the typical values of n_1 between 1.44 and 1.46, Δ will be calculated as a value typically between 0.001 and 0.02.

Step-index optical fibre is generally made by doping high-purity fused silica glass (SiO_2) with different concentrations of materials like titanium, germaniums or boron.

Step-index fibres are further classified into two types.

1. Single-mode fibres
2. Multimode fibres

1. Single-Mode Fibres

When the fibre core is so small that only light ray at 0° incident angle can stably pass through the length of the fibre without much loss, this kind of fibre is called single-mode fibre. The basic requirement for a single-mode fibre is that the core be small enough to restrict transmission to a singe mode. This lowest-order mode can propagate in all fibres with smaller cores.

Figure 10.18 shows the profile of a step-index fibre in single-mode operation.

Single-mode fibres propagate only one mode, because the core size approaches the operational wavelength. The diameter of single-mode fibres is small and is typically around 8 to 10 micrometres. A fibre core of this size allows only the fundamental or lowest-order mode to propagate around a 1300 nm wavelength.

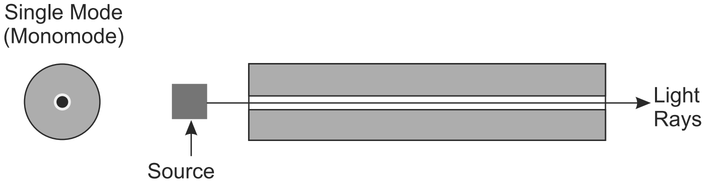


Fig. 10.18 Step-index fibre—single-mode operation

Generally, in order to relate the diameter of fibre core with mode of operation, a parameter called **normalised frequency parameter** '*V*' is to be used. In single mode fibres, *V* is less than or equal to 2.405 and they are used to propagate the fundamental mode down the fibre core, while high-order modes are lost in the cladding. For low *V* values, most of the power is propagated in the cladding material. Power transmitted by the cladding is easily lost at fibre bends. The value of *V* should remain near the 2.405 level.

Single-mode fibres have a lower signal loss and a wider bandwidth than multimode fibres. Single-mode fibres are capable of transferring higher amounts of data due to low fibre dispersion. In general, single-mode fibres are considered to be low-loss fibres, which increase system bandwidth and length.

Single-mode fibres are manufactured with the same materials as multimode fibres. Single-mode fibres are also manufactured by following the same fabrication process as multimode fibres.

Advantages of step-index single-mode fibre include the following.

1. Its core diameter is very small.
2. It has low attenuation.
3. It has very high bandwidth
4. It has low numerical aperture and hence it is used in long-distance communication

2. Multimode Fibres

When compared to single-mode fibres, multimode fibres can propagate more than one mode. Multimode fibres can propagate over 100 modes. The number of modes propagated depends

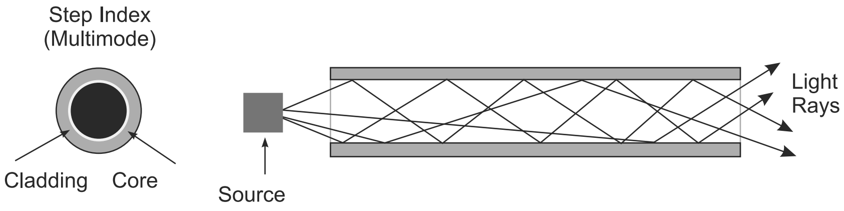


Fig. 10.19 Step-index fibre—multimode operation

on the core size and numerical aperture (NA). As the core size and NA increase, the number of modes increases. Figure 10.19 shows step-index fibre with multimode of operation.

In a multimode step-index fibre, a finite number of guided modes propagate. Number of modes is dependent on wavelength λ, core refractive index n_1, relative refractive index Δ, core radius a.

Number of modes (M) is normally expressed in terms of the normalised frequency V for the fibre, which is given as follows.

$$M = \frac{V^2}{2} = 4.9\left[\frac{d.\text{NA}}{\lambda}\right]^2 \tag{10.26}$$

where $$V = \frac{2\pi}{\lambda}a.\text{NA}$$

If a fibre has large core size and higher NA, it offers several advantages. They are as follows.

1. Light is launched into a multimode fibre with more ease.
2. The higher NA and the larger core size make it easier to make fibre connections.
3. Multimode fibres permit the use of Light-Emitting Diodes (LEDs) whereas single-mode fibres typically must use Laser diodes. Since LEDs are cheaper, less complex and last longer, they are preferred for most applications.

Multimode fibres also have some disadvantages. They are as follows.

1. As the number of modes increases, the effect of modal dispersion increases. Modal dispersion means modes arrive at the fibre end at slightly different times.
2. This time difference causes the light pulse to spread. Modal dispersion affects system bandwidth.
3. To maximise system bandwidth, fibre manufacturers have to adjust the core diameter, numerical aperture and index-profile properties of multimode fibres.

EXAMPLE 10.5

Calculate the total number of guided modes propagating in the multimode step-index fibre having a diameter of 50 μm and numerical aperture of 0.20 and operating at a wavelength of 1 μm.

Solution

$$\text{Number of modes} = M = 4.9\left[\frac{d.\text{NA}}{\lambda}\right]^2$$

$$= 4.9\left[\frac{50\times 10^{-6}\times 0.20}{1\times 10^{-6}}\right]^2 = 490$$

EXAMPLE 10.6

For Example 10.3, find the total number of guided modes propagating inside the graded-index fibre.

Solution

Total number of guided modes propagating in the multimode step-index fibre is calculated as 490.

$$M_{Graded} = \frac{M_{step}}{2}$$

$$= \frac{490}{2} = 245$$

EXAMPLE 10.7

For a fibre having a diameter of 5 µm, core refractive index of 1.45, cladding refractive index of 1.447 and wavelength of propagation as 1µm, find the number of modes propagated inside the fibre.

Solution:

$$M = 4.9\left[\frac{d.\text{NA}}{\lambda}\right]^2$$

$$= 4.9\left[\frac{5\times 10^{-6}.\left(\sqrt{1.45^2 - 1.447^2}\right)}{1\times 10^{-6}}\right]^2 = 1$$

EXAMPLE 10.8

Find the core radius necessary for single-mode operation at 850 nm of step-index fibre with refractive index of core as 1.480 and refractive index of cladding as 1.465, V = 2.405.

Solution

$$V = \frac{2\pi}{\lambda} a.n_1\sqrt{2\Delta}$$

$$\therefore \qquad a = \frac{2.405\lambda}{2\pi n_1\sqrt{2\Delta}}$$

$$\Delta = \frac{n_1^2 - n_2^2}{2n_1^2}$$

$$= \frac{1.48^2 - 1.465^2}{2 \times 1.48^2} = \frac{2.1904 - 2.1462}{2 \times 2.1904} = 0.01$$

$$\therefore \qquad a = \frac{2.405 \times 850 \times 10^{-9}}{2\pi \times 1.48 \times \sqrt{2 \times 0.01}} = 1.554\ \mu\text{m}$$

EXAMPLE 10.9

Calculate the refractive indices of the core and cladding material of a fibre if its NA is 0.22 and the relative refractive index is 0.010.

Solution

$$\Delta = \frac{n_1 - n_2}{n_1} = 0.012$$

$$NA = n_1\sqrt{2\Delta} = 0.22$$

$$\therefore \qquad n_1 = \frac{\text{NA}}{\sqrt{2\Delta}} = 1.42$$

$$0.012 = \frac{1.42 - n_2}{1.42}$$

$$\therefore \qquad n_2 = 1.40$$

EXAMPLE 10.10

A step-index fibre has a core diameter of 200 μm and NA of 0.3. Calculate the number of propagating modes at an operating wavelength of 850 nm.

Solution

$$\text{Number of modes} = \frac{V^2}{2}$$

$$V = \frac{2\pi}{\lambda} a . n_1 \sqrt{2\Delta}$$

$$\therefore \qquad \text{number of modes} = \frac{2\pi^2}{\lambda^2} a^2 \times NA^2$$

$$= \frac{2 \times (3.14)^2 \times 100^2 \times 10^{-12} \times (0.3^2)}{(850 \times 10^{-19})^2} = 24{,}589 \text{ modes}$$

EXAMPLE 10.11

Calculate the cut-off parameter and the number of modes supported by a fibre with refractive index of core as 1.54 and refractive index of cladding as 1.5. Core radius is 25 μm and operating wavelength is 1300 nm.

Solution

$$V = \frac{2\pi}{\lambda} a . \text{NA}$$

where $$\text{NA} = \sqrt{n_1^2 - n_2^2}$$

$$= \sqrt{1.54^2 - 1.5^2} = 0.349$$

$$\therefore \quad V = \frac{2\pi \times 25 \times 10^{-6} \times 0.349}{1300 \times 10^{-9}} = 42.16$$

$$\text{Number of modes} = \frac{V^2}{2} = \frac{42.16^2}{2} = 889$$

10.10.2 Graded-index Fibres

A graded-index fibre is an optical fibre whose core has a refractive index that decreases with increasing radial distance from the fibre axis. It can also propagate more than one mode. Multimode graded-index fibres typically have over one hundred propagating modes. Figure 10.20 shows the profile of a graded-index multimode fibre.

Because parts of the core closer to the fibre axis have a higher refractive index than the parts near the cladding, light rays follow sinusoidal paths down the fibre. The advantage of the graded-index fibre compared to multimode step-index fibre is the considerable decrease in modal dispersion.

A multimode graded-index fibre has a core of radius a. Unlike step-index fibres, the value of the refractive index of the core n_1 varies according to the radial distance r. The value of n_1 decreases as the distance from the centre of the fibre increases.

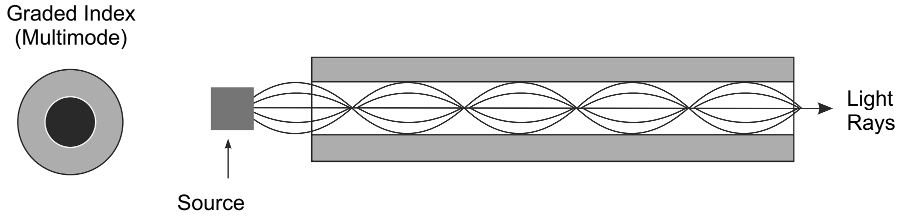


Fig. 10.20 A graded-index multimode fibre

The value of n_1 decreases until it approaches the value of the refractive index of the cladding 'n_2'. The value of n_1 must be higher than the value of n_2 to allow for proper mode propagation. Like the step-index fibre, the value of n_2 is constant and has a slightly lower value than the maximum value of n_1. The relative refractive index difference is determined using the maximum value of n_1 and the value of n_2.

The numerical aperture of a multimode graded-index fibre is at its maximum value at the fibre axis. This numerical aperture is approximately equal to

$$n_1\sqrt{2\Delta} \tag{10.27}$$

The refractive index of a graded-index fibre is given by

$$n(r) = n_1\left[1-2\Delta\left(\frac{r}{\alpha}\right)^d\right]^{1/2} \qquad \text{for} \quad 0 \le r \le \alpha \tag{10.28}$$

$$= n_1\,[1-2\Delta]^{1/2} \qquad \text{for} \quad r \ge \alpha$$

where r is the radial distance from the fibre axis,

d is the core radius,

n_1 is the refractive index of the core,

n_2 is the refractive index of the cladding,

Δ is the refractive index difference, and

α is the refractive index profile $\left[\dfrac{n_1^2 - n_2^2}{2n_1^2}\right]$.

For parabolic type graded-index fibre, $\alpha = 2$.

$$\text{Numerical Aperture (NA)} = [n(r)^2 - n_2^2] \tag{10.29}$$

$$= n_1\sqrt{2\Delta}\sqrt{1-\left(\frac{r}{\alpha}\right)^{\infty}} \tag{10.30}$$

$$= 0 \qquad \text{for } r > \alpha$$

In case of graded-index fibre, the number of modes propagated inside the fibre is ½ times the total number of guided modes propagating in the multimode step-index fibre ($M_{\text{Graded}} = M_{\text{step}}/2$).

Multimode graded-index fibres accept less light than multimode step-index fibres with the same core. The core's parabolic refractive index profile causes multimode graded-index fibres to have less modal dispersion.

Multimode graded-index fibres offer the following properties.

1. Relatively high source-to-fibre coupling efficiency

2. Low loss
3. Low sensitivity to microbending and macrobending
4. High bandwidth
5. Expansion capability

10.10.3 Differences between Step-index and Graded-index Fibres

S.No.	Step-index Fibres	Graded-index Fibres
1	The refractive index of the core is uniform throughout the core and it undergoes an abrupt or steep change at the cladding boundary.	The refractive index of the core is made to vary in the parabolic manner such that maximum refractive index is present at the centre of the core.
2	Step-index fibres in which the diameter of the core is about 50 to 200 μm in case of multimode operation and 10 μm in case of single-mode operation.	In graded-index fibres, the diameter of the core is about 50 μm in case of multimode fibre.
3	The light rays propagating through step-index fibres are in the form of meridional rays.	The light rays propagating through graded-index fibres are in the form of skew or helical rays.
4	The bandwidth of the fibre is about 50 MHz km for multimode step-index fibre and is greater than 100 MHz km for single-mode step-index fibre.	The bandwidth of the fibre is from 200 MHz km to 600 MHz km even though it has infinite bandwidth.
5	The attenuation is more for multimode fibres and less for single mode fibres.	The attenuation in a graded-index fibre is less.
6	The numerical aperture is more for multimode fibres and very less for single-mode fibres.	But in graded-index fibres, numerical aperture is less.
7	The signal distortion is more in multimode fibres and no distortion in single-mode fibres.	In graded-index fibres, signal distortion is very low.

10.10.4 Differences between Single-Mode and Multimode Fibres

S.No.	Single-Mode Fibres	Multimode Fibres
1	In a single-mode fibre, only one mode can propagate through the fibre.	In a multimode fibre, a large number of paths of modes for the light rays travelling through it.
2	Single-mode fibres have smaller core diameter 10 μm and the difference between refractive indices of core and cladding is very small.	In graded-index fibres, core diameter and the relative refractive index difference is large.

3	There is no degradation of signals during travelling through the fibre in case of single-mode type.	In multimode type, there is signal degradation due to multimode dispersion and material dispersion.
4	Due to high bandwidth, single-mode fibres are suitable for long-distance communication.	But due to large dispersion and attenuation, multimode fibres are less suitable for long-distance communication. They are suitable for local area network.
5	Launching of light into single-mode fibres and joining two fibres are very difficult.	It is easy in case of multimode fibres.
6	Single-mode fibres are difficult to manufacture and so are very costly.	Multimode fibres are not less difficult to fabricate and are not costly.

10.10.5 Types of Fibres Based on Material Composition

Fibres used for optical communication applications must guide light efficiently with low scattering, low absorption or attenuation and low dispersion. Materials satisfying these requirements are glasses and plastics.

There are three types of material that make up fibre-optic cables:

1. Glass
2. Plastic
3. Plastic-Clad Silica (PCS)

The above three types of cables differ with respect to attenuation. Attenuation is principally caused by two physical effects such as absorption and scattering. **Absorption** removes signal energy in the interaction between the propagating light and molecules in the core. **Scattering** redirects light out of the core to the cladding.

1. Glass Fibres

Glass fibre-optic cable has the lowest attenuation. A pure-glass, fibre-optic cable has a glass core and a glass cladding. This cable type has the most widespread usage. The glass used in a fibre-optic cable is ultra-pure, ultra-transparent, silicon dioxide or fused quartz. During the glass fibre-optic cable fabrication process, impurities are purposely added to the pure glass to obtain the desired indices of refraction needed to guide light. Germanium, titanium or phosphorous is added to increase the index of refraction. Boron or fluorine is added to decrease the index of refraction. Other impurities might somehow remain in the glass cable after fabrication. These residual impurities can increase the attenuation by either scattering or absorbing light.

Silica (SiO_2) glass fibres have very low loss and they are used in long-distance communication. These fibres will act as a transmission window at the wavelengths of 1.3 μm and 1.55 μm. There are also multicomponent glass fibres like sodium borosilicate glass fibre and soda lime silicate glass fibre. They have higher losses and so they are used in the endoscopic applications.

2. Plastic Fibres

Plastic fibre-optic cable has the highest attenuation among the three types of cable. Plastic fibre-optic cable has a plastic core and cladding. This fibre-optic cable is quite thick. The core generally consists of polymethylmethacrylate (PMMA) coated with a fluropolymer. Plastic fibre-optic cable was pioneered principally for use in the automotive industry. Plastic fibre-optic cable does have a problem with flammability. Because of this, it might not be appropriate for certain environments and care has to be taken when it is run through a plenum. Otherwise, plastic fibre is considered extremely rugged with a tight bend radius and the capability to withstand abuse.

The following are the examples of plastic fibres.

(i) *Core*: Polystyrene ($n = 1.6$)
Cladding: Methyl methacrylate ($n = 1.49$)

(ii) *Core*: Polymethyl methacrylate ($n = 1.49$)
Cladding: Its co-polymer ($n = 1.40$)

3. Plastic Clad Silica Fibre (PCS Fibre)

The attenuation of PCS fibre-optic cable falls between that of glass and plastic. PCS fibre-optic cable in which the core is glass of vitreous silica and the cladding is silicone elastomer plastic with a lower refractive index. Teflon is also used as the buffer coating material. PCS fabricated with a silicone elastomer cladding suffers from three major defects.

1. It has considerable plasticity, which makes connector application difficult.
2. Next, adhesive bonding is not possible.
3. Third, it is practically insoluble in organic solvents.

These PCS fibres have numerical aperture due to large difference between the refractive index of core and cladding.

4. Dopants Used for Cladding

Dopants are the materials which are added with SiO_2 (core) to produce similar material having slightly different refractive index called cladding. In some cases, doped silica acts as core and pure silica as cladding. The following gives some of the preferred composition of materials for a manufacturer core and cladding used in optical fibre.

Core	**–**	**Cladding**
SiO_2	–	B_2O_3– SiO_2
GeO_2–SiO_2	–	SiO_2
P_2O_5–SiO_2	–	SiO_2
GeO_2–B_2O_3–SiO_2	–	B_2O_3–SiO_2

Generally, SiO_2 has a refractive index of 1.46 at 850 nm. When TiO_2, Al_2O_3, GeO_2 AND P_2O_5 are added with silica, its refractive index increases whereas addition of B_2O_3 or fluorine with silica decreases its refractive index. The value of new refractive index of the doped material increases with the concentration of dopants.

5. Limitations of Optical-Fibre Cables

There are two basic limitations of an optical fibre.

1. The first is actual loss of light as it travels through the fibre.
2. The other is a maximum limitation of the bandwidth of the signals that can be carried.

Loss of light in an optical fibre is the result of absorption and impurities within the glass layer as well as losses caused by mechanical strains that bend the fibre at an angle that is so sharp that the light is actually able to "leak out" through the cladding region.

Losses are also dependent on the wavelength of the light employed since the amount of light absorbed by glass varies at different wavelengths. At 850 nanometres, a typical fibre has a loss of 4 to 5 dB per kilometre of length. At 1310 nanometres, this loss drops to under 3 dB per kilometre. The last two wavelengths are, therefore, obviously used for longer transmission distances. As a point of reference, typical well-designed fibre-optic transmission systems can sustain losses of anywhere from 10 to 30 dB.

Losses due to attenuation are independent of the frequency or data rate of the signals being transmitted. There is another loss factor. However, that is frequency related and is due to the fact that light can have many paths through a fibre. Figure 10.21 shows the mechanism of this loss through a step-index fibre.

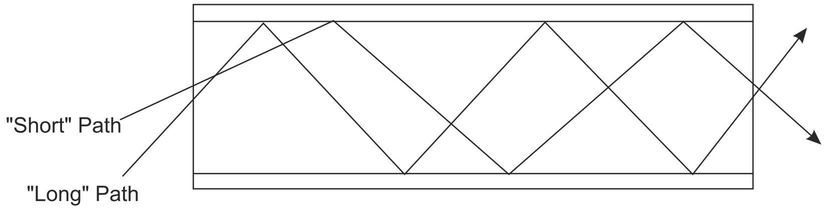


Fig. 10.21 Various light path lengths through a fibre

A light path nearly straight through a fibre is shorter than a light path with maximum bouncing. This means that for a fast rise-time pulse of light at the input to the fibre, some paths will result in light reaching the end of the fibre sooner than through other paths. This causes a spreading effect on the output rise and fall time of the light pulse which limits the maximum speed of light changes in a fibre and the maximum data rate of the fibre.

The disadvantages of fibre-optic cables and optical fibre communication system as a whole are listed as follows:

1. The termination equipment for fibre optics is costly compared to that of copper cable communication.

2. Repeated electrical to optical to electrical conversion is required as the whole communication is in optical domain.
3. Lack of clear-cut international specifications and guidelines for latest optical-communication-based systems.
4. Optical fibres are delicate and trained people are required to handle optical fibres.
5. Optical-fibre splicing and protection is still expensive and adds to the cost of the optical networks.
6. High cost of installation.

10.10.6 Multifibre Cable Systems

Multifibre cable systems are constructed with strength members that resist crushing during cable pulling and bends. The outer cable jackets are riser-rated, plenum-rated or low-smoke, zero-halogen-rated. The riser-rated outer jackets are composed of flame-retardant PVC or fluoropolymers. The plenum-rated jackets are composed of plenum PVC, whereas the zero-halogen-rated jackets are halogen-free and constructed out of polyolefin compounds.

Figure 10.22 shows a multiribbon (24-fibre) ribbon-cable system.

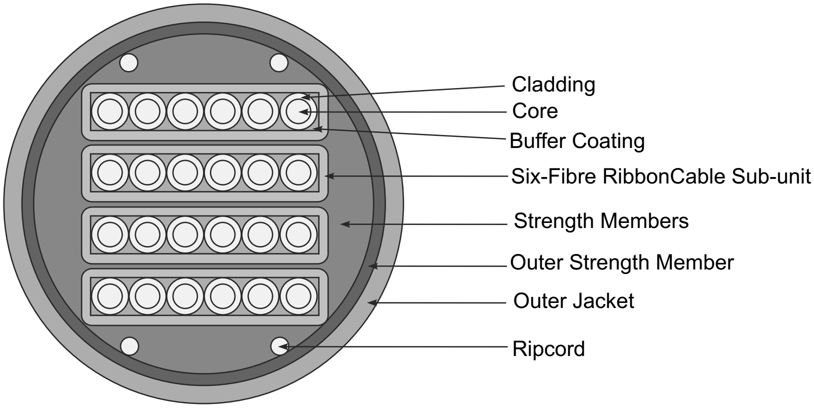


Fig. 10.22 A multiribbon cable system

Ribbon cables have a flat ribbonlike structure that enables installers to save conduit space as they install more cables in a particular conduit. Figure 10.23 shows a typical six-fibre, inside-plant cable system.

The central core is composed of a dielectric strength member with a dielectric jacket. The individual fibres are positioned around the dielectric strength member. The individual fibres have a strippable buffer coating. Typically, the strippable buffer is a 900 μm tight buffer. Each individual coated fibre is surrounded with a sub-unit jacket. Aramid yarn strength members surround the individual sub-units. Some cable systems have an outer strength member that

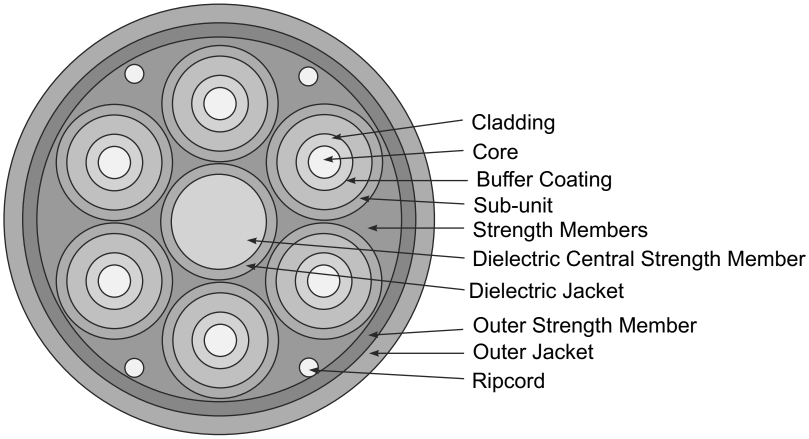


Fig. 10.23 A typical six-fibre inside-plant cable system

provides protection to the entire enclosed fibre system. Kevlar is a typical material used for constructing the outer strength member for premise cable systems. The outer jacket is riser-rated, plenum-rated or low-smoke, zero-halogen-rated.

Figure 10.24 shows a typical armoured outside-plant cable system.

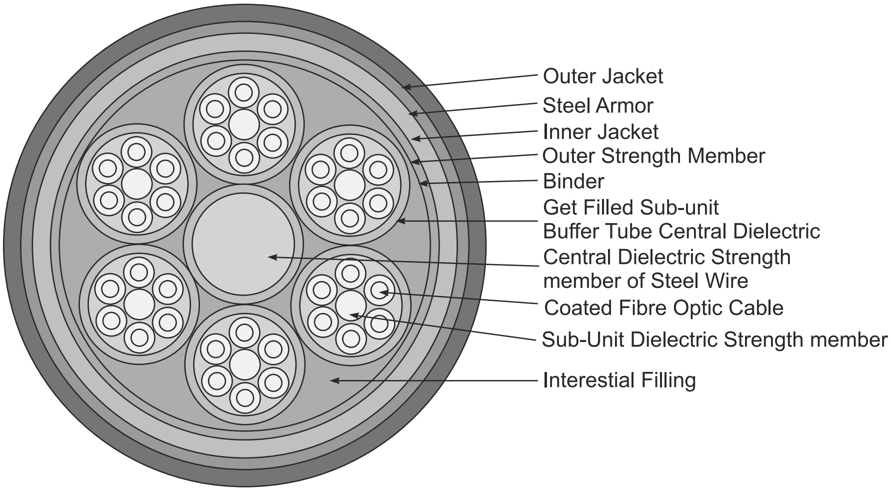


Fig. 10.24 A typical armoured outside-plant cable system

The central core is composed of a dielectric with a dielectric jacket or steel strength member. The individual gel-filled sub-unit buffer tubes are positioned around the central strength member. Within the sub-unit buffer tube, there are six fibres positioned around an optional dielectric strength member. The individual fibres have a strippable buffer coating.

All six sub-unit buffer tubes are enclosed within a binder that contains a water-blocking compound. An outer-strength member, typically constructed of strength members encloses the binder. The outer-strength member is surrounded by an inner Medium-Density Poly-Ethylene (MDPE) jacket. The corrugated steel armour layer between the outer High-Density Poly-Ethylene (HDPE) jacket and the inner MDPE jacket acts as an external strength member and provides physical protection. Conventional deep-water submarine cables use dual armour and a special hermetically sealed copper tube to protect the fibres from the effects of deep-water environments.

10.11 OPTICAL-FIBRE CONFIGURATIONS

By means of refraction and reflection, the light ray can be propagated down an optical fibre. The light propagation depends on the mode of propagation and the index profile of the fibre.

10.11.1 Mode of Propagation

A fibre-optic cable has two propagation modes, single mode and multimode. They perform differently with respect to both attenuation and time dispersion. The single-mode fibre-optic cable provides much better performance with lower attenuation.

When the light wave is guided down a fibre-optic cable, it exhibits certain modes. These are variations in the intensity of the light, both over the cable cross section and down the cable length. These modes are actually numbered from lowest to highest. In a very simple sense, each of these modes can be thought of as a ray of light. For a given fibre-optic cable, the number of modes that exist depends on the dimensions of the cable and the variation of the indices of refraction of both core and cladding across the cross section.

The following are the different modes of light propagation through an optical fibre.

1. Single-mode step-index
2. Single-mode dual-step-index
3. Multimode step-index
4. Multimode graded-index

1. Single-Mode Step-index Fibre

Figure 10.25 illustrates the single-mode propagation with a refractive index profile that is called step-index.

From the above figure, the diameter of the core is fairly small relative to the cladding. Because of this, when light enters the fibre-optic cable on the left, it propagates down toward the right in just a single ray, a single mode, which is the lowest-order mode. This lowest-order mode is confined to a thin cylinder around the axis of the core. The higher-order modes are absent.

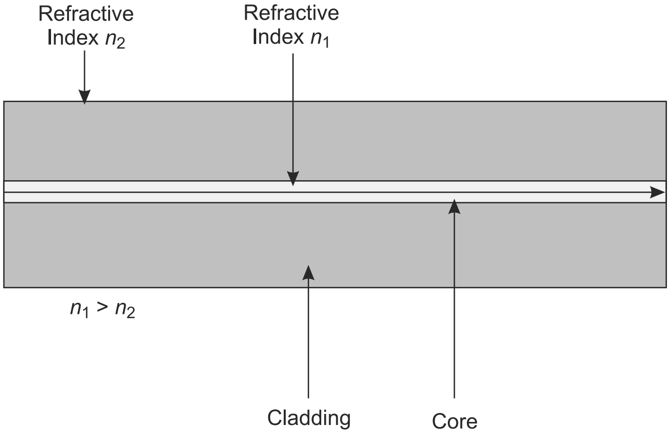


Fig. 10.25 Single-mode step-index fibre

Since the input signal is confined to a single ray path, that of the lowest-order mode, very little chromatic dispersion occurs. Single-mode propagation exists only above a certain specific wavelength called the **cut-off wavelength**. The cut-off wavelength is the smallest operating wavelength when single-mode fibres propagate only the fundamental mode. At this wavelength, the second-order mode becomes lossy and radiates out of the fibre core. As the operating wavelength becomes longer than the cut-off wavelength, the fundamental mode becomes increasingly lossy. The higher the operating wavelength is above the cut-off wavelength, the more power is transmitted through the fibre cladding. As the fundamental mode extends into the cladding material, it becomes increasingly sensitive to bending loss.

The core diameter for this fibre-optic cable is exceedingly small, ranging from 8 microns to 10 microns. The standard cladding diameter is 125 microns. Single-mode step-index fibres are manufactured using the Outside Vapour Deposition (OVD) process. OVD fibres are made of a core and cladding, each with slightly different compositions and refractive indices. The OVD process produces consistent, controlled fibre profiles and geometry. Fibre consistency is important, to produce seamless spliced interconnections using fibre-optic cable from different manufacturers. Single-mode fibre-optic cable is fabricated from silica glass. Because of the thickness of the core, plastic cannot be used to fabricate single-mode fibre-optic cable. Note that not all single-mode fibres use a step-index profile. Some of its variants use a graded-index method of construction to optimise performance at a particular wavelength or transmission band.

2. Single-Mode Dual-Step-index Fibre

These fibres are single-mode and have a dual cladding. Depressed-clad fibre is also known as doubly clad fibre. Figure 10.26 corresponds to single-mode propagation with a refractive index profile that is called dual-step-index.

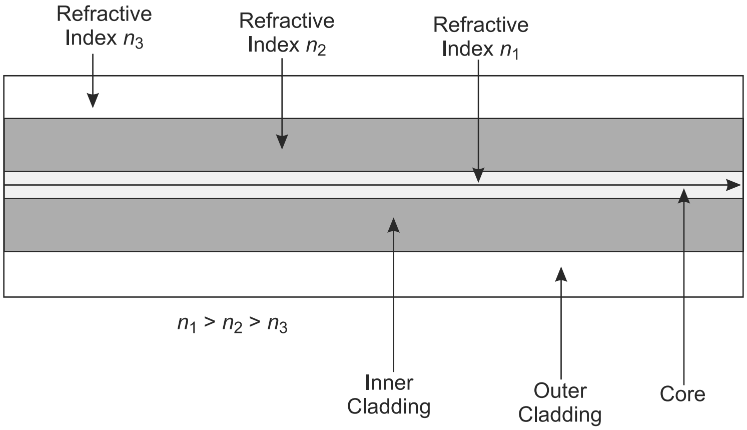


Fig. 10.26 Single-mode dual step-index fibre

A depressed-clad fibre has the advantage of very low macrobending losses. It also has two zero-dispersion points and low dispersion over a much wider wavelength range than a single-clad fibre. Single-mode depressed-clad fibres are manufactured using the Inside Vapour Deposition (IVD) process. The IVD or Modified Chemical Vapour Deposition (MCVD) process produces what is called depressed-clad fibre because of the shape of its refractive-index profile, with the index of the glass adjacent to the core depressed. Each cladding has a refractive index that is lower than that of the core. The inner cladding has lower refractive index than the outer cladding.

3. Multimode Step-index Fibre

Figure 10.27 illustrates the multimode propagation with a refractive index profile that is called step index. It corresponds to multimode propagation with a refractive index profile that is called step index. Here, the diameter of the core is fairly large relative to the cladding. There

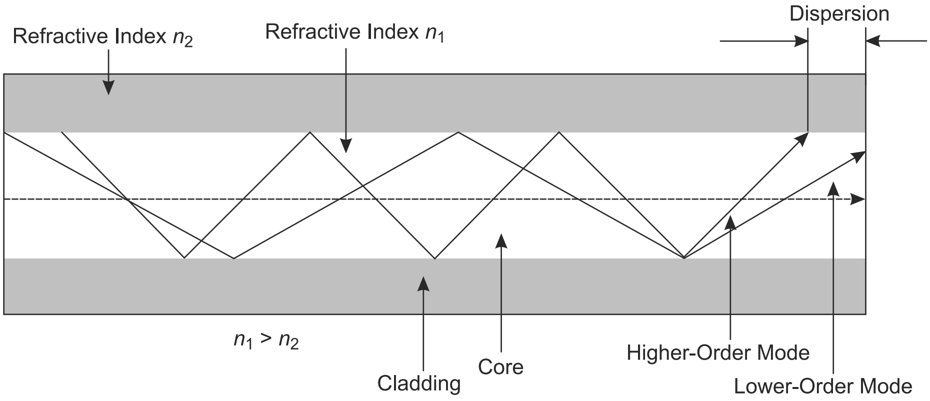


Fig. 10.27 Multi-mode step-index fibre

is also a sharp discontinuity in the index of refraction as you go from core to cladding. As a result, when light enters the fibre-optic cable on the left, it propagates down toward the right in multiple rays or multiple modes. This yields the designation multimode.

As indicated in the above figure, the lowest-order mode travels straight down the centre. It travels along the cylindrical axis of the core. The higher modes represented by rays, bounce back and forth, going down the cable to the left. The higher the mode, the more bounces per unit distance down to the right.

For the higher-order modes, the bouncing rays tend to leak into the cladding as they propagate down the fibre-optic cable and they lose some of their energy into heat which results in an attenuated output signal. The input pulse is split among the different rays that travel down the fibre-optic cable. For the lowest-order mode, the bouncing rays are all traversing paths of different lengths from input to output and they do not all reach the right end of the fibre-optic cable at the same time. When the output pulse is constructed from these separate ray components, the result is chromatic dispersion.

Fibre-optic cable that exhibits multimode propagation with a step-index profile is thereby characterised as having higher attenuation and more time dispersion than the other propagation methods. However, it is also the cheapest and widely used cables. These cables are especially attractive for link lengths up to 5 kilometres and they can be fabricated either from glass, plastic or PCS.

4. Multimode Graded-index Fibre

Generally, multimode graded-index fibre has a higher refractive index in the core that gradually reduces as it extends from the cylindrical axis outward. The core and cladding combination is essentially a single graded unit. Figure 10.28 shows a multimode graded-index fibre.

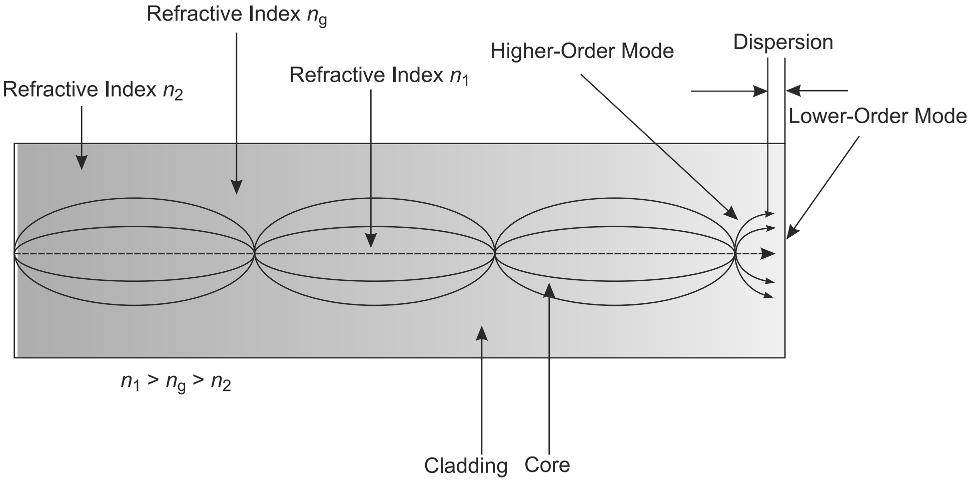


Fig. 10.28 Multimode graded-index fibre

This corresponds to multimode propagation with a refractive index profile that is called graded-index in which the variation of the index of refraction is gradual as it extends out from the axis of the core through the core to the cladding. There is no sharp discontinuity in the indices of refraction between core and cladding. The core here is much larger than in the single-mode step-index case.

Multimode propagation exists with a graded-index. However, the paths of the higher-order modes follow a series of ellipses. Due to their confined higher mode paths, the attenuation through them due to leakage is more limited than with a step index. The time dispersion is more limited than with a step index and therefore, attenuation and time dispersion are limitedly present. Glass is generally used to fabricate multimode graded-index fibre-optic cable.

10.11.2 Index Profile

The index profile of an optical fibre is a graphical representation of the refractive index across the fibre. The refractive index is plotted on the horizontal axis and the radial distance from the core axis is plotted on the vertical axis. Figure 10.29 shows the index profiles for different types of optical-fibre cables.

For a step-index fibre core, there is a central core with a uniform refractive index and an outside cladding that also has a uniform refractive index surrounds the core. But the refractive index of the cladding is lower than that of central core.

In graded-index fibre, it is noted that there is no cladding and the refractive index is also non-uniform. The refractive index of the core varies parabolically such that it is maximum at the core axis and minimum at the core-cladding boundary.

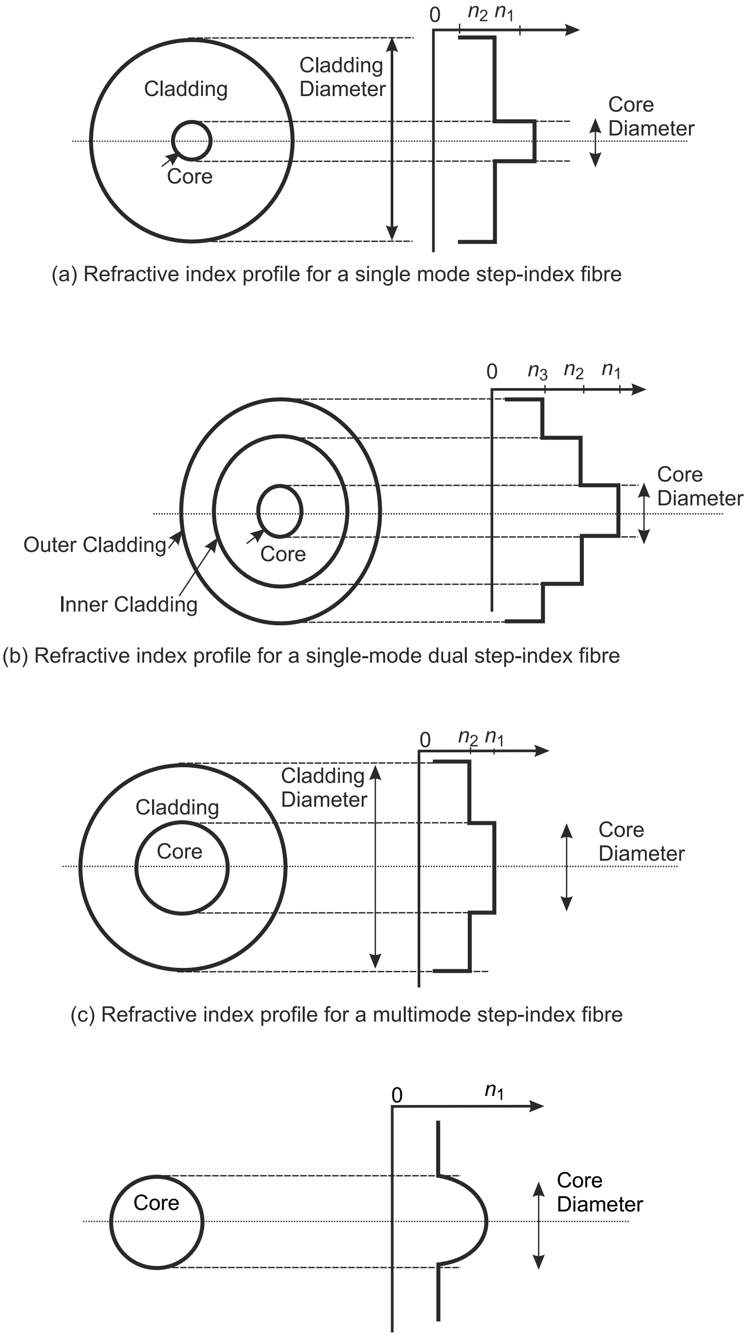
10.12 LOSSES IN OPTICAL FIBRES

There are several light losses which may occur during transmission of light signal inside the fibre or during the interconnection process of two fibres. They are listed as follows.

1. Absorption loss
2. Rayleigh scatter loss
3. Bending loss
4. Insertion loss
5. Return loss

10.12.1 Absorption Loss

Light travels best in clear substances. Impurities such as metal particles or moisture in the fibre can block some of the light energy. They absorb the light and dissipate it in the form of



(a) Refractive index profile for a single mode step-index fibre

(b) Refractive index profile for a single-mode dual step-index fibre

(c) Refractive index profile for a multimode step-index fibre

(d) Refractive-index profile for a multimode graded-index fibre

Fig. 10.29 Refractive-index profiles for different types of optical cables

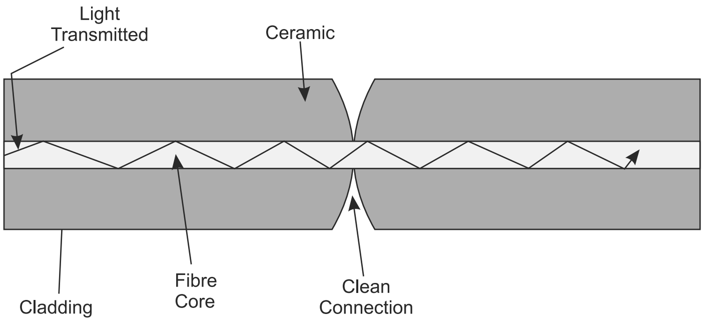


Fig. 10.30 Clean fibre without absorption loss

heat energy, which is cause for absorption loss. This loss can be reduced by using ultra-pure glass and dopant chemicals, especially to minimise impurities. Figure 10.30 shows the clean fibre without absorption loss.

10.12.2 Rayleigh Scatter Loss

Rayleigh scatter loss occurs at random when there are small changes in the refractive index of materials in which the light signal travels. In this case, it is the changes in the refractive index of the core and the cladding of the fibre optic cable. This loss is caused by the miniscule variation in the composition and density of the optical glass material itself, which is related to the fibre manufacturing process. Figure 10.31 shows the light scattering during transmission.

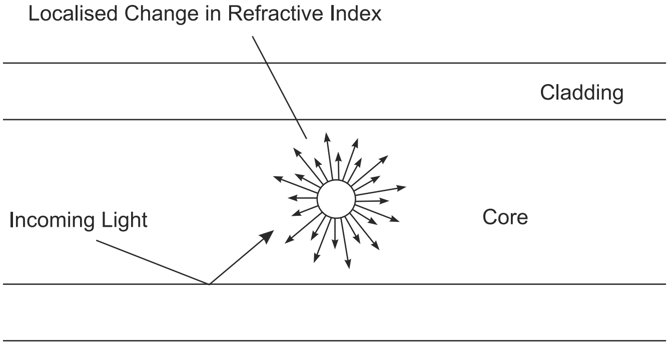


Fig. 10.31 Light scattering during transmission

10.12.3 Bending Loss

There are two forms of bending losses. They are as follows.

1. Macrobending
2. Microbending

When a cable is bent, it disrupts the path of the light signal. The tighter the bends of a cable, the greater it is of the light loss. Figure 10.32 shows the bending radii of the optical fibre.

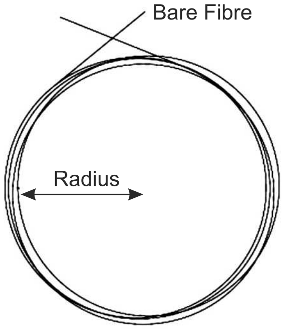


Fig. 10.32 Bending radii of the optical fibre

1. Macrobends

Macrobends means that the bending of the fibre optic cable in a tight radius. The bend curvature creates an angle that is too sharp for the light to be reflected back into the core which will cause an optical power loss. This optical power loss increases rapidly as the radius is decreased to an inch or less. Different fibre-optic cables have different specifications on how much the cable can bend without affecting the stated performance or loss. As an example, the recent G.657.B.3 fibre standard recommended by the International Telecommunication Union (ITU) will have the bending radius standardised as low as 5 mm.

2. Microbends

In comparison with macrobends, microbends refer to minute but sever bends in fibre that result in light displacement and increased loss. Pinching or squeezing the fibre is the cause of microbends and due to which there will be deformation in the fibre's core slightly, causing light to escape at these deflections. Microbending can be avoided by the correct selection of materials and proper cabling, handling and installation techniques.

10.12.4 Insertion Loss

An important and noticeable loss during the insertion of a fibre optic connector is the insertion loss. It is the loss of light signal and measured in decibels (dB). There are several causes of insertion losses. They are listed as follows.

1. The misalignment of ferrules during connection
2. The air gap between two mating ferrules
3. Absorption loss from impurities such as scratches and oil contamination

Insertion loss can be minimised by proper selection of interconnect materials, good polishing and termination process of fibre connectors.

10.12.5 Return Loss

Return loss, also known as **back-reflection loss**, is the loss of light signal that is reflected back to the original light source. This will occur when the light is reflected off the connector and travelled back along the fibre to the light source. This phenomenon is also known as the **Fresnel reflection**. It will also occur when there are changes in the refractive index of materials in which the light travels, such as the fibre core and the air gap between fibre interconnection. When light passes through these two different refractive indexes, some of the light signal is reflected back.

As a general rule, the greater the difference between two materials' refractive indices, the higher the loss. Regarding the return-loss figures, the higher the absolute value of the decibel unit means the better the performance of the interconnection.

10.13 OPTICAL CONNECTORS AND SPLICES

Fibre-optic connections permit the transfer of optical power from one component to another. Fibre-optic connections also permit fibre-optic systems to be more than just point-to-point data communication links. In fact, fibre-optic data links are often of a more complex design than point-to-point data links.

The main purpose of an optical fibre connector is to terminate the end of an optical fibre. It also enables quicker connection and disconnection than splicing. The optical fibre connector mechanically couples and aligns the cores of fibres so that light can pass.

There are many types of optical connectors. The usage of a particular connector depends on the particular equipment used it with and for the desired application. It is a mechanical device mounted on the end of a fibre-optic cable, light source, receiver or housing. The connector must direct light and collect light and must be easily attached and detached from equipment. A connector marks a place in the premises of fibre-optic data link where signal power can be lost and the BER can be affected by a mechanical connection.

10.13.1 Structure of the Optical-Fibre Connector

Figure 10.33 shows the structure of an optical-fibre connector.

The main component of the connector is the ferrule which can be constructed from a metallic or ceramic material and the more commonly used connectors such as FC/PC, ST, DIN and Diamond HMS-10 will have a diameter of 2.5 mm. The optical fibre runs along the length of the ferrule and exits centrally at its end face. It is here that the fibre and ferrule end face is cut to the connector's specification. Typical values of return loss for a connector

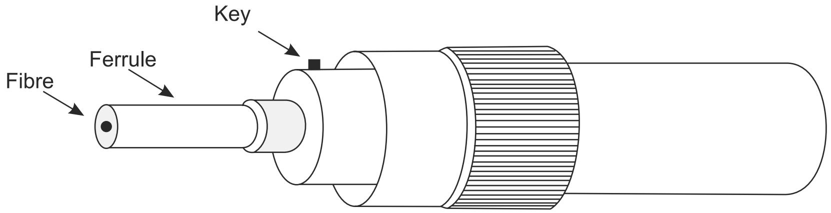


Fig. 10.33 Structure of an optical-fibre connector

can range from 14.6 dB for a straight connector end to even more than 60 dB for a slanted connector end. Next, key is a protrusion that ensures that the connector always aligns in the correct orientation when connected to its female counterpart.

Most optical-fibre connectors are spring-loaded. The fibre end faces of the two connectors are pressed together, resulting in a direct glass-to-glass or plastic-to-plastic contact, avoiding any glass-to-air or plastic-to-air interfaces which would finally result in higher connector losses and in higher reflection. Although there are many different optical-fibre connector types in use, the most common elements in a fibre connector can be shown in Figure 10.34.

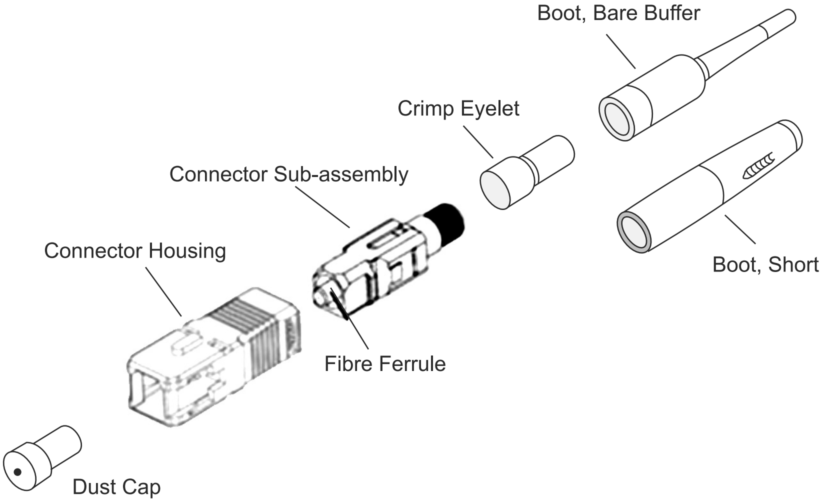


Fig. 10.34 Components of an optical-fibre connector

10.13.2 Optical-Fibre Connector Interface

Whenever the light from an optical fibre is required to pass into or out of an optical component or system, it is necessary to modify the divergent light that emits from the end of the fibre to match the characteristics of that component or system. Figure 10.35 shows such an interface in its simplest form. Here, a lens is used as the interface. The purpose of the lens is to collect the divergent beam of light that emits from the end of the optical fibre and convert it into a

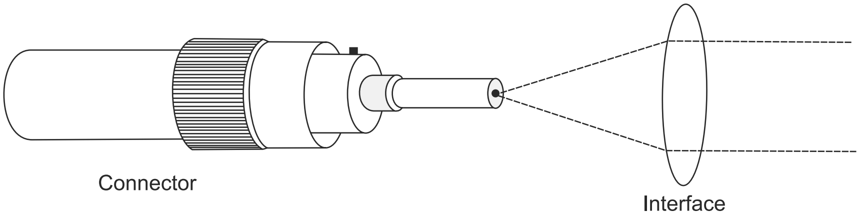


Fig. 10.35 Optical-fibre connector interface

collimated or parallel beam. The converse is also true for this type of interface in that the lens can take a collimated beam of light and focus it into an optical fibre.

10.13.3 Losses in Optical-Fibre Connectors

There are two losses to be considered in optical-fibre connectors.

1. Attenuation or insertion loss
2. Reflection or return loss

Attenuation loss can be minimised by providing the fibre connectors with grades like A, B, C and D for single-mode type fibre and M for Multimode type fibre. Among the above grades, A is the best and D is the worst. Similarly, return loss can be minimised by means of RL 1, 2, 3, 4 and 5 graded fibres where grade-1 gives the best performance and grade-5 gives the lowest performance.

There are a variety of optical-fibre connectors in use based on the dimensions and methods of mechanical coupling. They are used to join optical fibres where a connect/disconnect capability is required. The basic connector unit is a connector assembly. Generally, a connector assembly consists of an adapter and two connector plugs. For the user's convenience, optical-fibre connectors are generally assembled onto optical fibre in a supplier's manufacturing facility.

10.13.4 Applications of the Optical-Fibre Connectors

Optical-fibre connectors are popularly used in telephone exchanges, at installations on customer premises and in outside plant applications. Their uses specifically include the following.

1. To make the connection between equipment and the telephone plant in the central office
2. To connect the fibres to remote and outside plant electronics such as optical network units and digital loop carrier systems
3. For optical cross-connection in the central office
4. For patching the panels in the outside plant to provide architectural flexibility and to interconnect fibres belonging to different service providers
5. To connect the couplers, splitters to optical fibres
6. To connect the optical test equipment to fibres for testing and maintenance

10.13.5 Features of Optical-Fibre Connectors

In order to have a good optical-fibre-connector design, the following features are to be considered.

1. Low insertion loss
2. Low return loss
3. Ease of installation

4. Low cost
5. Reliability
6. Low environmental sensitivity
7. Ease of use

10.13.6 Optical-Fibre Splicing

Optical fibres may be connected to each other by connectors or by splicing. Splicing is joining two fibres together to form a continuous optical waveguide. The generally accepted splicing method is **arc-fusion splicing** in which the fibre ends are melted together with an electric arc. In order to get speedy splicing, a mechanical splice can be used.

Fusion splicing is done with a specialised instrument that typically operates as follows: The two cable ends are fastened inside a splice enclosure that will protect the splices, and the fibre ends are stripped of their protective polymer coating as well as the sturdier outer jacket. The ends are cut with a precision cutter to make them perpendicular, and are placed into special holders in the splicer. The splice is usually inspected via a magnified viewing screen to check the cuts before and after the splice. The splicer uses small motors to align the end faces together and emits a small spark between electrodes at the gap to burn off dust and moisture. Then the splicer generates a larger spark that raises the temperature above the melting point of the glass, fusing the ends together permanently. The location and energy of the spark is carefully controlled so that the molten core and cladding do not mix and this minimises optical loss.

A **splice-loss estimate** is measured by the splicer, by directing light through the cladding on one side and measuring the light leaking from the cladding on the other side. A splice loss under 0.1 dB is typical. The complexity of this process makes fibre splicing much more difficult than splicing a copper wire. Mechanical fibre splices are designed to be quicker and easier to install, but there is still the need for stripping, careful cleaning and precision cutting. The fibre ends are aligned and held together by a precision-made sleeve, often using a clear index-matching gel that enhances the transmission of light across the joint. Such joints typically have higher optical loss and are less robust than fusion splices, especially if the gel is used. All splicing techniques involve installing an enclosure that protects the splice.

10.14 OPTICAL SOURCES

Optical beams are generated by light sources for carrying information in an optical fibre system. The two most common sources of light are

1. Light-Emitting Diode (LED)
2. Light Amplification by Stimulated Emission of Radiation (LASER)

Both the sources work on the principle of electro-luminance. It is a phenomenon of emission of optical radiation by converting electrical energy into light.

10.14.1 Requirements for Light Sources

1. They must emit the required wavelengths of 1.3 μm and 1.55 μm in the case of silica fibres.
2. It is necessary to modulate the source at high speeds, more than several Gb/s.
3. The light source should have compact size and high efficiency.
4. They should be reliable, durable and inexpensive.
5. They must require small power for its operation.

10.14.2 LASER

In LASER, the basic light-emission mechanism is the same as that of LED. The growth of Laser technology has stimulated a broad range of scientific and engineering applications that exploit the unique properties of Laser light. These properties are derived from the distinctive way Laser light is produced in contrast to the generation of ordinary light. Lasers are often termed monochromatic, coherent and collimated sources of light.

1. Laser Operation

A Laser is a device that emits light through a process of optical amplification based on the stimulated emission of photons. The emitted light from the Laser is notable for its high degree of spatial and temporal coherence, unattainable using other technologies.

(a) Stimulated Absorption It is known that electrons exist at specified energy levels or states characteristic of a particular atom. These energy levels can be imagined as orbits around the nucleus of an atom. Usually, the atoms exist in the lowest energy state or ground state. Now, the electrons of the atoms from the ground state can be pumped to higher energy levels by providing energy in different ways. Figure 10.36 shows the states before and after the stimulated absorption of a photon by an atom.

When energy supplied to atoms is equal to the gap between the energy levels then only electrons from lower energy state can move to the higher energy state. Here the supplied energy (E_S) must be

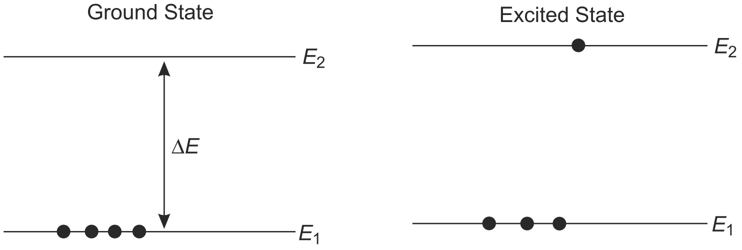


Fig. 10.36 States before and after the stimulated absorption of a photon by an atom

$$E_s = E_2 - E_1 \tag{10.31}$$

where E_2 is the higher energy level, and

E_1 is the lower energy level.

If energy pumping is done by light then depending on the material being used, specific wavelengths of light are absorbed to excite the electrons. This is known as stimulated absorption process.

(b) Spontaneous Emission In contrast to stimulated absorption, the emission process can occur in the following two different ways.

(i) If an electron spontaneously decays from higher to lower energy states, it emits a photon having energy equal to the energy difference between the two energy states. This is called spontaneous emission process. Figure 10.37 shows spontaneous emission of a photon.

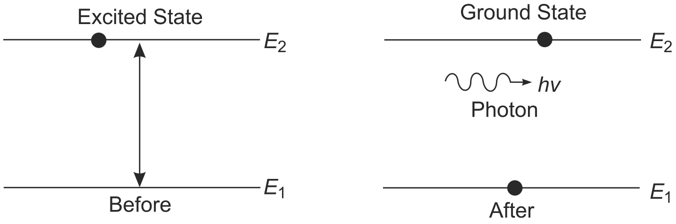


Fig. 10.37 Spontaneous emission of a photon

(ii) The frequency or wavelength of emitted radiation is related to the amount of energy released. So, depending upon the material being used, specific frequency (ν) or wavelength (λ), a photon is spontaneously emitted at the time of de-excitation. Therefore, the energy of the emitted radiation (E_e) must be

$$E_e = E_2 - E_1 = h\nu = \frac{hC}{\lambda} \tag{10.32}$$

where $$\nu = \frac{C}{\lambda}$$

(b) Stimulated Emission Practically, it is possible to force an emission process by means of a photon. It means that a photon of definite energy (E) can force an electron to move from higher to lower energy states having the same energy difference, yielding another photon, where

$$E = hv = E_2 - E_1 \tag{10.33}$$

This process results in the two photons (incoming and emitted) of the same energy and these two photons will be in the same phase. This is known as stimulated emission process. It is illustrated in Figure 10.38.

Therefore, an excited atom can relax to a stable state by relaxing a photon which is identical in energy, direction and phase with the incident photon. These two photons can interact with

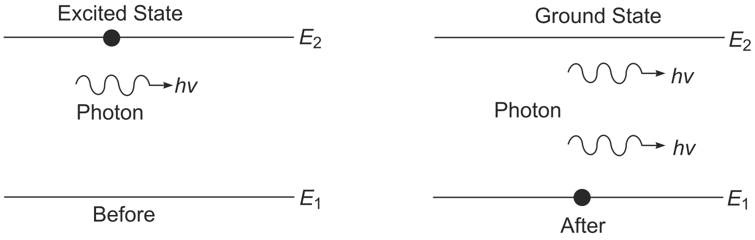


Fig. 10.38 Stimulated emission process

other excited atoms. Thus, started with one photon, it gets multiplied as two, four, eight photons, and so on. This amplification corresponds to a build-up of photons in the system as a result of the chain reaction of events.

2. Population Inversion

A population inversion occurs when a group of atoms exists in state with more members in an excited state than in lower energy states. This concept is fundamental in Laser science because the production of a population inversion is a necessary step in the workings of a standard Laser.

An incident photon can cause atomic transitions either upward (stimulated absorption) or downward (stimulated emission). When light is incident on a system of atoms, there is usually a net absorption of energy because there are many more atoms in the ground state than in the excited states in the case of thermal equilibrium.

There are a group of N atoms considered, each of which is capable of being in any one of the two following energy states such as,

1. The ground state, with energy E_1
2. The excited state with energy E_2, with $E_2 > E_1$

$$E = h\nu = E_2 - E_1$$

The number of these atoms which are in the ground state is given by N_1 and the number in the excited state is N_2. Since there are N atoms in total,

$$N_1 + N_2 = N \tag{10.34}$$

The energy difference between the two states is given by

$$\Delta E = E_2 - E_1 \tag{10.35}$$

The above expression determines the characteristic frequency v_{12} of light which will interact with the atoms and it is expressed as

$$\Delta E = E_2 - E_1 = h\nu_{12} \tag{10.36}$$

where h is being Planck's constant.

According to Boltzmann's law,

$$\frac{N_2}{N_1} = \exp^{\frac{-(E_2 - E_1)}{kT}} \tag{10.37}$$

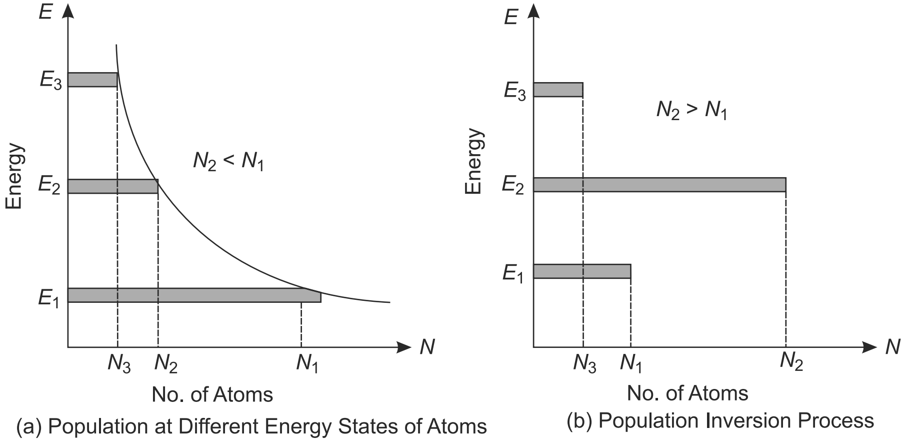
where

N_2 and N_1 are the populations in a given energy state E_2 and in the ground state E_1,

k is the Boltzmann's constant, and

T is the absolute temperature.

From the above expression, it is noted that the population is maximum in the ground state and decreases exponentially as one goes to higher energy states. It is also depicted in Figure 10.39.



(a) Population at Different Energy States of Atoms

(b) Population Inversion Process

Fig. 10.39 Population inversion process

In a normal situation, there are more atoms in the ground state ready to absorb photons than there are atoms in the excited state, ready to emit photons. However, if the situation is reverse, i.e. there are more atoms in an excited state than in the ground state, a net emission of photons can result. Such a condition is said to be population inversion. This can be done by providing an initial energy to the atoms by passing electrical current or illuminating it with a bright light pulse.

3. Types of Lasers

The following are the four different types of Lasers.

(a) Gas Lasers

(b) Solid Lasers

(c) Liquid Lasers

(d) Semiconductor Lasers

(a) Gas Lasers Gas Lasers consist of a mixture of helium and neon. This gaseous mixture is packed up into a glass tube and this packed mixture acts as an active medium. The pressure inside the tube is maintained at 1 torr for helium and 0.1 torr for neon. The length of the glass tube is approximately from 0.25m to 1m and diameter of the glass tube is nearly 1 cm. A typical gas Laser is shown in Figure 10.40.

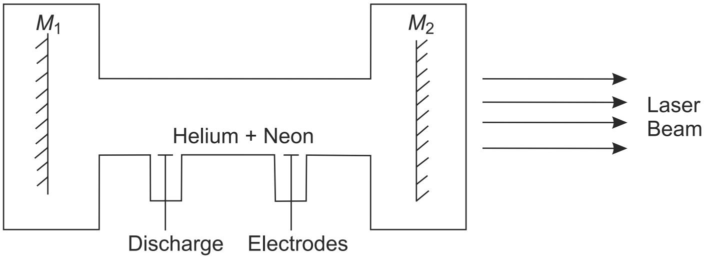


Fig. 10.40 Gas Laser

There are two electrodes present in the tube connected to a high-voltage dc source. This circuit results in the generation of the discharge inside the tube and this discharge works like a pump. There are two parallel mirrors also placed in front of each other and both the mirrors are present inside the tube. Only mirror M_1 shows the complete reflection and the mirror M_2 shows only partial reflection.

When the electric current is passed through the tube, a continuous light wave will start flowing inside the tube with constant frequency. It is also known as **coherent light wave**. It will come out from the side of the mirror M_2.

(b) Solid Lasers In solid Lasers, a rubylike crystal is used which acts as an active medium. It is basically cylindrical in shape. This crystal is surrounded by a xenon flash lamp *T*. This flash lamp is of helical shape. In this arrangement, this lamp acts as a pumping arrangement. Both the ends E_1 and E_2 of the crystal are properly polished. Similar to gas Lasers, the surface M_1 will do the complete reflection but on the other hand, M_2 will reflect partially. Whenever the current is passed through the arrangement, a Laser beam of red colour having large intensity will come out. Figure 10.41 shows a typical solid Laser.

(c) Liquid Lasers In liquid Lasers, organic dyes are used as active medium inside the glass tube. The complete circulation of dye is done in the tube with the help of a pump. From this organic dye, Laser light will emerge out.

(d) Semiconductor Lasers In semiconductor Lasers, junction diodes are used and in which the doping of a *PN* junction diode is done. Both the acceptors and donors are doped. These are known as ILD (Injection Laser Diodes). Whenever the current is passed, light modulation from the ILD can be seen. This is used in various electronic equipment.

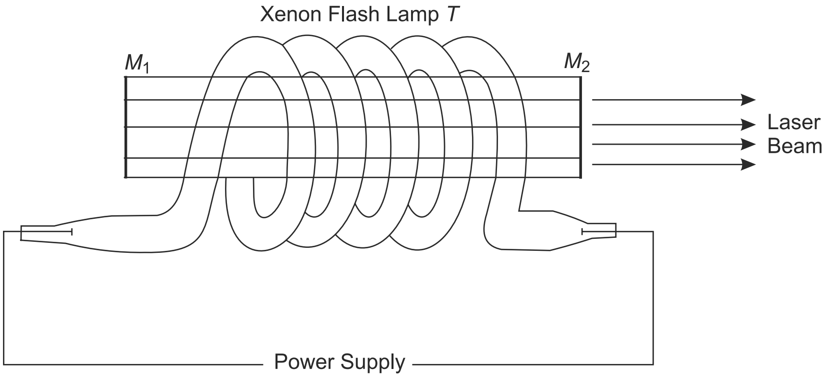


Fig. 10.41 Solid Laser

4. PN Junction Laser Diode

There are two basic types of *PN* junction laser diodes. They are listed as follows.

(a) Homojunction Laser

(b) Heterojunction Laser

(a) Homojunction Laser Homojunction Laser means that a *PN* junction is formed by a single crystalline material such that the basic material has been the same on both sides of the junction. For a GaAs laser, both *P* and *N* layers are made up of GaAs only.

But there are certain drawbacks of homojunction Lasers. They are listed as follows.

(i) Threshold current density is very large.

(ii) Only pulsed mode output is obtained.

(iii) Coherence and stability are very poor.

(iv) Electromagnetic field confinement is poor.

(b) Heterojunction Laser Heterojunction means that the material on one side of the junction differs from that on the other side of the junction. In a GaAs diode laser, a heterojunction is formed between GaAs and Ga Al As. These heterojunction Laser diodes are used as optical sources because of many advantages. They are listed as follows.

(i) Threshold current density is small.

(ii) Continuous-wave operation is also possible.

(iii) These are highly stable with longer life.

(iv) High output power can be achieved with low threshold current.

5. GaAs Laser Diode

A layer of GaAs is sandwiched between two layers of GaAlAs. GaAlAs has a wide energy gap and lower refractive index. Figure 10.42 shows the structure of a GaAs Laser diode.

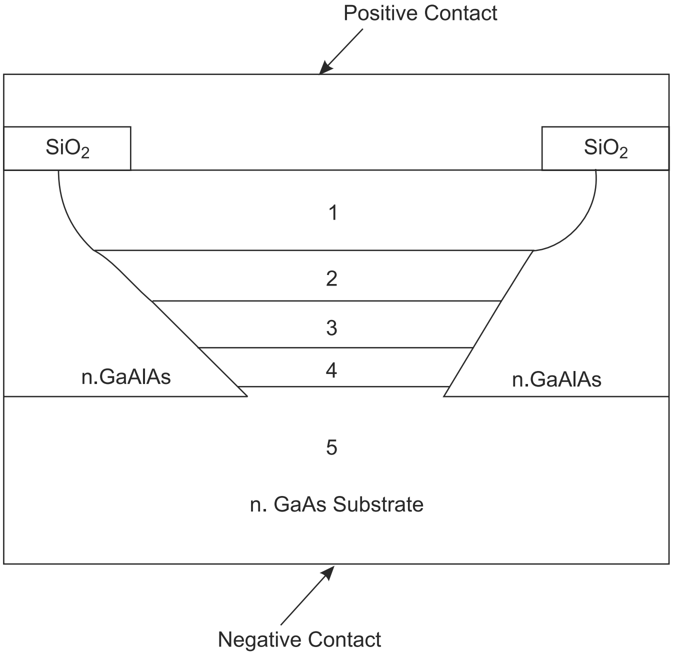


Fig. 10.42 Structure of a GaAs laser diode

In a GaAs diode Laser, a layer of GaAs is sandwiched between two layers of GaAlAs which have a wider energy gap and a lower refractive index than GaAs on GaAs substrate. The basic principle of working of a heterojunction laser diode is same with respect to homojunction Laser diode. The GaAs diode has *N-p-P* structure where *N* and *P* represent the wider bandgap semiconductors and *p* represents the narrow bandgap semiconductor.

Charge injection takes place into the active layer. As a result, spontaneous emission of photons is produced and some of the injected charges are stimulated to emit by other photons. A large number of injected charges are available for stimulated recombination and they create population inversion, provided the current density is sufficiently high and optical gain is also high. The bandgap difference prevents the diffusion of injected charge carriers from the active layer of a GaAs Laser. Stimulated radiation gives coherent Laser radiation.

The step change in refractive index provides an efficient waveguide structure. Thus, the injected charges and radiation are confined mainly to the active layer. This leads to get a lower value of threshold current. Further, the active region is few microns wide so that the threshold current is further reduced.

From the GaAs laser, the output wavelength is 0.8 μm when the bandgap of an active layer is 1.55 eV. The lower value of bandgap is due to narrow bandgap of the active layer. The front and back faces of the active layer are dielectric coated so that the reflections at the active-layer-to-air interface provide sufficient feedback for Laser oscillation.

The stripe geometry provides the confinement of charges in the lateral direction and longer life for the Laser diodes. Thus, high power output, narrow spectral width, high efficiency and high coherence can be achieved through the double heterojunction stripe Laser diodes.

6. Uses of Laser

There are a variety of applications of Lasers in various fields. Some of them are listed as follows.

(a) Scientific Applications

(i) Spectroscopy
(ii) Lunar Laser ranging
(iii) Material processing
(iv) Photochemistry
(v) Laser cooling
(vi) Nuclear fusion
(vii) Microscopy

(b) Military Applications

(i) Defensive countermeasures
(ii) Disorientation
(iii) Targeting
(iv) Firearms
(v) Eye-targeted lasers

(c) Medical Applications

(i) Cosmetic surgery
(ii) Removing scars, stretch marks, sunspots, wrinkles, birthmarks and hairs
(iii) Eye surgery and refractive surgery
(iv) Soft-tissue surgery
(v) General surgery
(vi) Gynaecological surgery
(vii) Urology
(viii) Laparoscopic surgery
(ix) Laser therapy
(x) Tooth whitening in dentistry

(d) Industrial and Commercial Applications

(i) Cutting and peeling of metals
(ii) Welding
(iii) Pollution monitoring
(iv) Range finders
(v) Barcode readers
(vi) Laser pointers

10.14.3 Light-Emitting Diodes (LEDs)

Light-emitting diode emits light by injection luminescence. Here, a *PN* junction diode is operated under forward bias. Under forward bias, majority carriers from both sides of the junction cross the internal potential barrier and enter the material at the other side where they are called **minority carriers** and cause the local minority-carrier population to be the larger than normal. This is called **minority-carrier injection**. The excess minority carriers diffuse away from the junction recombining with majority carriers as they do so. Figure 10.43 shows the recombination of the injected minority carriers with the majority carriers in a forward-biased *PN* junction.

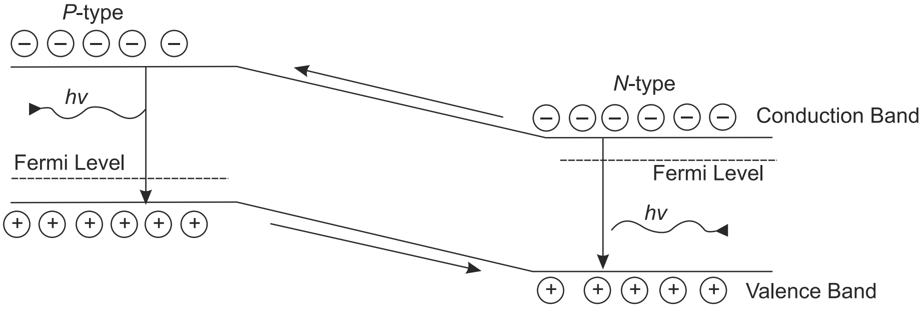


Fig. 10.43 Recombination of the injected minority carriers with the majority carriers in a forward biased PN junction

In LED, every injected electron takes part in a radiative recombination and hence gives rise to an emitted photon. In reverse bias, no carrier injection takes place and consequently no light is emitted.

The number of photons emitted is proportional to the carrier injection rate or the total current flowing. The wavelength of the emission is given by

$$\frac{hC}{\lambda} = E_C - E_\nu = E_g \tag{10.38}$$

$$\therefore \qquad \lambda = \frac{hC}{E_g}$$

By adding phosphorous with GaAs, the value of the bandgap is increased and the wavelength of the emitted radiation is in the visible range.

1. Requirements for a Suitable LED Material

For a suitable LED material, the following requirements are to be satisfied.

1. Energy gap should be lesser than 2 eV.
2. There should be both *P* and *N* types from that material.
3. There should have low resistivity.
4. Efficient radiative pathways must exist.

2. Structures of LED

There are high-radiance surface-emitting heterojunction LEDs and edge-emitting double heterojunction LEDs.

(a) Surface-Emitting LED Figure 10.44 shows the structure of a surface-emitting LED.

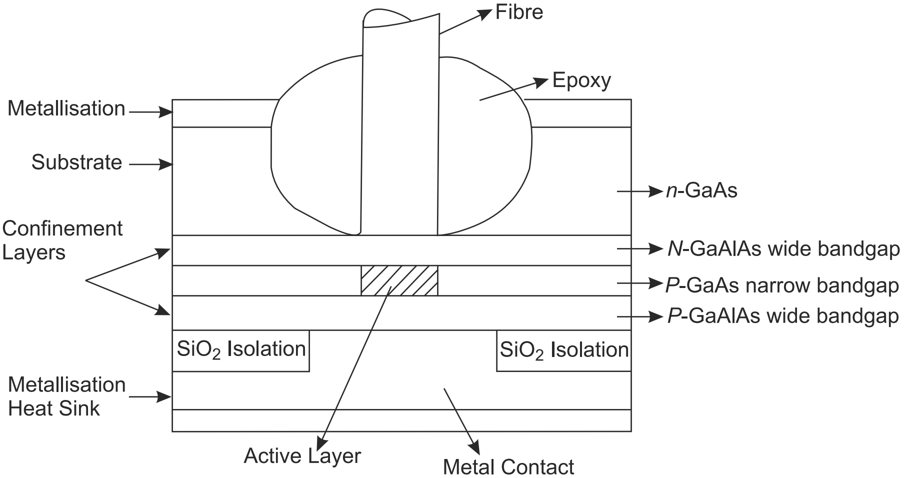


Fig. 10.44 Surface-emitting LED

In a surface-emitting LED, *N-p-P* structure forms a double heterojunction layer. A layer of GaAs is sandwiched between two layers of GaAlAs which has a wider bandgap and a lower refractive index. These two layers of GaAlAs form confinement layers. This dual confinement gives high efficiency and high radiance. The GaAs layer is the active layer or recombination region and is in the form of circular region. This area is typically 20 to 50 μm in diameter. In order to couple the fibres with the LED, the fibre-core diameter exceeding 100 μm and numerical aperture greater than 0.3 are highly preferred.

(b) Edge-Emitting LED Figure 10.45 shows the structure of an edge-emitting LED.

GaAs forms an active area which is in the form of a circular region at the middle of the active layer. GaAlAs layers form the optical confinement or light-guiding layers whose refractive index is lower than that of the active region.

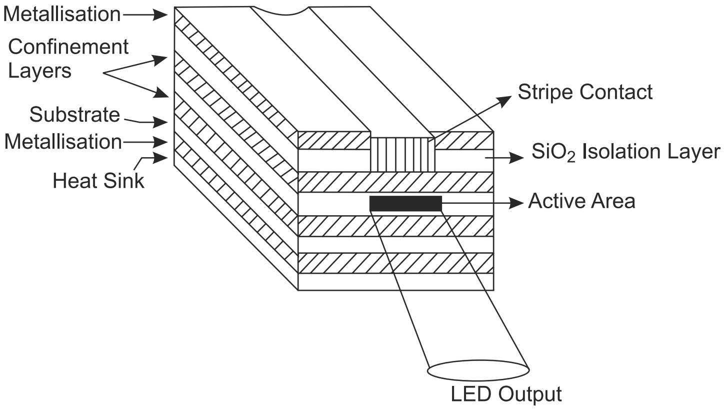


Fig. 10.45 Edge-emitting LED

The bandgap difference and refractive-index difference make the diode a waveguide. The output of the beam is highly incoherent. Lengths of the active regions range from 100 to 150 μm. So 50 to 100 μm core fibres can match these LEDs.

The radiance of the emitted beam is $B_\theta = B_0 \cos \theta$ where B_0 is the value of radiance at the centre of the beam. In the plane perpendicular to the junction, the half-power beam width is very small. The emission pattern of the edge-emitting diode is more directional than that of the surface-emitting LED.

3. General Properties of LEDs

The lifetime of an LED is about 10^5 hours. The edge emitters have lower drive current than the surface emitters. If the drive current is 100 mA, the optical output power is in between 0.5 μ*W* to 10 μ*W*. The manufacturing of an edge-emitting LED is more expensive than the surface-emitting LED. The modulation bandwidth of LED is comparatively low with rise time of 600 to 800 ns. But in GaAsP, edge-emitting LED at 1.3 μm; and GaAl, edge-emitting LED at 1.m; and GaAlAs, edge-emitting LED at 0.85 μm have smaller rise time. Thus, the bandwidth is greater than 200 MHz.

The spectral width of surface-emitting LEDs is about 1.3 nm and so the fibre bandwidth is reduced. For edge-emitting LEDs, the spectral width is more such that it is 50 nm at 0.85 μm wavelength and 70 nm at 1.3 μm wavelength.

10.15 OPTICAL DETECTORS

Photodetector is a device used to convert the light signals to electrical signals at the receiver end of the fibre link.

10.15.1 Requirements of Optical Detectors

A high-quality photodetector should satisfy the following requirements.

1. High quantum efficiency or conversion efficiency 'η' where η is the number of electrons produced per incident photon.
2. High spectral and frequency response in the range of operating wavelengths. Detector responsivity

$$R = \frac{I_p}{P_0} = \frac{\text{Photocurrent}}{\text{Input optical power}} = \eta \frac{q}{h\nu} \tag{10.39}$$

where $$\eta = \left[\frac{I_p / q}{P_0 / h\nu} \right] \tag{10.40}$$

3. Low dark current.
4. The signal dependent noise should be low.

10.15.2 *p-i-n* Photodiodes

A positive-intrinsic-negative (*p-i-n*) photodiode consists of *p* and *n* regions separated by a very lightly *n*-doped intrinsic region. Silicon pin diodes are used at 0.8 μm wavelength. In normal operation, the *pin* photodiode is under high reverse bias voltage. So the intrinsic region of the diode is fully depleted of carriers. When an accident photon has energy greater than or equal to the bandgap energy of the photodiode material, the electron-hole pair is created due to the absorption of photon. Such photon-generated carriers in the depleted intrinsic region, where most of the incident light photons are absorbed, are separated by the high electric field present in the depletion region and collected across the reverse-biased condition. This gives rise to a photocurrent flow in the external circuit. The *pin* photodiode acts as a linear device such that the photocurrent is directly proportional to incident optical power. Thus,

$$I = RP \tag{10.41}$$

I is photocurrent,

P is incident optical power, and

R is responsivity of the photodiode $R = \eta \frac{q}{h\nu}$

Thus, $$R = \frac{I}{P} \tag{10.42}$$

η is the quantum efficiency of the diode.

$$\eta = \frac{\text{No. of electron hole pair generated}}{\text{No. of incident photons}}$$

$$\eta = \left[\frac{I_p / q}{P_0 / h\nu} \right] \tag{10.43}$$

where $h\nu$ is the energy of the incident photon, and

q is the charge of electron.

The responsivity of the photodiode depends on the bandgap of the material operating wavelength, the doping and the thickness of the *p*, *I* and *n* regions of the diode. For example, to get high quantum efficiency and, hence, maximum sensitivity, the thickness of the depletion layer should be increased so that the absorption of photons will be maximum. But it reduces the response speed of the photodiode. Figure 10.46 shows *p-i-n* photodiode with reverse bias and its equivalent circuit.

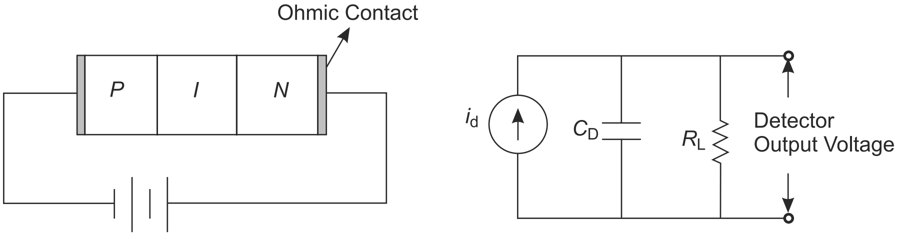


Fig. 10.46 *p-i-n* photodiode with reverse bias and its equivalent circuit

By analysing the above circuit, the bandwidth is inversely proportional to rise time 't_r'.

Thus, Bandwidth
$$=\frac{0.35}{t_r}=\frac{0.35}{2.19\,R_L C_D} \qquad (10.44)$$

For high-speed applications, C_D should be small. When t_r is very large, the speed of response is limited. For *p-i-n* diodes, t_r = 0.5 to 10 ns.

10.15.3 Avalanche Photodiodes (APDs)

Avalanche photodiode consists of four regions $p^+ - i - p - n^+$ in order to develop a very high electric field in the intrinsic region as well as to impart more energy to photoelectrons to produce new electron-hole pairs by impact ionisation. This impact ionisation leads to avalanche breakdown in the reverse-biased diode. So the avalanche photodiodes have high sensitivity and high responsivity over *p-i-n* diodes due to the avalanche multiplication.

This reach-through avalanche photodiode is shown in Figure 10.47. A high resistivity *p*-type material is deposited as an epitaxial intrinsic layer on a p^+ (heavily doped *p*-type) substrate. A *p*-type diffusion or ion-implant layer is then made in the high-resistivity intrinsic layer. A heavily doped n^+ layer is deposited on the *p* layer.

The term 'reach-through' arises from the photodiode operation. When a low reverse-bias voltage is applied, most of the potential drop is across the pn^+ junction. As shown in the above figure, the depletion layer widens with increasing bias until a certain voltage is reached at which the peak electric field at the p-n^+ junction is about 5 to 10% below that needed to cause

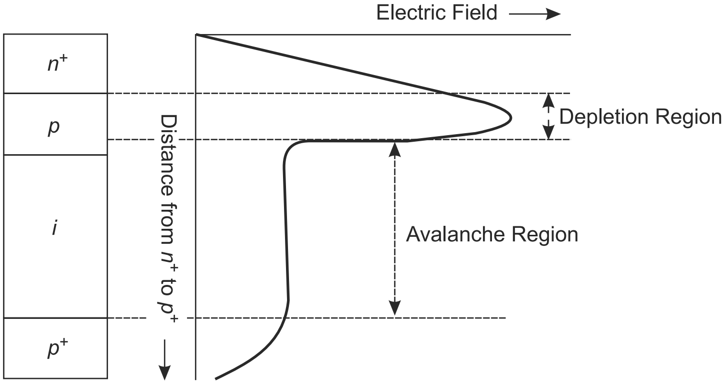


Fig. 10.47 Structure of reach-through avalanche photodiode

avalanche breakdown. At this point, the depletion layer just 'reaches through' to the nearly intrinsic region '*I*' of the diode. This reach-through APD is operated in the fully depleted mode.

Light enters the diode through p^+ region and is absorbed in the intrinsic region which also acts as the collection region for the photogenerated carriers. Further, the electron-hole pairs are separated by the electric field in the intrinsic region. The photogenerated electrons drift through the intrinsic region to the p-n^+ junction where a high electric field exists. In this high electric field region or multiplication region, the charge carrier multiplication takes place by impact ionisation or avalanche effect.

$$\text{Avalanche multiplication} = M = \frac{I_{\mathrm{m}}}{I_{\mathrm{p}}} \tag{10.45}$$

where I_m is total multiplied output current from diode, and

I_p is the primary photocurrent.

The value of M can be greater than 50. The responsivity of APD is given by

$$R = \eta \frac{q}{h\nu} M \tag{10.46}$$

10.15.4 Photodetector Materials

The responsivity of a photodetector depends on the type of material used. For long wavelengths, to get high responsivity and high quantum efficiency, the diode material should have a bandgap energy which is slightly lesser than the energy of the incident photons. Further, they should have low dark current.

Since the optical communication uses only long wavelengths, the detector material should respond the wavelength range from 1.3 μm to 1.7 μm.

(a) InGaAs It has high response for long wavelengths and optical absorption is very large. So this material is used for high-speed applications. It has low dark current, fast response and high quantum efficiency.

(b) Germanium Germanium photodetectors are used in high data transmission link operating at longer wavelengths. Since the bandgap is very narrow, the dark current is high. Further, avalanche multiplication noise is more due to high value of the electron and hole-ionisation rates ratio.

(c) There are GaAlSb, InGaSb, GaSb, GaAsSb and HgCd Te which are also used at longer wavelengths. These have moderate quantum efficiency, low dark current and moderate electron and hole-ionisation rates.

10.16 LIMITATIONS OF OPTICAL-CABLE COMMUNICATION

The following are some of the limitations of optical-cable communication.

1. When compared to copper-cable communication, the termination equipment for fibre optics is costly.
2. Repeated electrical to optical to electrical conversion is required.
3. Some of the components like amplifiers, splitters, multiplexers, de-multiplexers, etc. are still to be developed.
4. International specifications and guidelines for latest optical-communication-based systems are not in use.
5. Handling of optical fibres is difficult.
6. Optical-fibre splicing and protection is still expensive which increases the cost of the optical networks.

Even though there are disadvantages, as optical fibres offer almost enormous bandwidth, fibre-optic communication systems are being widely deployed all over the world.

10.17 APPLICATIONS OF OPTICAL FIBRES

Optical fibres are used today for a wide range of applications such as telecommunication system, data communication networking, medical applications, etc.

10.17.1 Telecommunication System

The fibre-optic technique has modernised the techniques of communication of information over very long distance in an economic as well as an efficient manner. It is very fast replacing

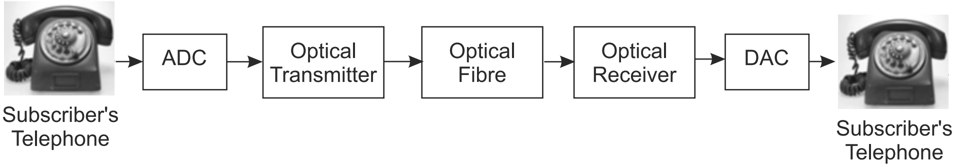


Fig. 10.48 A point-to-point telecommunication system

wire transmission on lines in telecommunications. Figure 10.48 shows a simple point-to-point telecommunication system between the subscribers using optical fibres. Communication is not only confined to a telephone call but instead provides numerous ways of transporting information like data, images, motion pictures from any part of the world to any other place at any time. Starting from long-haul communication systems on land and sea, optical fibres are carrying simultaneous telephone calls and other large signals between two exchanges while being used as interexchange junctions. The large information-carrying capacity of optical fibres also makes them attractive as an alternative to conventional copper twisted pair cables in a subscriber loop.

From Figure 10.34, in the subscriber's telephone, sound waves get converted into electrical signals and these electrical voice signals are changed into digital electrical pulses by means of Analog-to-Digital Converter (ADC). An optical transmitter consists of amplifier and LED or semiconductor laser, which will generate optical pulses as it is driven by the output from the ADC. Optical fibre is the heart of the system which is used to transmit the optical pulses. Next, there is the optical receiver consisting of a photodiode and an amplifier, whose outputs are electrical pulse coded signals. These digital signals are then converted back into continuous electrical signals by means of a Digital-to-Analog Converter (DAC). Finally, these electrical signals get converted into corresponding sound waves in the subscriber's telephone set.

Due to the exponential usage of telephone services, there is a need of development to make up the international telephone network. Each subscriber channel consists of a pair of optical fibres for transmission and reception purposes. Figure 10.49 illustrates a simple multipoint optical telecommunication system.

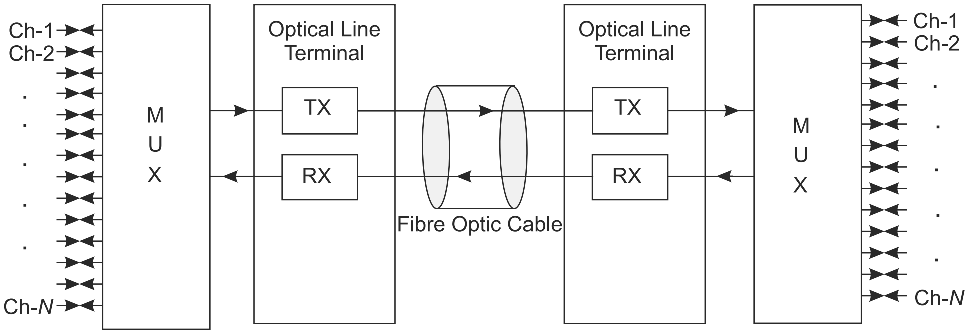


Fig. 10.49 A simple multipoint optical telecommunication system

10.17.2 Data-Communication Networks

Due to tremendous advantages of optical fibres for data transmission, it has already availed its place for data transmission over a distance from one computer to other. Figure 10.50 illustrates the block diagram of a point-to-point data-communication system that uses optical fibre as a communication medium.

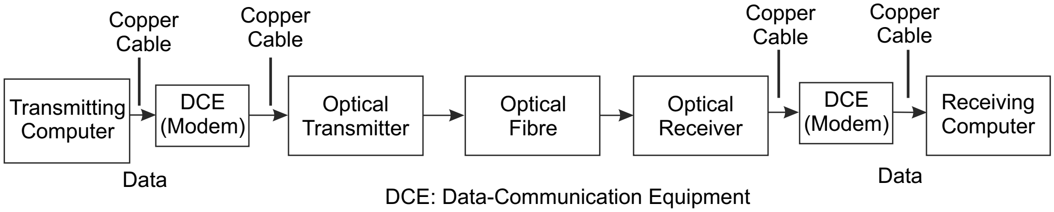


Fig. 10.50 Data-communication system using optical fibre

The optical fibre is widely used in computer networking like Local Area Networks (LAN), Metropolitan Area Networks (MAN) and Wide Area Networks (WAN). These communication networks are formed with computer systems, communication equipment, various terminals and peripheral devices connected to a common communication network. These three types of networks differ with respect to the size of the network and the way of operation of the communication systems. Local area networks usually occupy one building interconnecting various offices and departments. Metropolitan area networks covers numerous locations within a medium-size area such as a city or state. An example of MAN is automatic teller machines used by a local bank. Wide area networks extend coverage to national and international applications. They interconnect plants or offices dispersed throughout the world.

In order to send the data, video or motion pictures over a telephone line from a computer, it needs a conversion device in addition to the telephone line. It is called a Data-Communicating Equipment (DCE). An example of such equipment is a modem which converts a computer's digital form of electrical signal into analog signal for the analog telephone lines and vice versa at the receiving end.

Fibre-optic links are also useful for very short distances such as between large computer mainframes and their peripheral terminals and printers. Within computers, optical fibre is being used to carry signals between circuit boards.

10.17.3 Medical Applications

An endoscope is an instrument that is used to obtain a view of the interior of the body. Use of endoscopes has increased in recent years for both diagnostic and therapeutic procedures. Fibre-optic endoscopes are pliable, highly maneuverable instruments that allow access to channels in the body that older, semi-rigid instruments cannot access at all or can access only at great discomfort to the patient. Composed of multiple hairlike glass rods bundled together, these instruments can be more easily bent and twisted, and the intense light enables the endoscopist to see around.

Fibre-optic endoscopy is performed in order to evaluate areas of the head and neck that cannot otherwise be visualised. Whereas CT and MRI scans provide internal body information, they provide only a single snapshot in a single moment in time and not function such as how things move or work over a continuous period of time. These exams are performed without any sedation and are easily tolerated by patients as young as 5 years of age with their full cooperation.

Figure 10.51 shows a flexible fibre-optic endoscope which is an optical instrument that transmits light and carries images back to the operator via coherent bundles of very fine glass fibres, sheathed in an impervious protective covering of PVC, that forms a hermetic seal.

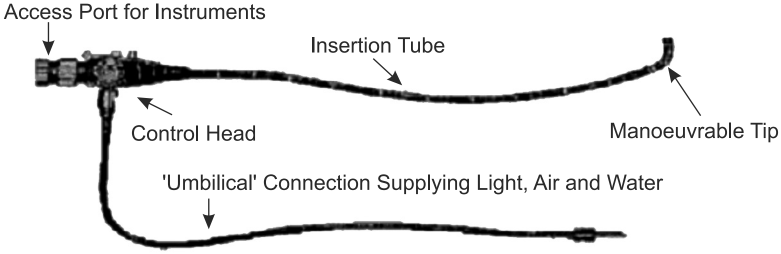

Access Port for Instruments
Insertion Tube
Control Head
Manoeuvrable Tip
'Umbilical' Connection Supplying Light, Air and Water

Fig. 10.51 A flexible fibre-optic endoscope

A typical endoscope consists of a control head, a flexible shaft and a manoeuvrable end tip. The instrument is often long and complicated in design with channels for the introduction of operating instruments used for the removal of tissue and suction channels used to clear the operative field. An 'umbilical' cord connects the head of the instrument to a light source. Air, water and suction channels are also contained within this cord. The complicated design poses significant challenges during use, decontamination and storage. Some designs are incompatible with the instrument being submerged during cleaning and all are intolerant of minimum temperatures required for thermal disinfection.

A rigid endoscope is also an optical instrument but it is nonflexible and made of surgical stainless steel. Figure 10.52 shows a rigid optical endoscope which can be designed for either high (e.g. steam) or low (e.g. chemical, ethylene oxide) temperature sterilisation methods.

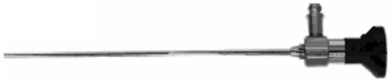


Fig. 10.52 Rigid optical endoscope

In rigid optical endoscope, there is a separate port which allows administration of drugs, suction and irrigation. This port may also be used to introduce small folding instruments such as forceps, scissors, brushes, snares and baskets for tissue excision (removal), sampling, or other diagnostic and therapeutic work. Endoscopes may be used in conjunction with a camera or video recorder to document images of the inside of the joint or chronicle an endoscopic procedure. New endoscopes have digital capabilities for manipulating and enhancing the video images.

Endoscopes are used routinely in many aspects of patient care, for example, in the gastro-intestinal tract for diagnosis and treatment of ulcers, cancers, strictures or bleeding sites. They are found in most clinical settings including operating theatres, endoscopy suites, outpatient clinics, wards, intensive-care units, accident and emergency departments, etc. Fibre-optic endoscopes now have widespread use in medicine and guide a myriad of diagnostic and therapeutic procedures. Different types of fibre-optic endoscopes according to their specific applications are listed as follows.

Arthroscopy Examination of joints for diagnosis and treatment, which is called arthroscopic surgery

Bronchoscopy Examination of the trachea and lung's bronchial trees to reveal abscesses, bronchitis, carcinoma, tumours, tuberculosis, alveolitis, infection, inflammation

Colonoscopy Examination of the inside of the colon and large intestine to detect polyps, tumours, ulceration, inflammation, colitis, diverticula and Chrohn's disease

Colposcopy Direct visualisation of the vagina and cervix to detect cancer, inflammation, and other conditions

Cystoscopy Examination of the bladder, urethra, urinary tract, uteral orifices and prostate (men) with insertion of the endoscope through the urethra

ERCP (Endoscopic Retrograde Cholangio-Pancreatography) Uses endoscopic guidance to place a catheter for X-ray fluoroscopy with contrast enhancement. This technique is used to examine the liver's biliary tree, the gall bladder, the pancreatic duct and other anatomy to check for stones, other obstructions and disease. X-ray contrast is introduced into these ducts via a catheter and fluoroscopic X-ray images are taken to show any abnormality or a blockage. If disease is detected, it can sometimes be treated at the same time or a biopsy can be performed to test for cancer or other pathology. ERCP can detect biliary cirrhosis, cancer of the bile ducts, pancreatic cysts, pseudocysts, pancreatic tumours, chronic pancreatitis and other conditions such as gall-bladder stones.

Endoscopic Biopsy Removal of tissue specimens for pathologic examination and analysis

Gastroscopy Examination of the lining of the asophagus, stomach, and duodenum. Gastroscopy is often used to diagnose ulcers and other sources of bleeding and to guide biopsy of suspect GI cancers

Laparoscopy Visualisation of the stomach, liver and other abdominal organs including the female reproductive organs, for example, the fallopian tubes

Laryngoscopy Examination of the larynx (voice box).

10.17.4 Fibre-Optic Lighting Applications

Fig. 10.53 Fibre-optic spray lighting

Fibre optics is perfect for creating that sensation of sleeping out under the stars, but they can also be used in a huge range of other innovative lighting designs, both indoors and out. Fibre-optic lighting can be used to dramatic effect in the bathroom or kitchen and the ability to separate the light output from the electrical supply has obvious safety applications. Figure 10.53 shows fibre-optic spray lighting.

Fibre-optic lighting presents a new way to create design with the utilisation of light. For this purpose, leaky mode fibres are used. A leaky mode fibre has an electric field that decays monotonically for a finite distance in the transverse direction but becomes oscillatory everywhere beyond that finite distance. Such a mode gradually leaks out of the waveguide as it travels down it and producing attenuation even if the waveguide is perfect in every respect. In a leaky mode, the leakage rate must be sufficiently small that the mode substantially maintains its shape as it decays.

Generally, the propagation of light through optical fibre can take place via meridional rays or skew rays. These skew rays suffer only partial reflection while meridional rays are completely guided. Thus, the modes allowing propagation of skew rays are called **leaky modes**. Some optical power is lost into clad due to these modes.

1. Architectural Lighting

In an architectural lighting, a bundle of fibres can be used to carry light from a remote light source to provide lighting in inaccessible parts of a building. Rows of leaky fibre-optic light spots also accentuate the architectural lines of a room or a building. This kind of fibre-optic lighting can be used to provide lighting in corridors or staircases with the replacement of bulbs and at low cost.

2. Display Lighting

Lighting of display items can be made by fibre optics with increased security and no added heat load. In particular, textiles, books, chocolates as well as jewellary can be displayed under fibre-optic lighting. If infrared and ultraviolet lighting is used instead of fibre-optic lighting, books become brittle and textiles fade in colour. As fibre-optic lighting is free from damaging ultraviolet light and infrared radiation, it is also used for lighting photosensitive paintings and museum exhibits.

3. Decorative Lighting

Optical fibres can also be used as a decorative piece. Attractive fibre-tree-based simple lighting systems are now available in the market at low prices. This kind of decorative lighting is also being used in swimming pools, nightclubs, aquarium lighting, and so on.

10.18 WAVELENGTH DIVISION MULTIPLEXING (WDM)

In fibre-optic communication, Wavelength Division Multiplexing (WDM) is a technology which multiplexes a number of optical carrier signals onto a single optical fibre by using different wavelengths of light. This technique enables two-way communications over single strand of fibre.

In a typical fibre-optic network, the data signal is transmitted using a single light pulse. In order to increase the capacity of a single fibre, the bit rate of the signal is to be increased such as 1 Mbps to 10 Mbps to 100 Mbps.

Wavelength division multiplexing is a cost-effective, flexible and scalable technology for increasing capacity of a fibre network. A conventional single-wavelength optical-fibre communication system, a simplex system, is shown in Figure 10.54.

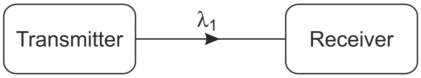


Fig. 10.54 A simplex wavelength optical-fibre communication system

In a duplex or bidirectional WDM, there are two different wavelength optical signals travelling in opposite directions providing bidirectional transmission. Information is travelling in one direction at a wavelength λ_1 and simultaneously the information in the opposite direction is travelling at a different wavelength λ_2 Figure 10.55 shows a duplex WDM.

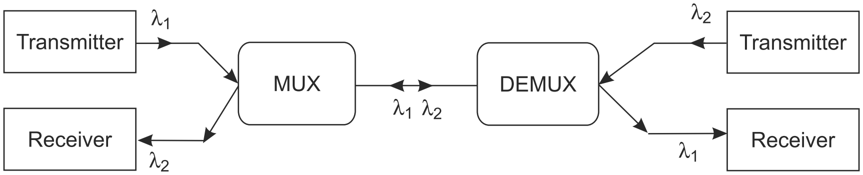


Fig. 10.55 A duplex WDM

In a multiplex WDM, various independent input signals of different wavelengths (λ_1, λ_2, ... λ_n) are combined at one end for transmission over a single fibre by means of using a multiplexer, and the output signals (λ_1, λ_2, ... λ_n) are separated by using a de-multiplexer at the other end. Figure 10.56 shows a multiplex WDM.

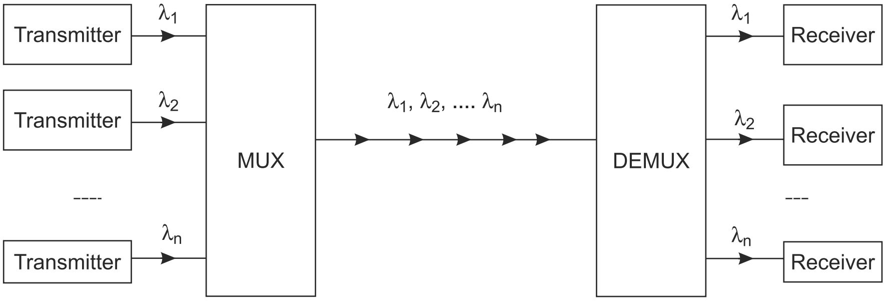


Fig. 10.56 A multiplex WDM

10.18.1 Architecture of WDM

The generalised architecture of WDM is illustrated in Figure 10.57. It is based on a simple concept of transmitting multiple signals, each with a different wavelength, instead of transmitting a single signal on a single wavelength. Each remains a separate data signal, at any bit rate with any protocol, unaffected by other signals on the fibre.

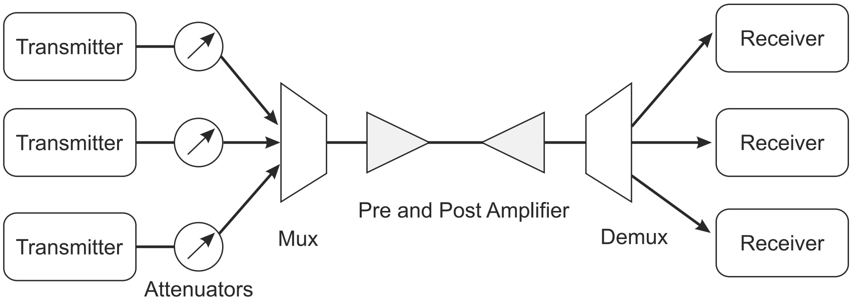


Fig. 10.57 Architecture of WDM

WDM system uses a multiplexer at the transmitter to join the signals together and a demultiplexer at the receiver to split them apart. With the right type of fibre, it is possible to have a device that does both simultaneously. Capacity of a given link can be expanded by simply upgrading the multiplexers and demultiplexers at each end.

10.18.2 Types of WDM

Wavelength division multiplexing is classified into two types based on different wavelength patterns.

1. Coarse Wavelength Division Multiplexing (CWDM)
2. Dense Wavelength Division Multiplexing (DWDM)

1. Coarse Wavelength Division Multiplexing (CDWM)

This method uses a wide spectrum and accommodates eight channels. This wide spacing of channels allows economic usage of optics with limited capacity. The main advantages of CWDM includes lower-cost, lower-capacity and shorter-distance applications

2. Dense Wavelength Division Multiplexing (DWDM)

This multiplexing method consists of 16 or more channels into a narrow spectrum very nearer to 1550 nm local attenuation minimum. Decreasing channel spacing requires the use of more precise and costly optics, but there is more scalability. Typical DWDM systems provide 1-44 channels of high capacity.

10.18.3 WDM using Filter Type Devices

To get wavelength division multiplexing, dichroic filters or a multilayer thin-film interference filter is designed to transmit light with a specific wavelength with all other wavelengths getting either absorbed or reflected. Reflection-type filters are normally used. They consist of a flat glass plate with multiple layers of different dielectric films deposited for the selection of different wavelengths. These types of filters can be used in series to separate a number of different wavelengths. Figure 10.58 shows a filter-type WDM for two wavelengths.

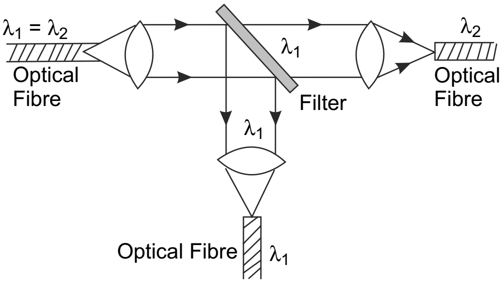


Fig. 10.58 A filter-type WDM for two wavelengths

10.18.4 WDM using Angular Dispersive Devices

Wavelength division multiplexing can be achieved by using angular dispersive devices like a prism or grating. Using these devices, a large number of channels can be combined or separated. Figure 10.59 shows WDM using angular dispersive optical system and Figure 10.60 shows WDM using reflection grating. Polychromatic light, light of different wavelengths emerging from the fibre is collimated by a lens and is passed through the angularly dispersive device which separates the various wavelengths into different spatially oriented parallel beams. The next focusing lens focuses the different parallel output beams into the separate receiving fibres.

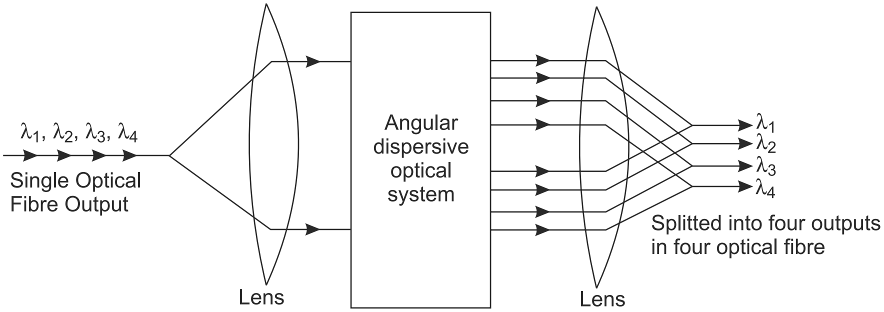

$\lambda_1, \lambda_2, \lambda_3, \lambda_4$
Single Optical Fibre Output
Lens
Angular dispersive optical system
Lens
λ_1
λ_2
λ_3
λ_4
Splitted into four outputs in four optical fibre

Fig. 10.59 WDM using angular dispersive optical system

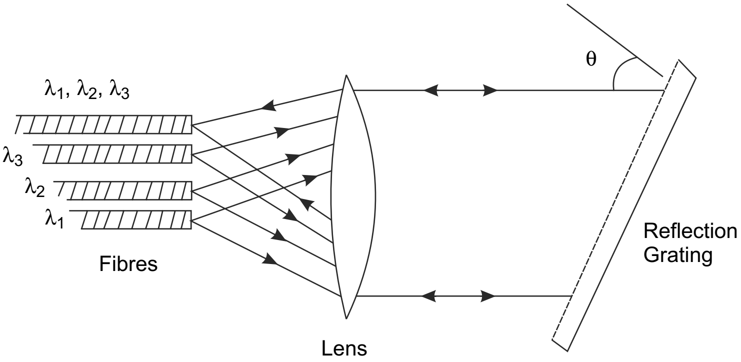


Fig. 10.60 Shows WDM using reflection grating

10.18.5 LAMDANET Star Network

Wavelength division multiplexing in optical-fibre systems can be implemented using either LED or injection laser sources with either multimode or single-mode fibre. In particular, the potential utilisation of the separate wavelength channels to provide dedicated communication services to individual subscriber terminals is an attractive concept within telecommunications.

For example, a multi-wavelength, single-mode optical 'star network, called LAMBDANET, has been developed using commercial components. This network, which is internally nonblocking, has been configured to allow the integration of point-to-point and point-to-multipoint wideband services, including video distribution applications. Figure 10.61 shows the block diagram of the LAMDANET star network.

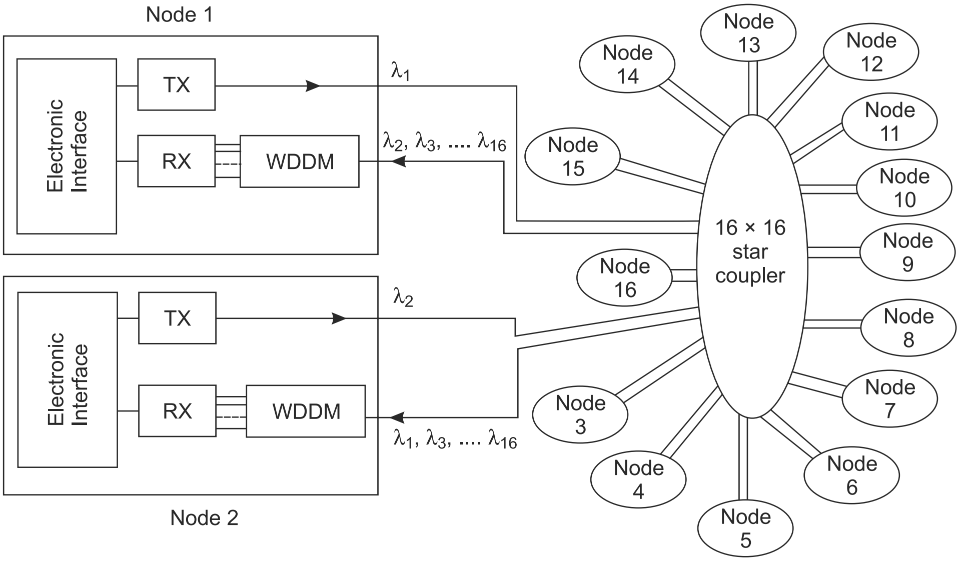


Fig. 10.61 The block diagram of the LAMDANET star network

The LAMBDANET star network incorporates a sixteen-port passive transmission star coupler. Each node was equipped with a single distributed feedback laser selected with centre wavelengths spaced at 2 nm intervals over the range 1527 to 1561 nm. Hence, each node transmits a unique wavelength, providing a communication capability to all other nodes. At the receive terminals every node could detect transmissions from all other nodes using a wavelength demultiplexer and sixteen optical receivers. This type of network has been found capable of operating at a transmission rate of 2Gbit/s over a distance of 40 km.

10.18.6 WDM using Spectral Slicing Method

Another method of WDM using LED source and spectral slicing method is illustrated in Figure 10.62.

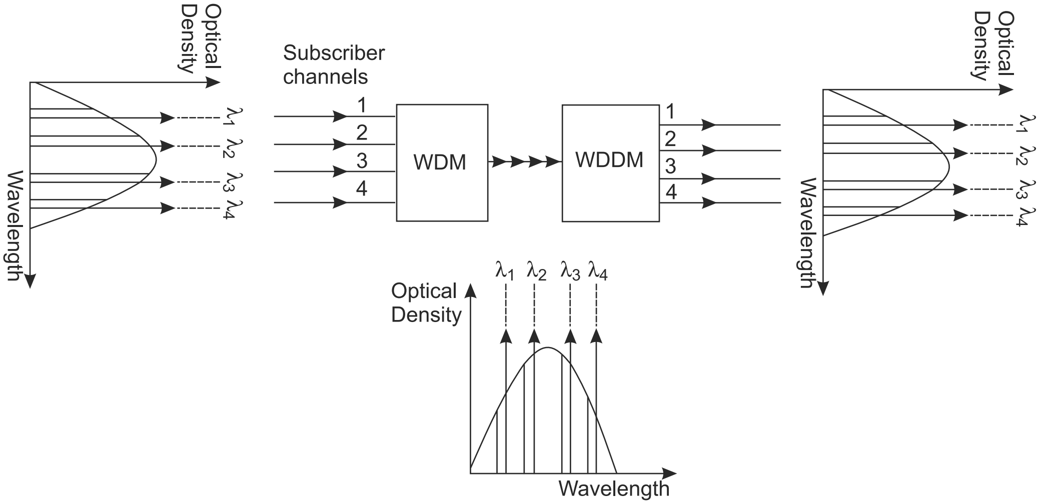


Fig. 10.62 WDM using LED source and spectral slicing method

In this method, wide spectral width (63 nm) edge-emitting LEDs is utilised to provide the multi-wavelength optical carrier signals which were transmitted on single-mode optical fibre. From edge-emitting LED, using the spectral slicing technique, a relatively narrow spectral width of 3.65 nm for each separate channel can be obtained for the optical WDM multiplexer device, prior to transmission down the optical link. A WDM demultiplexer (WDDM) device is also located at a distribution point in order to separate and distribute the different wavelength optical signals to the appropriate receiving terminals.

10.18.7 Advantages of WDM

The key features and benefits of WDM include the following.

1. In WDM method, wavelengths can accept virtually any services.

2. With fibre-capacity expansion, WDM adds up to 160X bandwidth to a single fibre.
3. CWDM and DWDM provide price performance for virtually any network.
4. It also provides the flexibility to change with changing network requirements

Summary

Fibre-optic communication is a method of transmission of information from one place to another place by means of light signals through an optical fibre. The light signal forms an electromagnetic carrier signal that is modulated to carry the information.

The refractive index of a medium is defined as the ratio of velocity of light in vacuum to the velocity of light in the medium. Refraction results in a change of direction for a light ray. When a ray is incident on the interface between two media of different indices, refraction will take place.

For the phenomenon of total internal reflection to take place, the following two conditions are to be satisfied.

- The glass at around the centre of the fibre (core) should have higher refractive index (n_1) than that of the material (cladding) surrounding the fibre (n_2).
- The light should be incident at an angle of θ which will be greater than the critical angle θ_c.

Critical angle (θ_c) of a medium is defined as the value of the incident angle at which the angle of refraction is 90°. It is the angle of incidence above which total internal reflection occurs.

Acceptance angle (θ_a) is defined as the maximum angle to the axis at which the light may enter the fibre in order to be propagated, at which the angle of incidence at the core-cladding boundary is equal to the critical angle of the core medium.

Numerical Aperture (NA) of the fibre is the light-collecting efficiency of the fibre and it is the measure of the amount of light rays that can be accepted by the fibre. This factor gives the relationship between the acceptance angle and the refractive indices of the three media involved, namely the core, cladding and the air.

Skew rays are the light rays propagated through graded-index fibres. They are the rays which describe angular helices as they progress along the fibre. They follow helical path around the axis of the fibre and these rays do not cross the axis of the fibre.

An optical fibre has a cylindrical shape which consists of three sections.

- Core
- Cladding
- Protective enclosure

Fibres used in communication are classified into two major types.

- Step-index fibres
- Graded-index fibres

Optical beams are generated by light sources for carrying information in an optical-fibre system. The two most common sources of light are

- Light-Emitting Diode (LED)
- Light Amplification by Stimulated Emission of Radiation (LASER)

Photodetector is a device used to convert the light signals to electrical signals at the receiver end of the fibre link. Two of the popular photodetectors discussed are

- *p-i-n* photodiode
- Avalanche photodiode

Wavelength Division Multiplexing (WDM) is a technology which multiplexes a number of optical carrier signals onto a single optical fibre by using different wavelengths of light. This technique enables two-way communications over a single strand of fibre.Wavelength division multiplexing is a cost-effective, flexible and scalable technology for increasing capacity of a fibre network.

REVIEW QUESTIONS

PART-A

1. What are the uses of optical fibres?
2. What are the parts of an optical fibre?
3. Draw the structure of an optical fibre and mark its parts.
4. State the principle used in the working of fibres as light guides.
5. What are the different types of optical fibres?
6. Define total internal reflection.
7. What are the conditions for having total internal reflection for a fibre?
8. What is critical angle?
9. Define acceptance cone angle.
10. Define numerical aperture.
11. What is relative refractive-index difference?
12. How will you calculate the number of propagating modes in a step-index fibre?
13. How will you calculate the number of propagating modes in a graded-index fibre?
14. Differentiate between single-mode and multimode fibres.
15. Draw the index profile of a step-index fibre.
16. Draw the index profile of a graded-index fibre.
17. What are meridional rays?
18. What are skew rays?
19. Give the relation between numerical aperture of skew rays and meridional rays.
20. What are the various losses in optical fibres?

21. Mention the significance of optical-fibre connectors.
22. What is meant by optical splicing?
23. What are the required properties of light sources used in optical communication?
24. State the drawbacks of homojunction laser diodes.
25. Mention the types of LED structures.
26. Define internal quantum efficiency of an LED.
27. What are the required properties of a photodetector?
28. Mention any two types of photodetectors used in optical communication.
29. State the limitations of OFC.
30. What is the significance of WDM?

PART-B

1. Discuss the different types of fibres along with their diagrams for refractive-index profile and light-ray propagation.
2. Distinguish between single-mode fibres and multimode fibres and also between step-index fibres and graded-index fibres.
3. Discuss the propagation of meridional rays along with the acceptance angle and numerical aperture of a step-index fibres.
4. Derive an expression for the amount of transmitted energy for a given incidental energy when meridional rays are propagating through the step-index fibre.
5. What are skew rays? Derive the numerical aperture and compare them with meridional rays.
6. Explain the principle of operation of different types of *p-n* junction diode lasers with their suitable diagrams.
7. Describe any two types of photosources used in optical communication.
8. What is the principle of LED used as an optical source in optical communication? Discuss its structures.
9. With a neat diagram, explain the principle of operation of a *p-i-n* photodiode used in optical communication.
10. With neat diagram, explain the principle of operation of an avalanche photodiode used in optical communication.

11

SATELLITE COMMUNICATION

Objectives

- To know about the features and different satellite-frequency bands
- To discuss the details of the satellite communication systems
- To discuss the types, various orbits, earth-station subsystem and transponder subsystem
- To provide details about satellite-launching procedures, satellite antennas and radiation patterns
- To provide details about radio-wave transmission

11.1 INTRODUCTION

A satellite is any natural or artificial body moving around a celestial body. As a satellite is used as a repeating station, the data originating and terminating point can be anywhere on the earth. It could be used as a sort of repeater station in reference communication satellites or as a photographer taking pictures of regions of interest during its periodic motion or even taking photographs of clouds and monitoring other atmospheric parameters and thus assisting in weather forecasting.

A satellite is a specialised wireless receiver/transmitter that is launched by a rocket and placed in orbit around the earth. There are hundreds of satellites currently in operation. They are used for such diverse purposes as weather forecasting, television broadcast, amateur radio communications, Internet communications and the Global Positioning System (GPS).

The first artificial satellite, launched by Russia in the late 1950s, was used to transmit a simple Morse-code signal over and over. Modern satellites can now receive and re-transmit thousands of signals simultaneously.

Satellites are also used to carry the required instrumentation to provide intercontinental communication services. Though the idea of satellites originated from the desire to put an object in space that would stand still with respect to the earth, thus making possible a host of communication services, there are many more situations where the satellites need not be stationary with respect to the earth to perform the intended function and hence have much lower orbital height.

A satellite enables communications to be established over long distances well beyond the line of sight. Communication satellites are used for many applications including relaying telephone calls, providing communications to remote areas of the earth, providing satellite communications to ships, aircraft and other mobile vehicles and there are many more ways in which communication satellites can be used.

11.2 FEATURES OF SATELLITE COMMUNICATION

Due to the following features, satellite communication is preferred widely

1. Highly survivable (physical survivability and robustness)
2. Independent of terrestrial infrastructure
3. Able to provide the load-sharing and surge-capacity solution for larger sites
4. Best for redundancy as they add a layer of path diversity and link availability

Satellite systems perform effectively when

1. Terrestrial infrastructure is damaged, destroyed, or overloaded
2. Interconnecting widely distributed networks
3. Providing interoperability between disparate systems and networks
4. Providing broadcasting services over very wide areas such as a country, region, or an entire hemisphere
5. Providing connectivity for the "last mile" in cases where fibre networks are simply not available
6. Providing mobile/transportable wideband and narrowband communications
7. Natural disasters or terrorist attacks occur. Satellites are the best and most reliable platform for communications in such situations when fibre networks or even terrestrial wireless can be disrupted by tsunamis, earthquakes, or hurricanes. Satellites are an instant infrastructure.

11.3 SATELLITE-FREQUENCY BANDS

The frequency bands commonly used for satellite communication for different applications are listed in Table 11.1.

Table 11.1 Frequency bands used for satellite communication

Uplink Frequencies	Downlink Frequencies	Band	Applications
5.925–6.425 GHz 6.725–7.025 GHz 14.25–14.50 GHz 27.5–30.0 GHz	3.700–4.200 GHz 4.5–4.8 GHz 11.45–11.7 GHz 17.7–21.2 GHz	C Ext. C Ku (IN INSAT–3B) Ka	Commercial Applications
292–312 MHz 7.9–8.4 GHz 43.5–45.5 GHz	250–270 MHz 7.25–7.75 GHz 20.2–21.2 GHz	Microwave X Ka	Military Use

11.4 ADVANTAGES OF USING SATELLITE COMMUNICATION

There are several advantages of satellite communication. Some are listed as follows.

1. Ubiquitous Coverage

A group of satellites can cover virtually the earth's entire surface.

2. Instant Infrastructure

Satellite service can be offered in areas where there is no terrestrial infrastructure and the costs of deploying a fibre or microwave network are prohibitive. It can also support services in areas where existing infrastructure is outdated, insufficient or damaged.

3. Independent of Terrestrial Infrastructure

Satellite service can provide additional bandwidth to divert traffic from congested areas, provide overflow during peak usage periods, and provide redundancy in the case of terrestrial network outages.

4. Temporary Network Solutions

For applications such as news gathering, homeland security, or military activities, satellites can often provide the only practical, short-term solution for getting necessary information in and out.

5. Rapid Provisioning of Services

Since satellite solutions can be set up quickly, communications networks and new services can be quickly recovered and reconfigured. In addition, you can expand services electronically without traditional terrestrial networks. As a result, you can achieve a high level of communications rapidly without high budget expenditures.

11.5 BASIC ELEMENTS OF SATELLITE COMMUNICATION

Satellite communications are comprised of two main components. They are

1. The satellite
2. The ground station

11.5.1 The Satellite

The satellite itself is also known as the **space segment** and is composed of three separate units. They are represented as

1. Fuel system
2. Satellite and telemetry controls
3. Transponder

The **transponder** includes the receiving antenna to pick up signals from the ground station, a broadband receiver, an input multiplexer and a frequency converter which is used to reroute the received signals through a high-powered amplifier for downlink.

The primary role of a satellite is to reflect electronic signals. In the case of a telecom satellite, the primary task is to receive signals from a ground station and send them down to another ground station located a considerable distance away from the first. This relay action can be two-way, as in the case of a long-distance phone call.

Another use of the satellite is when, as is the case with television broadcasts, the ground station's uplink is then downlinked over a wide region, so that it may be received by many different customers possessing compatible equipment. If the satellite is equipped with cameras or various sensors, it merely downlinks any information it picks up from its vantage point.

11.5.2 The Ground Station

This is the **earth segment**. The ground station's job is twofold. In the case of an uplink, or transmitting station, terrestrial data in the form of baseband signals, is passed through a baseband processor, an up converter, a high-powered amplifier, and through a parabolic dish antenna up to an orbiting satellite. A downlink, or receiving station, works in the reverse fashion as the uplink, ultimately converting signals received through the parabolic antenna to a baseband signal.

11.6 APPLICATIONS OF SATELLITE COMMUNICATIONS

There are several applications of satellite communications over the world. Some important applications are listed below.

1. Telecommunications

Since the beginnings of long-distance telephone network, there has been a need to connect the telecommunications networks of one country to another. This has been accomplished in several ways. Submarine cables have been used most frequently. However, there are many occasions where a large long-distance carrier will choose to establish a satellite-based link to connect to transoceanic points, geographically remote areas or poor countries that have little communications infrastructure. Groups like the international satellite consortium, Intelsat, have fulfilled much of the world's need for this type of service.

2. Cellular Communication

Satellites are mainly utilised to increase the bandwidth available to ground-based cellular networks. Every cell in a cellular network divides up a fixed range of channels which consist of either frequencies, as in the case of FDMA systems, or time slots, as in the case of TDMA. Since a particular cell can only operate within those channels allocated to it, overloading can occur.

By using satellites which operate at a frequency outside those of the cell, we can provide extra satellite channels on demand to an overloaded cell. These extra channels can just as easily be, once free, used by any other overloaded cell in the network, and are not bound by bandwidth restrictions like those used by the cell. In other words, a satellite that provides service for a network of cells can allow its own bandwidth to be used by any cell that needs it without being bound by terrestrial bandwidth and location restrictions.

3. Television Signals

Satellites have been used to transmit broadcast television signals between the network hubs of television companies and their network affiliates. In some cases, an entire series of programming is transmitted at once and recorded at the affiliate, with each segment then being broadcast at appropriate times to the local viewing places.

4. Marine Communications

In the maritime community, satellite communication systems such as Inmarsat provide good communication links to ships at sea. These links use a VSAT type device to connect to geosynchronous satellites, which in turn link the ship to a land-based point of presence to the respective telecommunications system.

5. Space-Borne Land Mobile

Along the same lines as the marine-based service, there are VSAT devices which can be used to establish communication links even from the world's most remote regions. These devices can be handheld, or fit into a briefcase.

6. Satellite Messaging for Commercial Jets

Another service provided by geosynchronous satellites is the ability of a passenger on an airborne aircraft to connect directly to a land-based telecom network.

7. Global Positioning Services

This is Another VSAT oriented service, in which a small apparatus contains the ability to determine navigational coordinates by calculating a triangulating of the signals from multiple geosynchronous satellites.

11.7 BLOCK DIAGRAM OF A SATELLITE-COMMUNICATION SYSTEM

The essential components of a satellite communication system include the following:

1. A **satellite** capable of receiving signals beamed at it from the earth station, amplifying them and doing frequency translation before transmitting the same back towards the earth for intended users.
2. **Transmission path** comprising both uplink from the earth station to the designated satellite as well as the downlink from satellite to the intended users.
3. **Earth station** equipped suitably to transmit and receive signals to and from the designated satellite.

Components 1 and 2 together are grouped under the heading of **space segment,** while the component 3 is called the **earth segment**. Figure 11.1 shows different components of a satellite communications set-up. The space segment and the earth segment are described as follows.

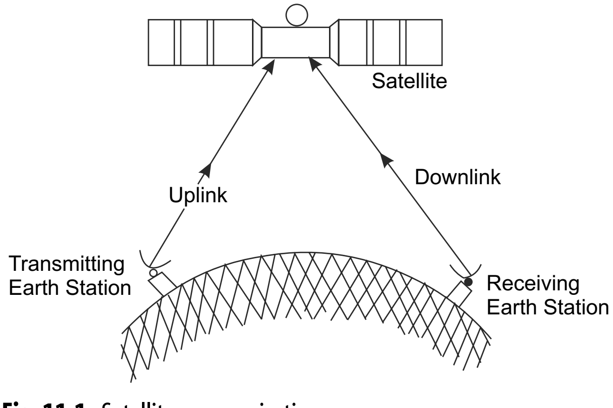


Fig. 11.1 Satellite communication

11.7.1 Space Segments

The space segment comprises the satellite and the transmission path. The satellite does the seemingly simple task of relaying the signal received by it after carrying out necessary amplification and frequency translation. The important functional subsystems of a communication satellite, shown in Figure 11.2, include the following.

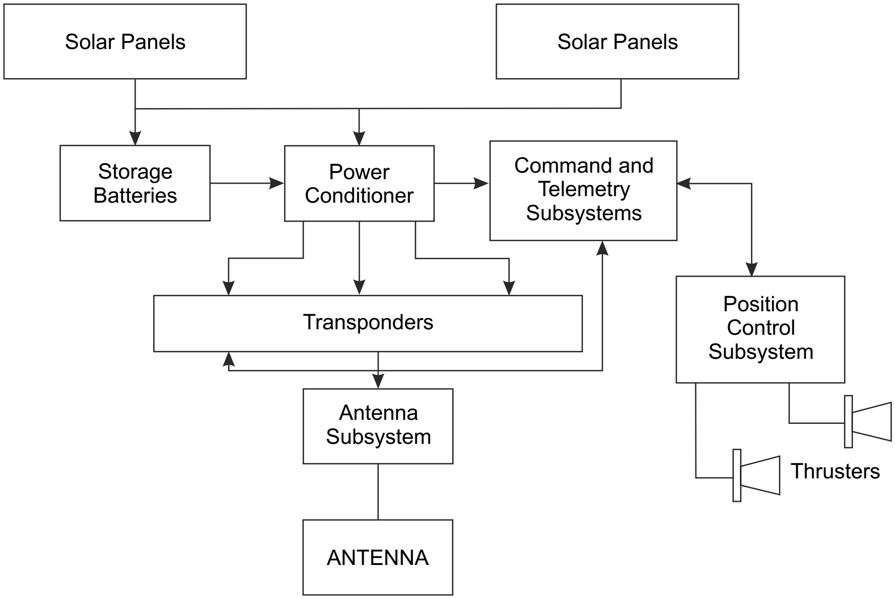


Fig. 11.2 A typical block diagram of a satellite system

1. Source of energy
2. Power generation and distribution subsystem
3. Stabilisation subsystem
4. Antenna subsystem
5. Thrust subsystem
6. Temperature-control subsystem
7. Transponders
8. Telemetry, tracking and command subsystem

The antenna subsystem, transponders and associated electronics together constitute the communications payload.

1. Source of Energy

The satellite derives the electrical power required for operation of various subsystems from solar energy. An array of solar cells distributed either around the body of the satellite or on the solar panels provides the required opto-electronic conversion.

2. Power Generation and Distribution Subsystem

The subsystem generates electric power in terms of required voltages and load currents for various functional subsystems from available solar energy. It also contains charging units for the storage batteries. The storage batteries are continuously charged from the solar-energy-driven chargers and are then used to provide electric power to satellite components during the periods when solar energy is not available.

3. Stabilisation Subsystem

The stabilisation subsystem ensures that the satellite remains in a fixed orientation with reference to its orbit and that its antenna subsystem always points in the right direction. Satellites are spin-stabilised.

4. Antenna Subsystem

The satellite may use the same or different antennae for reception and retransmission. One of the important requirements of a satellite antenna is to produce the desired shape of the illuminated area on the earth beneath the satellite, known as **satellite footprint**. Most communication satellites nowadays use several antennae of various sizes, shapes and configurations, so as to produce one or more of the radiated beams, namely omnidirectional, global and spot beams. All satellites use an omnidirectional antennae following injection of the satellite into the parking orbit before it is positioned in the final designated slot. A nondirectional antenna such as omnidirectional antennae provides communication between the satellite and the control station until the satellite's directional antennae are fully deployed and oriented in the correct direction.

5. Thrust Subsystem

The thrust subsystem is used to carry out adjustments to the satellite's orbit, orbital position and altitude. It comprises a set of on-board propulsion units along with their fuel tanks and associated firing and control circuits.

6. Temperature-Control Subsystem

The purpose of the temperature-control subsystem is to optimise heat dissipation by balancing the temperature distribution across the exterior and within the interior of the satellite.

7. Transponders

The transponder subsystem receives the signal beamed at it from the transmitting earth station on the uplink, does the required input filtering, down converts to a different frequency and amplifies it to the desired level before it is retransmitted to the intended receiving station. Based on the uplink/downlink frequency being handled, transponders are categorised as C-band transponders, ku-band transponders, ka-transponders, and so on.

8. Telemetry, Tracking and Command (TTC) Subsystem

The purpose of the Telemetry, Tracking and command (TTC) subsystem is to continuously transmit information on vital parameters such as those related to satellite's orbital position and attitude and also those related to the performance specifications of various subsystems to the ground stations. In addition, this subsystem also serves the purpose of receiving commands from ground stations for carrying out certain corrective action. The TTC system is very important both during orbital injection and positioning phase and subsequently throughout the operational life of the satellite. The subsystem has its own dedicated radio link and its own omnidirectional antenna. The omnidirectional antenna is important during the launch and positioning phase. The satellite may continue to use the same link even during the operational life though there is an increasing tendency to switch the TTC operations to one of the transponders and its associated antennae after the initial orbital injection and positioning phase is over.

9. Transmission Path

The transmission path can also be considered as a part of the space segment in addition to the satellite.The transmission path, due to the attenuation it imparts to the signal and the interference that it causes to the signal, has a great bearing on the performance of the overall satellite communication system. The strength of a radiated signal is inversely proportional to the square of the distance. As a result, in a geostationary satellite that is about 36,000 km from the earth's surface, the received signal at the earth station may be only a few picowatts.

The attenuation of the signal is because of two main reasons. The first source of loss is the space loss which is caused by the antennas not being 1005 efficient. The second source, which is of course the major source of loss, is the attenuation of signal in the atmosphere. The main factors responsible for atmospheric losses include presence of oxygen in free molecular form, uncondensed water, vapour, rain, fog, snow, hail and free electrons in the atmosphere.

11.7.2 Earth Segments

The earth station and associated infrastructure constitute the earth segment. The earth station transmits to and receives from a satellite. An earth station may be located on a ship at sea or even on an aircraft. Major subsystems of an earth station include the following.

1. Transmitter
2. Receiver
3. Antenna
4. Tracking equipment
5. Terrestrial interface

There are three main categories of earth stations, namely

1. Transmit-and-receive earth stations
2. Transmit-only earth stations
3. Receive-only earth stations

Transmit-and-receive earth stations are encountered in two-way communication systems. **Receive-only stations** are used mostly in direct TV broadcast satellite systems and CATV systems. **Transmit-only earth stations** are used in data-collection systems.

The transmitter converts the signal to be transmitted to the uplink frequency with proper encoding and modulation. The signal is then amplified and directed to the appropriate polarisation port of the antenna feed.

Different components of the earth station receiver include Low Noise Amplifier (LNA), down converter, demodulator, decoder and baseband signal-processing circuitry. It is important to achieve a low-system noise temperature in the receiving channel.

The antenna is the most visible part of the earth station. The earth station's performance to a great extent depends upon its antenna characteristics. The antenna should have a high directive gain, low noise temperature and a radiation pattern that does not exhibit a large-side lobe level. The commonly used antenna types are the parabolic antenna with focal-point feed and the cassegrain antenna. Phased-array antennas and horn antennas are also used.

The tracking-and-pointing subsystem is used to track the satellite and point the antenna beam accurately to the satellite in both transmit and receive modes. The basic earth station comprising the transmitter, the receiver and the antenna merely provides an access to the high-frequency satellite link. For the link to be usable by the ultimate end users and for the satellite capacity to be utilised in a cost-effective and efficient manner, the earth station must be equipped with necessary infrastructure to provide an interface with terrestrial communication services.

11.8 TYPES OF SATELLITES

There are two types of satellites based on their performance. They are

1. Passive satellite
2. Active satellite

11.8.1 Passive Satellite

A satellite that is used only as a reflector of signals transmitted from the earth station is called a passive satellite as it does not carry any equipment for receiving and transmitting the signal. It is a mere reflector and, therefore, a large amount of transmitter power is required to get a signal of suitable strength after reflection at the receiving point on the earth.

11.8.2 Active Satellite

A satellite that carries equipment for receiving earth signals, processing them and retransmitting them towards the earth is called an active satellite. Such a satellite also has to carry electrical power. Satellites presently in use are active satellites.

According to the applications, satellites are of four different types, namely

1. Communications satellites
2. Weather satellites
3. Remote-sensing satellites
4. Science-research satellites
5. GPS satellites

11.9 THEORY OF SATELLITE COMMUNICATION

Satellites can be placed in orbits around the earth at different altitudes. Depending on the heights above the earth' surface, orbits are classified as follows.

1. Low-earth orbit
2. Medium-earth orbit
3. Geosynchronous orbit

A satellite remains in the sky in the circular orbit if its linear velocity is so adjusted that the resulting centrifugal force caused by its rotation around the earth is equal and opposite to the earth's gravitational force. By equating the two forces,

$$\frac{GM \times m}{r^2} = \frac{mV^2}{r} \tag{11.1}$$

$$\frac{GM}{r} = V^2$$

$$\therefore V = \sqrt{\frac{GM}{r}} \text{ m/s} \tag{11.2}$$

where M is the mass of the earth, and

r is the radius of the circular orbit in km.

The orbital period of the satellite is calculated as follows.

$$T = \frac{2\pi r}{V} = 2\pi\sqrt{\frac{r^3}{GM}} \tag{11.3}$$

$$= 1.66 \times 10^{-4} \times r \text{ minutes}$$

The height of the orbit $h = (r - 6370)$ km

where 6370 km is the radius of the earth.

EXAMPLE 11.1

Calculate the radius of the circular orbit if the period is one day.

Solution

$$\omega = \frac{2\pi}{\text{One day}} = 7.272 \times 10^{-5} \text{ rad/s}$$

Using $\mu = 3.986005 \times 10^{14}$ m^3/s^2

$$\therefore a = \left(\frac{\mu}{\omega^2}\right)^{1/3} = 42241 \text{ km}$$

11.9.1 Low-Earth Orbit (LEO)

When a satellite circles close to the earth, it is said to be in Low-Earth Orbit (LEO). Satellites in LEO are just 200–500 miles (320–800 kilometres) high. Low-earth orbit refers to satellites in orbit at less than 22300 miles above the earth. To obtain a low-earth orbit, the speed of a satellite should be high. Low-earth-orbit satellites travel at a speed of 28 × 10^3 km/h (17,000 miles per hour) and complete one rotation around the earth in 1.5 hours.

This type of an orbit reduces transmission times as compared to a GEO. A LEO orbit can also be used to cover a polar region, which the GEO cannot accomplish. Since it does not appear stationary to earth stations, however, earth stations need an antenna assembly that will track the motion of the satellite.

A low-earth orbit is useful because its nearness to the earth gives it spectacular views. Satellites that observe our planet, like remote-sensing and weather satellites, often travel in LEOs because from this height, they can capture very detailed images of the earth's surface.

11.9.2 Medium-Earth Orbit (MEO)

Medium-earth orbit systems operate at about 10,000 kilometres above the earth, which is lower than the GEO orbit and higher than most LEO orbits. The MEO orbit is a compromise between the LEO and GEO orbits. Compared to LEOs, the more distant orbit requires fewer satellites to provide coverage than LEOs because each satellite may be in view of any particular location for several hours. Compared to GEOs, MEOs can operate effectively with smaller, mobile equipment and with less latency (signal delay).

Although MEO satellites are in view longer than LEOs, they may not always be at an optimal elevation. To combat this difficulty, MEO systems often feature significant coverage overlap from satellite to satellite and this in turn requires more sophisticated tracking and switching schemes than GEOs.

11.9.3 Geosynchronous Orbit (GEO)

If the orbit is at a height of 35,860 km above the surface of the earth, it is called a geosynchronous orbit. This is in the equatorial plane. So a satellite travelling at a speed of 11,000 km/h will complete one rotation around the earth in 24 hours. This equals the period of revolution of the earth around its own axis. Consequently, a satellite rotating in the direction of rotation of the earth appears stationary over a point on the surface of the earth. Such an orbit is called a geostationary orbit or geosynchronous orbit. Figure 11.3 shows a geosynchronous orbit.

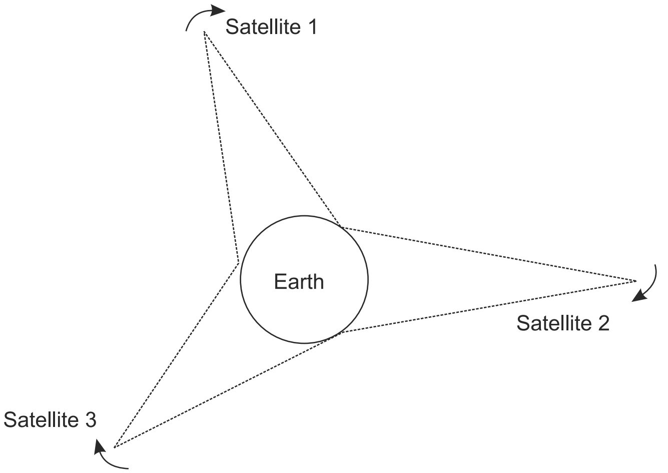


Fig. 11.3 A geosynchronous orbit

11.10 SATELLITE ORBITS

A satellite orbiting the earth stays in position because the centrifugal force on the satellite balances the gravitational attractive force of the earth. In addition, atmospheric drag must be negligible and this requires the satellite to be at a height greater than about 600 km. The choice of orbit is of fundamental importance, as it determines the transmission path loss and delay time, the earth coverage area and the time period the satellite is visible from any given area.

For satellite communication purposes, orbits are classified as follows.

1. Circular polar orbit
2. Elliptical orbit
3. Geostationary orbit

A typical communication satellite with various orbits is shown in Figure 11.4.

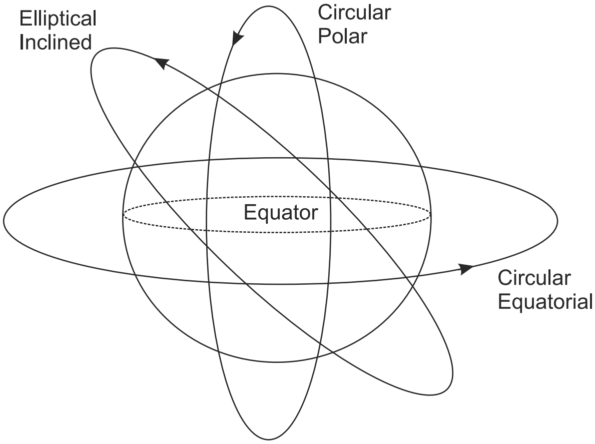


Fig. 11.4 A typical communication satellite with various orbits

11.10.1 Circular Polar Orbit

Polar orbits are useful for viewing the planet’s surface and also for special purposes such as navigational satellites. As a satellite orbits in a north-south direction, the earth spins beneath it in an east-west direction. A satellite in polar orbit can eventually scan the entire surface of the earth and the time taken for one complete orbit and a synchronous orbit is one for which the periodic time is an integer multiple or submultiple of the earth’s rotational period. No other orbit gives such thorough coverage of the earth. Satellites that monitor the global environment, like remote-sensing satellites and certain weather satellites, are almost always in polar orbit and generally, this orbit is not preferred for communication satellites.

A polar orbit has an inclination, or angle, of 90 degrees. It is perpendicular to an imaginary line that slices through the earth at the equator as shown in Figure 11.5.

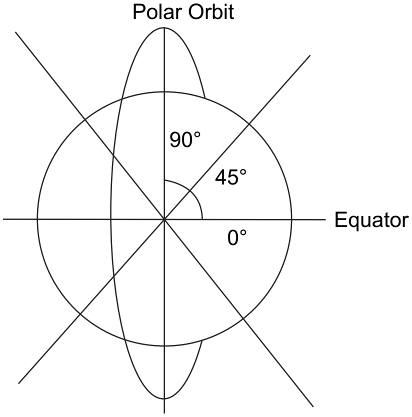


Fig. 11.5 Inclination of polar orbit

11.10.2 Elliptical Orbit

A satellite in elliptical orbit follows an oval-shaped path. The main advantage of an inclined elliptical orbit is that it provides coverage of the polar region. The highest point of the orbit is arranged to occur over the region requiring most coverage. A satellite in this orbit takes about 12 hours to circle the planet. Like polar orbits, elliptical orbits move in a north-south direction. Figure 11.6 shows an elliptical orbit.

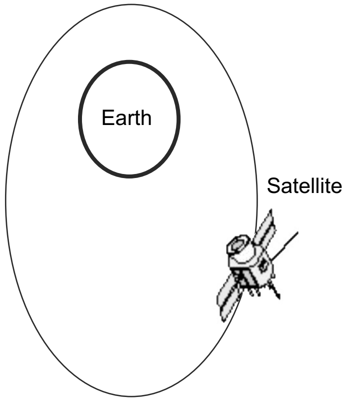


Fig. 11.6 Elliptical orbit

This puts the satellite at its greatest height and, therefore, gives the greatest earth coverage in this region. Also, the transit time is largest at the highest point of the orbit making the satellite visible for a relatively long period of time over these regions. The inclined elliptical orbit does not permit continuous contact with the satellite from a fixed spot on the earth.

11.10.3 Geostationary Orbit

It is the synchronous orbit which is most widely used. A satellite in the geosychonous equatorial orbit is located directly above the equator, exactly 22,300 miles out in space. The rotational period of the earth about its own axis is 23 hours and 56 minutes and a satellite in geostationary orbit travelling in the same direction as the earth's rotation completes one revolution about the earth's axis in this time. Due to the rotational speed of the earth, almost 24 hours to spin on in its axis, the satellite and the earth move together. So, a satellite in GEO always stays directly over the same spot on the earth.

Since the satellite appears stationary to an observer on the earth, it is called a geostationary orbit and hence keeping track of a geostationary satellite is relatively easy and the satellite is continually visible from within its service area on the earth. Another advantage of the geostationary orbit is that the Doppler shift of frequency is negligible.

A geostationary orbit offers the following advantages.

1. The satellite remains almost stationary relative to the earth antennas so that the cost of computer-controlled tracking of the satellite is avoided. A fixed antenna is satisfactory.
2. It is not necessary to switch from one satellite to another as one disappears over the horizon.
3. There are no breaks in transmission. A geosynchronous satellite is permanently in view.
4. Because of its distance, a geosynchronous satellite is in line of sight from 42.4% of the earth's surface. A large number of earth stations may thus intercommunicate the information.
5. These satellites give global coverage with the exception of the polar regions.
6. There is no Doppler shift of frequency.

There are a few drawbacks of geostationary orbit.

1. Latitudes greater than 81.25° north and south are not covered.
2. Because of the distance of the satellite, the received signal power, which is inversely proportional to the square of the distance, is weak and the signal propagation delay is 270 milliseconds.

Many communications satellites travel in geosynchronous orbits, including those that relay TV signals into our homes.

11.10.4 Orbital Adjustments or Station Keeping

Geosynchronous satellites must have their orbits adjusted occasionally to keep them in position. Even if the satellites were launched into a perfect orbit, natural forces introduce a slight drift of the orbit with a good launch and, however, the movement of the satellite relative to an earth station will be very slow. Here, the drift is caused by

1. Minor gravitational perturbances of the orbit due to the sun, the moon and the oblateness of the earth, and
2. By the pressure of solar radiation.

Drift caused by the pressure of solar radiation will vary with the size of the satellite but it is small when compared with gravitational drift. The orbit of the satellite needs adjusting periodically. This can be done by the release of gas under pressure or in a layer satellite by small rockets. Figure 11.7 illustrates the orbital adjustment.

The gravitational pull of the sun and moon can cause the orbit of a geostationary to incline as shown above in the figure. The angle of inclination in degrees is the specified inclination error. This can be kept within ±0.1° by means of command controls from the earth. The control routine necessary to keep the satellite in position is referred to as station keeping.

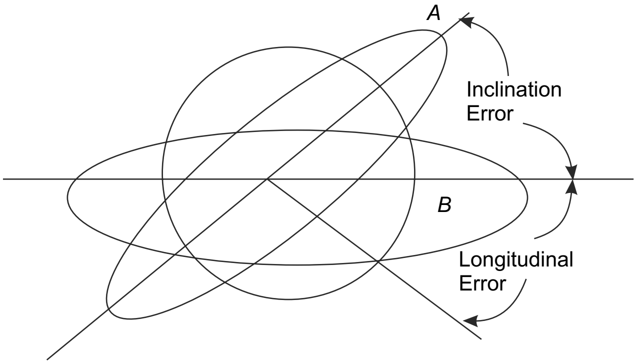


Fig. 11.7 Orbital adjustment

11.10.5 Laws Governing Satellite Motion

There are two basic laws which define the governing of satellite motion. They are

1. Kepler's laws
2. Newton's laws

1. Kepler's Laws

Satellite motion is defined by three laws of Kepler which are stated as follows.

(a) First Law The path followed by the satellite (earth) around the primary (sun) is an ellipse. The ellipse has two focal points. The centre of mass **(barycentre)** is always centred at one of the foci. Figure 11.8 shows the first law of Kepler.

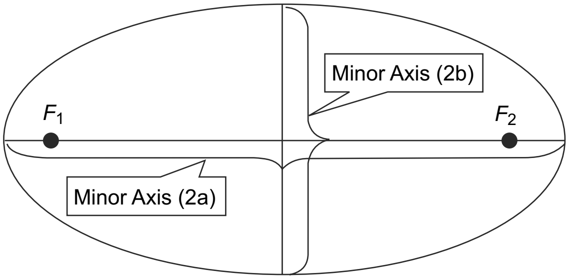


Fig. 11.8 First Law of Kepler

(b) Second Law For equal time intervals, the satellite sweeps out equal areas in its orbital plane, focused at the barycentre. Figure 11.9 shows the second law of Kepler.

Assuming the satellite travels distances S_1 and S_2 in 1 second then the areas A_1 and A_2 will be equal.

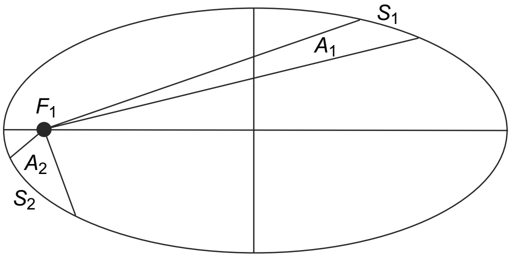


Fig. 11.9 Second law of Kepler

(c) Third Law

The square of the periodic time of orbit is proportional to the cube of the mean distance between the two bodies, namely satellite and the primary.

The mean distance of the satellite is equal to the semi-major axis *a*.

$$a^3 = \frac{\mu}{\omega^2} \tag{11.4}$$

where μ is the earth's gravitational constant and is given by 3.986005×10^{14} m^3/s^2, and ω is the mean angular velocity of the satellite, which is related to the period by

$$\omega = \frac{2\pi}{T} \tag{11.5}$$

$$T^2 = \left(\frac{4\pi^2}{\mu}\right)a^3 \tag{11.6}$$

2. Newton's Laws

Newton's laws of motion characterised the forces that give rise to Kepler's laws.

(a) First Law Every body continues in its state of rest or uniform motion in a straight line unless it is compelled to change that state by forces impressed on it.

(b) Second Law The rate of change of momentum of a body is proportional to the force impressed on it and is in the same direction as that force. Mathematically,

$$\sum F = m\ddot{r} \tag{11.7}$$

(c) Third Law To every action, there is an equal and opposite reaction.

Newton also postulated the law of gravitation, which states that any two bodies attract one another with a force proportional to the product of their masses and inversely proportional to the square of the distance between them.

Mathematically,

$$F = \frac{GMm}{r^2}\hat{r} \tag{11.8}$$

G is the gravitational constant = 6.67×10^{-8} dyn.cm^2/g^2

Newton's and Kepler's laws completely explain the motion of planets around the sun (and satellites around the earth).

11.10.6 Orbital Parameters

The satellite orbit could be circular or elliptical. Its characteristic parameters are governed by Kepler's laws. Based on these laws, the important orbital parameters of satellites orbiting the earth are mathematically derived.

1. Geocentre

A satellite rotates in an orbit that forms a plane passing through the centre of gravity of the earth called the geocentre. Figure 11.10 shows the orbital plane passing through the geocentre.

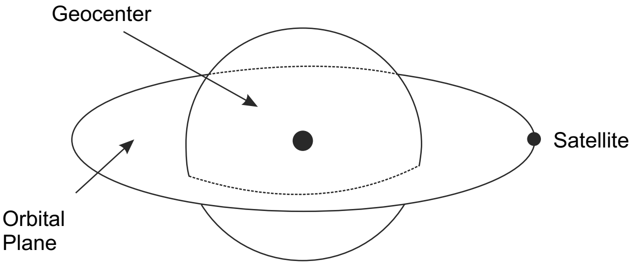


Fig. 11.10 The orbital plane passes through the geocentre

2. Posigrade Orbit

Posigrade orbit is an orbit in which the projection of the satellite's position on the earth's equatorial plane revolves in the direction of the rotation of the earth. The inclination of a posigrade orbit is less than 90°. Most orbits are posigrade.

3. Retrograde Orbit

Retrograde orbit is an orbit in which the projection of the satellite's position on the earth's equatorial plane revolves in the direction opposite to that of the rotation of the earth. The inclination of a retrograde orbit is greater than 90°.

4. Perigee

When a satellite follows a noncircular orbit around the earth, the satellite's path is an ellipse with the centre of the earth at one focus. Such a satellite has variable altitude and variable orbital speed. The point of lowest altitude is called perigee. The term also applies to the minimum distance in kilometres or miles between the satellite and the centre of the earth. (Perigee can be measured between the satellite and the earth's surface, although this is a less precise specification because the earth is not a perfect sphere. The difference is approximately 4,000 miles or 6,400 kilometres.) Figure 11.11 shows the calculation of perigee.

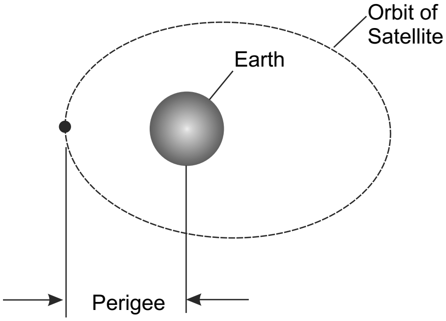


Fig. 11.11 Calculation of perigee

5. Apogee

Apogee is the point in the satellite orbit farthest from the centre of the earth. The term also applies to the maximum distance in kilometres or miles between the satellite and the centre of the earth. Apogee can also be measured between the satellite and the earth's surface, although this is a less precise specification because the earth is not a perfect sphere. Figure 11.12 shows the calculation of apogee.

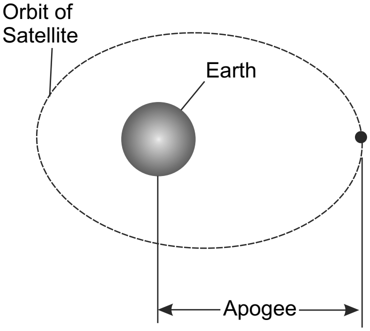


Fig. 11.12 Calculation of apogee

6. Orbital Eccentricity

Orbit eccentricity, represented by e, is a measure of the displacement of the centre of the orbit from the centre of the earth. It is defined as the ratio of the distance between the centre of the ellipse and the centre of the earth to the semimajor axis of the ellipse. Orbit eccentricity can also be expressed in terms of apogee and perigee distances as

$$\text{Eccentricity } (e) = \frac{\text{Apogee} - \text{Perigee}}{\text{Apogee} + \text{Perigee}} \tag{11.9}$$

Here, apogee and perigee represent the distances of apogee and perigee points from the centre of the earth.

Similarly,

$$\text{Eccentricity } (e) = \frac{\text{Apogee} - \text{Perigee}}{2\alpha} \tag{11.10}$$

where α is the semimajor axis of the ellipse.

$$\alpha = \frac{\text{Apogee} - \text{Perigee}}{2} \tag{11.11}$$

$$\text{Apogee} = \alpha\,(1 + e) \tag{11.12}$$

$$\text{Perigee} = \alpha\,(1 - e) \tag{11.13}$$

EXAMPLE 11.2

A satellite moving in an elliptical eccentric orbit has the semimajor axis of the orbit equal to 16,000 km. If the difference between the apogee and the perigee is 30,000 km, determine the orbit eccentricity.

Solution

$$\text{Apogee} = \alpha(1 + e)$$

$$\text{Perigee} = \alpha(1 - e)$$

$$\text{Eccentricity } (e) = \frac{\text{Apogee} - \text{Perigee}}{2\alpha} \qquad \text{(Eq. 11.10)}$$

$$\text{Apogee} - \text{Perigee} = \alpha\,(1 + e) - \alpha(1 - e) \quad = 2\alpha\, e$$

$$\therefore \quad \text{Eccentricity } (e) = \frac{30000}{2 \times 16000}$$

$$= \frac{30000}{32000} = 0.93$$

EXAMPLE 11.3

A satellite in an elliptical orbit has an apogee of 30,000 km and a perigee of 1000 km. Determine the semimajor axis of the elliptical orbit.

Solution

$$\text{Semimajor axis } \alpha = \frac{\text{Apogee} - \text{Perigee}}{2}$$

$$= \frac{30000 + 1000}{2} \quad 15{,}500 \text{ km}$$

EXAMPLE 11.4

The farthest and the closest points in a satellite's elliptical eccentric orbit from the earth's surface are 30,000 km and 200 km respectively. Determine the apogee, the perigee and the orbit eccentricity. Assume radius of earth to be 6370 km.

Solution

$$\text{Eccentricity} = \frac{\text{Apogee} - \text{Perigee}}{2\alpha}$$

$$\text{Semimajor axis } \alpha = \frac{\text{Apogee} + \text{Perigee}}{2}$$

$$\therefore\ 2\alpha = \text{Apogee} + \text{Perigee}$$

$$\text{Eccentricity } (e) = \frac{\text{Apogee} - \text{Perigee}}{\text{Apogee} + \text{Perigee}}$$

$$= \frac{36370 - 6570}{36370 + 6570} = \frac{29800}{42940} = 0.693$$

EXAMPLE 11.5

Determine then apogee and perigee distances if the orbit eccentricity is 0.5 and the distance from the centre of the ellipse to the centre of the earth is 14000 km.

Solution

The distance from the centre of the ellipse to the centre of the earth is given by $\alpha \times e$

$$\therefore\ \alpha \times e = 14000$$

$$\therefore\ \alpha = \frac{14000}{e} = \frac{14000}{05} = 28000 \text{ km}$$

$$\text{Apogee} = \alpha(1 + e)$$

$$= 28000\,(1 + 0.5) \quad = 42000 \text{ km}$$

$$\text{Perigee} = \alpha\,(1 - e)$$

$$= 28000\,(1 - 0.5) \quad = 14000 \text{ km}$$

EXAMPLE 11.6

A satellite is moving in a highly eccentric orbit having the farthest and closest points as 35,000 km and 500 km respectively from the surface of the earth. Determine the orbital timeperiod and the velocity at the apogee and perigee points. Assume the earth's radius is 6360 km.

Solution

Apogee distance = 35000 + 6360 = 41360 km

Perigee distance = 500 + 6360 = 6860 km

$$\text{Semimajor axis } \alpha = \frac{\text{Apogee + Perigee}}{2}$$

$$= \frac{41360 + 6860}{2} = 24110 \text{ km}$$

$$\text{Orbital time period } T = 2\pi \times \sqrt{\frac{\alpha^3}{\mu}}$$

$$\mu = \text{GM}$$

$$= 6.67 \times 10^{-11} \times 5.98 \times 10^{24} \quad = 39.8 \times 10^{13}$$

$$\therefore \quad T = 2\pi \times \sqrt{\frac{\alpha^3}{\mu}} = 2\pi \times \sqrt{\frac{(24110 \times 10^3)^3}{39.8 \times 10^{13}}} = 37200 \text{ seconds}$$

$$\text{Velocity at any point on the orbit is given by } V = \sqrt{\mu\left(\frac{2}{r} - \frac{1}{\alpha}\right)}$$

$$\mu = \text{GM}$$

$$= 6.67 \times 10^{-11} \times 5.98 \times 10^{24}$$

$$= 39.8 \times 10^{13}$$

At apogee point, $r = 41360$ km

$$V = \sqrt{39.8 \times 10^{13}\left(\frac{2}{41360 \times 10^3} - \frac{1}{24110 \times 10^3}\right)}$$

$$= \sqrt{39.8 \times 10^{13}\left(\frac{48220 - 41360}{41360 \times 10^3 \times 24110 \times 10^3}\right)}$$

$$= \sqrt{\frac{39.8 \times 10^{13} \times 6860}{41360 \times 24110 \times 10^{3}}} = 523 \text{ m/s}$$

At perigee point, $r = 6860$ km

$$V = \sqrt{39.8 \times 10^{13} \left(\frac{2}{6860 \times 10^{3}} - \frac{1}{24110 \times 10^{3}} \right)}$$

$$= \sqrt{39.8 \times 10^{13} \left(\frac{48220 - 6860}{6860 \times 24110 \times 10^{3}} \right)}$$

$$= \sqrt{\frac{39.8 \times 10^{13} \times 41360}{6860 \times 24110 \times 10^{3}}} = 9.976 \text{ km/s}$$

EXAMPLE 11.7

The sum of apogee and perigee distances of a certain elliptical satellite orbit is 50,000 km and the difference of apogee and perigee distances is 30,000 km. Determine the target eccentricity.

Solution

$$\text{Eccentricity } e = \left(\frac{r_a - r_p}{r_a + r_p} \right) = \frac{\text{Apogee} - \text{Perigee}}{\text{Apogee} + \text{Perigee}}$$

$$= \left(\frac{30000}{50000} \right) = 0.6$$

EXAMPLE 11.8

The semimajor axis and the semiminor axis of an elliptical satellite orbit are 20,000 km and 16,000 km respectively. Determine the apogee and perigee distances.

Solution

$$\text{Semimajor axis} = \frac{\text{Apogee} + \text{Perigee}}{2} = \frac{r_a + r_p}{2}$$

$$\text{Semimin or axis} = \sqrt{\text{Apogee} \times \text{Perigee}} = \sqrt{r_a \times r_p}$$

$$\frac{r_a + r_p}{2} = 20000 \text{ km}$$

$$\therefore \quad r_a + r_p = 2 \times 20000 = 40000 \text{ km}$$

$$\therefore \qquad r_\mathrm{p} = 40000 - r_\mathrm{a}$$

$$\sqrt{r_a \times r_p} = 16000$$

$$\therefore \qquad r_\mathrm{a} \times r_\mathrm{p} = 16000^2 = 256000000$$

Substituting the value of r_p,

$$r_\mathrm{a}\,(40000 - r_\mathrm{a}) = 256000000$$

$$r_\mathrm{a}^{\;2} - 40000 r_\mathrm{a} + 256000000 = 0$$

$$\therefore r_a = \frac{40000 \pm \sqrt{16\times10^8 - 10.24\times10^8}}{2}$$

$$= \frac{40000 \pm \sqrt{5.76\times10^8}}{2} = \frac{40000 \pm 2.4\times10^4}{2}$$

$$= 3.2 \times 10^4,\ 1.6 \times 10^4 = 32000 \text{ km},\ 16000 \text{ km}$$

r_a = 32,000 km, because it cannot be 16,000 km if the semimajor axis is 20,000 km.

$$\therefore\ r_\mathrm{p} = 40000 - r_\mathrm{a}$$

$$\therefore\ r_\mathrm{p} = 40000 - 32000 = 8000 \text{ km}$$

7. Orbital Period

Orbital period (T) is defined as the time taken to complete one rotation. It is given by

$$T = 2\pi \times \sqrt{\frac{\alpha^3}{\mu}} \tag{11.14}$$

where $\mu = \mathrm{GM}$ (11.15)

G is gravitation constant = 6.67×10^{-11} N-m^2/kg^2

M is mass of the earth = 5.98×10^{24} kg

EXAMPLE 11.9

Satellite-1 in an elliptical orbit has the orbit semimajor axis equal to 18,000 km, and satellite-2 in an elliptical orbit has the semimajor axis equal to 24,000 km. Determine the relationship between their orbital periods.

Solution

$$T = 2\pi \times \sqrt{\frac{\alpha^3}{\mu}}$$

$$\mu = GM$$

If α_1 and α_2 are the values of the semimajor axis of the elliptical orbits of the satellites 1 and 2, and T_1 and T_2 are the corresponding orbital periods then

$$T_1 = 2\pi \times \sqrt{\frac{\alpha_1^3}{\mu}}$$

$$T_2 = 2\pi \times \sqrt{\frac{\alpha_2^3}{\mu}}$$

$$\therefore \frac{T_2}{T_1} = \sqrt{\frac{\alpha_2^3}{\alpha_1^3}}$$

$$= \sqrt{\frac{24000^3}{18000^3}} = \left(\frac{4}{3}\right)^{3/2} = 1.54$$

Thus, orbital period of Satellite-2 is 1.54 times the orbital period of satellite-1.

EXAMPLE 11.10

Calculate the orbital period of a satellite in an eccentric elliptical orbit if the distance from the centre of the ellipse to the centre of the earth is 25,000 km. Gravitation constant is 6.67×10^{-11} Nm^2/kg^2 and mass of the earth is 5.98×10^{24}kg.

Solution

$$T = 2\pi \times \sqrt{\frac{\alpha^3}{\mu}}$$

$$\mu = GM$$

$$= 6.67 \times 10^{-11} \times 5.98 \times 10^{24} \qquad = 39.8 \times 10^{13}$$

$$\therefore T = 2\pi \times \sqrt{\frac{(25000 \times 10^3)^3}{39.8 \times 10^{13}}}$$

$$= 2\pi \times \sqrt{\frac{15625 \times 10^{18}}{39.8 \times 10^{13}}} = 39250 \text{ seconds}$$

8. Orbital Velocity

A satellite in orbit moves faster when it is close to the planet or other body that it orbits, and slower when it is farther away. When a satellite falls from high altitude to lower altitude, it gains speed, and when it rises from low altitude to higher altitude, it loses speed.

A satellite in circular orbit has a constant speed which depends only on the mass of the planet and the distance between the satellite and the centre of the planet.

Orbital velocity (V) can be computed as follows.

$$V = \sqrt{\mu\left(\frac{2}{r} - \frac{1}{\alpha}\right)} \tag{11.16}$$

where r is the distance of the satellite from the centre of the earth.

From the above expression, it is clear that the instantaneous orbital velocity keeps changing in case of an elliptical orbit due to continuous change in the distance of the satellite from the centre of the earth. This is a general expression that can be used to compute the satellite velocity at any given point in its elliptical orbit. For a circular orbit, the semimajor axis is the same as the radius R_e which reduces the above expression to

$$V = \sqrt{\frac{\mu}{r}} = \sqrt{\frac{\mu}{R_e + H}} \tag{11.17}$$

EXAMPLE 11.11

Determine the orbital velocity of a satellite moving in a circular orbit at a height of 150 km above the surface of the earth given that gravitation constant $G = 6.67 \times 10^{-11}$ N-m^2/kg^2, mass of the earth $M = 5.98 \times 10^{24}$ kg, radius of earth R_e is 6370 km.

Solution

$$V = \sqrt{\frac{\mu}{R_e + H}}$$

$$\mu = GM$$

$$= 6.67 \times 10^{-11} \times 5.98 \times 10^{24} \qquad = 39.8 \times 10^{13} \text{ Nm}^2/\text{kg}$$

$$\therefore V = \sqrt{\frac{39.8 \times 10^{13}}{(6370 + 150) \times 10^3}} \quad = 7.813 \text{ km/s}$$

EXAMPLE 11.12

Determine the escape velocity for an object to be launched from the surface of the earth from a point where the earth's radius is 6360 km. Gravitation constant is 6.67×10^{-11} Nm^2/kg^2 and mass of the earth is 5.98×10^{24}kg.

Solution

$$\text{Escape velocity} = \sqrt{\frac{2\mu}{r}}$$

$$\mu = GM$$

$$= 6.67 \times 10^{-11} \times 5.98 \times 10^{24} = 39.8 \times 10^{13}$$

$$\therefore \text{ Escape velocity} = \sqrt{\frac{2 \times 39.8 \times 10^{13}}{6360 \times 10^{3}}} = 11.2 \text{ km/s}$$

9. Satellite Height

In a circular orbit, the height is simply the distance of the satellite from the earth. The height is really the distance between the centre of the earth and the satellite. By considering the radius of the earth as 3960 miles and a satellite that is 5000 miles above the earth, the satellite height is considered to be about 8960 miles from the centre of the earth. Figure 11.13 shows satellite height in circular orbit.

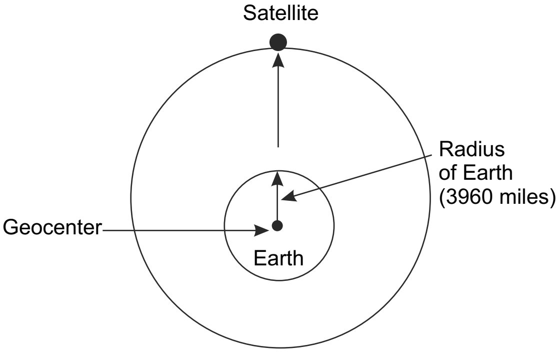


Fig. 11.13 Satellite height in circular orbit

10. Latitude and Longitude Drift in Inclined Orbits

In case of inclined synchronous orbits, both latitude and longitude undergo a drift which is a function of the inclination angle. This orbit inclination in effect gives the satellite an apparent movement in the form of 'figure eight' with the maximum deviation in latitude from the equator given by

$$\lambda_{\max} = i \tag{11.18}$$

where $\lambda_{\max}$ is the maximum latitude deviation in degrees, and

i is the angle of inclination in degrees.

Maximum deviation in longitude from the ascending node for $i < 5°$ is given by

$$\Psi_{\max} = \frac{i^2}{228} \tag{11.19}$$

where $\Psi_{\max}$ is the maximum longitude deviation from the ascending node.

The orbital inclination can be corrected by applying a velocity impulse perpendicular to the orbital plane when the satellite passes through the nodes. For a given inclination angle, the required impulse amplitude is given by

$$\Delta V = \sqrt{\frac{\mu}{\alpha}} \tan i \qquad (11.20)$$

EXAMPLE 11.13

Determine the magnitude of velocity impulse needed to correct the inclination of 2° in the satellite orbit 35,800 km above the surface of the earth, given that radius of the earth is 6364 km, mass of the earth is 5.98 × 10^{24} kg and gravitation constant is 6.67 × 10^{-11} N-m^2/kg^2.

Solution

The magnitude of the velocity impulse is given by

$$= \sqrt{\frac{\mu}{r}} \tan i$$

$$\mu = GM$$

$$= 6.67 \times 10^{-11} \times 5.98 \times 10^{24} \quad = 39.8 \times 10^{13}$$

$$r = (6364 + 35800)\text{ km} = 42164 \times 10^{3}\text{ m}$$

$$\therefore \text{ Magnitude of the velocity impulse} = \sqrt{\frac{39.8 \times 10^{13}}{42164 \times 10^{3}}} \tan 2° \quad = 107 \text{ m/s}$$

11. Azimuth and Elevation Angles

The azimuth angle is defined as the angle produced by the intersection of the local horizontal plane and the plane formed by the satellite, earth station and the earth's centre with the true north. Depending upon the location of the earth's station, the azimuth angle can be computed from the northern and southern hemispheres.

(a) For Northern Hemisphere If the earth station is towards the west of the satellite then

$$A = 180° - A' \qquad (11.21)$$

If the earth station is towards the east of the satellite then

$$A = 180° + A' \qquad (11.22)$$

(b) For Southern Hemisphere If the earth station is towards the west of the satellite then

$$A = A' \qquad (11.23)$$

If the earth station is towards the east of the satellite then

$$A = 360° - A' \qquad (11.24)$$

$$\text{where } A' = \tan^{-1}\left[\frac{\tan|\theta_s - \theta_L|}{\sin\theta_l}\right] \qquad (11.25)$$

θ_S is the satellite longitude,

θ_L is the earth-station longitude, and

θ_l is the earth-station latitude.

The elevation angle (E) is defined as the angle produced by the intersection of the local horizontal plane and the plane constituted by the satellite, centre of the earth and earth station with the line of sight between the satellite and the earth station. It can be computed as follows.

$$E = \tan^{-1}\left[\frac{r - R_e \cos\theta_l \cos|\theta_s - \theta_L|}{R_e \sin\left[\cos^{-1}(\cos\theta_l \cos|\theta_S - \theta_L|)\right]}\right] - \left[\cos^{-1}\left(\cos\theta_l \cos|\theta_S - \theta_L|\right)\right] \qquad (11.26)$$

where R_e is the radius of the earth.

12. Earth-Coverage Angle and Slant Angle

Figure 11.14 shows the earth-coverage angle. The earth-coverage angle (2α) is a function of the elevation angle (E).

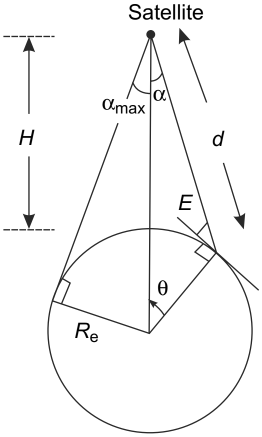


Fig. 11.14 Earth-coverage angle

The earth-coverage angle can be computed as follows.

$$2\alpha = 2\sin^{-1}\left[\frac{R_e}{R_e + H}\cos\ E\right] \qquad (11.27)$$

where R_e is radius of the earth, and

H is the height of the satellite above the earth's surface.

Maximum earth-coverage angle will be for $E = 0$ which gives

$$2\alpha_{\text{max}} = 2\sin^{-1}\left[\frac{R_e}{R_e + H}\right] \tag{11.28}$$

The slant range (d) can be computed from the following expression.

$$d^2 = (R_e + H)^2 + R_e^2 - 2R_e\,(R_e + H).\sin\left[E + \sin^{-1}\left(\frac{R_e}{R_e + H}\cos E\right)\right] \tag{11.29}$$

EXAMPLE 11.14

The basis of a satellite orbiting around the earth is the centripetal force (F_1) due to the earth's gravitation acting towards the centre of the earth balancing the centrifugal force (F_2) acting away from the centre. Calculate the centrifugal force for a satellite of 100 kg mass orbiting with a velocity of 8 km/s at a height of 200 km above the surface of the earth. Assume mean radius of the earth to be 6370 km.

Solution

$$\text{Centripetal force} = \frac{mV^2}{(R + H)}$$

The centripetal force balances the centrifugal force.

$$\therefore \text{ Centrifugal force} = \frac{mV^2}{(R + H)}$$

$$= \frac{100 \times (8000)^2}{(6370 + 200) \times 10^3} = \frac{64 \times 10^8}{6570 \times 10^3} = 974 \text{ newtons}$$

EXAMPLE 11.15

A geosynchronous satellite moving in an equatorial circular orbit at a height of 35,800 km above the surface of the earth gets inclined at an angle of 2° due to some reasons. Calculate the maximum deviation in latitude and also the maximum deviation in longitude. Also determine maximum displacements in km caused by latitude and longitude displacements. Assume radius of the earth is 6364 km.

Solution

Height of orbit = 35,800 km

Radius of earth = 6364 km

Orbit radius r = 35 800 + 6364 = 42,164 km

Angle of inclination = 2°

Maximum latitude deviation from equator due to inclination is given by

$$\lambda_{\max} = i = 2°$$

Maximum longitude deviation from ascending node is given by

$$\psi_{\max} = \frac{i^2}{228}$$

$$= \frac{i^2}{228} = \frac{2^2}{228} = 0.0175°$$

Maximum displacement (in km) due to $\lambda_{\max}$ is given by

$$D\lambda_{\max} = \alpha\, i\left(\frac{\pi}{180}\right) \qquad = \frac{42164 \times 2 \times \pi}{180} = 1471\,\text{km}$$

Maximum displacement (in km) due to $\psi_{\max}$ is given by

$$D_{\psi} = D_{\lambda}\left(\frac{\psi_{\max}}{\lambda_{\max}}\right)$$

$$= 1471 \times \frac{0.0175}{2} = 12.9\ \text{km}$$

EXAMPLE 11.16

A geosynchronous satellite orbiting at 42,164 km from the earth's centre has a circular equatorial orbit. The orbit gets inclined due to some reason and it is observed that the maximum displacement due to latitude deviation is 500 km. Determine angle of inclination between the new orbital plane and the equatorial plane.

Solution

$$D\lambda_{\max} = r\,\lambda_{\max}$$

$$\lambda_{\max} = i \text{ (angle of inclination)}$$

$$D\lambda_{\max} = r \times i$$

$$\therefore i = \frac{D\lambda_{\max}}{r} = \frac{500}{42164} = 0.68°$$

EXAMPLE 11.17

A geostationary satellite moving in an equatorial circular orbit is at a height of 35,786 km from earth's surface. If the earth's radius is taken as 6378 km, determine the theoretical maximum coverage angle. Also determine the maximum slant angle.

Solution

For theoretical maximum coverage angle, elevation angle $E = 0$.

Maximum coverage angle,

$$2\alpha_{\max} = 2\sin^{-1}\left[\frac{R_e}{R_e + H}\cos E\right]$$

$$2\alpha_{\max} = 2\sin^{-1}\left[\frac{6378}{6378 + 35786}\cos 0°\right] = 17.4°$$

The slant range (d),

$$d^2 = (R_e + H)^2 + R_e^2 - 2R_e\,(R_e + H).\sin\left[E + \sin^{-1}\left(\frac{R_e}{R_e + H}\cos E\right)\right]$$

$$= (6378)^2 + (42164)^2 - 2 \times 6378 \times 42164 \times \sin 8.7°$$

$$= 40678884 + 1777802896 - 537843984 \times 0.1512 = 1737139041$$

$\therefore\ d = 41679$ km

11.11 SATELLITE EARTH-STATION SUBSYSTEM

Figure 11.15 shows the block diagram of a satellite earth-station subsystem in which there are four major subsystems. They are

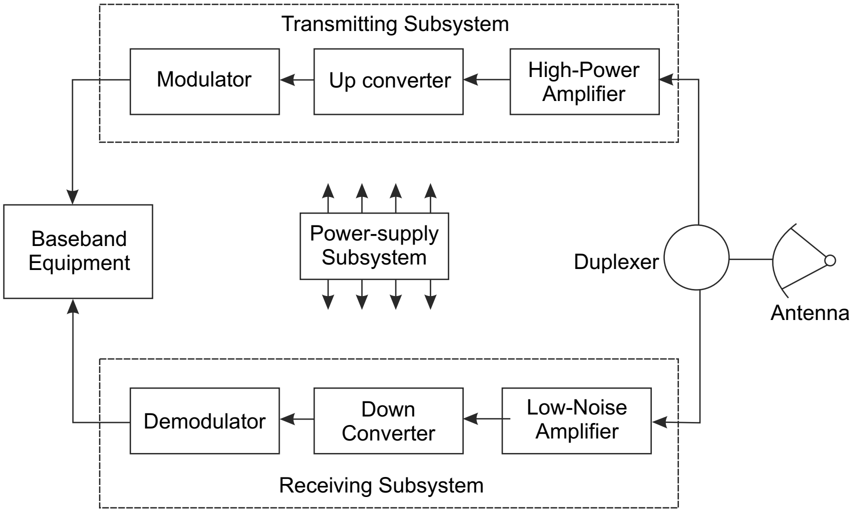


Fig. 11.15 Block diagram of a satellite earth-station subsystem

1. Power-supply subsystem
2. Transmitting subsystem
3. Receiving subsystem
4. Antenna subsystem

11.11.1 Power-Supply Subsystem

In the power-supply subsystem, a feasible commercial supply is used as the prime power source. Two standby generators will be installed to provide emergency power during the commercial out stages. To provide uninterrupted power supply, a rectifier, a storage battery and static inverter are to be used. Under normal conditions, commercial ac power is fed into the rectifier and the output of the rectifier maintains a float charge on the battery and provides the dc supply.

11.11.2 Transmitting Subsystem

The transmitting subsystem, also called an **uplink system**, includes the baseband processing stages, modulators, up-converters and high-power amplifiers. The block diagram of a transmitting subsystem is shown in Figure 11.16.

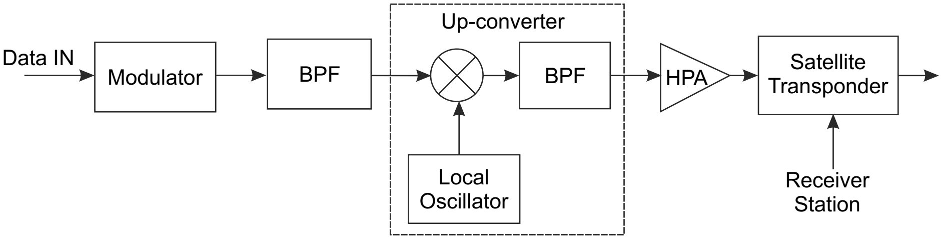


Fig. 11.16 Transmitting subsystem of a satellite earth station

The baseband could be multiplexed telephone channels or video and TV sound. The up-converter converts 70 MHz to 6 GHz signals. The up-converter could be either a frequency-synthesiser type or frequency-changeable crystal type.

The high-power amplifier is preferably a low-power klystron power amplifier for each carrier individually or high-power travelling wave-tube amplifier for several carriers collectively.

11.11.3 Receiving Subsystem

The receiving subsystem, also called **downlink system**, includes the baseband processing stages, demodulators, down-converters and low-noise amplifiers. The block diagram of a transmitting subsystem is shown in Figure 11.17.

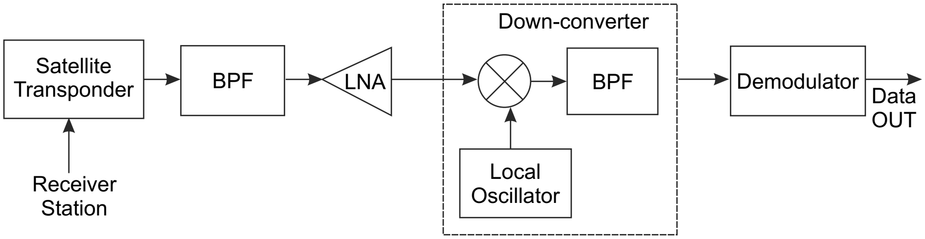


Fig. 11.17 Receiving subsystem of satellite earth station

The low-noise amplifier could be a parametric amplifier either cooled or uncooled. Most of the small earth stations employed in domestic satellite communication systems use a 40 K uncooled pre-amplifier. In contrast to an up-converter in the transmitting subsystem, a down-converter is used in a receiving subsystem which is used to recover the original information.

11.11.4 Antenna Subsystem

The main requirement of transmitting at a ground station is that it should concentrate all the RF power towards the satellite, and the power in the minor lobe represents wastage of the signal power and this may cause interference to other services too. The receiving antenna should pick up signals, both communications and tracking, coming from the satellite and from no other directions. Radiations picked up from other directions are unwanted and those are added to the noise and the signal-to-noise ratio of the desired signal. For these reasons, both the receiving and transmitting antenna have to be directional and always pointed at the satellite.

Generally, parabolic horn and cassegrain antennas have been used. The horn antenna is more efficient and less noisy but large in size and more expensive. The cassegrain antenna, which is shown in Figure 11.18, appears to be ideal for fixed ground stations. It has a main parabolic reflector which is illuminated by a hyperbolic subreflector. The feed to the subreflector is at the back of the main reflector.

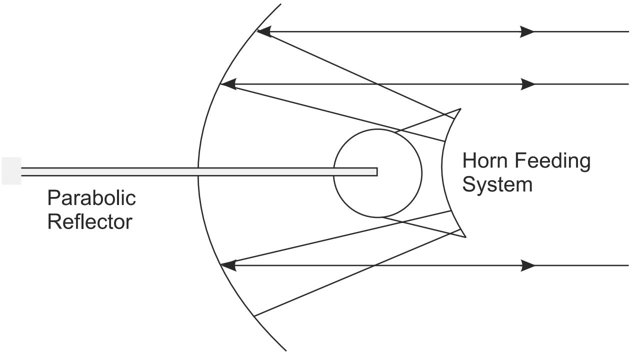


Fig. 11.18 Cassegrain antenna for earth station

The subreflector is placed at the focal point at the main reflector so that the radiation from the main reflector is in the form of a parallel beam. This antenna has the advantage of having low spillover, more flexibility in design and more mechanical stability of the feed system.

11.12 SATELLITE TRANSPONDER SUBSYSTEM

A satellite transponder is equipment which receives the signal beamed at it from the transmitting earth station on the uplink, does the required input filtering, down-converts to a different frequency and amplifies it to the desired level before it is retransmitted to the intended receiving station. The block diagram of a satellite transponder is shown in Figure 11.19.

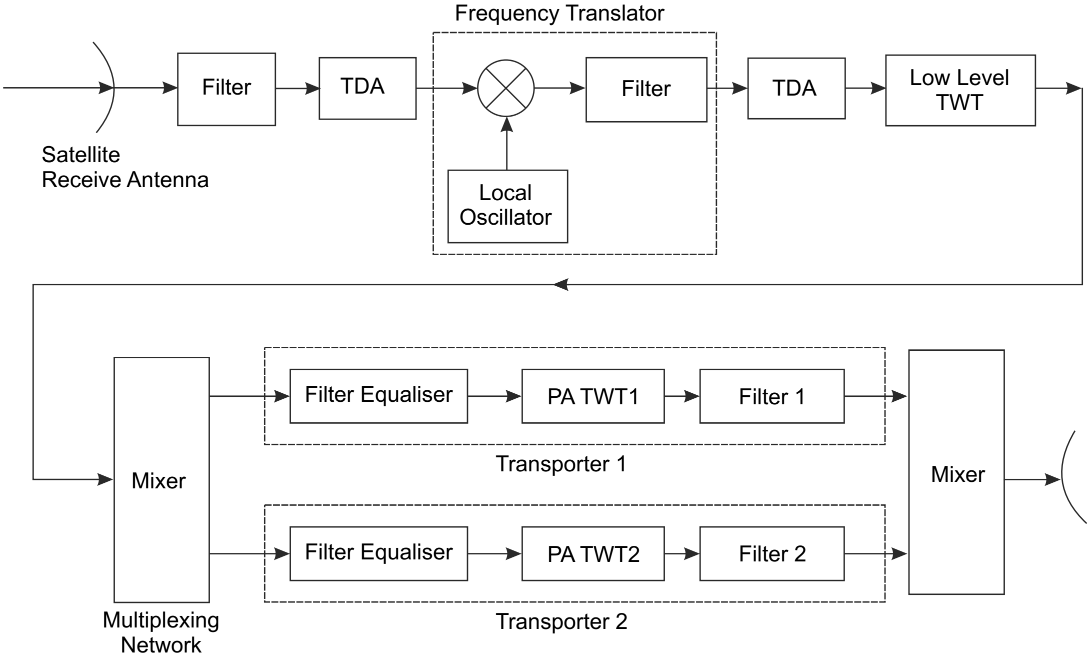


Fig. 11.19 Block diagram of a satellite transponder

Based on the uplink/downlink frequency being handled, transponders are categorized as C-band transponders, ku-band transponders, ka-transponders, and so on. Most satellites have more than one transponder. Bandwidth handled by a transponder differs from one satellite design to another, but most satellites have a frequency of 36 MHz.

11.13 SATELLITE LAUNCHING

In order to place a satellite into geosynchronous orbit, a large amount of energy is required. The satellite-launching process can be divided into two phases: They are as follows.

1. Launch phase
2. Orbit-injection phase

11.13.1 The Launch Phase

During the launch phase, the launch vehicle places the satellite into the transfer orbit, which is an elliptical orbit that has at its farthest point from the earth, or apogee, the geosynchronous elevation of 22,238 miles, and at its nearest point, or perigee, an elevation of usually not less than 100 miles. This is depicted in Figure 11.20.

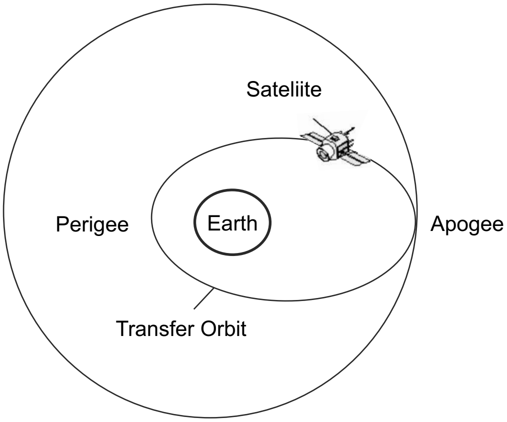


Fig. 11.20 The elliptical transfer orbit

11.13.2 The Orbit-Injection Phase

The energy required to move the satellite from the elliptical transfer orbit into the geosynchronous orbit needs to be supplied. This is known as the orbit-injection phase.

11.13.3 Launch Vehicles

Basically, there are two types of launch vehicles. They are as follows.

1. Expendable rockets which are destroyed while completing their mission
2. Space shuttles which are reusable

1. Expendable Rockets

There are three stages of expendable rockets used for communication satellites. They are as follows.

(a) In the first stage, there are several hundred thousand pounds of a kerosene/liquid oxygen mixture and a number of solid fuel rocket boosters that produce a tremendous display of flame—and ear-splitting noise—as the rocket lifts off the pad. It raises the satellite to an elevation of about 50 miles.

(b) In the second stage, the satellite rises to 100 miles.

(c) In the third stage, the satellite is placed into the transfer orbit. After the satellite is placed in its transfer orbit, the rocket's mission is complete, and its remnants fall on the earth.

The satellite is placed in its final geosynchronous orbital slot, which is fired on-command while the satellite is at the apogee of its elliptical transfer orbit. Figure 11.21 shows a picture of an expendable satellite-launch vehicle.

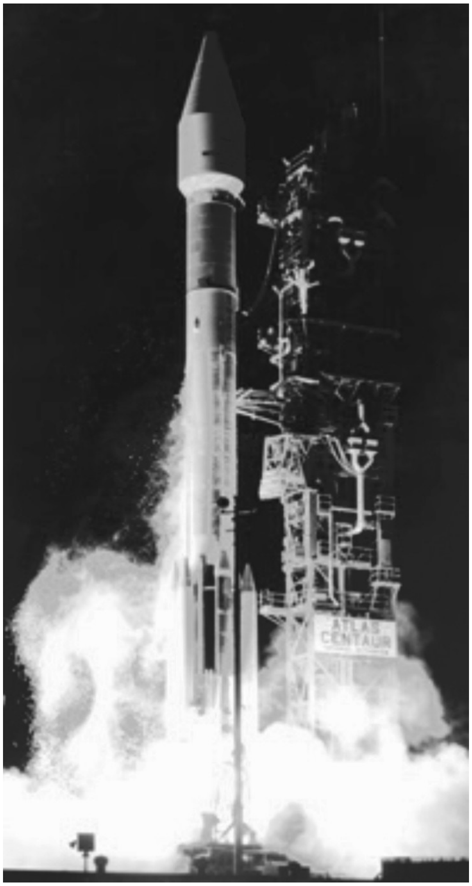


Fig. 11.21 Expendable satellite-launch vehicle [Courtesy: Atlas IIAS]

2. Space Shuttle

Satellites are carried into orbit by shuttles, where they go through a process of being launched. Shuttles flying into an orbit are inclined at 28.5 degrees to the equator. Space shuttles can carry a maximum of 3 or 4 satellites. These satellites are either pushed into orbit by a space arm or can be gently pushed out into orbit while spinning. The space shuttle shown performs the functions of the first two stages of an expendable launch vehicle. The satellite together with the third stage is mounted in the cargo bay of the shuttle. When the shuttle reaches its orbital elevation of 150 to 200 miles, the satellite and third-stage assembly are ejected from the cargo compartment. Then the third stage is fired, placing the satellite into the elliptical transfer orbit. At the apogee of the transfer orbit, the satellite is moved into its designated geosynchronous orbital slot. After all of its cargo has been jettisoned, the shuttle returns to the earth and is reused.

11.14 RADIO-WAVE TRANSMISSION

After the radiation of radio signals by the antenna, they will propagate through space and will reach the receiving antenna. For this reason, the energy level of the signal decreases rapidly as the distance from the transmitting antenna is increased and also the electromagnetic signal can take one or more of several different paths to the receiving antenna. The path that a radio signal takes depends upon many factors including the frequency of the signal, atmospheric conditions, and the time of day.

There are three basic paths of radio-signal transmission through space. They are as follows.

1. Ground-wave or surface-wave transmission
2. Sky-wave transmission

.3. Space-wave transmission.

11.14.1 Ground-Wave Transmission

The ground or surface waves leave the antenna and remain close to the earth. Figure 11.22 shows the ground-wave transmission. The ground waves will follow the curvature of the earth and they can travel at distances beyond the horizon.

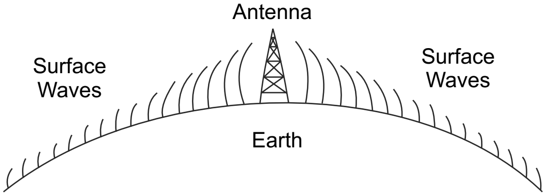


Fig. 11.22 Ground-wave transmission

At the low-and medium-frequency ranges, there is a strong propagation of ground waves. Ground waves are the main signal path for radio signals in the 30 kHz to 3 MHz range. These signals can propagate for hundreds and sometimes thousands of miles at these low frequencies. At the higher frequencies beyond 3 MHz, the earth begins to attenuate the radio signals.

Objects on the earth and terrain features become the same order of magnitude in size as the wavelength of the signal and will, therefore, absorb and otherwise affect the signal. For this reason, the ground-wave propagation of signals above 3 MHz is insignificant except within several miles of the antenna.

11.14.2 Sky-wave Transmission

Sky-wave signals are radiated by the antenna into the upper atmosphere where they will be bent or reflected back to earth. This bending of the signals is caused by a region in the upper atmosphere known as the ionosphere. Sky-wave propagation is illustrated in Figure 11.23.

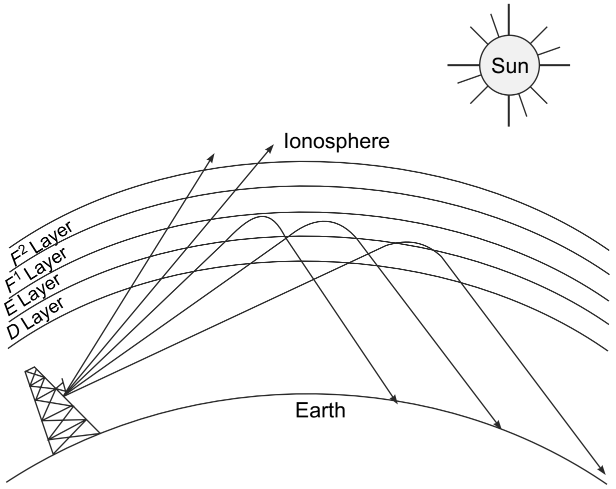


Fig. 11.23 Sky-wave propagation

Ultraviolet radiation from the sun causes the upper atmosphere to ionise. The atoms take on extra electrons or lose electrons to become positive and negative ions, respectively. This results in a relatively thick but invisible layer that exists above the earth. The ionosphere is generally considered to be divided into three basic layers. They are listed as follows.

1. D layer
2. E layer
3. F layer

Since the D and E layers are the farthest from the sun, these layers are considered weakly ionised areas. They exist only during daylight hours and during daylight hours they tend to absorb radio signals in the medium-frequency range from 300 kHz to 3 MHz.

The effect of refraction with different angles of radio signals entering the ionosphere is different for different angles. When the angle is large with respect to the earth, the radio signals are bent slightly and pass on through the ionosphere and are lost in space. If the angle of entry is smaller, the radio wave will actually be bent and sent back to the earth. Because of this effect, it actually appears as though the radio wave has been reflected by the ionosphere. The radio waves are sent back to the earth with minimum signal loss. The result is that the signal is propagated over an extremely long distance.

11.14.3 Space-Wave Transmission

This is the next method of radio signal propagation performed by space waves. A space wave travels in a straight line directly from the transmitting antenna to the receiving antenna. This is illustrated in Figure 11.24.

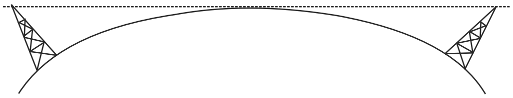


Fig. 11.24 Space-wave transmission

Direct or space waves are not refracted and also they do not follow the curvature of the earth. Because of their straight-line nature, direct waves will be blocked at some points because of the curvature of the earth. The signals will travel horizontally from the antenna until they reach the horizon at which point they are blocked. If the signal is to be received beyond the horizon then the antenna must be high enough to intercept the straight-line radio waves.

11.15 TROPOSPHERIC SCATTERING

Whenever there is turbulence during the travel of a radio wave through the troposphere, it makes an abrupt change in velocity. This causes a small amount of the energy to be scattered in a forward direction and returned to the earth at distances beyond the horizon. This phenomenon is repeated as the radio wave meets other turbulences in its path. The total received signal is an accumulation of the energy received from each of the turbulences.

This type of propagation enables VHF and UHF signals to be transmitted far beyond the normal line of sight. When the space wave is transmitted, it undergoes very little attenuation within the line-of-sight horizon. When it reaches the horizon, the wave is diffracted and follows the earth's curvature. Beyond the horizon, the rate of attenuation increases very rapidly and signals soon become very weak and unusable.

The magnitude of the received signal depends on the number of turbulences causing scatter in the desired direction and the gain of the receiving antenna. The scatter area used for tropospheric scatter is known as the **scatter volume**. The angle at which the receiving antenna must be aimed to capture the scattered energy is called the **scatter angle**. The scatter volume and scatter angle are shown in Figure 11.25.

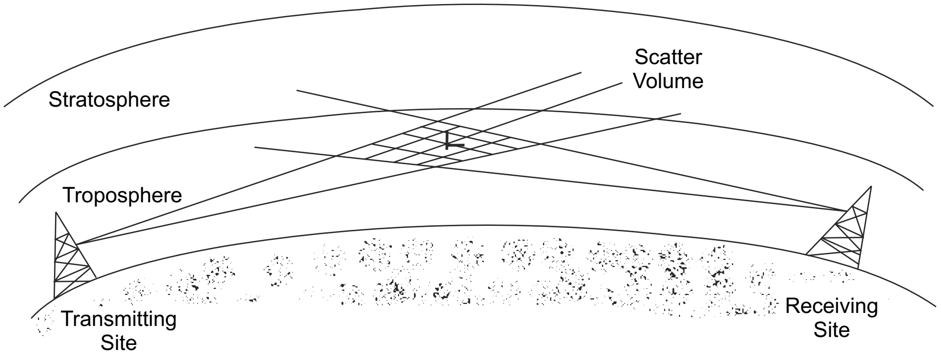


Fig. 11.25 Tropospheric scattering propagation

The angle of radiation of a transmitting antenna determines the height of the scatter volume and the size of the scatter angle. A low signal take-off angle produces a low scatter volume, which in turn permits a receiving antenna that is aimed at a low angle to the scatter volume to capture the scattered energy. As the signal take-off angle is increased, the height of the scatter volume is increased. Due to this situation, the amount of received energy decreases.

As the distance between the transmitting and receiving antennas is increased, the height of the scatter volume must also be increased. Therefore, the received signal level decreases as the distance is increased.

The tropospheric region that contributes most strongly to tropospheric scatter propagation lies near the midpoint between the transmitting and receiving antennas and just above the radio horizon of the antennas.

11.16 SATELLITE ANTENNAS

An antenna is one or more electrical conductors of a specific length that radiates radio waves generated by a transmitter or that collects radio waves at the receiver. Whenever voltage is applied to the antenna, an electric field will be set up and this voltage will cause current to flow in the antenna. This current flow will produce a magnetic field. The electric and magnetic fields are emitted from the antenna and propagate through space over very long distances.

It is important to consider the transmission and reception of radio waves with their orientation of the magnetic and electric fields with respect to the earth. The direction of the electric field specifies the polarisation of the antenna. If the electric field is parallel to the earth, the electromagnetic wave is said to be horizontally polarised and if the electric field is perpendicular to the earth then the wave is vertically polarised.

11.16.1 Dipole Antenna

The satellite antenna may be a length of wire, a metal rod or a piece of tubing. Satellites with many different sizes and shapes are used in satellite communication. The length of the conductor is dependent upon the frequency of transmission. Antennas radiate most effectively when their length is directly related to the wavelength of the transmitted signal. The most common lengths are one-half and one-quarter wavelengths. One of the most widely used antenna type is the half-wave dipole shown in Figure 11.26

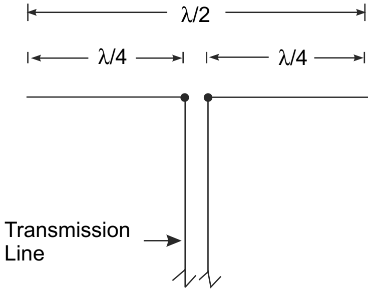


Fig. 11.26 A half-wave dipole antenna

Figure 11.27 shows the radiation pattern of a half-wave dipole. The dipole is at the centre hole of the doughnut shape, and the doughnut itself represents the radiated energy.

If the dipole receiving antenna is pointed towards the transmitter or vice versa, it must be broadside to the direction of the transmitter. If the antenna is at some angle, the maximum signal will not be received.

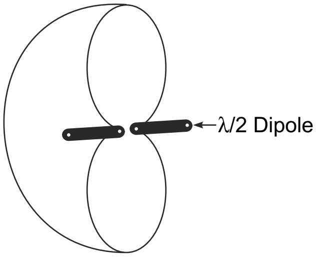


Fig. 11.27 Radiation pattern of a half-wave dipole

11.16.2 Folded Dipole

Another type of the half-wave dipole is the folded dipole. To construct this folded dipole antenna, a piece of twin lead is cut to a length of one-half wavelength and the two ends are soldered together as shown in Figure 11.28.

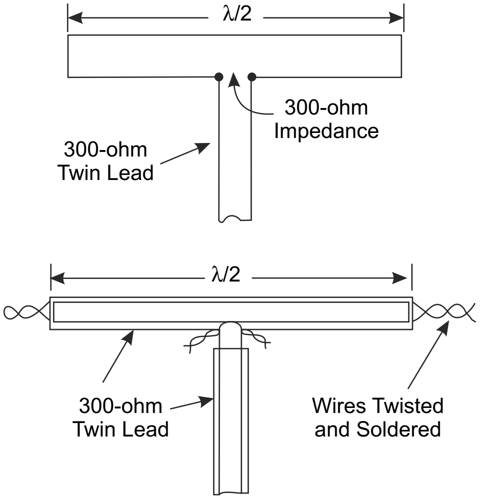


Fig. 11.28 Folded dipole

11.16.3 Directional Antennas

The major advantage of a vertical antenna with omnidirectional characteristics is that it can send message in any direction or receive them from any direction. But this requires an antenna with directivity which is the ability of an antenna to send or receive signals over a narrow horizontal directional range. In other words, the antenna has a response or directivity curve that makes it highly directional based on its physical orientation. The advantage of a directional antenna is that it eliminates interference from other signals being received from all directions except the direction of the desired signal. It provides selectivity based on the direction of the station to be received, and thereby effectively rejects signals from transmitters in all other directions.

11.16.4 Parasitic Arrays

A parasitic array consists of a basic antenna connected to a transmission line and one or more additional conductors that are not connected to the transmission line. These additional conductors are referred to as **parasitic elements**. The basic antenna itself is referred to as the **driven element**. Typically, the driven element is a half-wave dipole or some variation. The parasitic elements are slightly longer than and slightly less than one-half wavelength.

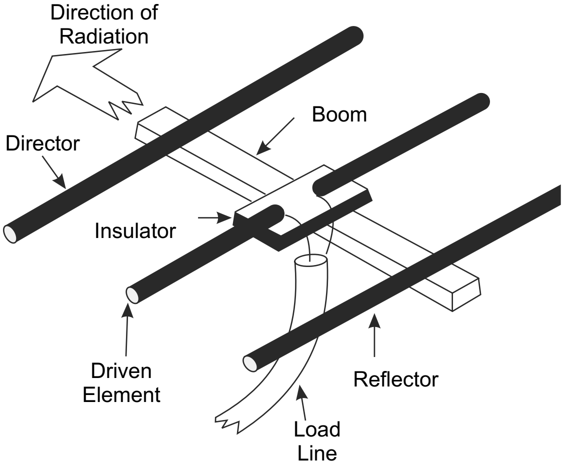


Fig. 11.29 A parasitic array

These parasitic elements are placed in parallel with and near the driven element. A common arrangement is illustrated in Figure 11.29

11.16.5 Horn Antenna

A horn antenna is used to transmit radio waves from a metal pipe or a waveguide used to carry radio waves out into space, or collect radio waves into a waveguide for reception. It typically consists of a short length of rectangular or cylindrical metal tube closed at one end, flaring into an open-ended conical or shaped horn on the other end. The radio waves are usually introduced into the waveguide by a coaxial cable attached to the side, with the central conductor projecting into the waveguide. The waves then radiate out of the horn end in a narrow beam. However, in some equipment, the radio waves are conducted from the transmitter or to the receiver by a waveguide, and in this case the horn is just attached to the end of the waveguide. Figure 11.30 shows the photographic view of a horn antenna.

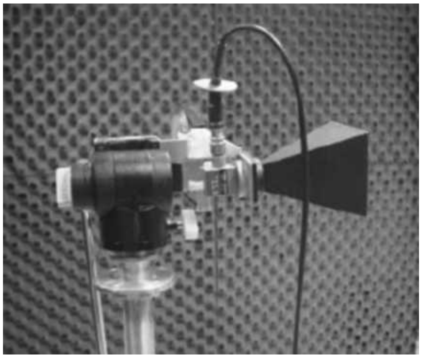


Fig. 11.30 Photographic view of a horn antenna

11.16.6 Helical Antenna

A helical antenna consists of a conducting wire wound in the form of a helix. And they are mounted over a ground plane. The feed line is connected between the bottom of the helix and the ground plane. They can operate in one of two principal modes: normal mode or axial mode.

In the **normal mode**, the dimensions of the diameter and the pitch are small compared to the wavelength. The antenna acts similarly to an electrically short dipole or monopole, and the radiation pattern, similar to these antennas, is omnidirectional, with maximum radiation at right angles to the helix axis. The radiation is linearly polarised parallel to the helix axis.

In the **axial mode**, the dimensions of the helix are comparable to a wavelength. The antenna functions as a directional antenna radiating a beam off the ends of the helix, along the antenna's axis. It radiates circularly polarised radio waves. Figure 11.31 shows the photographic view of a helical antenna.

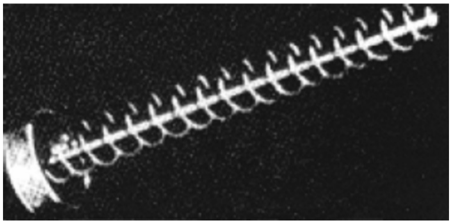


Fig. 11.31 Photographic view of a helical antenna

11.17 RADIATION PATTERN

The radiation pattern is a graphical representation of the field magnitude at a fixed distance from an antenna as a function of direction, i.e. angular variation of the test antennas radiation. It defines the variation of the power radiated by an antenna as a function of the direction away from the antenna. This power variation as a function of the arrival angle is observed in the antenna's far field.

Antenna radiation pattern is three-dimensional, but can be described on two-dimensional paper. The most popular technique is to record signal level along a great circle or conical cuts through the radiation pattern. In other words, one angular coordinate is held fixed, while the other varies. Radiation pattern is referred as the function of azimuth/elevation angles. Figure 11.32 shows an example of a doughnut-shaped or toroidal radiation pattern.

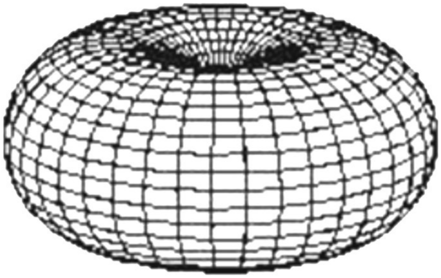


Fig. 11.32 Toroidal radiation pattern

In this case, along the *z*-axis, which would correspond to the radiation directly overhead the antenna, there is very little power transmitted. In the *x*-*y* plane (perpendicular to the *z*-axis), the radiation is maximum. These plots are useful for visualising which directions the antenna radiates. Figure 11.33 shows a polar radiation pattern.

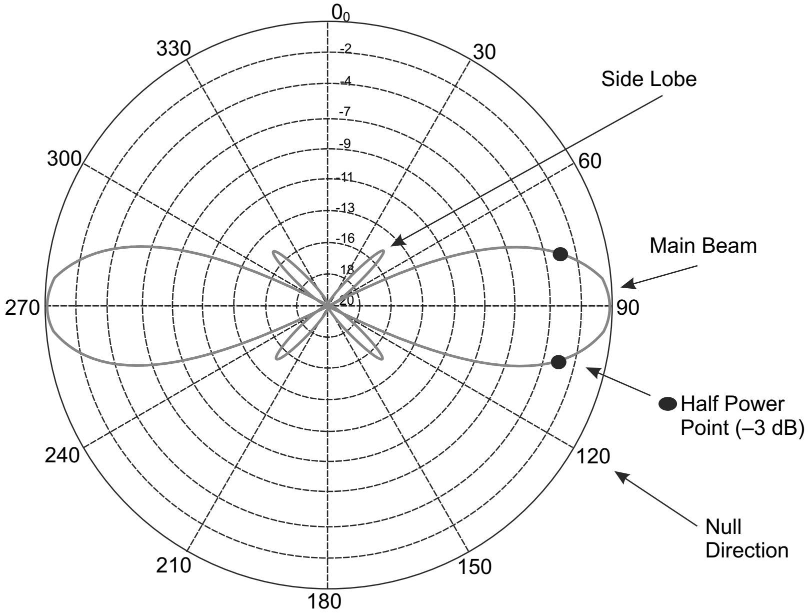


Fig. 11.33 Polar radiation pattern

There are several radiation pattern parameters. They are listed as follows.

1. Half-power Beam Width

The Half-Power Beam width (HPBW) is the angular separation in which the magnitude of the radiation pattern is decreased by 50% (or –3 dB) from the peak of the main beam. From Figure 11.33, the pattern decreases to –3 dB at 77.7 and 102.3 degrees. Hence, the HPBW is 102.3 – 77.7 = 24.6 degrees.

2. Main Lobes

The main beam is the region around the direction of maximum radiation that is within 3 dB of the peak of the main beam. The main beam in Figure 11.33 is centred at 90 degrees.

3. Side Lobes

The side lobes are smaller beams that are away from the main beam. These side lobes are usually radiation in undesired directions which can never be completely eliminated. The side lobes in Figure 11.33 occur at roughly 45 and 135 degrees

4. Antenna Directivity

It is a measure of how ‘directional’ an antenna’s radiation pattern is. An antenna that radiates equally in all directions would have effectively zero directionality, and the directivity of this type of antenna would be 1 (or 0 dB).

5. Gain

The gain of an antenna in any given direction is defined as the ratio of the power gain in a given direction to the power gain of a reference antenna in the same direction. It is important to state that an antenna with gain doesn’t create radiated power. The antenna simply directs the way the radiated power is distributed relative to radiating the power equally in all directions and the gain is just a characterisation of the way the power is radiated.

6. Polarisation

The polarisation of an antenna is the polarisation of the radiated fields produced by an antenna, evaluated in the far field. Hence, antennas are often classified as **linearly Polarised or a circularly polarised antenna**. A horizontally polarised antenna will not communicate with a vertically polarised antenna, and a vertically polarised antenna transmits and receives vertically polarised fields only. Consequently, if a horizontally polarised antenna is trying to communicate with a vertically polarised antenna, there will be no reception.

11.17.1 Types of Antenna Radiation Patterns

There are many types of antenna radiation patterns. The most common types are listed as follows.

1. Omnidirectional beam
2. Pencil beam
3. Fan beam
4. Shaped beam

1. Omnidirectional Beam

The omnidirectional beam is most popular in communication and broadcast applications. The azimuth pattern is circular but the elevation pattern will have some directivity to increase the gain in the horizontal directions. Figure 11.34 shows an omnidirectional beam.

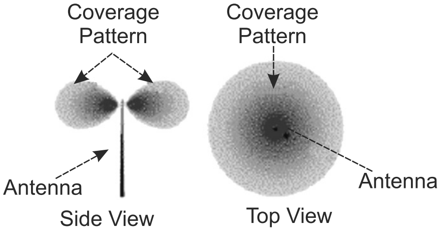


Fig. 11.34 Omnidirectional beam

2. Pencil Beam

A pencil beam is applied to a highly directive antenna pattern consisting of a major lobe contained within a cone of small solid angle. Usually, the beam is circularly symmetric about the direction of peak intensity. Figure 11.35 shows a pencil beam.

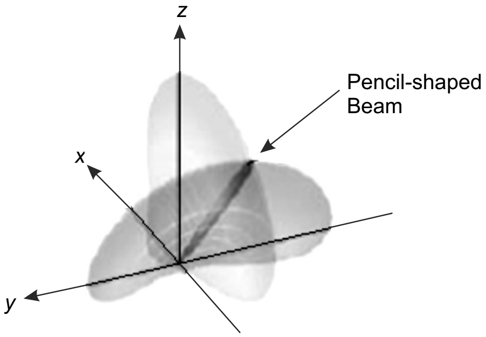


Fig. 11.35 A pencil beam

3. Fan Beam

A fan beam is narrow in one direction and wide in another direction. A typical use of a fan beam would be in a search or in surveillance radar. Figure 11.36 shows a fan beam.

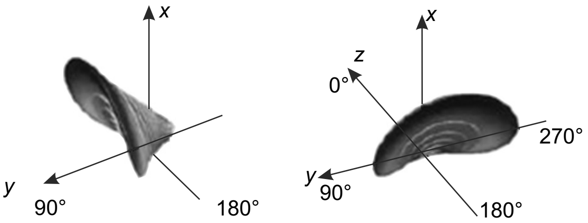


Fig. 11.36 A fan beam

4. Shaped Beam

Shaped beams are also used in search and surveillance.

Radiation patterns are generally defined as the far-field power or field strength produced by the antenna as a function of the direction (azimuth and elevation) measured from the antenna position. The behaviour of the field is changed with the distance from the antenna and there are three regions to be considered.

(a) Reactive Near-field Region It is the region in the space immediately surrounding the antenna in which the reactive field dominates the radiating field.

(b) Radiating Near-field Region It is found beyond the former region and is also called **Fresnel region**. In this region, the radiating field begins to dominate.

(c) Far-field Region Beyond this region, the reactive field and also the radial part of the fields becomes negligible. This region is also called **Fraunhofer region**. Generally, measurements are taken in the far-field region. In case of large planar antennas, it is more convenient to make near-field measurements and to calculate the far field.

11.17.2 Antenna Radiation-Pattern Lobes and Nulls

A radiation lobe can be defined as a portion of the radiation pattern bounded by regions of relatively weak radiation intensity. The main lobe is a high radiating energy region. Other lobes are called **side lobes** and the lobe radiating in the counter direction to the desired radiation is called **back lobe**. Regions for which radiation is very weak are called **nulls**. Figure 11.37 shows radiation-pattern lobes.

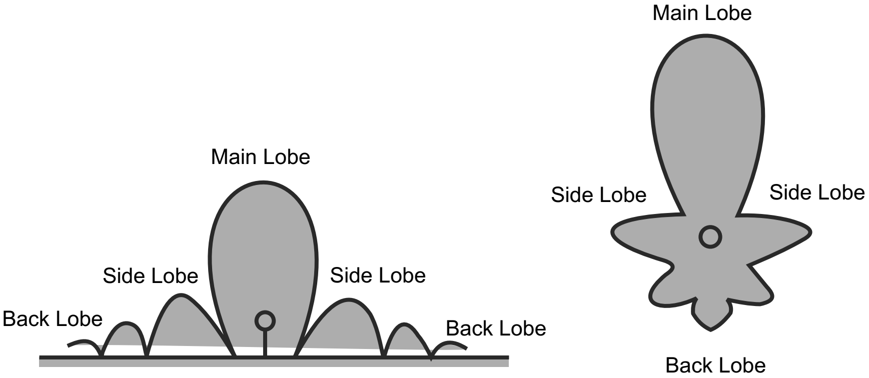


Fig. 11.37 Radiation-pattern lobes

11.17.3 Footprints

The footprint of a communications satellite is the ground area that its transponders offer coverage to and determines the satellite-dish diameter required to receive each transponder's signal. There is usually a different map for each transponder as each may be aimed to cover different areas of the ground. Footprint maps usually show either the estimated minimal satellite-dish diameter required or the signal strength in each area measured in dBW.

In general, footprints are the geographical representation of satellite antenna-radiation pattern. It is the area on the earth's surface that the satellite can receive from or transmit to. The shape of a footprint depends on the path of the orbit, height and the type of antenna used. Figure 11.38 shows the footprint of a satellite.

11.17.4 Categories of Radiation Patterns

Radiation patterns from a satellite antenna are classified as follows.

1. Spot beam
2. Zonal beam

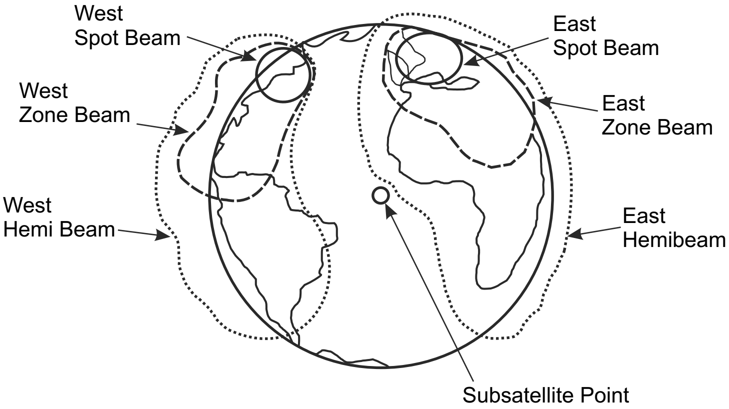


Fig. 11.38 Satellite footprint

3. Hemispherical beam
4. Earth (global) beam

1. Spot and Zonal Beams

Spot and zonal beams are the smallest beams. Spot beams concentrate their power to very small geographical areas. Therefore, they have higher Effective Isotropic Radiated Power (EIRP) than those targeting much larger areas. These beams are capable of covering less than 10% of the earth' surface. Higher the downlink frequency, the more easily a beam can be focused into a smaller spot pattern.

2. Hemispherical Beam

The hemispherical beams blanket 20% of the earth's surface and they have EIRPs 3 dB lower than those transmitted by spot beams.

3. Earth (Global) Beam

The radiation patterns of earth beams are capable of covering approximately 42% of the earth's surface, which is the maximum view of any one geosynchronous satellite. Power levels are also considerably lower with earth beams than with spot, zonal or hemispherical beams.

Summary

A satellite is a specialised wireless receiver/transmitter that is launched by a rocket and placed in orbit around the earth. They are used for such diverse purposes as weather forecasting, television broadcast, amateur radio communications, Internet communications and the Global Positioning System (GPS). The first artificial satellite, launched by Russia in the late 1950s, was used to transmit a simple Morse code signal over and over. Modern satellites can now receive and re-transmit thousands of signals simultaneously.

There are two main components of satellite communication. They are as follows.

- The satellite
- The ground station

There are three major components of satellite communication. They are as follows.

- A satellite capable of receiving signals beamed at it from the earth station, amplifying them and doing frequency translation before transmitting the same back towards the earth for intended users.
- Transmission path comprising both uplink from the earth station to the designated satellite as well as the downlink from satellite the to intended users.
- Earth station equipped suitably to transmit and receive signals to and from the designated satellite.

Among the above three components, 1 and 2 together are grouped under the heading of space segment, and component 3 is called the earth segment.

A satellite that is used only as a reflector of signals transmitted from an earth station is called a passive satellite as it does not carry any equipment for receiving and transmitting the signal. A satellite that carries equipment for receiving earth signals, processing them and retransmitting them towards the earth is called an active satellite. Such a satellite also has to carry electrical power. Satellites presently in use are active satellites.

Satellites can be placed in orbits around the earth at different altitudes. Depending on the heights above the earth' surface, orbits are classified as follows.

- Low-earth orbit
- Medium-earth orbit
- Geosynchronous orbit

Low-Earth Orbit (LEO) refers to satellites in orbit at less that 22,300 miles above the earth. To obtain a low-earth orbit, the speed of the satellite should be high. Medium-earth orbit (MEO) systems operate at about 10,000 kilometres above the earth, which is lower than the GEO orbit and higher than most LEO orbits. If the orbit is at a height of 35,860 km above the surface of earth, it is called geosynchronous orbit. This is in the equatorial plane.

The major subsystems of a satellite earth station are

- Power-supply subsystem
- Transmitting subsystem
- Receiving subsystem
- Antenna subsystem

A transmitting subsystem, also called an uplink system, includes the baseband processing stages, modulators, up-converters and high-power amplifiers. A receiving subsystem, also called downlink system, includes the baseband processing stages, demodulators, down-converters and low-noise amplifiers. The receiving antenna should pick up signals, both communications and tracking, coming from the satellite and from no other directions. Generally, parabolic horn and cassegrain antennas have been used.

A satellite transponder is equipment which receives the signal beamed at it from the transmitting earth station on the uplink, does the required input filtering, down-converts to a different frequency and amplifies it to the desired level before it is retransmitted to the intended receiving station.

Radiation pattern is graphical representation of the field magnitude at a fixed distance from an antenna as a function of direction, i.e. angular variation of the test antenna's radiation. It defines the variation of the power radiated by an antenna as a function of the direction away from the antenna. It is also referred as the function of azimuth/elevation angles.

REVIEW QUESTIONS

PART-A

1. Define a satellite.
2. Mention the applications of satellites.
3. What are the special features of satellite communication?
4. List the frequency ranges of satellite communication for various applications.
5. What are the advantages of satellite communication?
6. Name the basic elements of a satellite.
7. What is the purpose of a thrust system?
8. What is the role of transponders in satellite communication?
9. What are the earth segments?
10. Differentiate between passive and active satellites.
11. Define low-earth orbit.
12. Define medium-earth orbit.
13. Define geostationary orbit.
14. What is a geosynchronous orbit?
15. Write the expression for calculating the orbit velocity.
16. What are the types of satellite orbits?
17. What is an inclined elliptical orbit?
18. What is a circular polar orbit?
19. Draw the diagram of a typical communication satellite with various orbits.
20. List the advantages and disadvantages of geostationary orbit.
21. What do you mean by orbital adjustment?

22. Define Kepler's laws for satellite motion.
23. What are Newton's laws for satellite motion?
24. Define geocenter.
25. What is posigrade orbit?
26. What is retrograde orbit?
27. Define perigee and apogee of a satellite.
28. What is orbital eccentricity?
29. Give the expressions of perigee and apogee of a satellite.
30. Define orbital period.
31. Define orbital velocity.
32. How will you calculate the satellite height?
33. Define azimuth angle and elevation angle.
34. Give the expression for calculating the slant angle.
35. What are the various substations of a satellite earth-station subsystem?
36. What is the significance of radiation patterns?
37. Classify satellite footprints.

PART-B

1. What are the basic elements of a satellite system? Explain them in detail.
2. With a neat block diagram, explain the functioning of each block of a satellite communication system.
3. What are the types of satellite orbits based on their heights above the surface of the earth? Explain them in detail.
4. What are the types of satellite orbits based on satellite communication purposes? Explain them in detail.
5. Explain the various orbital parameters of satellite communication system in detail.
6. Describe and derive the expressions of various orbital parameters of satellite communication.
7. With a neat block diagram, describe the working of a satellite earth-station subsystem.
8. What are the various subsystems of a satellite earth station? Explain them in detail.
9. Draw the block diagrams of uplink model and downlink model of a satellite earth station and explain their functioning.
10. Draw the block diagram of a satellite transponder and explain the functioning of its various blocks.
11. A satellite is orbiting in a near-earth circular orbit at a distance of 640 km. Determine its orbital period. Assume the radius of the earth is 6360 km.

12. For an eccentric elliptical satellite orbit, the apogee and perigee points are at a distance of 50,000 km and 8000 km respectively from the centre of the earth. Determine the semi-major axis, semi-minor axis and orbit eccentricity.
13. A geostationary satellite moving in an equatorial circular orbit is at a height of 35,786 km from the earth's surface. If the earth's radius is taken as 6378 km, determine the maximum coverage angle, if the minimum possible elevation angle is 5°.
14. An earth station is located at 30°W longitude and 60°N latitude. Determine the earth station azimuth and elevation angles with respect to a geostationary satellite located at 50°W longitude.
15. Explain the satellite-launching process in detail.
16. Describe various satellite-radiation patterns in detail.

12

RADAR PRINCIPLES

Objectives

- To know about the principles and functions of a radar system
- To discuss the pulse radar system and various antennas used in a radar system in detail
- To discuss the different types of radars and radar displays
- To provide details about moving-target indicator, continuous-wave radar and frequency-modulated CW radar and tracking radar

12.1 INTRODUCTION

Radar is an electromagnetic system for the detection and location of objects. It operates by transmitting a particular type of waveform, a pulse-modulated sine wave for example, and detects the nature of the echo signal. Radar is used to extend the capability of one's senses for observing the environment, especially the sense of vision. The value of radar lies not in being a substitute for the eye, but in doing what the eye cannot do.

Radar can be designed to see through those conditions which are capable of affecting normal human vision, such as darkness, haze, fog, rain and snow. In addition, radar has the advantage of being able to measure the distance or range of the object.

Radar uses include meteorological detection of precipitation, measuring ocean-surface waves, air-traffic control and civilian police detectors. Radar's importance in military applications remains and it is the primary asset in active electromagnetic emissions in battlespace.

12.2 WHAT IS RADAR?

RADAR is an acronym for Radio Detection and Ranging. It is a stand-alone active system having its own transmitter and receiver that is used for detecting the presence and finding the exact location of a far-off target. It does so by transmitting electromagnetic energy in the form of short bursts in most cases, and then detecting the echo signal returned by the target. The ranging is computed from the time that elapses between the transmission of energy and reception of echo. It is shown in Figure 12.1. The location of the target can be determined from the angle or direction of arrival of the echo signal by using a scanning antenna preferably transmitting a very narrow beamwidth.

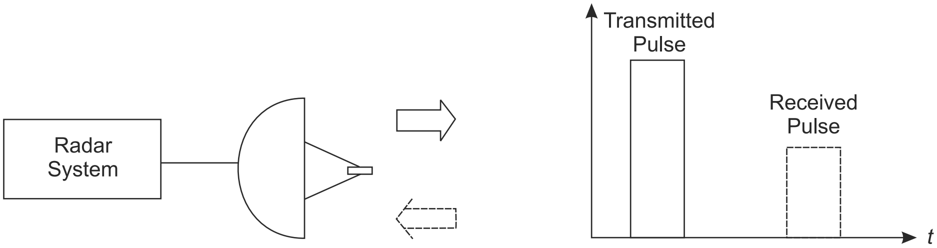


Fig. 12.1 A radar

The radar can be used to determine the velocity of a moving target, track the target and even determine some of the physical features of the target. It is also a principal source of navigational aid to aircraft and ships. It forms a vital part of an overall weapon guidance or a fire-control system. Most of the radar functions lies in its capability to detect a target, find its range and determine its velocity.

12.3 PRINCIPLES OF RADAR MEASUREMENT

Radar measurement of range, or distance, is made possible because of the properties of radiated electromagnetic energy.

1. The electromagnetic waves are reflected if they meet an electrically leading surface. If these reflected waves are received again at the place of their origin, that means an obstacle is in the propagation direction
2. Electromagnetic energy travels through air at a constant speed, at approximately the speed of light,
   - 300,000 kilometres per second, or
   - 186,000 statute miles per second, or
   - 162,000 nautical miles per second.

This constant speed allows the determination of the distance between the reflecting objects (airplanes, ships or cars) and the radar site by measuring the running time of the transmitted pulses.

3. This energy normally travels through space in a straight line, and will vary only slightly because of atmospheric and weather conditions. By using special radar antennas, this energy can be focused into a desired direction. Thus, the direction of the reflecting objects can be measured.

These principles can basically be implemented in a radar system, and allow the determination of the distance, the direction and the height of the reflecting object.

12.4 BASIC RADAR SYSTEM

The basic components of a radar system are shown in Figure 12.2. The radar antenna illuminates the target with a microwave signal, which is then reflected and picked up by a receiving device. The electrical signal picked up by the receiving antenna is called **echo** or **return**. The radar signal is generated by a powerful transmitter and received by a highly sensitive receiver.

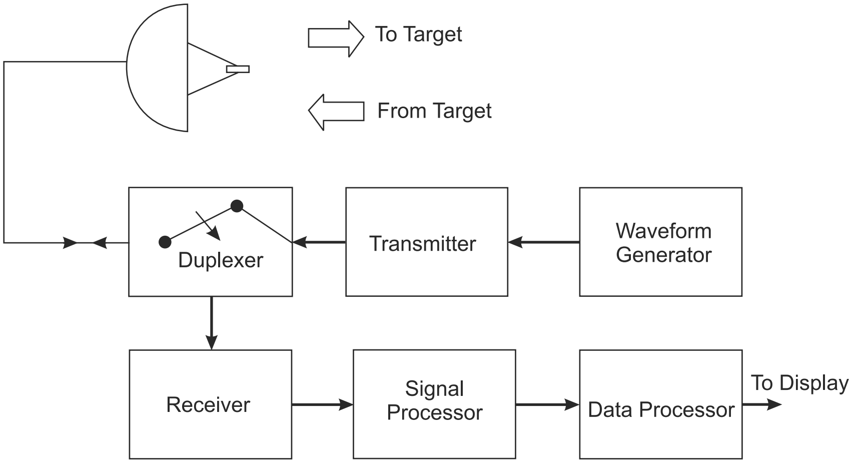


Fig. 12.2 Basic radar system

The radar signal waveform, as generated by the waveform generator, modulates a high-frequency carrier and the modulated signal is raised to the desired power level in the transmitter portion. The transmitter could be a power amplifier employing any of the microwave tube amplifiers such as Klystron, Travelling Wave Tube (TWT), Crossed Field Amplifier (CFA) or even a solid-state device. The radar waveform is generated at a low power level which makes it far easier to generate different types of waveforms required for different radars. The most common radar waveform is a repetitive train of short pulses. CW is employed to determine the radial velocity of the moving target from Doppler frequency shift.

12.4.1 Signal Routing

The radar transmitter produces short-duration high-power RF pulses of energy. The duplexer alternately switches the antenna between the transmitter and receiver so that only one antenna need be used. This switching is necessary because the high-power pulses of the transmitter would destroy the receiver if energy were allowed to enter the receiver. The antenna transfers the transmitter energy to signals in space with the required distribution and efficiency. This process is applied in an identical way on reception.

The transmitted pulses are radiated into space by the antenna as an electromagnetic wave. This wave travels in a straight line with a constant velocity and will be reflected by an aim. The antenna receives the back-scattered echo signals. During reception, the duplexer leads the weak echo signals to the receiver. The hypersensitive receiver amplifies and demodulates the received RF signals. The receiver provides video signals on the output.

The indicator should present to the observer a continuous, easily understandable, graphic picture of the relative position of radar targets. All targets produce a diffuse reflection, i.e. it is reflected in a wide number of directions. The reflected signal is also called **scattering**. **Backscatter** is the term given to reflections in the opposite direction to the incident rays.

Radar signals can be displayed on the traditional Plan Position Indicator (PPI) or other more advanced radar display systems. A PPI has a rotating vector with the radar at the origin, which indicates the pointing direction of the antenna and hence the bearing of targets. It shows a maplike picture of the area covered by the radar beam.

12.4.2 Radar Signal Timing

Generally, most functions of a radar set are time-dependent. Time synchronisation between the transmitter and receiver of a radar set is required for range measurement. Radar systems radiate each pulse during transmit time (or pulse width τ), wait for returning echoes during listening or rest time, and then radiate the next pulse, as shown in Figure 12.3.

A synchroniser coordinates the timing for range determination and supplies the synchronising signals for the radar. It sends signals simultaneously to the transmitter, which sends a new pulse, and to the indicator, and other associated circuits. The time between the beginning of one pulse and the start of the next pulse is called Pulse Repetition Time (PRT) and is equal to the reciprocal of PRF.

$$\mathrm{PRT} = \frac{1}{\mathrm{PRF}} \tag{12.1}$$

The Pulse Repetition Frequency (PRF) of the radar system is the number of pulses transmitted per second. The frequency of pulse transmission affects the maximum range that can be displayed.

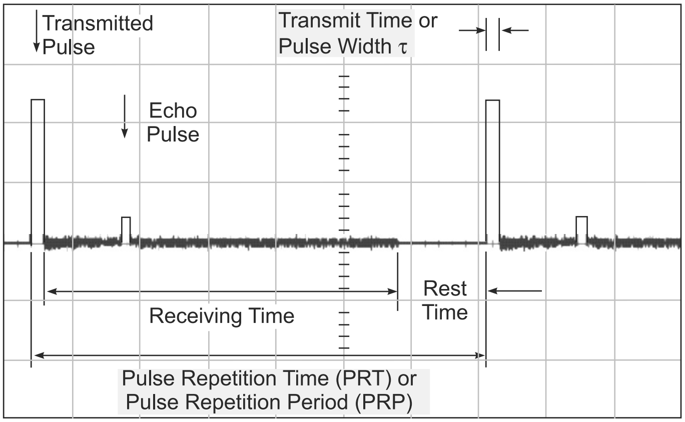

Transmitted Pulse
Transmit Time or Pulse Width τ
Echo Pulse
Receiving Time
Rest Time
Pulse Repetition Time (PRT) or Pulse Repetition Period (PRP)

Fig. 12.3 A typical radar timeline

12.4.3 Ranging

The distance of the aim is determined from the running time of the high-frequency transmitted signal and the propagation c_0. The actual range of a target from the radar is known as **slant range**. Slant range is the line-of-sight distance between the radar and the illuminated object. **Ground range** is the horizontal distance between the emitter and its target and its calculation requires knowledge of the target's elevation. Since the waves travel to a target and back, the round trip time is dividing by two in order to obtain the time the wave took to reach the target. Therefore, the following formula arises for the slant range:

$$R = \frac{c_0 \times t}{2} \tag{12.2}$$

where c_0 is the speed of light = 3×10^8 m/s,

t is the measured running time (s), and

R is the slant range of antenna (m).

Range is the distance from the radar site to the target measured along the line of sight.

$$\nu = \frac{s}{t} \tag{12.3}$$

$$c_0 = \frac{2R}{t} \tag{12.4}$$

where c_0 is the speed of light = 3×10^8 m/s, at which all electromagnetic waves propagate.

If the respective running time t is known then the distance R between a target and the radar set can be calculated by using this equation.

The maximum unambiguous range for a given radar system can be determined by using the formula:

$$R_{\text{unamb}} = \frac{(\text{PRT} - \tau).c_0}{2} \tag{12.5}$$

While determining the maximum range, the pulse repetition time of the radar is important because target return times that exceed the PRT of the radar system appear at incorrect ranges on the radar screen. Returns that appear at these incorrect ranges are referred as **ambiguous returns** or **second-time around echoes**. The pulse width τ in this equation indicates that the complete echo impulse must be received.

The maximum measuring distance R_{max} of a radar unit is not orientated only at the value determined in the radar equation but also on the duration of the receiving time. The maximum range at which a target can be located so as to guarantee that the leading edge of the received backscatter from that target is received before transmission begins for the next pulse. This range is called **maximum unambiguous range**.

The Pulse Repetition Frequency (PRF) determines this maximum unambiguous range of the given radar before ambiguities start to occur. This range can be determined by using the following equations:

$$R_{\text{max}} = \frac{c_0\,.(\text{PRT} - P_{\text{W}})}{2} \tag{12.6}$$

$$R_{\text{max}} \approx \frac{(\text{PRT} - P_{\text{W}})\ \text{in}\ \mu\text{s}}{6.66\,\text{in}\ \mu\text{s}}\ \text{in km} \tag{12.7}$$

where c_0 is the speed of light = 3×10^8 m/s.

The pulse width (P_{W}) in these equations indicates that the complete echo impulse must be received. If the transmitted pulse is very short (1 ms), it will be ignored. But some radars use very long pulses (up to 800 microseconds) and the backscattered signal must be compressed in the receiver.

When the leading edge of the echo pulse falls inside the transmitting pulse, it is impossible to determine the **round-trip time**, which means that the distance cannot be measured. The minimum detectable range R_{min} depends on the transmitters pulse with τ, and the recovery time t_{recovery} of the duplexer.

$$R_{\text{min}} = \frac{(\tau + t_{\text{recovery}}).c_0}{2} \tag{12.8}$$

EXAMPLE 12.1

A pulsed radar with a PRF of 1 kHz receives an echo pulse exactly 0.15 ms after it transmits. What should be the target range in km? Also determine the maximum unambiguous range of the radar.

Solution

$$\text{Range } R = \frac{c_0 \times t}{2}$$

$$= \frac{3\times10^{8}\times0.15\times10^{-3}}{2}$$

$$= 0.225\times10^{5} = 22.5 \text{ km}$$

Maximum unambiguous range will be

$$R_{\text{unamb}} = \frac{c_0}{2}\times \text{PRF}$$

$$= \frac{3\times10^{8}\times1000}{2} = 150\,\text{km}$$

12.4.4 Determination of the Direction

The angular determination of the target is determined by the directivity of the antenna. **Directivity**, sometimes known as the **directive gain**, is the ability of the antenna to concentrate the transmitted energy in a particular direction. An antenna with high directivity is also called a **directive antenna**. By measuring the direction in which the antenna is pointing when the echo is received, both the azimuth and elevation angles from the radar to the object or target can be determined. The accuracy of angular measurement is determined by the directivity, which is a function of the size of the antenna.

The **true bearing** (referenced to true north) of a radar target is the angle between true north and a line pointed directly at the target. This angle is measured in the horizontal plane and in a clockwise direction from true north, as shown in Figure 12.4.

In order to have an exact determination of the bearing angle, a survey of the north direction is necessary. Nowadays, modern radar sets use GPS satellites to determine the north direction independently.

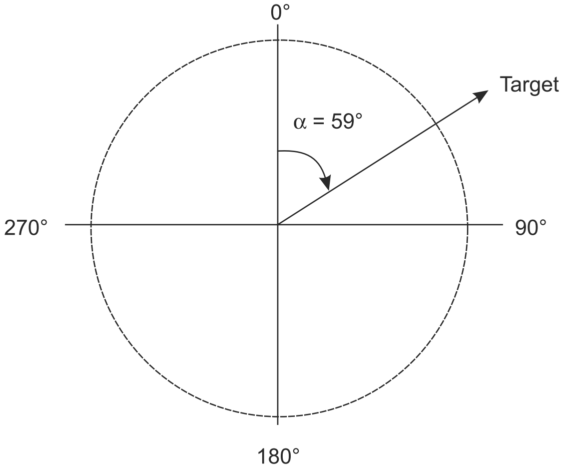


Fig. 12.4 Measurement of angle

12.4.5 Frequencies and Powers used in Radars

Table 12.1 shows various frequencies and powers used in radars. Most radar equipments operate in the frequency range of 100 to 250,000 MHz with concentrations in a few bands such as L, S, X, etc. Long-range radars operate in S and L bands.

Table 12.1 Frequencies and powers used in radars

Name of the Band	Frequency range in GHz	Maximum allowable power (mW)
UHF	0.3–1.0	5.0
L	1.0–1.5	30.0
S	1.5–3.9	25.0
C	3.9–8.0	15.0
X	8.0–12.5	12.0
Ku	12.5–18.0	2.0
K	18.0–26.5	0.6

1. UHF-Band Radar

There are some specialised radar sets developed for this frequency band (300 MHz to 1 GHz). It is a good frequency for the operation of radars for the detection and tracking of satellites and ballistic missiles over a long range. These radars operate for early warning and target acquisition like the surveillance radar for the Medium Extended Air Defense System.

2. L-Band Radar

This frequency band (1 to 1.5 GHz) is preferred for the operation of long-range air-surveillance radars out to 250 NM (≈400 km). They transmit pulses with high power, broad bandwidth and often an intrapulse modulation often. Due to the curvature of the earth, the achievable maximum range is limited for targets flying with low altitude. These objects disappear very fast behind the horizon.

3. S-Band Radar

The atmospheric attenuation is higher than in the L-Band. Radar sets need a considerably higher transmitting power than in lower frequency ranges to achieve a good maximum range. In this frequency range, the influence of weather conditions is higher than in the L-band. Therefore, a couple of weather radars work in the S-Band, but more in subtropic and tropic climatic conditions, because here the radar can see beyond a severe storm.

4. C-Band Radar

In C-Band, there are many mobile military battlefield surveillance, missile control and ground surveillance radar sets with short or medium range. The size of the antennas provides excellent accuracy and resolution, but the relatively small-sized antennas don't bother about a fast relocation. The influence of bad weather conditions is very high. Therefore, air-surveillance radars often use an antenna feed with circular polarisation. This frequency band is predetermined for most types of weather radar used to locate precipitation in temperate zones.

5. X- and Ku-Band Radars

In this frequency band (8 and 18 GHz), the relationship between used wavelength and size of the antenna is considerably better than in lower frequency bands. The X-Band is a relatively popular radar band for military applications like airborne radars for performing the roles of interceptor, fighter, attack of enemy fighters and of ground targets. A very small antenna size provides a good performance. Missile guidance systems at the X-Band are of a convenient size and are, therefore, of interest for applications where mobility and light weight are important and very long range is not a major requirement.

This frequency band is widely used for maritime civil and military navigation radars. Very small and cheap antennas with high rotation speed are adequate for a fair maximum range and good accuracy. Slotted waveguides and small patch antennas are used as radar antennas, mostly under a protective radom.

This frequency band is also popular for space-borne or airborne imaging radars based on Synthetic Aperture Radar (SAR), both for military electronic intelligence and civil geographic mapping. A special Inverse Synthetic Aperture Radar (ISAR) is in use as a maritime airborne instrument for pollution control.

6. K-Band Radar

The higher the frequency, the higher is the atmospheric attenuation. Otherwise, the achievable accuracy and the range resolution too rise. Radar applications in this frequency band provide short range, very high resolution and high data-renewing rate.

12.4.6 Radar-Range Equation

The radar equation relates the range of radar to the characteristics of the transmitter, receiver, antenna, target and environment. It is useful not just as a means for determining the maximum distance from the radar to the target, but it can serve both as a tool for understanding radar operation and as a basis for radar design.

If the power of the radar transmitter is denoted by P_T and if an isotropic antenna is used, which radiates uniformly in all directions, the power density S_t at a distance d from the radar is equal to the transmitter power divided by the surface area $4\pi\, d^2$ of an imaginary sphere of radius d, i.e.

Power density from isotropic antenna (S_t) = $\dfrac{P_T}{4\pi d^2}$ (12.9)

Radars employ directive antennas to channel, or direct, the radiated power P_T into some particular direction. The gain G of an antenna is a measure of the increased power radiated in the direction of the target as compared with the power that would have been radiated from an isotropic antenna. It may be defined as the ratio of the maximum radiation intensity from the subject antenna to the radiation intensity from a lossless, isotropic antenna with the same power input. The power density at the target from an antenna with a transmitting gain G is

Power density from directive antenna = $\dfrac{P_T G}{4\pi d^2}$ (12.10)

The target intercepts a portion of the incident power and reradiates it in various directions. The measure of the amount of incident power intercepted by the target and reradiated back in the direction of the radar is denoted as the radar cross section σ, and is defined by the relation

Power density of echo signal at radar = $\dfrac{P_T G}{4\pi d^2}\dfrac{\sigma}{4\pi d^2}$ (12.11)

The radar cross section σ has units of area. It is a characteristic of the particular target and is a measure of its size as seen by the radar. The radar antenna captures a portion of the echo power. If the effective area of the receiving antenna is denoted A_e, the power P_R, received by the radar, is

$$P_R = \frac{P_T G}{4\pi d^2}\frac{\sigma}{4\pi d^2}.A_e$$

$$= \frac{P_T G.\sigma.A_e}{(4\pi)^2\, d^4} \qquad (12.12)$$

The maximum radar range R_{max} is the distance beyond which the target cannot be detected. It occurs when the received echo signal power P_R, just equals the minimum detectable signal S_{min}. Therefore,

$$R_{max} = \left[\frac{P_T G.\sigma.A_e}{(4\pi)^2\, S_{min}}\right]^{1/4} \qquad (12.13)$$

This is the fundamental form of the radar equation. Note that the important antenna parameters are the transmitting gain and the receiving effective area.

The relationship between the transmitting gain and the receiving effective area of an antenna is

$$G = \frac{4\pi A_e}{\lambda^2} \tag{12.14}$$

from which A_e is derived as

$$A_e = \frac{G\lambda^2}{4\pi} \tag{12.15}$$

By substituting Equation (12.15) into Equation (12.13),

$$R_{max} = \left[\frac{P_T G^2 \, .\lambda^2}{(4\pi)^3 \, S_{min}}\right]^{1/4} \tag{12.16}$$

Next, by substituting Equation (12.14) into Equation (12.16),

$$R_{max} = \left[\frac{P_T A_e^2 \, .\sigma}{4\pi\lambda^2 \, S_{min}}\right]^{1/4} \tag{12.17}$$

These simplified versions of the radar equation do not adequately describe the performance of practical radar. Many important factors that affect range are not explicitly included. In practice, the observed maximum radar ranges are usually much smaller than what would be predicted by the above equations, sometimes by as much as a factor of two.

12.5 CLASSIFICATION OF RADAR SYSTEMS

12.5.1 Classification According to Specific Functions

Depending on the desired information, radar sets must have different qualities and technologies. Figure 12.5 shows the classification of radar systems according to specific functions, different qualities and techniques.

1. Imaging Radar

Imaging radar sensors measure two dimensions of coordinates of a picture of the area covered by the radar beam. An imaging radar forms a picture of the observed object or area. Imaging radars have been used to map the earth, other planets, asteroids, other celestial objects and to categorise targets for military systems.

2. Non-Imaging Radar

Non-imaging sensors take measurements in one linear dimension, as opposed to the two-dimensional representation of imaging sensors. Typically, implementations of a non-imaging radar system are speed gauges and radar altimeters. These are also called **scatter meters** since

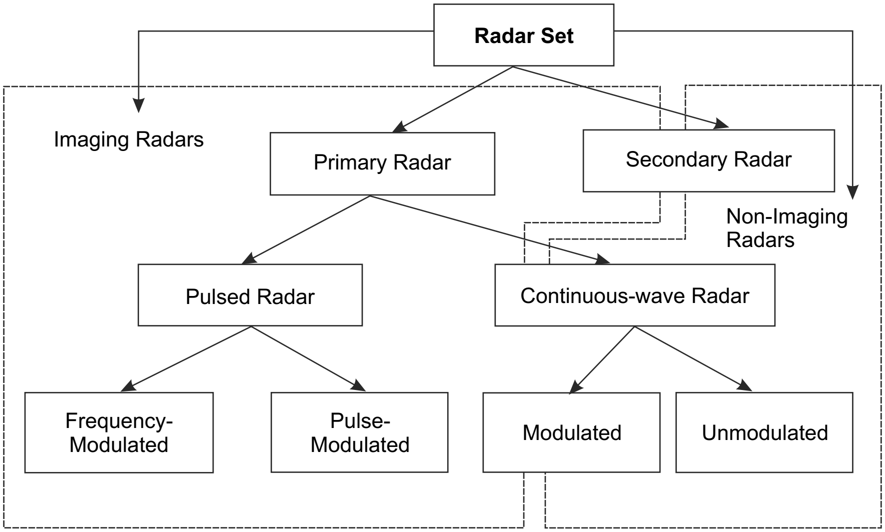


Fig. 12.5 Radar systems classified according to specific function

they measure the scattering properties of the object or region being observed. Non-imaging secondary radar applications are immobiliser systems in some recent private cars.

3. Primary Radar

A primary radar transmits high-frequency signals which are reflected at targets. The arisen echoes are received and evaluated. This means, unlike secondary radar sets, a primary radar unit receives its own emitted signals as an echo again. Primary radar sets are fitted with an additional interrogator as secondary radar mostly, to combine the advantages of both systems.

4. Secondary Radar

At secondary radar, a transponder, **trans**mitting res**ponder,** is placed on board and this transponder responds to interrogation by transmitting a coded reply signal. This response can contain much more information than a primary radar unit is able to acquire.

5. Continuous-Wave Radar

CW radar sets transmit a high-frequency signal continuously. The echo signal is received and processed permanently too. The transmitted signal of these equipment is constant in amplitude and frequency. These equipment are specialised in speed measuring, e.g. these equipment are used as speed gauges of the police. One has to resolve two problems with this principle. They are

- To prevent a direct connection of the transmitted energy into the receiver
- To assign the received echoes to a time system to be able to do runtime measurements

A direct connection of the transmitted energy into the receiver can be prevented by means of the following.

- Spatial separation of the transmitting antenna and the receiving antenna, e.g. the aim is illuminated by a strong transmitter and the receiver is located in the missile flying direction towards the aim.
- Frequency-dependent separation by the Doppler frequency during the measurement of speeds.

6. Pulse Radar

Pulse radar sets transmit a high-frequency impulse signal of high power. After this impulse signal, a longer break follows in which the echoes can be received, before a new transmitted signal is sent out. Direction, distance and sometimes, if necessary, the height or altitude of the target can be determined from the measured antenna position and propagation time of the pulse signal. In order to get a good range resolution and a good maximum range, these radar sets transmit a very short pulse with an extremely high pulse power.

Pulse radar transmits a relatively weak pulse with a longer pulse width. It modulates the transmitting signal to obtain a distance resolution within the transmitting pulse.

12.5.2 Classification According to Usage

Radar systems may also be divided into types based on the designed use. Figure 12.6 shows the classification of radar systems according to its use in various fields.

Air-defense radars can detect air targets and determine their position, course and speed in a relatively large area. The maximum range of air-defense radar can exceed 300 miles, and the bearing coverage is a complete 360-degree circle. Air Traffic Control (ATC) surveillance radars are commonly used in Air Traffic Management (ATM).

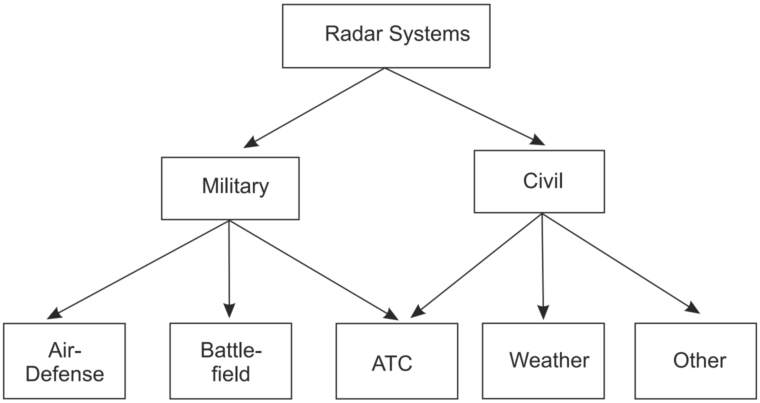


Fig. 12.6 Radar systems classified according to use

12.6 BASIC PULSE RADAR SYSTEM

Pulse radar sets transmit a high-frequency impulse signal of high power. After this impulse signal, a longer break follows in which the echoes can be received, before a new transmitted signal is sent out. Direction, distance and sometimes, if necessary, the height or altitude of the target can be determined from the measured antenna position and propagation time of the pulse signal. These classical radar sets transmit a very short pulse to get a good range resolution with an extremely high pulse power to get a good maximum range. Figure 12.7 shows the block diagram of typical pulsed radar of high power.

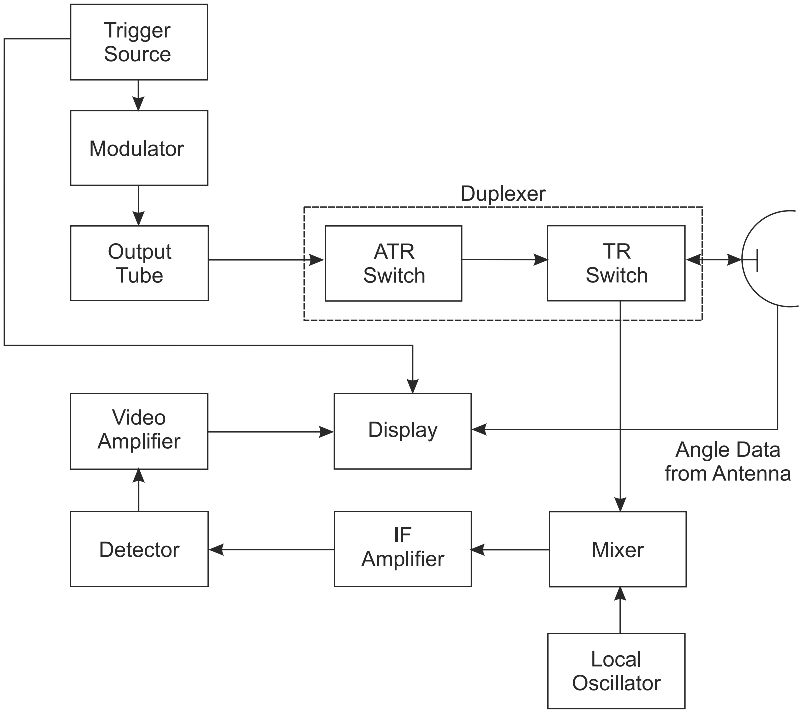


Fig. 12.7 Block diagram of a high-power pulsed radar

The triggering circuit provides trigger pulses to the modulator. The modulator stage switches ON/OFF the supply voltage to the output tube. Thus, the output tube is switched ON/OFF and this switching is carried out by the trigger source through a modulator stage. The output tube is usually a magnetron oscillator. Sometimes, it may be multicavity klystron or a travelling-wave tube in which case, a microwave source has to be used to supply input signal drive to these amplifier tubes. The high-power pulse is given to the duplexer.

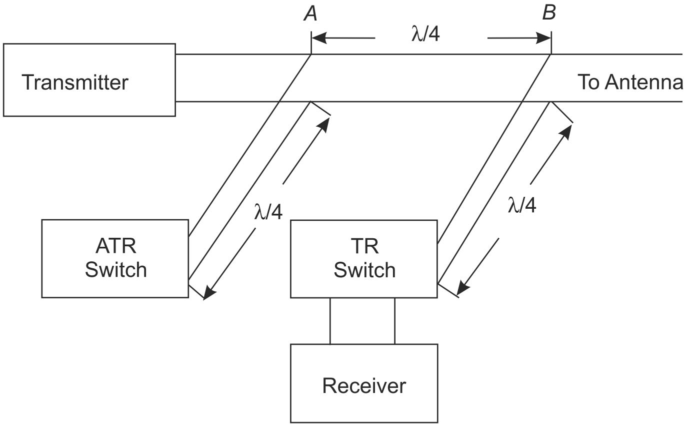


Fig. 12.8 Duplexer in pulsed radar

A duplexer switches ON/OFF the antenna between the receiver and the transmitter. It has two switches, the TR and ATR, as shown in Figure 12.8. These switches are so arranged that the transmitter and receiver are alternately connected to the antenna but they are never connected to each other. These switches are basically gas tubes which act as a short circuit when a voltage appears across them and remain open circuit in the absence of voltage.

With the transmitter output pulse, both these switches become short circuited. Thus, the two λ/4 waveguide sections joining these switches to the main waveguide have short-circuited terminations and offer infinite impedance at their junctions to the main waveguide. Therefore, the transmitter energy travels freely to the antenna and the receiver input terminals are short circuited. When the transmitted pulse ends, the TR and ATR switches become open. With the ATR switch offering open circuit load, the input impedance of the λ/4 section becomes zero and thus, the waveguide is shorted at the point *A*. Impedance of the waveguide at point *B* is infinite and, therefore, no energy can travel from point *B* towards *A*. With the TR switch acting as open circuit, the receiver is now connected through a λ/4 section to the radar antenna.

The received echo pulse is given to the mixer. A diode mixer is commonly used for this purpose, since it has a low noise figure. The mixer heterodynes the local oscillator output and the signal input and produces an intermediate frequency lying between 30 to 60 MHz at its output. The IF signal is amplified by a high-gain low-noise IF amplifier. Usually, transistor amplifier stages with cascade connections are used for this purpose. The IF stage is a broadband amplifier, which is used for the amplification of fairly narrow pulses. To meet large bandwidth requirements, stagger tuning may be used.

A Schottky barrier diode is often used as a detector and its output is amplified by the video amplifier. This amplifier has the same bandwidth as the IF amplifier. The output of the video amplifier is fed to a display unit. CRT is commonly used for this purpose.

The output of a radar receiver may be displayed by any one of the three ways. They are

1. Deflection modulation of a CRT screen or A-scope

2. Intensity modulation of a CRT or Plan Position Indicator (PPI)
3. Feeding the signal to a computer

EXAMPLE 12.2

A pulsed radar is transmitting 1 μs pulses at a repetition rate of 1 kHz. Assume that no pulse compression technique has been used while processing the echo pulses. Determine whether two targets separated in range by 500 m but having same angular position can be resolved by this radar on the basis of range.

Solution

$$\text{Range resolution} = \frac{c_0 \times \tau}{2}$$

$$= \frac{3 \times 10^8 \times 10^{-6}}{2} = 150 \text{ m}$$

This means that radar can resolve up to an intertarget separation in range of 150 m. Therefore, the given radar will be able to resolve the targets.

EXAMPLE 12.3

Determine the centre of the frequency spectrum, interline spacing of the spectrum and the matched bandwidth of a pulse that is 10 μs wide and has an RF signal frequency of 10 GHz. Also determine the frequencies of the spectral lines closest to the centre if the PRF is 1 kHz.

Solution

$$\text{Pulse width} = 10\ \mu\text{s}$$

$$\text{Matched bandwidth} = \frac{1}{10 \times 10^{-6}} = 100 \text{ kHz}$$

$$\text{Centre frequency} = 10 \text{ GHz}$$

$$\text{Modulating frequency} = 1 \text{ kHz}$$

∴ the two closest frequencies to the centre of the spectrum are 1, 00, 00,001 kHz and 9, 99, 99,999 kHz.

EXAMPLE 12.4

The transmitted pulse has a 4 μs wide envelope. The frequency across this width is swept linearly from 495 to 505 MHz. Determine the centre of spectrum, matched bandwidth and the compressed pulse width.

Solution

$$\text{Centre of spectrum} = \frac{495+505}{2} = 500 \text{ MHz}$$

$$\text{Bandwidth} = 505 - 495 = 10 \text{ MHz}$$

$$\text{Pulse width} = \frac{1}{10} = 0.1\ \mu\text{s}$$

12.7 ANTENNAS USED IN RADAR SYSTEM

The antenna is one of the most important parts of a radar system. The essential functions performed by a radar system are listed as follows.

1. It transfers the transmitter energy to signals in space with the required distribution and efficiency. This process is applied in an identical way during reception.
2. It ensures that the signal has the required pattern in space. Generally, this has to be sufficiently narrow in azimuth to provide the required azimuth resolution.
3. It has to provide the required frequency of target position updates. In the case of a mechanically scanned antenna, this equates to the revolution rate. A high revolution rate can be a significant mechanical problem given that a radar antenna in certain frequency bands can have a reflector with immense dimensions and can weigh several tons.
4. It must measure the pointing direction with a high degree of accuracy.

The antenna structure must maintain the operating characteristics under all environmental conditions. The basic performance of a radar can be shown to be proportional to the product of the antenna area or aperture and the mean transmitted power. There are two types of antenna often used in a radar system. They are

1. The parabolic dish antenna
2. The phased-array antenna

12.7.1 Parabolic Dish Antenna

The parabolic dish antenna is the most frequently used antenna type in radar systems. Figure 12.9 shows a parabolic dish antenna used for the purpose of weather forecasting.

A dish antenna consists of one circular parabolic reflector and a point source situated in the focal point of this reflector. This point source is called **primary feed** or simply **feed**. The circular parabolic reflector is constructed of metal, usually a frame covered by

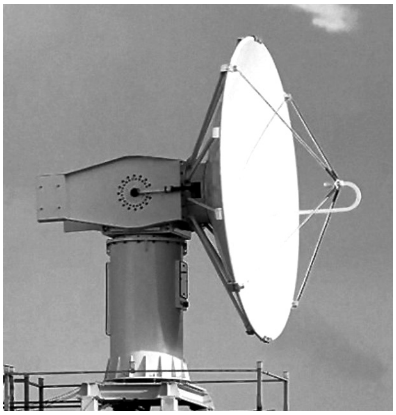


Fig. 12.9 A parabolic dish antenna for weather forecasting

metal mesh at the inner side. The width of the slots of the metal mesh has to be less than λ/12. This metal covering forms the reflector, acting as a mirror for the radar energy.

According to the laws of optics and analytical geometry, for this type of reflector, all reflected rays will be parallel to the axis of the parabolic reflector which gives us ideally one single reflected ray parallel to the main axis with no side lobes. The field leaves this feedhorn with a spherical wavefront. As each part of the wavefront reaches the reflecting surface, it is shifted 180° in phase and sent outward at angles that cause all parts of the field to travel in parallel paths. This is an idealised radar antenna and produces a pencil beam. If the reflector has an elliptical shape then it will produce a fan beam. Figure 12.10 illustrates the principle of a parabolic reflector in an ideal case.

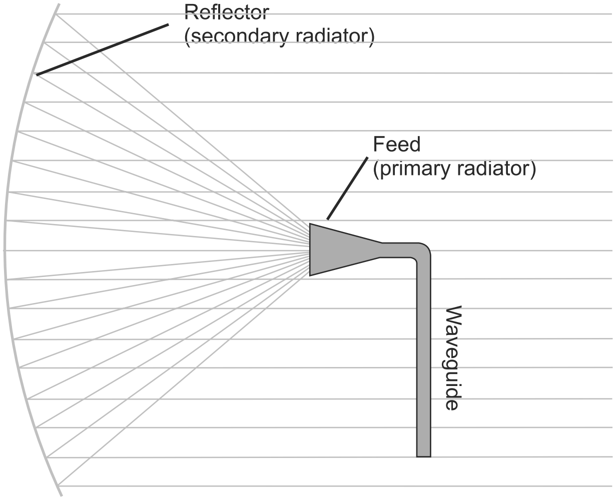


Fig. 12.10 Principle of a parabolic reflector

The real parabolic-antenna pattern has a conical form because of irregularities in the production. This main lobe may vary in angular width from one or two degrees in some radar sets to 15° to 20° in other radars. The radiation pattern of a parabolic antenna contains a major lobe, which is directed along the axis of propagation, and several small minor lobes.

The gain *G* of an antenna with a parabolic reflector can be determined as follows:

$$G = \frac{160^2}{\theta_{AZ} \,.\, \theta_{EL}} \tag{12.18}$$

where θ_{AZ} is the beamwidth in azimuth angle, and

θ_{EL} is the beamwidth in elevation angle.

12.7.2 Phased-Array Antenna

A phased-array antenna is composed of lots of radiating elements, each with a phase shifter. Beams are formed by shifting the phase of the signal emitted from each radiating element, to provide constructive/destructive interference so as to steer the beams in the desired direction. Figure 12.11 illustrates electronic beam deflection in the case of phased-array antenna.

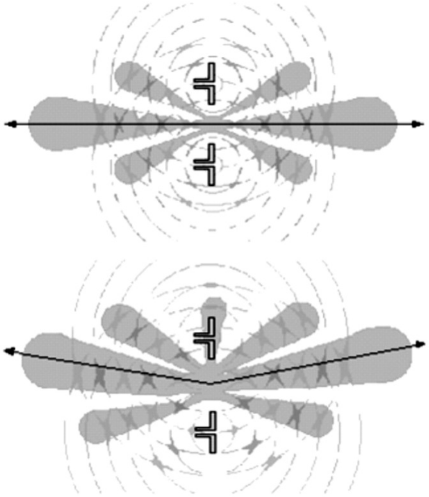


Fig. 12.11 Electronic beam deflection

In Figure 12.11, by considering the upper case, both radiating elements are fed with the same phase. The signal is amplified by constructive interference in the main direction. The beam sharpness is improved by the destructive interference. By considering the lower case, the signal is emitted by the lower radiating element with a phase shift of 10° earlier than of the upper radiating element. Because of this, the main direction of the emitted sum signal is moved upwards.

The main beam always points in the direction of the increasing phase shift. If the signal to be radiated is delivered through an electronic phase shifter giving a continuous phase shift, the beam direction will be adjustable. However, this cannot be extended more. The highest value, which can be achieved for a phased-array antenna is 120° (60° left and 60° right).

12.7.3 Antenna Scanning Pattern

Most radiators radiate stronger radiation in one direction than in another. A radiator such as this is referred to as **anisotropic**. However, a standard method allows the positions around a source to be marked so that one radiation pattern can easily be compared with another. The energy radiated from an antenna forms a field having a definite radiation pattern. A radiation pattern is a way of plotting the radiated energy from an antenna. This energy is measured at various angles at a constant distance from the antenna. The shape of this pattern depends on the type of antenna used.

To plot this pattern, two different types of graphs, rectangular-and polar-coordinate graphs are used. The **polar-coordinated graph** has proved to be of great use in studying radiation patterns. In the polar-coordinate graph, points are located by projection along a rotating axis to an intersection with one of several concentric, equally spaced circles. The polar-coordinate graph of the measured radiation is shown in Figure 12.12.

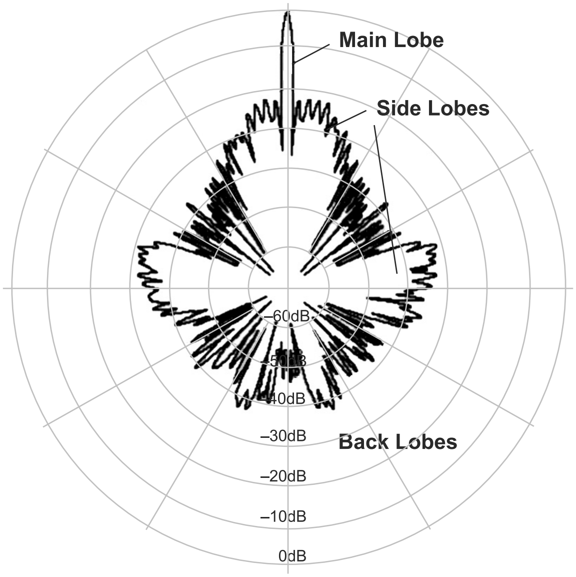


Fig. 12.12 Antenna pattern in a polar-coordinate graph

From Figure 12.12, we can observe the following:

- The **main lobe** is the region around the direction of maximum radiation (usually the region that is within 3 dB of the peak of the main beam). The main lobe in the figure is northbound.
- The **side lobes** are smaller beams that are away from the main beam. These side lobes are usually radiation in undesired directions which can never be completely eliminated. The **side-lobe level** is an important parameter used to characterise radiation patterns.
- One side lobe is called **back lobe**. This is the portion of radiation pattern that is directed opposing the main beam direction.

The graph in Figure 12.13 shows the rectangular-coordinated graph for the same source.

In the rectangular-coordinate graph, points are located by projection from a pair of stationary, perpendicular axes. The horizontal axis on the rectangular coordinate graph corresponds to the circles on the polar-coordinate graph. The vertical axis on the rectangular-coordinate graph corresponds to the rotating axis (radius) on the polar-coordinate graph.

The measurement scales in the graphs can have linear as well as logarithmic steps. From a plotted antenna pattern, some important characteristics of an antenna are to be measured. They are the following:

The **front-to-back ratio**, the ratio of power gain between the front and rear of a directional antenna. The value of the side lobe in Figure 12.13 is 180°.

The **side lobe ratio**, the maximum value of the side lobes away from the main beam. In Figure 12.13, the value of the side lobe is in + 6 degrees.

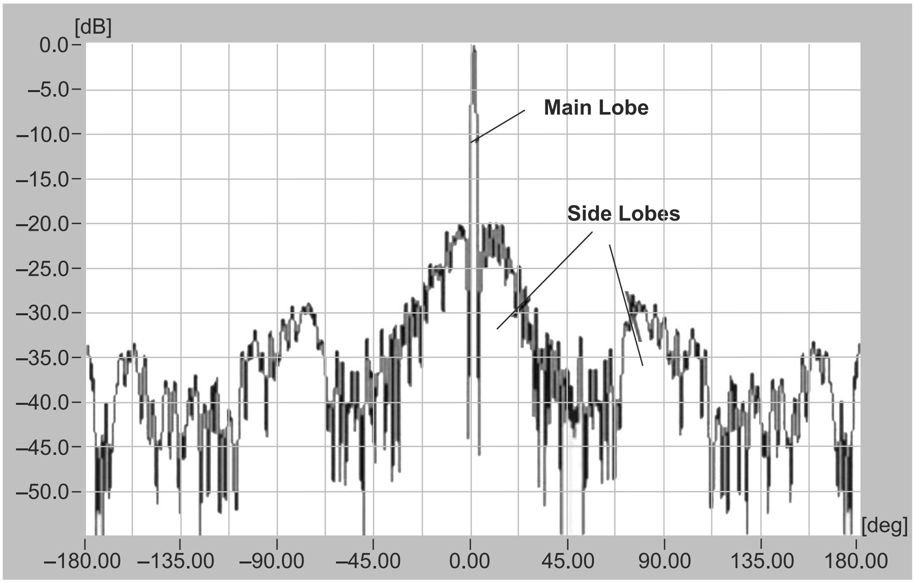


Fig. 12.13 Antenna pattern in a rectangular coordinate graph

12.7.4 Analysis of an Antenna Pattern

For the analysis of an antenna pattern, the following terms are to be considered.

1. Beamwidth

The angular range of the antenna pattern in which at least half of the maximum power still emitted is described as 'beamwidth'. Bordering points of this major lobe are, therefore, the points at which the field strength has fallen in the room around 3 dB regarding the maximum field strength. This angle is then described as beam width or aperture angle or half power (–3 dB) angle represented by θ or φ. The beamwidth θ is exactly the angle between the two blackmarked power levels in Figure 12.14.

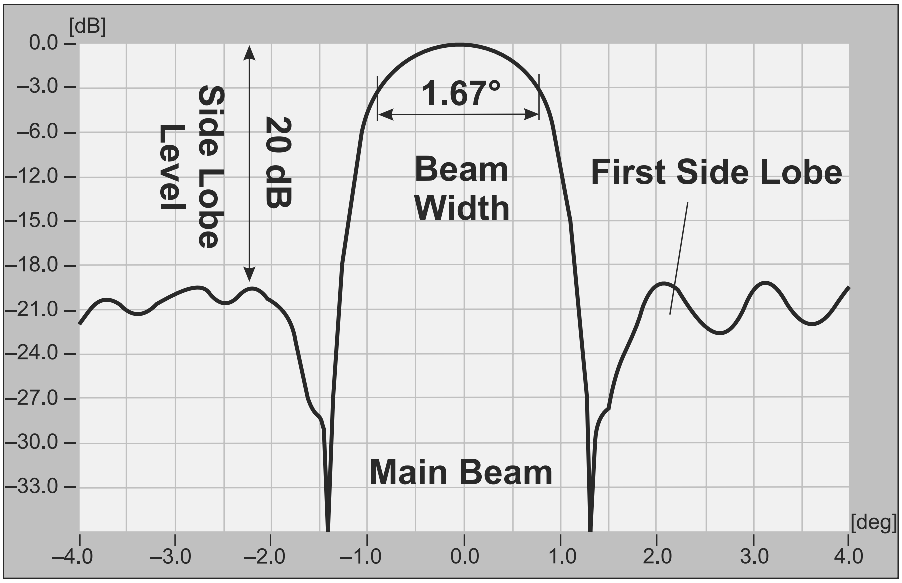


Fig. 12.14 Antenna pattern in a rectangular-coordinate graph with narrower scale

2. Aperture

An isotropic radiator disperses all energy at a surface of a sphere. The power has a defined density in a given distance. A directive antenna concentrates the energy in a smaller area. The power density is higher than by an isotropic radiator. The density can be expressed as power per area unit too. The received power can be compared with a related surface. This area is called **effective aperture**.

The effective aperture of an antenna A_e is the surface presented to the radiated or received signal. It is a parameter that governs the performance of the antenna. The antenna gain is related to the effective area by the following relationship.

$$G = \frac{4\pi . A_e}{\lambda^2}; \qquad A_e = K_a . A \tag{12.19}$$

where λ is the wavelength,

A_e is effective antenna aperture,

A is the physical area of the antenna, and

K_a is the antenna aperture efficiency.

The aperture efficiency depends on the distribution of the illumination across the aperture. If this is linear then $Ka = 1$. This high efficiency is offset by the relatively high level of side lobes obtained with linear illumination. Therefore, antennas with more practical levels of side lobes have an antenna aperture efficiency less than one ($A_e < A$).

3. Major and Minor Lobes

The antenna pattern has radiation concentrated in several lobes. The radiation intensity in one lobe is considerably stronger than in the other. The strongest lobe is called **major lobe** and the others are minor or **side lobes**. Generally, major lobes are those in which the greatest amount of radiation occurs. Side or minor lobes are those in which the radiation intensity is least.

4. Front-to-back Ratio

The front-to-back ratio is the ratio of power gain between the front and rear of a directional antenna. In most cases, there is a distinctive back lobe in the antenna-pattern diagram. Sometimes it is not possible to find a lobe exactly opposite to the main beam and in which case, the front-to-back ratio refers to the largest side lobe in the area of ±10° to ±30° around the opposite direction of the main beam. A high front-to-back ratio is desirable because this means that a minimum amount of energy is radiated in the undesired direction.

12.8 RADAR DISPLAYS

The purpose of the radar display is to visually present the information contained in the radar echo signal in a form suitable for operator interpretation and action. When the display is connected directly to the video output of the receiver, the information displayed is called **raw video.** This is the 'traditional' type of radar presentation. When the receiver video output is first processed by an automatic detector or automatic detection and tracking processor (ADT), the output displayed is sometimes called **synthetic video**.

The Cathode Ray Tube (CRT) has been almost universally used as the radar display. There are two basic cathode-ray tube displays. One is the deflection-modulated CRT**,** such as the **A-scope**, in which a target is indicated by the deflection of the electron beam. The other is the intensity-modulated CRT, such as the PPI, in which a target is indicated by intensifying the electron beam and presenting a luminous spot on the face of the CRT.

In general, deflection-modulated displays have the advantage of simpler circuits than those of intensity-modulated displays, and targets may be more readily discerned in the presence of noise or interference. On the other hand, intensity-modulated displays have the advantage of presenting data in a convenient and easily interpreted form. The deflection of the beam or the appearance of an intensity-modulated spot on a radar display caused by the presence of a target is commonly referred to as a **blip**.

The focusing and deflection of the electron beam may be accomplished electrostatically, electromagnetically, or by a combination of the two. **Electrostatic deflection CRTs** use an electric field applied to pairs of deflecting electrodes, or plates, to deflect the electron beam. Such tubes are usually longer than magnetic tubes, but the overall size, weight and power dissipation are less. Electromagnetic deflection CRTs require magnetic coils, or deflection yokes, positioned around the neck of the tube. They are relatively lossy and require more drive power than electrostatic devices. **Deflection-modulated CRTs**, such as the A-scope, generally employ electrostatic deflection. **Intensity-modulated CRTs**, such as the PPI, generally employ electromagnetic deflection.

The ability of an operator to extract information efficiently from a CRT display will depend on such factors as the brightness of the display, density and character of the background noise, pulse-repetition rate, scan rate of the antenna beam, signal clipping, decay time of the phosphor, length of time of blip exposure, blip size, viewing distance, ambient illumination, dark adaptation, display size and operator fatigue. Empirical data derived from experimental testing of many of these factors are available.

There are various types of CRT displays used with radar systems. Some of them are common while others are used for specific applications. Some of the more commonly used radar displays include

1. A-scope or A-scan
2. B-scope
3. F-scope
4. Plan Position Indicator (PPI)

12.8.1 A-Scope

The A-Scope represents an oscilloscope like display where the horizontal coordinate represents the range and the vertical coordinate represents the target echo amplitude. It is shown in Figure 12.15. It is the most commonly used display.

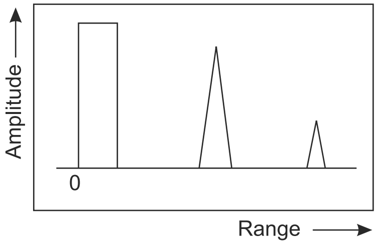


Fig. 12.15 A-Scope

Horizontal sweep is triggered every time a pulse is transmitted providing a reference point. The end of sweep, the right extreme of the display, represents the maximum range capability of the radar. The echo signal causes a deflection in the vertical direction. The separation between the starting reference and the echo deflection represents the target range. The deflection is either linearly or logarithmically proportional to target amplitude.

A slight variation of A-scope is the **A/R scope**. Here, any desired segment of time base can be expanded. It is illustrated in Figure 12.16. It is commonly used in tracking radars.

Another variation of the A-scope popular with tracking radars is the R-scope. It is illustrated in Figure 12.17.

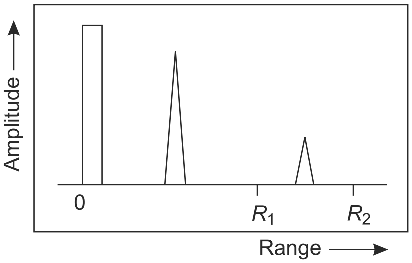


Fig. 12.16 A/R-Scope

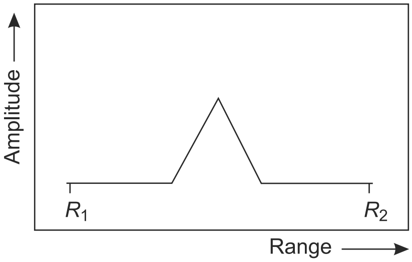


Fig. 12.17 R-Scope

In this, a limited range segment around the centre, that is adjustable, is adjustable. The range segment is usually the tracking range gate.

12.8.2 B-Scope

It is an intensity-modulated display with horizontal and vertical axes respectively representing azimuth angle and range. It is illustrated in Figure 12.18. The entire lower edge of the display is the radar location. It is commonly used in airborne radar, particularly when the aircraft is on an intercept mission. It shows the true range. The cross-range dimension, however, gets distorted on this display. Even if two targets are at a constant cross range, they appear at different separations at different ranges.

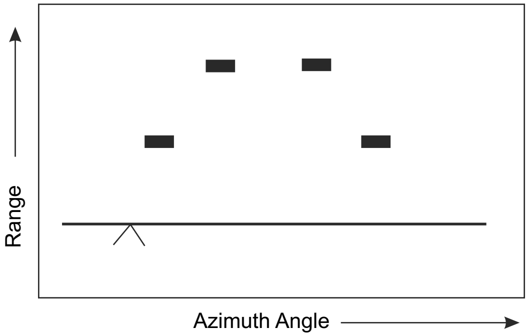


Fig. 12.18 B-Scope

In another operational mode of a B-scope, called **B-prime scope**, the vertical axis represents the target's radial velocity rather than its range. It is illustrated in Figure 12.19. The velocity is zero along a horizontal line in the centre. Targets above this line are those which are closing on to the radar and those below this line are those which are moving away.

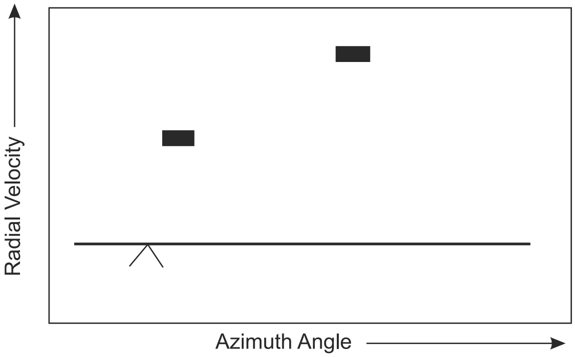


Fig. 12.19 B-prime scope

12.8.3 F-Scope

Horizontal and vertical axes of an F-scope display represent azimuth and elevation track error. It is illustrated in Figure 12.20.

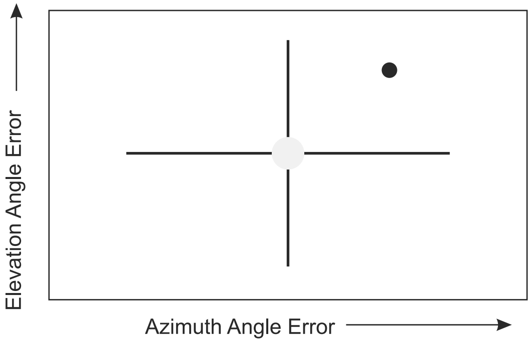


Fig. 12.20 F- scope

The centre of the display indicates the antenna's beam-axis location. The blip's displacement from the centre indicates the target's position with respect to the antenna beam axis.

12.8.4 Plan Position Indicator (PPI)

This display shows an intensity-modulated maplike circular display that gives the target location in polar coordinates. It is illustrated in Figure 12.21. The radar location is in the centre of the display. The target range is represented by the radial distance from the centre and the target's azimuth angle is given by the angle from the top of the display, usually north, clockwise. In some types of PPI display called **Offset** or **Sector PPI**, the radar location is offset from the centre of the display, as shown in the Figure 12.22. It is commonly used in search radars.

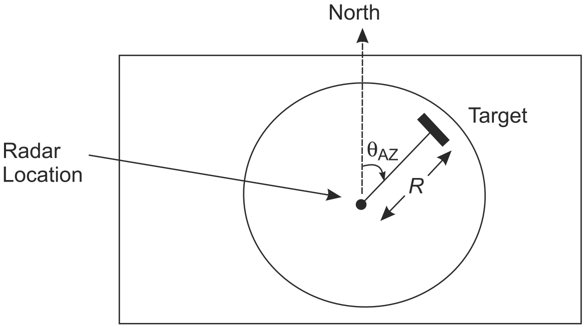


Fig. 12.21 Circular display in PPI

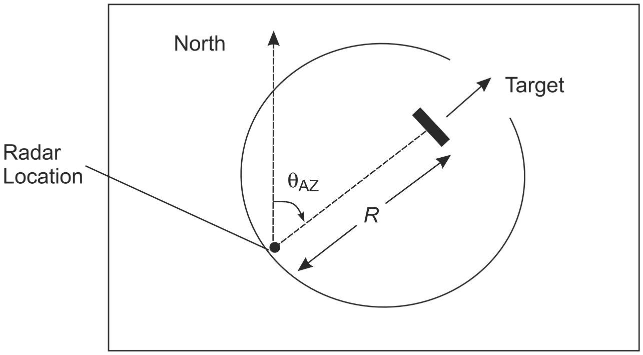


Fig. 12.22 Circular display in Sector PPI

The signal output of the receiver is given to the control grid of the CRT so that the beam gets intensity-modulated. The grid is biased beyond cut-off and thus screen areas corresponding to targets are brightened up. The deflection of the beam is achieved with application of sawtooth current to horizontal and vertical deflection coils. The deflection yoke, which is similar to that of a deflection yoke of a TV picture, is rotated in synchronism with the radar antenna. As a result, the beam is not only deflected radially outwards, starting from the centre, but also rotates continuously across the CRT screen, thereby covering the entire screen. The presence of any object is indicated by a bright spot on the screen which shows the position of the object on the target area. The range is measured radially out from the centre to the brightened spots.

CRT employed for this radar display has a phosphor with long persistence. This ensures that the screen of the PPI display does not flicker because the scanning speed is quite low as compared to the TV field frequency of 52 Hz. The resolution of the display is dependent upon the beamwidth of the transmitted energy, pulse width, the transmitted frequency and area of the CRT screen. For this reason, CRTs with large screens are employed.

The PPI is particularly suitable for search radars, especially when conical scanning is employed.

12.9 SEARCH RADAR

When radar is used to cover an all-round area and search for an unknown target, it is called search radar. Such a radar should be capable of scanning a large volume of space to acquire a target. Thus, scanning must be done rapidly. For this, an antenna with a beamwidth which is too narrow is used since a narrow beam would take a long time to scan a large volume of space. Once the target is approximately located then a narrow beam of electromagnetic energy is sent to find its exact location. This can be done either by reducing the beamwidth of the same transmitting antenna or by passing the information of the target from their radar to another radar which has a narrow beamwidth. Such a radar is termed **tracking radar**.

12.10 TRACKING RADAR

Once a target has been located by the search radar, it may then be tracked. For this purpose, radars with pencil-beam radiation are used. Radars used purely for tracking may employ conical scan or monopulse system. Radar that provides angular information of the target accurately is said to be tracking in range.

12.11 MOVING TARGET INDICATOR (MTI)

The Moving Target Indicator (MTI) radar system effectively handles moving targets such as aircraft and is capable of measuring their range and radial-velocity component in the presence of strong clutter due to stationary and even slow-moving undesired objects such as buildings, clouds, rain, etc. It is again based on the Doppler shift imparted to the transmit signal by the moving target to determine the target's radial-velocity component. The range is measured from the time lapse between the transmit signal and the received echo.

12.11.1 Doppler Effect

The apparent frequency of electromagnetic wave is dependent upon the relative motion of the source and the observer. An observer is considered to be standing on a platform approaching a fixed source of radiation with a relative velocity + ν_r. If both the observer and the source of energy are stationary and the frequency of radiation is f_r, it would be noted that f_r crests of wave per second are passing beyond the observer. If the observer is moving forward at a velocity ν_r, the observer will come across more than f_r crests per second. The number of crests observed under this condition is given by

$$f_{\mathrm{r}} + f_{\mathrm{d}} = f_{\mathrm{r}}\left(1 + \frac{\nu_{\mathrm{r}}}{\nu_{\mathrm{c}}}\right) \tag{12.20}$$

where ν_{c} is the velocity of the wave, and

f_{d} is the Doppler frequency difference or shift.

In radar, the signal undergoes the Doppler frequency shift when impinging upon a moving target. As this target reflects the waves, it is considered a moving source, transmitting energy towards a stationary observer. Then there is another Doppler shift. Hence, Doppler frequency in radar is given as

$$f_{\mathrm{d}} = 2\,f_d' = \frac{2\,\nu_{\mathrm{r}}\,f_{\mathrm{r}}}{\nu_{\mathrm{c}}}$$

$$= \frac{2\,\nu_r}{\lambda} \quad \left[\because \frac{f_{\mathrm{r}}}{\nu_{\mathrm{c}}} = \lambda\right] \tag{12.21}$$

It should be noted that this Doppler frequency shift will take place only if the target moves radially and not in tangential motion. The Doppler frequency shift may be used to determine the relative velocity of the target. Thus, moving targets can be distinguished from stationary targets on the basis of Doppler frequency shift.

EXAMPLE 12.5

A vehicle is moving towards a stationary CW Doppler radar transmitting at 10 GHz along the axis of the radar with a speed of 108 km/h. Determine the Doppler shift and the frequency of the received signal. What would be the received signal frequency if the vehicle was moving away from the radar along the same axis?

Solution

$$\lambda = \frac{c_0}{f}$$

$$= \frac{3 \times 10^8}{10 \times 10^9} = 0.03 \text{ m}$$

$$\text{Doppler shift } f_{\mathrm{d}} = \frac{2\,\nu_{\mathrm{r}}}{\lambda}$$

$$= \frac{2 \times 30}{0.03} = 2000 \text{ Hz}$$

Received frequency

$$= 10 \text{ GHz} + 2 \text{ kHz} = 9.00002 \text{ GHz}$$

If the vehicle is moving away from the radar, the received frequency is

$$10 \text{ GHz} - 2 \text{ kHz} = 9.999998 \text{ GHz}$$

EXAMPLE 12.6

A pulse doppler radar emitting at 10 GHz has PRF of 2 kHz. Determine if this radar is capable of measuring the radial velocity of 165 m/s of a closing target without any ambiguity. What would be the desired pulse-repetition rate to achieve that?

Solution

$$\lambda = \frac{c_0}{f}$$

$$= \frac{3 \times 10^8}{10 \times 10^9} = 0.03 \text{ m}$$

Maximum unambiguous doppler shift $= \frac{\text{PRF}}{2} = 1 \text{ kHz}$

Doppler shift $f_\text{d} = \frac{2\nu_\text{r}}{\lambda}$ which gives V_r (unambiguous)

$$= \frac{\lambda f_\text{d}}{2}$$

$$= \frac{0.03 \times 1000}{2} = 15 \text{ m/s}$$

$$f_\text{d} = \frac{2\nu_\text{r}}{\lambda}$$

$$= \frac{2 \times 165}{0.03} = 11000 \text{ Hz}$$

The desired PRF = 2 × 11000 = 22000 Hz

EXAMPLE 12.7

A CW radar waveform is modulated by 100 Hz sinusoidal signal and the amplitude of the modulating signal is such that it causes a maximum frequency deviation of 500 Hz. Determine the approximate bandwidth of the spectrum of the waveform. Also, determine the range-resolving capability of the radar.

Solution

Bandwidth $B = 2(Mf + 1).f_\text{m}$

$$Mf = \frac{\text{Maximum frequency deviation}}{\text{Modulating frequency}}$$

$$= \frac{500}{100} = 5$$

$$\therefore \quad B = 2(5 + 1).100 = 1200 \text{ Hz}$$

$$\text{Range resolution} = \frac{c_0}{2B} = \frac{3\times 10^8}{2\times 1200} = 12500\text{ m}$$

12.11.2 MTI Principle

An MTI, being a pulse system, relies on the phase difference between the transmitted signal and the corresponding echo to compute the Doppler. This phase difference for successive transmit pulses of RF energy and their corresponding echoes changes in case of moving targets at a rate equal to the Doppler frequency shift. The phase difference, however, remains the same in case of stationary targets and changes at a very small rate in case of slow-moving targets so as to be easily distinguishable from the phase-difference information produced by relatively much faster desired targets. The principle of operation of echo-signal processing is shown in Figure 12.23.

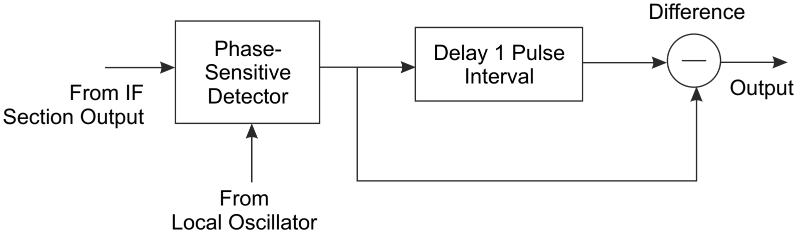


Fig. 12.23 Principle of operation of echo-signal processing

Each echo from a given range gate is subtracted coherently from a delayed version of the previous echo from the range gate. If the target is stationary, both echoes would produce the same phase difference and there would be complete cancellation provided the noise is absent. If the echo changes phase slightly due to target motion, the cancellation would be incomplete. For a target in uniform motion, there would be constant change in phase from pulse to pulse and there is no cancellation.

12.11.3 Block Diagram of MTI Radar

In principle, a moving-target indicator system compares a set of received echo pulses with those received during the previous sweep. The echoes belonging to the stationary targets cancel out while those corresponding to moving targets do not cancel and show only a phase change. Thus, the clutter is completely removed from the display and it reduces the time taken by the operator to observe the target. It allows easy detection of moving targets whose echoes are hundreds of times smaller than those of nearby stationary targets. This would not been possible without the use of MTI. Figure 12.24 shows the block diagram of MTI radar.

The radar-transmitter frequency in the MTI system is given by the sum of two oscillators produced at the output of mixer 2. The first of these oscillators is the STALO which stands

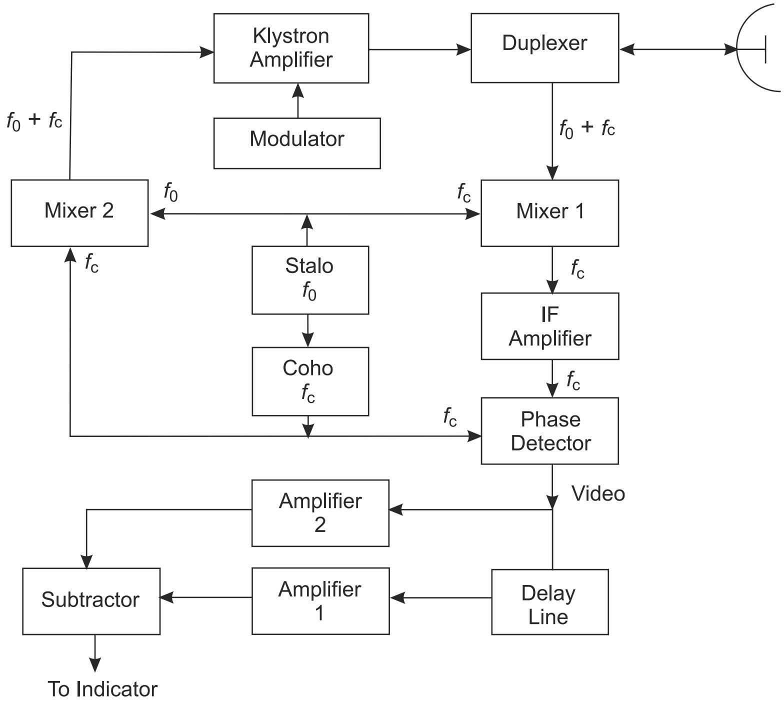


Fig. 12.24 Block diagram of MTI radar

for Stable Local Oscillator and oscillates at frequency f_0. The other oscillator is the COHO or Coherent Oscillator, oscillating at a frequency f_c. This frequency is the same as the intermediate frequency of the receiver and for this reason it is termed the coherent frequency. The sum frequency ($f_0 + f_c$) is given as input signal to the output tube which is a multicavity klystron amplifier in this case. This amplifier amplifies the signal and provides a high-power pulse when the modulator switches on this tube. The transmitter output pulse is passed on to the antenna through the duplexer.

The transmitted pulse is received back by the radar antenna after its reflection from the target. In case of a moving target, the received pulse undergoes a Doppler frequency shift. The received pulses are passed on to the mixer 1 of the receiver. The mixer heterodynes the received signal of frequency ($f_0 + f_c$) with the output of the STALO at f_0 and gives the output at the difference frequency f_c. The stages mixer 1 and mixer 2 are similar in all respects except that the output frequencies are different. It is the difference frequency in mixer 1 and sum frequency in mixer 2.

The difference frequency signal present at the output of the mixer is amplified by the IF amplifier and given to the phase-sensitive detector. This detector compares this IF signal with the reference signal obtained from the COHO stage and gives an output depending upon the phase difference between the two signals. Since all received signal pulses will have a phase difference compared with the transmitted pulse, the phase detector gives output for stationary as well as moving targets. While the phase shifts for the stationary targets remain constant, for moving targets, phase shifts are changing. This happens because of Doppler effect in moving targets. A change of half-cycle in the Doppler frequency shift would cause an output of opposite polarity in the phase-detector output. Thus, the output of the phase detector will have an output that has different magnitudes and polarities for successive pulses in case of a moving target, whereas in fixed targets the magnitude and polarity of the output will remain the same for all the transmitted pulses.

12.11.4 Blind Speeds in MTI

In MTI radar, Doppler shift is recovered by measuring the frequency displacement of the echo spectrum from the transmit spectrum [Figure 12.25(a)]. The process is simple because the transmit spectrum is a single line. In pulsed systems, the transmit spectrum comprises an infinite number of spectral lines separated by the PRF. In such a case, if the target does not produce any Doppler shift, the situation is shown in Figure 12.25(b). This is the case when the target is either stationary or is moving at the same rate as the radar. The echo and the transmit signal are at the same frequency. If the target's radial component of velocity is such that the Doppler shift produced is less than half the PRF, the Nyquist criterion is satisfied and the Doppler information is extracted without any ambiguity. This situation is shown in Figure 12.25(c).

It may be mentioned that the frequency of the sampled wave is recovered as the smallest frequency span from the received spectrum line to the closest transmit spectral line. The sampling rate is the same as the PRF. The target's radial component of velocity vector is such that the Doppler shift produced as a result is more than half the sampling rate, which is shown in Figure 12.25 (d). The Doppler measurement is now ambiguous as the Doppler measurement still reads the Doppler shift as the location of the received spectral line to the nearest transmit line. The ambiguity in Doppler measurement arising out of undersampling, also called **aliasing,** leads to the apparent Doppler shift being different from the true Doppler shift.

The true and apparent Doppler shifts are interrelated by

$$f_A = [(f_d = \text{MOD PRF}) - \text{PRF}] \text{ MOD PRF} \tag{12.22}$$

or

$$f_A = [(f_d = \text{MOD PRF}) + \text{PRF}] \text{ MOD PRF} \tag{12.23}$$

whichever gives a smaller absolute value. The MOD operator is the remainder of the division process.

Another problem is that of 'blind speeds'. When the Doppler shift equals an integer multiple of PRF, the moving target echo signal's spectral lines coincide with the spectral lines of the transmit signal and so are the spectral lines of the stationary target echoes [Figure 12.25(e)]. When then target's radial-velocity component is such that it travels a distance of $n.\lambda/2$ along

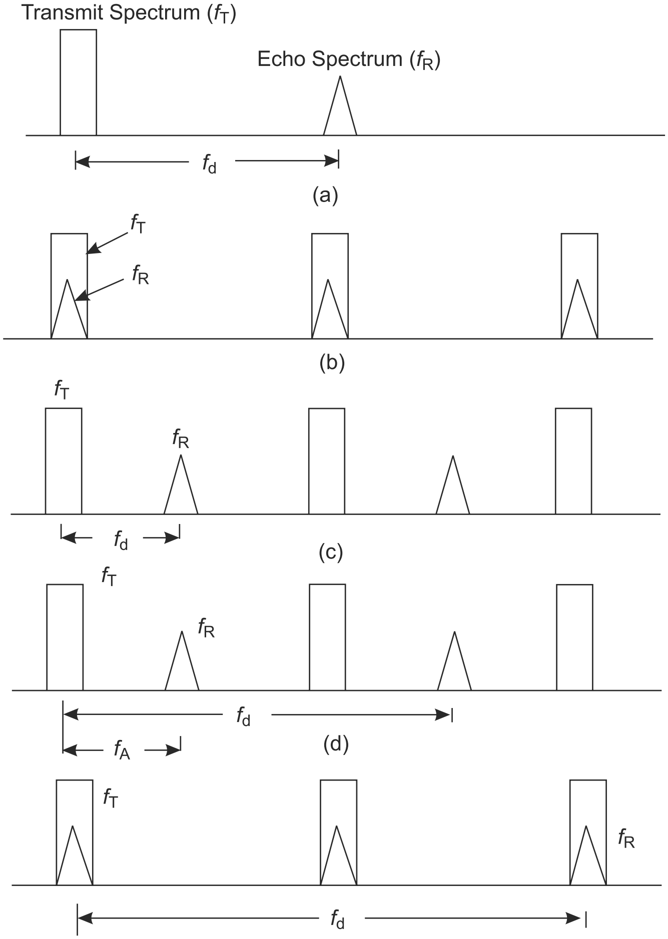


Fig. 12.25 Phase-detector outputs

the radar axis during the time between successive transmit pulses then the phase difference between the corresponding echo pulses would be 2nπ radians which is equivalent to no-phase change or a stationary target. Though use of Doppler filters effectively attenuates echoes at zero Doppler shifts and at integer multiples of PRF for clutter rejection, a moving target producing these Doppler shifts cannot be detected. Such Doppler shifts and the associated radial-velocity components are called **blind dopplers** and **blind speeds** respectively.

The Doppler shifts and blind speeds can be computed from

$$f_B = n.\ \mathrm{PRF} \tag{12.24}$$

$$V_B = [n.c.\ \mathrm{PRF}/2f] \tag{12.25}$$

where $n = \pm1, \pm2, \pm3....$, and

f is the operating frequency.

EXAMPLE 12.8

An MTI radar system operating at 10 GHz and a repetition rate of 1000 Hz receives echoes from an aircraft approaching the radar with a radial-velocity component of 1 km/s. Determine the radial-velocity component as measured by the radar.

Solution

If f_d is the true Doppler shift then f_d can be expressed as

$$f_d = \frac{2\nu_r \cdot f}{c}$$

$$= \frac{2\times 1000\times 10\times 10^{10}}{3\times 10^{8}} = 66.67 \text{ kHz}$$

If f_A is the apparent Doppler shift then

$$f_A = [(f_d \text{ MOD PRF}) - \text{PRF}] \text{ MOD PRF}$$

or

$$f_A = [(f_d \text{ MOD PRF}) + \text{PRF}] \text{ MOD PRF}$$

whichever gives the smaller absolute value.

From the first equation

$$f_A = [(66670 \text{ MOD } 1000) - 1000] \text{ MOD } 1000$$

$$= [670 - 1000] \text{ MOD } 1000$$

$$= -330 \text{ MOD } 1000 = -330 \text{ Hz}$$

From the second equation

$$f_A = [(66670 \text{ MOD } 1000) + 1000] \text{ MOD } 1000$$

$$= [670 + 1000] \text{ MOD } 1000$$

$$= 1670 \text{ MOD } 1000 = 1670 \text{ Hz}$$

The first equation has given lower absolute value. Therefore, the apparent Doppler shift is –330 Hz.

Radial velocity corresponding to the Doppler shift is given by

$$\nu_r = \frac{c \cdot f_d}{2f}$$

$$= \frac{(3\times 10^{8}).(-330)}{2\times 10\times 10^{10}} = -4.95 \text{ m/s}$$

EXAMPLE 12.9

A 3 cm MTI is operating at a PRF of 2000 Hz. Find the lowest blind speed.

Solution

Operating wavelength = 3 cm = 0.03 m

PRF = 2000 Hz

Lowest blind speed = $\left(\frac{n\lambda}{2}\right)$PRF for $n = 1$

$$= \left(\frac{0.03}{2}\right) \times 2000$$

= 30 m/s = 108 km/h

EXAMPLE 12.10

Two MTI radar systems are operating at the same PRF but have different operating frequencies. If the third blind speed of one is equal to fifth blind speed of the other, find the ratio of their operating frequencies.

Solution

Let the operating frequency of first radar = f_1

The operating frequency of second radar = f_2

The third blind speed of first radar = $\left(\frac{3c}{2f_1}\right)$PRF

Also, fifth blind speed of the second radar = $\left(\frac{5c}{2f_2}\right)$PRF

$$\therefore \quad \left(\frac{3c}{2f_1}\right)\text{PRF} = \left(\frac{5c}{2f_2}\right)\text{PRF}$$

$$\left(\frac{3}{f_1}\right) = \left(\frac{5}{f_2}\right)$$

$$\therefore \quad \left(\frac{f_1}{f_2}\right) = \frac{3}{5}$$

12.12 CONTINUOUS-WAVE (CW) RADAR

Continuous-Wave (CW) radars transmit electromagnetic waves continuously towards the target and there is a continuous reflection of these waves from the targets. It is possible to use a single antenna for transmission and reception in pulsed radars and this was achieved with the duplexer switch. In CW radars, the transmitted and reflected waves propagate simultaneously for transmission and reception.

The CW radar makes use of Doppler effect for speed measurement of targets. Figure 12.26 shows the block diagram of a CW Doppler radar.

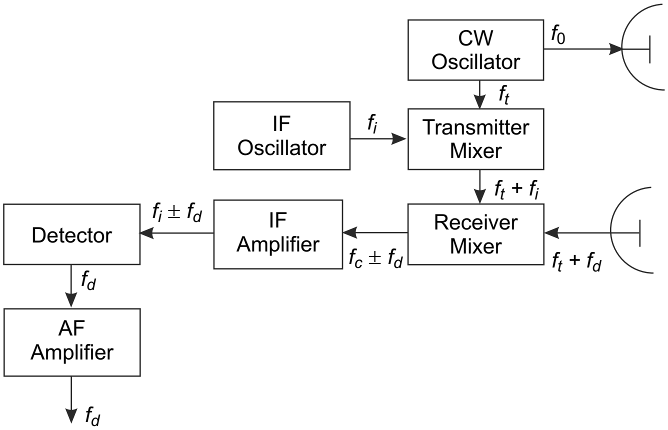


Fig. 12.26 Block diagram of a CW Doppler radar

The transmitter section is a low-power microwave oscillator such as reflex klystron that generates sinusoidal signals in the microwave range. This signal is transmitted by the transmitting antenna. A small fraction of the transmitter signal is fed to the IF signal generated by the IF oscillator. Sum of the transmitter signal frequency (f_t) and the IF signal (f_c) is selected at the output of the transmitter mixer.

The receiver antenna picks up the waves reflected from the target and for moving targets, the received signal frequency equals $f_t \pm f_d$. The signal is given to the receiver mixer where it mixes with the output of the transmitter mixer. At the output of the receiver mixer, the difference frequency signal is obtained at $f_c \pm f_d$. This signal is amplified by the IF amplifier and given to the detector stage. The detector circuit recovers the Doppler frequency from the IF signal and passes it to the AF amplifier where it is amplified. The amplified signal is given to a frequency counter. Since the Doppler frequency shift f_d is proportional to the velocity of the target, the output of the counter gives an indication of the target speed. The frequency counter is so designed that at its output, the target speed is displayed directly in kilometers/hour than showing the Doppler frequency. However, the display does not give indication as to whether the target is approaching or receding, because the sign of the Doppler frequency shift is lost. The CW Doppler radar is not capable of giving the range of the target.

1. Advantages of CW Doppler Radar

(a) Low transmitting power
(b) Low power consumption
(c) Simple circuitry
(d) Small size
(e) Mobile in nature
(f) Used to give accurate measurement of relative velocity of the target and the reading is unaffected by stationary objects
(g) Capable of measuring a large range of target velocities

2. Drawbacks of CW Doppler Radar

(a) It is limited in the maximum power it transmits and this places a limit on its maximum range.
(b) It is not capable of indicating the range of the target and can show its velocity only.
(c) If a large number of targets are present, it gets confused rather easily.

12.13 FREQUENCY-MODULATED CW RADAR

A continuous-wave radar cannot give the range of a target, because the transmitted signal is unmodulated. As a result, the receiver cannot sense which particular cycle of oscillations is being received at a moment. If the transmitted carrier is frequency-modulated then it should be possible to eliminate this main drawback. Using frequency modulation will, however, increase the bandwidth and thus, it is seen that for conveying more information, more bandwidth is required.

The bandwidth can be determined as follows.

$$\text{Bandwidth } B = 2(Mf + 1).\, f_{\text{m}} \tag{12.26}$$

where Mf is modulation index, and

f_{m} is the modulating frequency.

Figure 12.27 shows the block diagram of frequency-modulated CW radar used in aircraft for measurement of their altitudes. For this reason, it is commonly referred as **airborne altimeter**. Here, a sawtooth wave is used for frequency-modulating a CW carrier. The other types of waveforms might also be used as modulating signals but the sawtooth waveform gives the simplest circuit arrangement. Thus, the frequency of the transmitted signal increases

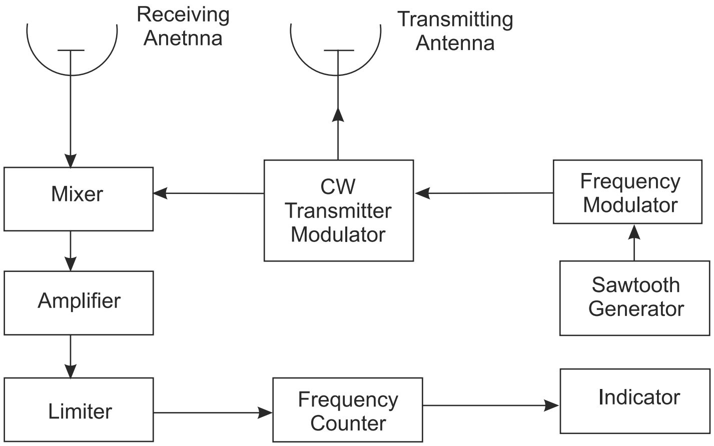


Fig. 12.27 Block diagram of frequency-modulated CW radar

linearly with the increasing amplitude of the modulating signal. In this case, the target is the earth which is stationary with respect to the aircraft.

Since increase in the amplitude of the modulating signal is uniform with time, the rate of increase in frequency in the transmitted signal brought by the modulating signal is also uniform with time. For a given height of aircraft, a known time will be required for the waves to travel from the earth to the aircraft. Thus, during this time, a definite change in signal frequency will take place. If it is possible to measure the frequency change in the signal, it will give an indication of the height of the aircraft. Thus, the frequency counter is switched on just when the modulating signal has zero frequency, i.e. the transmitted frequency equals f_c. This signal frequency is picked up after its reflection from the earth and may be used to switch off the counter. The final reading of the counter will give an indication of the change in the frequency and, hence, about the height of the aircraft. This is indicated by the indicator connected in the receiver.

For a case, when the relative velocity of the aircraft and the earth is not zero, another frequency shift will be produced due to Doppler effect and this frequency shift will be superimposed on the frequency difference. This frequency shift can now be used to measure the relative velocity of the aircraft in the same way as in a Doppler radar. However, the time difference between the transmission and reception of a particular cycle of the signal will be constant and, hence, the average frequency difference will also be constant. Therefore, correct height measurement can still be made on the basis of average frequency difference.

FM-CW radar is used as an altimeter in aircraft and because of the short range involved, it is used in preference to pulsed radars. It has quite a low power requirement as compared to pulsed radars. The size of this equipment is small and quite suitable for aircraft installations. Because reflection has to take place from the earth, which has a large size compared to the aircraft, a small size can be used. The transmitting powers used are quite small.

12.14 TRACKING RADAR

The primary function of tracking radar is the automatic tracking of moving targets. It is usually a ground-based system used to track airborne targets. The tracking radar antenna sends out a very narrow beam whose width could be anywhere between a fraction of a degree to a degree or so in both azimuth and elevation to get the desired resolution for tracking purpose. It is necessary to acquire the target with a search radar having a beam of relatively much larger width before a track action is initiated. In the track mode, whenever the target tends to move away from the radar-beam axis, an error signal is generated which in the closed loop is used to steer the radar antenna either mechanically or electronically to keep the target always illuminated by the radar beam.

12.14.1 Track Modes

Tracking can be carried out using range (called **range tracking**), Doppler (called Doppler or **velocity tracking**) and angular (**angle tracking**) information. This allows the radar to follow the motion of a target in azimuth and elevation due to angle tracking, range due to range tracking and Doppler due to Doppler tracking. However, not all radars track in all dimensions. Different track modes include

1. Single-Target Track (STT)
2. Spotlight track
3. Multitarget track
4. Track-While-Scan (TWS)

In the single-target track (STT), the radar tracks a single target. It is continuously dedicated to a single moving target. Such a radar samples the target information at the radar PRF. Single-target trackers are capable of tracking targets with great accuracy.

In the **spotlight track**, the radar sequentially dwells upon various targets spending a certain specified time on each target. It is not as accurate as the single-target track due to the fact that a given target is likely to undergo a change in its coordinates during the time between two successive dwell periods.

A **multitarget track** mode is capable of simultaneously tracking multiple targets with an accuracy matching that of a single-target track. In a **Track-While-Scan** (TWS) system, the radar samples the position of several targets once per scan and then with the help of certain samples. TWS is truly a search radar's operational mode. It is not essentially a tracking operation because for true multitarget tracking, each target must be sampled at the Nyquist rates corresponding to the radar servoloop and target-maneuvering bandwidths. The required sampling rate may typically be 10 to 20 samples per second for each target. In a TWS process, the target may be sampled once every 10 to 15 seconds.

12.14.2 Block Diagram of Tracking Radar

Figure 12.28 shows the basic block diagram of a tracking radar. These radars use angular information as the basis for tracking operation. But for accurate tracking, it is important that the radar concentrates on one target at a time. If there are more targets in the radar antenna's beam, techniques should be used to ignore other returns from other targets.

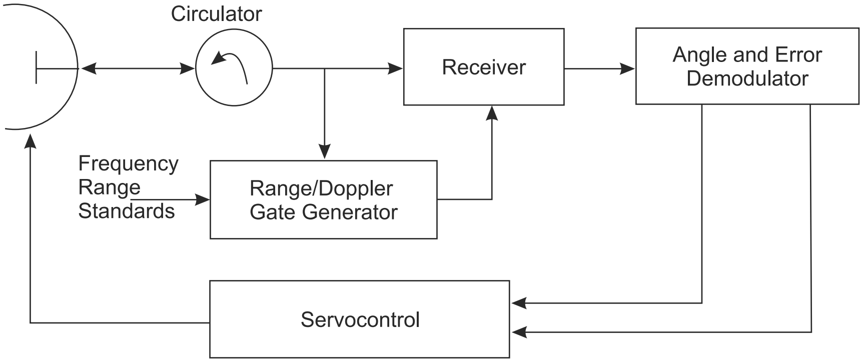

Circulator
Receiver
Angle and Error Demodulator
Frequency Range Standards
Range/Doppler Gate Generator
Servocontrol

Fig. 12.28 Basic block diagram of a tracking radar

Range gating, a part of range tracker and Doppler gating, which is a part of Doppler Tracker can be used for the purpose. Time and frequency control for range and Doppler gating is done in range and Doppler trackers respectively. The angular error signals for the desired target to be tracked are developed in the error-demodulator block which is also controlled by range/Doppler gate generation block and then fed back to the steerable antenna in a closed loop for tracking.

12.14.3 Types of Tracking Radar

Tracking radars are classified based on the methodology used to develop angular errors. The commonly used tracking methodologies are as follows.

1. Lobe switching
2. Sequential lobing
3. Conical scan
4. Amplitude comparison monopulse
5. Phase comparison monopulse

1. Lobe Switching

In the lobe-switching tracking technique, the antenna beam is rapidly switched between two positions around the antenna-beam axis in a single plane. The amplitudes of the echoes from the target to be tracked for the two lobe positions are compared. The difference between the two amplitudes indicates the location of the target with reference to the antenna axis. When the target is on the axis, the difference is zero as the echo amplitudes for the two lobe positions are identical. The lobe-switching technique has the disadvantage that it loses its effectiveness if the target cross section changes between different returns in one scan.

2. Sequential Lobing

In sequential lobing, a squinted radar beam is sequentially placed in discrete angular positions, usually four around the antenna axis. It is schematically shown in Figure 12.29.

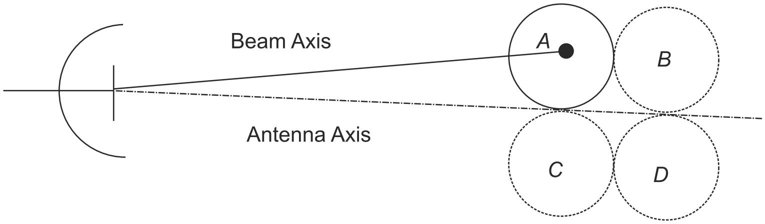


Fig. 12.29 Sequential lobing

The angular information on the target is determined by processing several target echoes. The tracking-error information is contained in the target-signal amplitude variations. The squinting and squinted beam switching between different positions is done electrically in modern radars using this tracking methodology. Since the beams can be switched very rapidly using electronic means, the transmitted beam is usually not scanned. The lobing is on receive only. Also, virtually any scanning pattern can be used. The scan pattern can be changed from scan to scan. It is because of this reason that this type of tracking radar is less affected by amplitude-modulated jamming.

3. Conical Scanning

It is similar to sequential scanning except for the difference that in case of a conical scan, the squinted beam is scanned rapidly and continuously in a circular path around the axis, as shown in Figure 12.30.

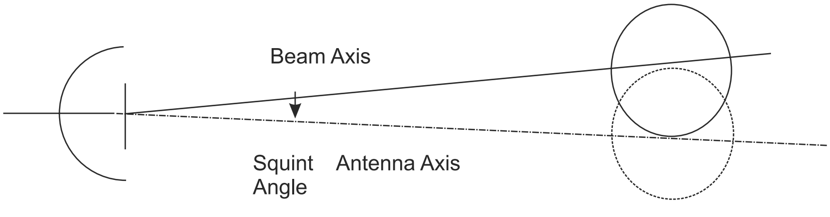


Fig. 12.30 Conical scanning

If the target to be tracked is off the antenna axis, the amplitude of the target echo signal varies with the antenna's scan position. The tracking system senses these amplitude variations as a function of scan position to determine the target's angular coordinates. The error information is then used to steer the antenna axis so as to coincide with the target location. For true tracking, the scan frequency must be such that Nyquist criterion for the sampling rate is satisfied.

The conical scanning technique of tracking is highly vulnerable to amplitude-modulated jamming, particularly gain-inversion jamming.

4. Amplitude-Comparison Monopulse Tracking

One of the drawbacks of the above-mentioned tracking techniques, including lobe switching, sequential lobing and conical scanning, is that their tracking accuracy gets severely affected if the target's radar cross section changes during the time when the beam is being switched or scanned, as the case may be, to get the desired number of samples. In addition, these techniques also suffer from their vulnerability to AM jamming. Monopulse tracking overcomes these shortcomings by generating all the required angle-error information from one pulse only.

5. Phase-Comparison Monopulse Tracking

In phase comparison monopulse tracking, it is the phase difference between the received signals in different antenna elements that contains information on angle errors. In all, at least two antenna elements are required each for azimuth and elevation-error detection. When the target is on the axis, the magnitude of phase difference would be zero. If it is off-axis then magnitude and sense of phase difference would determine the magnitude and sense of the off-axis angle.

The phase difference produced per unit angular error increases if the elements are wide apart. But if they are too far apart, an off-axis signal may produce identical phases at the antenna elements. This gives rise to ambiguity. A practical system could have two pairs of antenna elements each for azimuth and elevation. The outer pair gives the desired sensitivity while the inner pair resolves ambiguity.

6. Range Tracking

Range tracking is the process of tracking a moving target based on its range coordinates. Even though the commonly used tracking methodology in tracking radars is angle tracking, a range tracker forms a part of the angle tracker also. A range tracker in that case continuously measures the target range and based on the range data, generates a range gate so that the target is at the centre of the gate. Range tracking, thus, provides an effective means of distinguishing the desired target to be tracked by using angular means from the other targets within a radar beam.

The first step in any tracker is target acquisition which provides an idea of target coordinates so that the radar beam could be pointed in that direction. A range tracker could do the job of target acquisition very well. Typically, the range tracker divides the minimum to maximum range into small range increments and as the antenna scans a given region, it examines each of the range increments in a given direction simultaneously for presence of target. The antenna is made to scan slowly enough for the target to remain within the radar bandwidth as different range increments are being examined in a given direction.

Range tracker is a closed-loop system. The error corresponding to deviation of target's range location from the centre of the range gate is sensed and fed back to the range gate generating circuitry to reposition the gate in such a way that the target is at the centre. The commonly used technique of sensing range-tracking error is that of using split gate comprising an **early gate** and a **late gate** as shown in Figure 12.31.

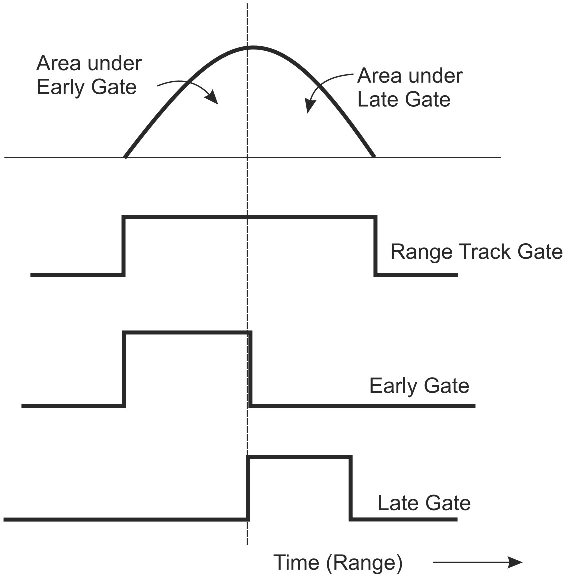


Fig. 12.31 Range tracker

When the target is at the centre, the area under the echo pulse when the early gate is open is same as the area under the echo pulse when the late gate is open. If the signals under the two gates are integrated and a difference is taken, it would be zero. If the target is off-centre, one signal would be greater than the other. The magnitude and sense of the difference signal can be used to reposition the gate.

7. Velocity Tracking

Velocity tracking is the process that makes use of Doppler shift information. A Doppler tracks error using split-filter error detection. The track error is represented by the difference between the target IF and the receiver's normal IF. The error after filtering is used to change the receiver

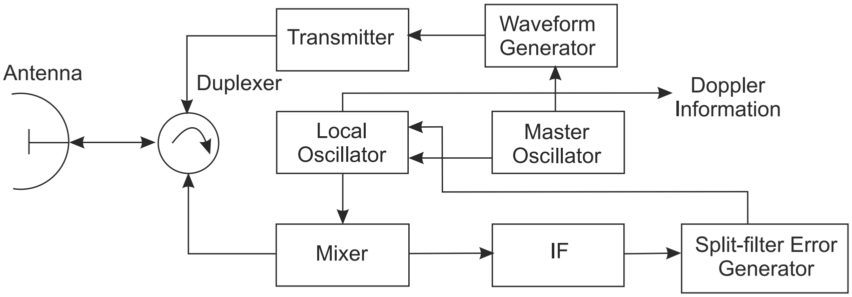


Fig. 12.32 Velocity tracker

local oscillator frequency until the Doppler shifted signal is the nominal IF. Figure 12.32 shows the block diagram of a velocity tracker.

Summary

Radio Detection and Ranging is abbreviated as RADAR. It is an electromagnetic system for the detection and location of objects. A radar operates by transmitting a particular type of waveform and detects the nature of the echo signal. It can be designed to see through those conditions impervious to normal human vision, such as darkness, haze, fog, rain and snow.

Radar is a stand-alone active system having its own transmitter and receiver that is used for detecting the presence and finding the exact location of a far-off target. In a radar system, the transmitter could be a power amplifier employing any of the microwave tube amplifiers such as Klystron, Travelling Wave Tube (TWT), Crossed Field Amplifier (CFA) or even a solid-state device. The hypersensitive receiver amplifies and demodulates the received RF signals. The receiver provides video signals on the output. Radar signals can be displayed on the traditional Plan Position Indicator (PPI) or other more advanced radar display systems.

The time between the beginning of one pulse and the start of the next pulse is called Pulse-Repetition Time (PRT) and is equal to the reciprocal of PRF. The Pulse-Repetition Frequency (PRF) of the radar system is the number of pulses transmitted per second.

The radar equation relates the range of radar to the characteristics of the transmitter, receiver, antenna, target and environment. It is useful for determining the maximum distance from the radar to the target and it can serve both as a tool for understanding radar operation and as a basis for radar design.

Pulse radar sets transmit a high-frequency impulse signal of high power. These classically radar sets transmit a very short pulse to get a good range resolution with an extremely high pulse power to get a good maximum range. When radar is used to cover an all-round area and search for an unknown target, it is called search radar. Such a radar should be capable of scanning a large volume of space to acquire a target. A radar that provides angular information of the target accurately is said to be tracking in range.

The antenna system transfers the transmitter energy to signals in space with the required distribution and efficiency and this process is also applied in an identical way during reception. There are two types of antenna often used in a radar system. They are

- The parabolic dish antenna
- The phased-array antenna

The Moving Target Indicator (MTI) radar system is based on the Doppler shift imparted to the transmit signal by the moving target to determine the target's radial-velocity component and it effectively handles moving targets such as aircraft and is capable of measuring their range and radial-velocity component in the presence of strong clutter due to stationary and even slow-moving undesired objects such as buildings, clouds, rain, etc.

Continuous-Wave (CW) radars transmit electromagnetic waves continuously towards the target and there is a continuous reflection of these waves from the targets. It is possible to use a single antenna for transmission and reception in pulsed radars and this was achieved with the duplexer switch. In CW radars, the transmitted and reflected waves propagate simultaneously for transmission and reception. But a continuous-wave radar cannot give the range of a target, because the transmitted signal is unmodulated. Then the receiver cannot sense which particular cycle of oscillations is being received at a moment. If the transmitted carrier is frequency-modulated then it should be possible to eliminate this main drawback.

A tracking radar is used for automatic tracking of moving targets. It is usually a ground-based system used to track airborne targets. The tracking-radar antenna sends out a very narrow beam whose width could be anywhere between a fraction of a degree to a degree or so in both azimuth and elevation to get the desired resolution for tracking purpose. Tracking radars are classified based on the methodology used to develop angular errors.

Range tracking is the process of tracking a moving target based on its range coordinates. It provides an effective means of distinguishing the desired target to be tracked by using angular means from the other targets within a radar beam. Velocity tracking is a process that makes use of Doppler shift information.

REVIEW QUESTIONS

PART-A

1. What is RADAR?
2. What is radar ranging?
3. What are the basic components of a radar system?
4. List the radar transmitters used in communication.
5. State the purpose of PPI.

6. Define pulse-repetition time.
7. Define pulse-repetition frequency.
8. What do you mean by slant range? Give its expression.
9. What is maximum unambiguous range?
10. Give the expression for radar-range equation.
11. How will you classify a radar based on its specific functions?
12. How will you classify a radar based on its designed use?
13. What is pulse radar?
14. Mention the various functions performed by the radar antenna.
15. What are the types of antenna used in a radar system?
16. How do you calculate the gain of the parabolic antenna?
17. Define beamwidth of an antenna pattern.
18. Give the relationship between the antenna gain and the effective area.
19. What are major and minor lobes of an antenna pattern?
20. What is front-to-back ratio of antenna pattern?
21. What are the different radar displays available?
22. State the purpose of A-Scope.
23. Mention the purpose of B-Scope.
24. What are PPI display and sector PPI?
25. What is the role of a search radar?
26. What is the purpose of a tracking radar?
27. State the advantages of MTI radar.
28. State the principle of MTI.
29. What is a continuous-wave radar?
30. List the advantages and disadvantages of CW radar.
31. What is the main drawback of a CW radar? How will you overcome it?
32. What are the different track modes available?
33. List the commonly used tracking methodologies.
34. What is lobe switching?
35. What is sequential lobing?
36. What is conical scanning?
37. Define range tracking.
38. Define velocity tracking.

PART-B

1. With a neat block diagram of a basic radar system, explain the functioning of its blocks.
2. Explain radar ranging in detail.
3. What are frequencies and powers used in radars? Explain.
4. How will you derive radar-range equation? Describe in detail.
5. List various classifications of a radar system. Explain them in brief.
6. Draw the block diagram and explain the functioning of a high-power pulsed radar.
7. What are the essential functions of antennas used in radar communication? Explain also about its types.
8. What are the different antenna-scanning patterns? Explain in detail.
9. What are the different parameters to be considered in the analysis of an antenna pattern? Explain them in brief.
10. Explain various radar displays available in brief.
11. What is MTI radar? State its principle. Explain its functioning with a neat block diagram.
12. With a neat block diagram, describe the functioning of a CW Doppler radar.
13. State the purpose of tracking radar. What are the different types of track modes? Briefly explain them.
14. Explain various tracking methodologies available in brief.
15. Mention the significance of range tracking and velocity tracking in radars.

13

WIRELESS COMMUNICATION

Objectives

- To know the needs, examples, media and applications of wireless communication
- To discuss details about mobile communication and advanced mobile communication
- To discuss wireless LAN, PAN, Bluetooth and Zigbee
- To provide the details, the needs and principle of caller ID
- To provide the details of cordless telephones, and pager and facsimile systems.

13.1 INTRODUCTION

Wireless communication is the transfer of information over a distance without the use of wires. The distances involved may be short or long, i.e. from a few metres to thousands or millions of kilometres.

Wireless operations permit services, such as long-range communications, that are impossible to implement with the use of wires. Telecommunication, for example, uses some form of energy like radio frequency, acoustic energy, etc. to transfer information without the use of wires.

13.2 NEEDS OF WIRELESS COMMUNICATION

Wireless communication is preferably used in order to meet the following requirements.

1. To span a distance beyond the capabilities of typical cabling,

2. To provide a backup communications link in case of normal network failure,
3. To link portable or temporary workstations,
4. To overcome situations where normal cabling is difficult or financially impractical, or
5. To remotely connect mobile users or networks.

13.3 EXAMPLES OF WIRELESS EQUIPMENT

The following are some commonly used examples for wireless communication.

1. Mobile radio service used by business, industrial and public safety entities
2. Cellular telephones and pagers, which provide connectivity for portable and mobile applications, both personal and business
3. Global Positioning System (GPS) which allows drivers of cars and trucks, captains of boats and ships, and pilots of aircraft to ascertain their location anywhere on the earth
4. Cordless telephone sets which are limited-range devices
5. Satellite television which allows viewers in almost any location to select from hundreds of channels

13.4 TRANSMISSION MEDIA FOR WIRELESS COMMUNICATION

Wireless communication can be performed by any of the following.

1. Radio Frequency (RF) communication
2. Microwave communication, for example long-range line-of-sight via highly directional antennas, or short-range communication
3. Infrared (IR) short-range communication

13.5 APPLICATIONS OF WIRELESS TECHNOLOGY

The following are some of the popular applications using wireless technology.

1. Security Systems Wireless technology may replace hard-wired implementations in security systems for homes or office buildings.

2. Television Remote Control Latest televisions use wireless remote control units. Now, radio waves are also used.

3. Cellular Telephone These instruments use radio waves to enable the operator to make phone calls from many locations worldwide. They can be used anywhere there is a cellular telephone site to house the equipment required to transmit and receive the signal that is used to transfer both voice and data to and from these instruments.

4. Wi-Fi It is a wireless local area network that enables portable computing devices to connect easily to the Internet.

5. Wireless Energy Transfer It is a process whereby electrical energy is transmitted from a power source to an electrical load that does not have a built-in power source, without the use of interconnecting wires.

6. Computer Interface Cards In order to answer the call of customers, many manufactures of computer peripherals turned to wireless technology to satisfy their consumer needs. Initially, these units used bulky, highly limited transceivers to mediate between a computer and a keyboard and mouse. However, more recent generations have used small, high-quality devices.

13.6 RANGE OF WIRELESS SERVICES

Radio spectrum is used for a wide range of services. These can be broken into the following broad classes:

1. **Broadcasting services** which include short wave, AM and FM radio as well as terrestrial television.
2. **Mobile communications of voice and data** including maritime and aeronautical mobile for communications between ships, airplanes and land; land mobile for communications between a fixed base station and moving sites such as a taxi fleet and paging services, and mobile communications either between mobile users and a fixed network or between mobile users, such as mobile telephone services;
3. **Fixed services** may be either point-to-point or point-to-multipoint services.
4. **Satellites** used for broadcasting, telecommunications and Internet, particularly over long distances
5. **Amateur radio** and other uses including military, radio astronomy, meteorological and scientific uses.

13.7 HISTORY OF MOBILE TELEPHONES

The history of mobile telephones can be broken into four periods.

1. The **first period** involved mobile telephones that exclusively used a frequency band in a particular area. These telephones had severe problems with congestion and call completion. If one customer was using a particular frequency in a geographic area, no other customer

could make a call on that same frequency. Further, the number of frequencies allocated to mobile telephone services was small, limiting the number of simultaneous calls.

2. Next, the introduction of cellular technology greatly expanded the efficiency of frequency use of mobile phones. Rather than exclusively allocating a band of frequency to one telephone call in a large geographic area, a cell telephone breaks down a geographic area into small areas or cells. Different users in different, i.e. non-adjacent cells, are able to use the same frequency for a call without interference.
3. First-generation cellular mobile telephones developed around the world using different, incompatible analogue technologies. For example, in the 1980s in the US, there was the Advanced Mobile Phone System (AMPS) and the UK had the Total Access Communications System (TACS). The result was a wide range of largely incompatible systems, particularly in Europe, although the single AMPS system was used throughout the US.
4. **Second generation (2G)** mobile telephones used digital technology. The adoption of second-generation technology differed substantially between the United States and Europe and reverses the earlier analogue mobile experience. In Europe, a common standard was adopted, partly due to government intervention. Global System for Mobile communication was first developed in the 1980s and was the first 2G system which allows full international roaming, automatic location services, common encryption and relatively high-quality audio. GSM is now the most widely used 2G system worldwide, in more than 130 countries, using the 900 MHz frequency range.
5. The final stage in the development of mobile telephones is **third-generation (3G)** technology. These systems will allow for significantly increased speeds of transmission and are particularly useful for data services. For example, 3G phones can more efficiently be used for e-mail services, and downloading music and videos from the Internet. They can also allow more rapid transmission of images, for example from camera phones.

13.8 PRINCIPLES OF MOBILE COMMUNICATION

Each mobile uses a separate, temporary radio channel to talk to the cell site. The cell site talks to many mobiles at once, using one channel per mobile. Channels use a pair of frequencies for communication in which one frequency is for the forward link for transmitting from the cell site and the other frequency is used for reverse link for the cell site to receive calls from users. Radio energy dissipates over distance, so mobiles must stay near the base station to maintain communications.

The basic structure of mobile networks includes telephone systems and radio services. Where a mobile radio service operates in a closed network and has no access to the telephone systems, the mobile telephone service allows interconnection to the telephone network. Figure 13.1 shows the structure of a mobile network.

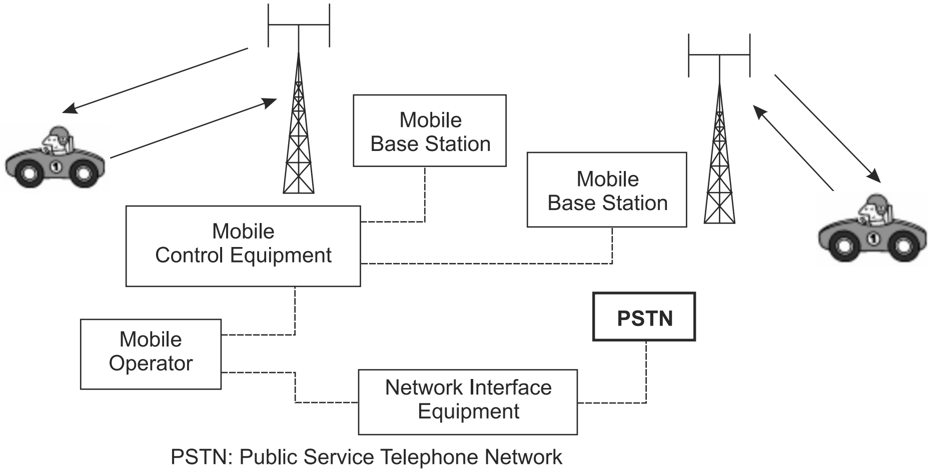


Fig. 13.1 Basic structure of a mobile network

In earlier days, mobile service was structured similar to television broadcasting in which one very powerful transmitter located at the highest spot in an area used to broadcast in a radius of up to 50 km. Figure 13.2 shows a metropolitan area configured as a traditional mobile telephone network with one high-power transmitter.

Now a days, the cellular concept has structured the mobile telephone network in a different way. Instead of using one powerful transmitter, many low-power transmitters are placed throughout the coverage area. For example, by dividing a metropolitan region into one hundred different cells with low-power transmitters using 12 channels each, the system capacity theoretically could be increased from 12 channels to 1200 channels using 100 low-power transmitters.

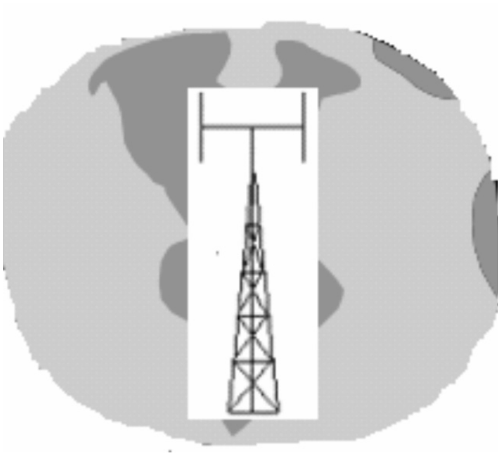


Fig. 13.2 Earlier mobile system

13.8.1 Cellular Concept

The cellular concept is a major breakthrough in solving the problem of spectral congestion and user capacity. It offers very high capacity in a limited spectrum allocation without any major technological changes.

The cellular concept has the following system-level ideas.

1. Replacing a single, high-power transmitter with many low-power transmitters, each providing coverage to only a small area.
2. Neighbouring cells are assigned different groups of channels in order to minimise interference.
3. The same set of channels is then reused at different geographical locations.

In a mobile telephone system, all the channels could not be reused in every cell due to interference problems caused by mobile units using the same channels in adjacent areas. Areas had to be skipped before the same channel could be reused. These interference effects were not due to the distance between areas, but to the ratio of the distance between areas to the transmitter radius of the areas. By reducing the radius of an area by 50%, service providers could increase the number of potential customers in an area. Systems based on areas with a 1 km radius would have 100 times more channels than systems with areas of 10 km in radius. It is to be noted that by reducing the radius of areas to a few hundred metres, millions of calls could be served.

The cellular concept employs variable low-power levels, which allows cells to be sized according to the subscriber density and demand of a given area. As the population grows, cells can be added to accommodate that growth. Frequencies used in one cell cluster can be reused in other cells. Channels can be handed off from cell to cell to maintain constant phone service as the user moves between cells as shown in Figure 13.3.

The cellular radio equipment, or **base station**, can communicate with mobiles as long as they are within range. Radio energy dissipates over distance, so the mobiles must be within

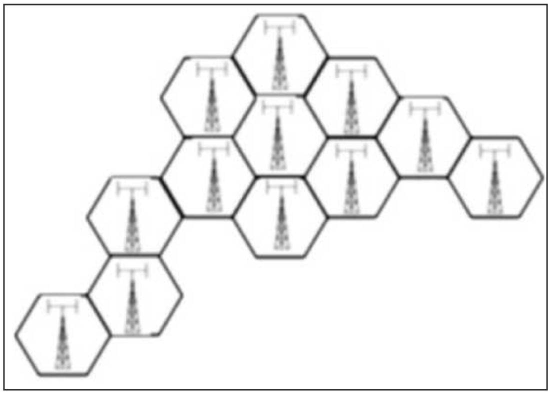


Fig. 13.3 Cellular architecture

the operating range of the base station. Like the early mobile radio system, the base station communicates with mobiles via a channel. The channel is made of two frequencies, one for transmitting to the base station and one to receive information from the base station.

13.8.2 Cells

A cell is the basic geographical unit of a cellular system. The actual radio coverage of a cell is known as the **cell footprint**. The coverage area is divided into honeycomb-shaped areas from which the term 'cellular' is derived. Cells are base stations transmitting over geographic areas represented as hexagons. Each cell size varies depending on the landscape.

Irregular cell structure and irregular placing of the transmitter may be acceptable in the initial system design. However, as traffic grows, where new cells and channels need to be added, it may lead to inability to reuse frequencies because of co-channel interference. For systematic cell planning, a regular shape is assumed for the footprint. Coverage contour should be circular. However, it is impractical because it provides ambiguous areas with either multiple or no coverage. Due to economic reasons, the hexagon has been chosen due to its maximum area coverage and a conventional cellular layout is often defined by a uniform grid of regular hexagons.

13.8.3 Clusters

Any group of cells in smaller areas is known as a cluster. No channels are reused within a cluster. Figure 13.4 illustrates a seven-cell cluster.

Each cluster utilises the entire available radio spectrum. Due to clustering, the adjacent cells cannot use the same frequency spectrum because of interference. So the frequency bands have to be split into chunks and distributed among the cells of a cluster.

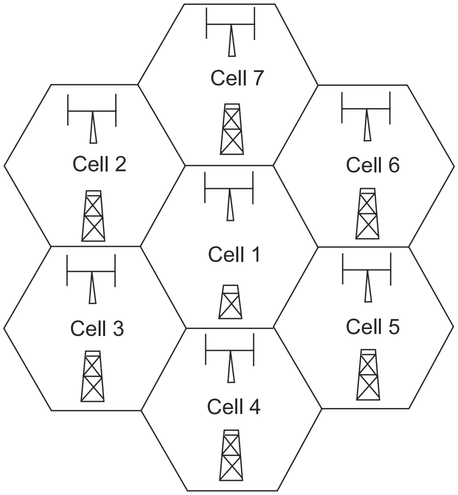


Fig. 13.4 Seven-cell cluster

13.8.4 Frequency Reuse

Due to the availability of a small number of radio-channel frequencies for mobile systems, it is necessary to find a way to reuse radio channels in order to carry more than one channel at a time. The solution to the above problem is called **frequency planning** or frequency reuse.

The concept of frequency reuse is based on assigning to each cell a group of radio channels used within a small geographical area. Cells are assigned a group of channels that is completely different from neighbouring cells. The coverage area of cells is called the footprint. The footprint is limited by a boundary so that the same group of channels can be used in different cells that are far enough away from each other so that their frequencies do not interfere. Figure 13.5 illustrates the concept of frequency reuse.

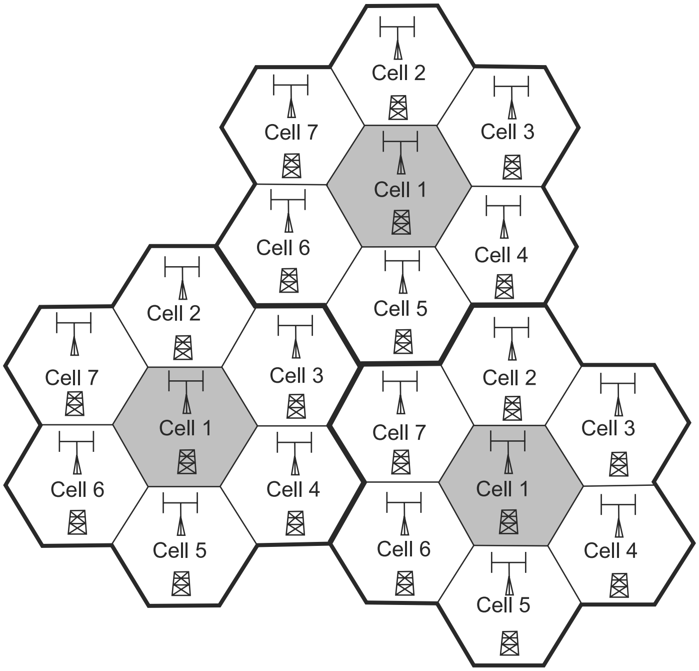


Fig. 13.5 Concept of frequency reuse

Cells with the same number have the same set of frequencies. Here, because the number of available frequencies is 7, the frequency reuse factor is 1/7. That means, each cell is using 1/7 of available cellular channels.

Mathematically, the frequency-reuse concept is illustrated as follows.

Let a system with a fixed number of full-duplex channels available in a given area be denoted as F. Each service area is divided into clusters and allocated a group of channels, which is divided among N cells in a group where all cells have the same number of channels but do not necessarily cover the same size area. Each cell is allocated a group of G channels and F channels are divided among N cells into channel groups, each having the same number of channels.

Thus, the total number of cellular channels available in a cluster can be expressed as

$$F = GN \tag{13.1}$$

where F is the number of full-duplex channels available in a cluster,

G is the number of channels in a cell, and

N is the number of cells in a cluster.

The N cells which collectively use the complete set of available channel frequencies make up the cluster. When a cluster is duplicated M times within a given service area, the total number of full-duplex channels can be expressed as

$$\begin{aligned} C &= MGN \qquad (13.2) \\ &= MF \end{aligned}$$

where C is the total channel capacity in a given area,

M is the number of clusters in a given area,

G is the number of channels in a cell, and

N is the number of cells in a cluster.

From Equation (13.2), the capacity of a cellular system is directly proportional to the number of times a cluster is replicated in a given service area. The factor N is called the **cluster size**.

When the cluster size is reduced and the cell size held constant, more clusters are required to cover a given area and the total channel capacity increases. A large cluster size indicates that the ratio between the cell radius and the distance between co-channel cells is small. The frequency-reuse factor of a cellular telephone system is inversely proportional to the number of cells in a cluster (($1/N$). Each cell within a cluster is assigned $1/N^{\text{th}}$ of the total available channels in the cluster. The value for N is a function of how much interference a mobile or base station can tolerate while maintaining a sufficient quality of communications.

The number of subscribers who can use the same set of channels in non-adjacent cells at the same time in a small area is dependent on the total number of cells in the area. The number of simultaneous users is generally four, but in densely populated areas, that may be significantly higher. The number of users is called the Frequency-Reuse Factor (FRF). It is mathematically defined as follows.

$$\text{FRF} = \frac{W}{C} \tag{13.3}$$

where FRF is the frequency reuse factor,

W is the total number of full-duplex channels in an area, and

C is the total number of full-duplex channels in a cell.

13.8.5 Cell Splitting

It is impractical to apply the cellular concept of creating full systems with many small areas. To overcome this problem, the idea of cell splitting is to be implemented. As a service area becomes full of users, this approach is used to split a single area into smaller ones. In this way, urban centres can be split into as many areas as necessary in order to provide acceptable service levels in heavy traffic regions, while larger, less expensive cells can be used to cover remote rural regions. It is illustrated in Figure 13.6.

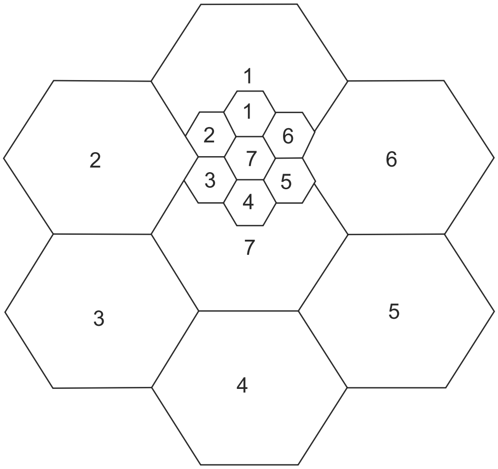


Fig. 13.6 Cell splitting

The purpose of cell splitting is to increase the channel capacity and improve the availability and reliability of a cellular telephone network. The point when a cell reaches maximum capacity occurs when the number of subscribers needing to place a call at any given time equals the number of channels in the cell. This is called **maximum traffic load** of the cell. Splitting cell areas creates new cells, providing an increase in the degree of frequency reuse, thus increasing the channel capacity of a cellular network.

When traffic density starts to build up and the frequency channels *F* in each cell *C* cannot provide enough mobile calls, the original cell can be split into smaller cells. Usually, the new size is one half the older size.

$$\text{New cell area} = \frac{\text{Old cell area}}{4} \tag{13.4}$$

Let each cell carry the same maximum traffic load of the old cell. Then

$$\frac{\text{New traffic load}}{\text{Unit area}} = 4 \times \frac{\text{Traffic load}}{\text{Unit area}} \tag{13.5}$$

Cell splitting increases the number of channels in order to increase capacity and provides an orderly growth in a cellular system. There will be a corresponding reduction in antenna height and transmitter power. Cell splitting accommodates a modular growth capability. This in turn leads to capacity increase, essentially via a system re-scaling of the cellular geometry without any changes in frequency planning.

Small cells lead to more cells/area which in turn leads to increased traffic capacity. For new cells to be smaller in size, the transmit power must be reduced. If $n = 4$ then with a reduction of cell radius by a factor of 2, the transmit power should be reduced by a factor of 2^4.

Theoretically, cell splitting could be repeated indefinitely. But in practice, it is limited

1. By the cost of base stations
2. Handover (fast and low speed traffic)
3. Not all cells are split at the same time: practical problems of channels sites, such as co-channel interference exist
4. Innovative channel-assignment schemes must be developed to address this problem for practical systems

13.8.6 Cell Sectoring

Cell sectoring is another means of increasing the channel capacity of a cellular telephone system to decrease the co-channel reuse ratio while maintaining the same cell radius. Capacity improvement can be achieved by reducing the number of cells in a cluster, thus increasing the frequency reuse. To accomplish this, the relative interference must be reduced without decreasing the transmit power.

In a cellular telephone system, co-channel interference can be decreased by replacing a single omnidirectional antenna with several directional antennas, each radiating within a smaller area. These smaller areas are called sectors, and decreasing co-channel interference while increasing capacity by using directional antenna is called sectoring.

Using this cell-sectoring technique, the signal-to-interference ratio is reduced and this reduces the cluster size, thereby increasing the capacity. Directional antennas are used for the reduction of co-channel interference that results by focusing the radio propagation in only the direction where it is required. A lower frequency reuse factor allows a larger number of channels per cell increasing the overall capacity of the cellular network.

13.8.7 Handover/Handoff

Handover or handoff is a problem in cellular systems and it occurs as a mobile moves into a different cell during an existing call, or when going from one cellular system into another. As adjacent areas do not use the same radio channels, a call must either be dropped or transferred from one radio channel to another when a user crosses the line between adjacent cells. Because dropping the call is unacceptable, the process of handoff was created. Handoff occurs when the mobile telephone network automatically transfers a call from radio channel to radio channels as the mobile crosses adjacent calls.

1. Dwell Time

The time over which a user remains within one cell is called the dwell time. The statistics of the dwell time are important for the practical design of handover algorithms. The statistics of the dwell time vary greatly, depending on the speed of the user and the type of radio coverage.

2. Handover indicator

In the cellular system, each channel constantly monitors the signal strengths of all of its reverse voice channels to determine the relative location of each mobile user with respect to the channel. This information is forwarded to one channel that makes decisions regarding handover. There is a Mobile Assisted Hand Over (MAHO) in which the mobile station measures the received power from surrounding channels and continually reports the results of these measurements to the serving channel.

3. Prioritising Handover

A dropped call is considered a more serious event than call blocking. Therefore, channel-assignment schemes must give priority to handover requests. A fraction of the total available channels in a cell is reserved only for handover requests. However, this reduces the total carried traffic. Dynamic allocation of the channels can solve this problem. Queuing of handover requests is another method to decrease the probability of forced termination of a call due to a lack of available channel. The time span over which a handover is usually required leaves room for queuing handover request.

4. Practical Handover

High-speed users and low-speed users have vastly different dwell times which might cause a high number of handover requests for high-speed users. This will result in interference and traffic-management problems. The umbrella cell approach helps solve these problems. High-speed users are serviced by larger cells, while low-speed users are handled by smaller cells.

13.9 ADVANCED MOBILE PHONE SERVICE (AMPS)

Advanced Mobile Phone Service (AMPS) is released using 800 MHz to 900 MHz frequency band and 30 kHz bandwidth for each channel as a fully automated mobile telephone service. It is the first standardised cellular service in the world and is currently the most widely used standard for cellular communications. It maximises the cellular concept of frequency reuse by reducing radio power output but it includes the following limitations.

1. Low calling capacity
2. Limited spectrum
3. No room for spectrum growth
4. Poor data communications
5. Minimal privacy

Figure 13.7 shows a general block diagram of a typical AMPS unit.

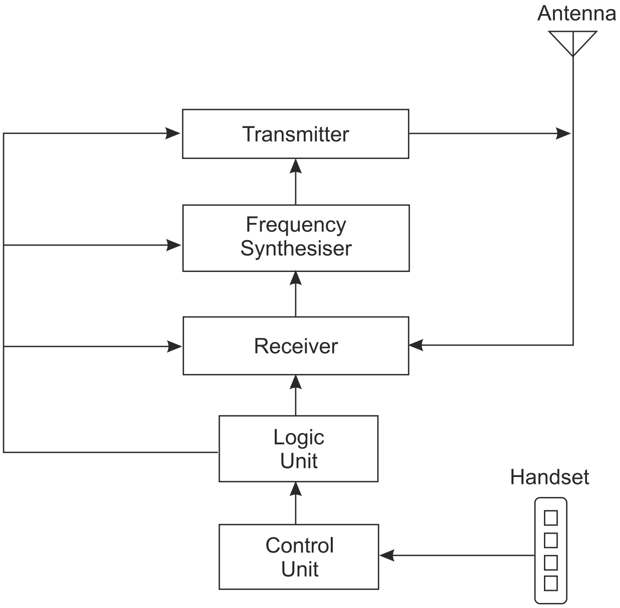


Fig. 13.7 Block diagram of a typical AMPS unit

There are five major sections available in AMPS. They are as follows.

1. Transmitter
2. Receiver
3. Synthesiser
4. Logic unit
5. Control unit

13.9.1 Transmitter

The block diagram of the transmitter of an AMPS is shown in Figure 13.8. It is a low-power FM unit operating in the frequency range of 825 to 845 MHz. Channel 1 is 825.03 MHz, Channel 2 is 825.06 MHz, and so on.

The carrier furnished by a frequency synthesiser is phase-modulated by the voice signal. The phase modulator produces a deviation of ±12 kHz. Pre-emphasis is used to help minimise

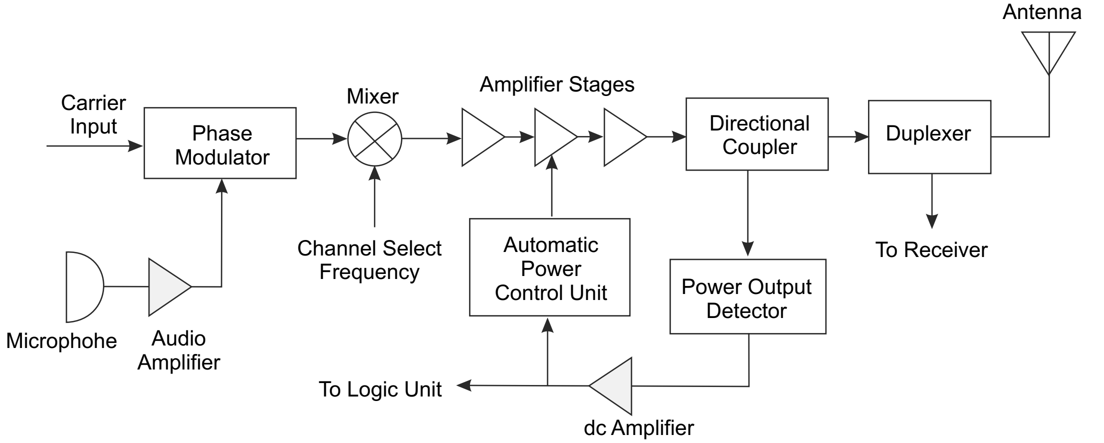


Fig. 13.8 Block diagram of the transmitter of AMPS

noise. The modulator output is translated up to the final transmitter frequency by a mixer whose second input also comes from the frequency synthesiser. The mixer output is fed to Class C power amplifier stages where the output signal is developed. The final amplifier stage is designed to supply about 600 mW to the antenna.

In this transmitter, the output power is controllable by the cell site and Mobile Telephone Switching Office (MTSO). Special control signals picked up by the receiver are sent to an Automatic Power Control (APC) circuit that sets the transmitter to one of the eight power-output levels. The APC circuit can introduce power attenuation by controlling the supply voltage to one of the immediate power-amplifier stages.

The output power of the transmitter is monitored internally by built-in circuits. A microstrip directional coupler taps off an accurate sample of the transmitter output power and rectifies it into a proportional dc signal. This signal is used in the APC circuit and is transmitted back to the cell site, permitting the MTSO to know the current power level. The APC permits optimum cell-site reception with minimal power. It also helps minimise interference from other stations on the same or adjacent cells.

The transmitter output is fed to a duplexer circuit that allows the transmitter and receiver to share the same antenna. Since cellular telephone units use full-duplex operation., the transmitter and receiver operate simultaneously. The transmit and receive frequencies are spaced 45 MHz apart to minimise interference. The duplexer consists of two very sharp bandpass filters, one for the transmitter and one for the receiver. The transmitter output passes through this filter to the antenna.

13.9.2 Receiver

The receiver is typically a dual-conversion superheterodyne. A Radio Frequency (RF) amplifier boosts the level of the received cell-site signal. The receiver frequency range is

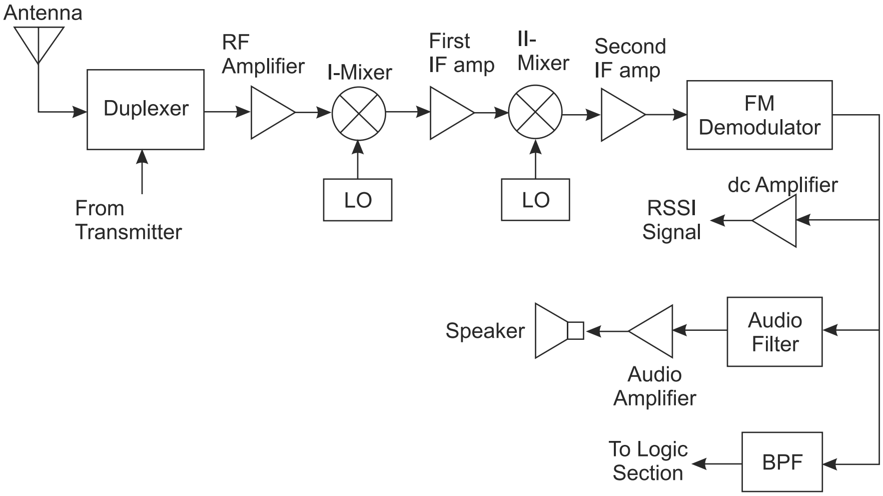


Fig. 13.9 Block diagram of the receiver of AMPS

870.03 to 889.98 MHz. The receive channels are spaced 30 kHz apart. The block diagram of the receiver of an AMPS is shown in Figure 13.9.

From Figure 13.9, the first mixer translates the incoming signal down to a first Intermediate Frequency (IF) of 82.2 MHz. The local oscillator frequency sets the receive channel. The signal passes through IF amplifiers and filters to the second mixer, which is driven by a crystal-controlled local oscillator. The second IF is usually either 10.7 MHz or 455 kHz. The signal is then demodulated, de-emphasised, filtered and amplified before it is applied to the output speaker in the handset.

The output of the demodulator is also fed to the other filter circuits that select out the control audio tones and digital control data stream sent by the cell site to set and control both the transmitter and receiver. The demodulator output is also filtered into a dc level whose amplitude is proportional to the strength of the received signal. This is the ‘Receive Signal Strength Indicator (RSSI) signal’ that is sent back to the cell site so that MTSO can monitor the received signal from the cell and make decisions about switching to another cell.

13.9.3 Frequency Synthesiser

The frequency synthesiser shown in Figure 13.10 develops all the signals used by the transmitter and receiver. It uses standard Phase-Locked Loop (PLL) circuits and a mixer. A crystal-controlled oscillator provides the reference for the PLLs. One PLL incorporates a Voltage-Controlled Oscillator (VCO) whose output frequency is used as the local oscillator for the first mixer in the receiver. This signal is mixed with the output of a second PLL VCO to derive the transmitter output frequency.

The output VCO frequency is determined by the frequency-division ratio of the divider in the feedback path between the VCO and the phase detector. In a cellular radio, this frequency-

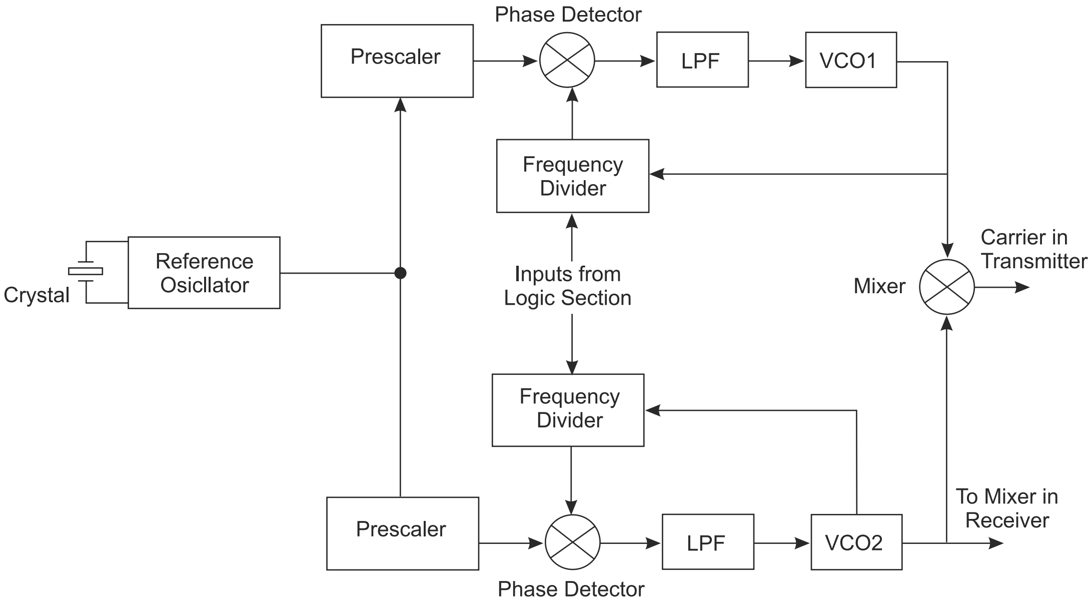


Fig. 13.10 Block diagram of frequency synthesiser

division ratio is supplied by the MTSO via the cell site. When a mobile unit initiates or is to receive a call, the MTSO selects an unused channel. It then transmits a digitally coded signal to the receiver containing the frequency-division ratios for the transmitter and receiver PLLs. This sets the transmit-and-receive channel frequencies.

13.9.4 Logic Unit

The logic unit shown in Figure 13.11 contains the master control circuitry for the cellular radio. It is made up of an embedded microcontroller with RAM and ROM plus additional circuitry used for interpreting signals from the MTSO and cell site and generating control signals for the transmitter and receiver.

All cellular radios contain a Programmable Read-Only Memory (PROM) chip called the Number Assignment Module (NAM). The NAM contains the Mobile Identification Number (MIN), which is the telephone number assigned to the unit. The NAM PROM is burned when the cellular radio is purchased and the MIN assigned. This chip allows the radio to identify itself when a call is initiated or when the radio is interrogated by the MTSO.

All cellular mobile radios are fully under the control of the MTSO through the cell site. The MTSO sends a serial data stream at 10 kbps through the cell site to the radio to control the transmit-and-receive frequencies and transmitter power. The MTSO monitors the received cell-signal strength at the cellular radio by the way of the RSSI signal and monitors the transmitter power level. These are transmitted back to the cell site and MTSO. Audio tones are used for signalling purposes.

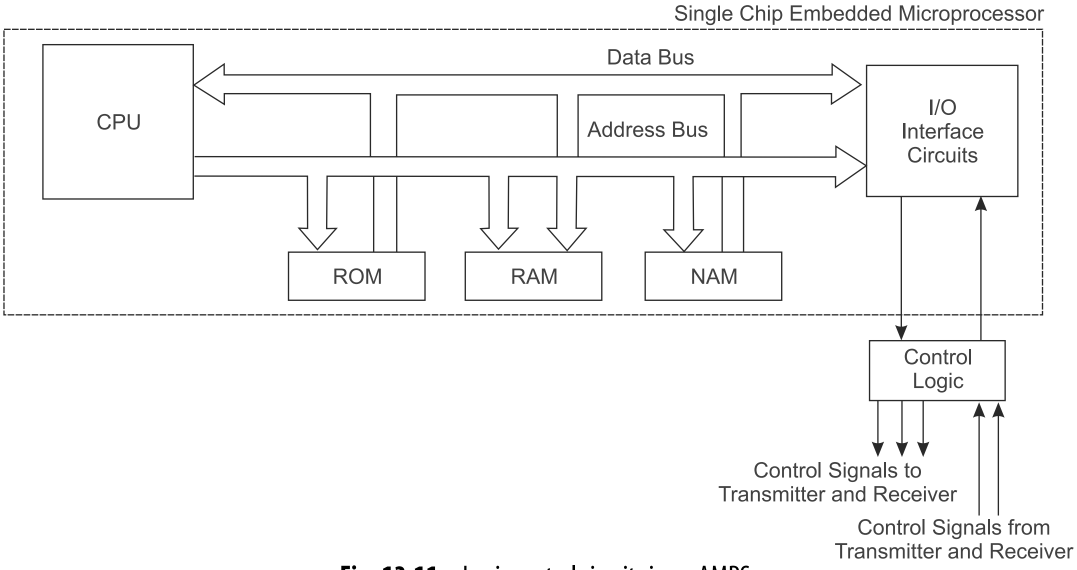


Fig. 13.11 Logic control circuits in an AMPS

13.9.5 Control Unit

The control unit contains the handset with speaker and microphone. This may be a standard handset as used in a regular telephone on a mobile unit. However, these circuits are built into the handheld units. The main control unit contains a complete touchtone dialing circuit as shown in Figure 13.12. The control unit is operated by a separate microprocessor that drives the LCD display and other indicators. It also implements all manual control functions. The microprocessor memory permits storage of often-called numbers and autodial feature.

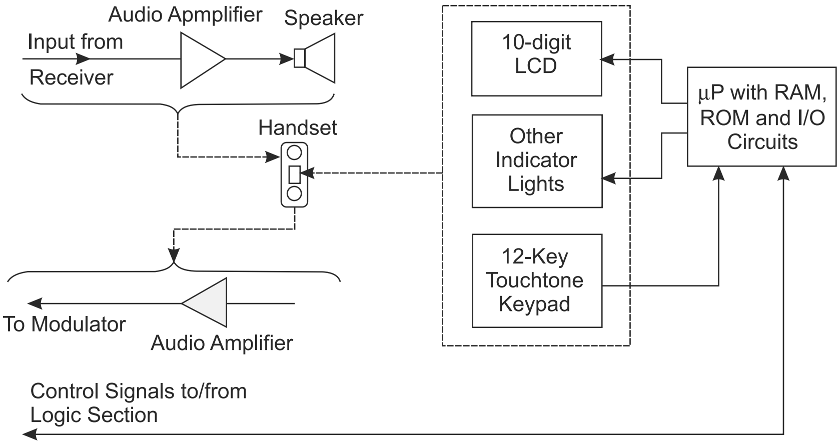


Fig. 13.12 Control unit with handset

13.10 DIGITAL CELLULAR SYSTEM

Due to increasing demand for mobile telephone service, service providers are in need of an alternate to landline networks. While the average landline phone call lasts at least ten minutes, mobile calls usually run 90 seconds and with this capacity, the quality of the service has decreased rapidly. Figure 13.13 shows the components of a typical digital cellular system.

The advantages of digital cellular technologies over analog cellular networks include increased capacity and security, increased digital processing over analog processing, increased integration of circuitry on a few chips and multimode phones.

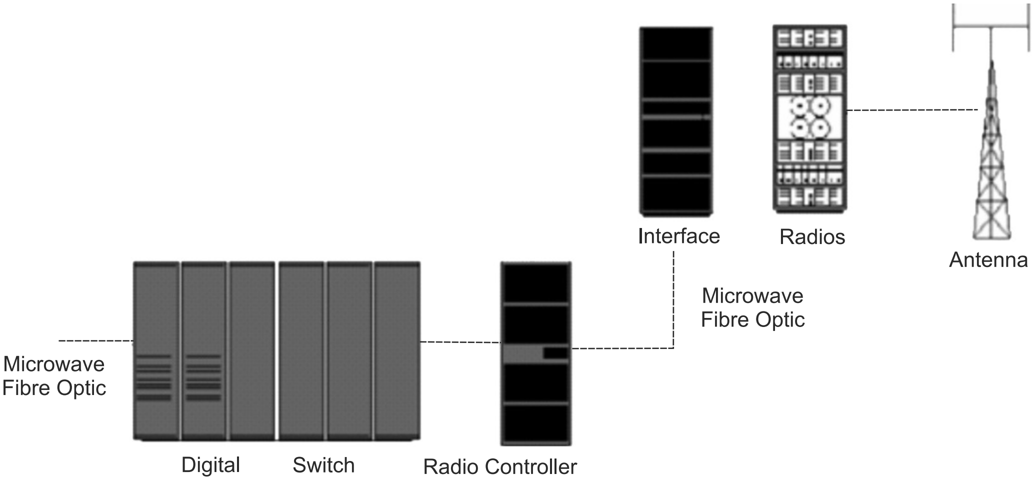


Fig. 13.13 Components of a typical digital cellular system

13.11 MULTIPLE-ACCESS TECHNIQUES

A multiple-access system can be designed by using digital-modulation techniques at the transmitter and the corresponding signal-processing techniques at the receiver. Multiple-access scheme techniques are more tolerant to interference. These are used to allow many mobile users to share a common bandwidth simultaneously. The techniques include

1. Frequency-Division Multiple Access (FDMA)
2. Time-Division Multiple Access (TDMA)
3. Code-Division Multiple Access (CDMA)
4. Spatial-Division Multiple Access (SDMA)

13.11.1 Frequency-Division Multiple Access (FDMA)

Frequency-division multiple-access method assigns individual channels to individual users. Each user is allocated a unique frequency band. These bands are assigned on demand to users who request service. During the period of the call, no other user can share the same frequency band. The bandwidths of FDMA channels are relatively narrow (25–30 kHz) as each channel supports only one call per carrier. Figure 13.14 shows the frequency-division multiple access method.

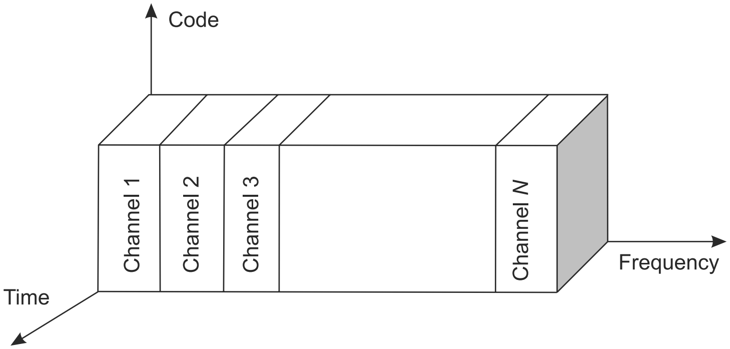


Fig. 13.14 Frequency-division multiple-access method

In this technique, disjoint sub-bands of frequencies are allocated to the different users on a continuous time basis. Guard bands are used to reduce the interface between users allocated adjacent bands. Guard bands act as buffer zones. The fixed assignment of a frequency to a sender makes the scheme very inflexible and limits the number of senders. After the assignment of a voice channel, the base station and the mobile transmit simultaneously and continuously. FDMA is usually implemented in narrowband systems. If an FDMA channel is not in use, it cannot be used by other users to increase the system capacity.

In an FDMA, all users can transmit signals simultaneously and they are separated from one another by their frequency of operation. During the period of the call, no other user can share the same channel and so the channel carries only one phone circuit at a time. Basically, it is built upon Frequency-Division Multiplexing (FDM). FDM is mostly a suitable multiplexing method and it is easier to implement when FDM is used for channel access.

The FDMA mobile unit uses duplexers since both the transmitter and receiver operate at the same time which is known as Frequency-Division Duplexing (FDD). This results in an increase in the cost of FDMA subscriber units and base stations. In FDD systems, the users are assigned a channel as a pair of frequencies in which one frequency is used for forward channel and the other frequency is used for the reverse channel. In FDMA, senders using a certain frequency band can use this band continuously. Also, a guard space is required to avoid frequency-band overlapping and to avoid adjacent channel interference. In both forward and reverse channels, the signal transmitted must be confined within its assigned band. Similar to a wired FDM system, in forward wireless channels, the signal received by all mobile terminals has the same received power and the interference is controlled by adjusting the sharpness of the transmitter and receiver filters for the separate carrier frequencies. On reverse channel, mobile terminals will be operating at different distances from the base stations.

The Receive Signal Strength (RSS) at the base station of a signal transmitted by a mobile terminal close to the base station and the RSS at the base station of a transmission by a mobile terminal at edges of the cell are often substantially different. This problem is said to be **near-far problem**.

The FDMA/FDD system was commonly used in first-generation analog cellular systems like AMPS and some cordless telephones. In this system, forward and reverse channels use different carrier frequencies and a fixed subchannel pair is assigned to a user terminal during the communication session.

13.11.2 Time-Division Multiple Access (TDMA)

Time-division multiple-access systems divide the transmission time into time slots and every user is assigned one or a set of well-defined time slots within a **time frame**. A transmitting user sends its own data only in the designated time slot and waits for the remaining time-frame duration till it gets another time slot in the next frame. Precise time synchronisation among all users is an important and necessary feature of TDMA multiple-access method. Usually, a central unit controls the synchronisation and the assignment of time slots. Figure 13.15 shows the time-division multiple-access system.

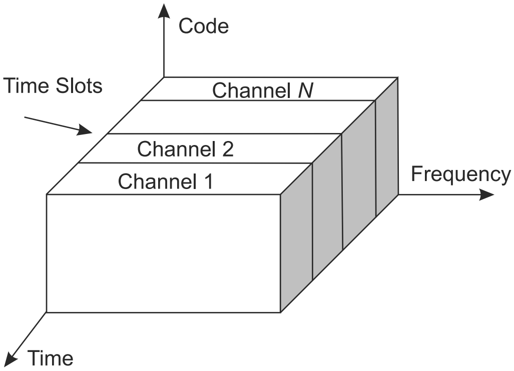


Fig. 13.15 Time-division multiple-access method

In TDMA too, guard spaces are required which represent time gaps, and have to separate the different periods when the senders use the medium. TDMA systems basically transmit the data as a buffer and burst method. So, transmissions for any user are noncontinuous.

In TDMA, a frame consists of a number of slots. A touch frame is made up of preamble information and trail bits. Figure 13.16 shows the bit format of TDMA.

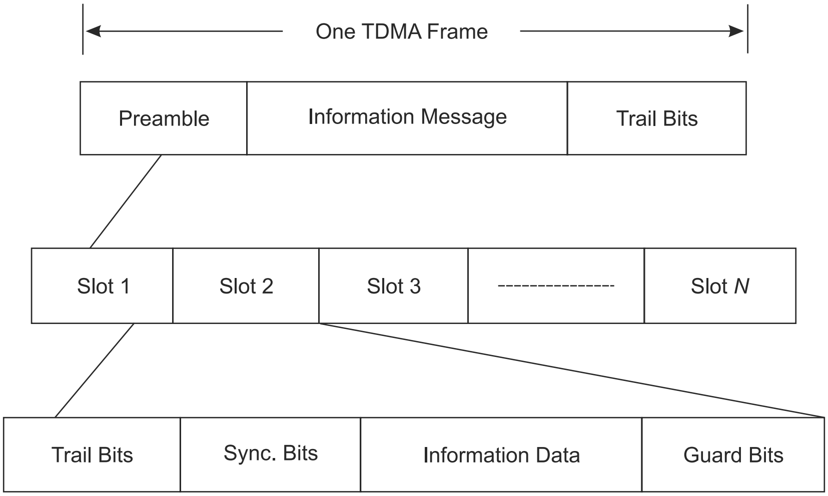


Fig. 13.16 Bit format of TDMA

In a TDMA frame, the preamble contains the address and synchronisation information that both the base station and the subscribers use to identify each other. Guard times are utilised to allow synchronisation of the receivers between different slots and frames. Different TDM wireless standards have different TDMA frame structures.

The following are some of the features of the TDMA method.

1. Data transmission for users of a TDMA system is not continuous, but occurs in bursts. Due to this facility, subscriber transmitter can be turned off when not in use. This results in low battery consumption.
2. The handoff process is much simple for a subscriber unit due to discontinuous transmission in TDMA. During the idle time slots, it is able to listen to offers from other base stations.
3. In TDMA, the number of time slots per frame depends on modulation technique, available bandwidth and some other factors.
4. A duplexer is not required in TDMA, because different time slots are used for transmission and reception.
5. Transmission rates of a TDMA system are high when compared to FDMA channels.
6. In TDMA, the guard time should be minimised. The transmitted signal at the edges of a time slot are suppressed sharply in order to shorten the guard time.

In some cellular systems, digital packets of information are sent during each time slot and reassembled by the receiving equipment into the original voice components. TDMA uses the same frequency band and channel allocations as AMPS. It provides three to six time channels in the same bandwidth as a single AMPS channel. Using digital voice encoders, TDMA is able to use up to six channels in the same bandwidth where AMPS uses one channel.

13.11.3 Code-Division Multiple Access (CDMA)

In Code-Division Multiple-Access (CDMA) systems, the narrowband message signal is multiplied by a very large bandwidth signal called the spreading signal which is a pseudo-noise code sequence. CDMA can accommodate various wireless users with different bandwidth requirements, switching method and technical characteristics without any need for coordination. But each user signal contributes to the interference seen by other users. Power-control techniques are essential in the efficient operation of a CDMA system.

Generally, CDMA is an interference-limited system. It has a soft capacity limit. That means, each user is a noise source on the shared channel and the noise contributed by users accumulates. It will limit to how many users a system will sustain. Mobiles that transmit excessive power increase interference to other mobiles. For CDMA, precise power control of mobiles is critical in maximising the system's capacity and increasing battery life of the mobiles. The goal is to keep each mobile at the absolute minimum power level necessary to ensure acceptable service quality.

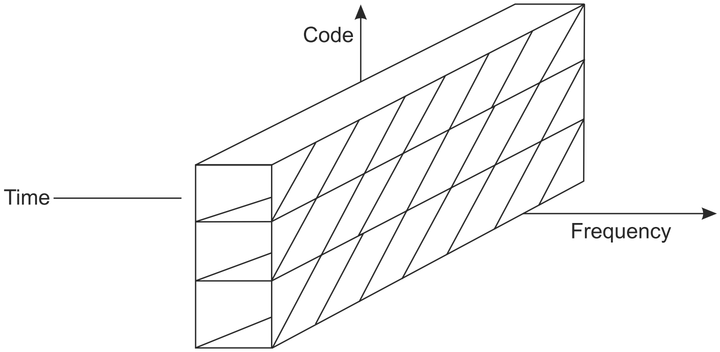


Fig. 13.17 Code-division multiple-access system

13.11.4 Spatial-Division Multiple Access (SDMA)

This type of access method is an extension of frequency reuse. It uses highly directional antennas to pinpoint users and reject others on the same frequency. Figure 13.18 shows the concept of spatial-division multiple-access and in which very narrow antenna beams at the cell-site base station are able to lock in on one subscriber but block another while both subscribers are using the same frequency.

Modern antenna technology using adaptive phased arrays is making this possible. Such antennas allow cellphone carriers to expand the number of subscribers by more aggressive frequency reuse because finer discrimination can be achieved with the antennas. SDMA is also widely used in WLANS and other broadband wireless applications

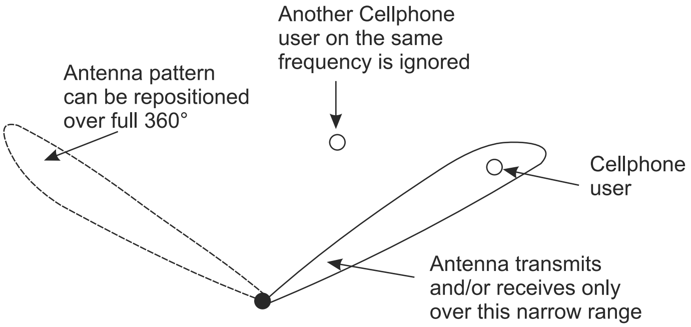

Another Cellphone user on the same frequency is ignored
Antenna pattern can be repositioned over full 360°
Cellphone user
Antenna transmits and/or receives only over this narrow range

Fig. 13.18 Concept of spatial-division multiple access

13.12 WIRELESS LANS

One of the most complex and expensive parts of any Local Area Network (LAN) is the cabling. The cables themselves, as well as their installation and maintenance, are expensive. In addition, LAN needs change regularly due to growth in the organisations. If new users are to be added, the LAN must be reconfigured during expansion, reorganisations and moves. This kind of problem will be solved by using wireless LAN.

Wireless LAN is one in which a mobile user can connect to a Local Area Network (LAN) through a wireless (radio) connection. This technology allows the users to connect several computers together wirelessly, without an entire jungle of cables running everywhere. Due to the lack of physical transmission lines, it is relatively easy for outsiders to tap the network. Therefore, a few simple steps should be taken to protect a WLAN network against unauthorised access. Figure 13.19 shows the difference between wired LAN and wireless LAN working together.

In a large office, access points, switches and routers are stand-alone products. An access point is like a cellphone tower, but its signal distance is measured in feet, not miles. In a large building, users can roam between access points without losing a connection. Wireless LANs have become popular in homes due to ease of installation and the increasing popularity of laptop computers. Public businesses such as coffee shops and malls have begun to offer wireless access to their customers.

13.12.1 Architecture

Wireless LAN architecture includes the following hardware elements.

1. Stations
2. Basic service net
3. Extended service net
4. Distribution system

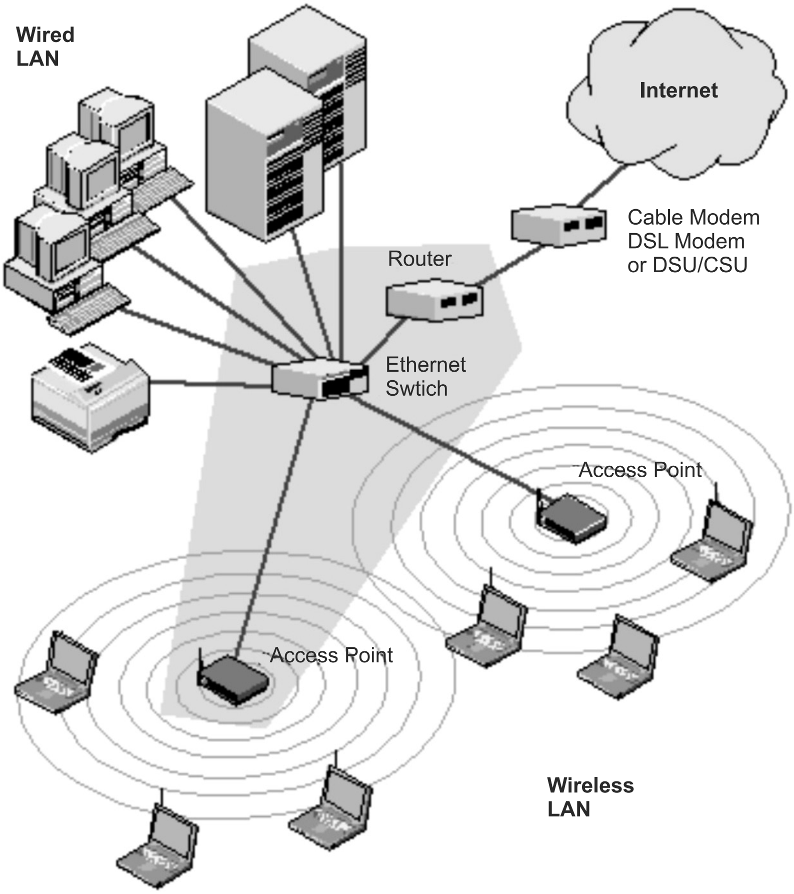


Fig. 13.19 Wired LAN and wireless LAN

1. Stations

All components that can connect into a wireless medium in a network are referred to as stations. All stations are equipped with Wireless Network Interface Cards (WNICs). Wireless stations fall into one of two categories: Access Points (AP) and clients.

(a) **Access Points (APs)** are base stations (routers) for the wireless network. They transmit and receive radio frequencies for wireless-enabled devices to communicate with.

(b) **Wireless clients** can be mobile devices such as laptops, Personal Digital Assistants (PDA), IP phones or fixed devices such as desktops and workstations equipped with a wireless network interface.

2. Basic Service Set (BSS)

The Basic Service Set (BSS) is a set of all stations that can communicate with each other. There are two types of BSS: independent BSS and infrastructure BSS.

(a) An **independent BSS** is an ad-hoc network that contains no access points, which means it cannot connect to any other basic service set.

(b) An **infrastructure BSS** can communicate with other stations not in the same basic service set by communicating through access points.

3. Extended Service Set (ESS)

An Extended Service Set (ESS) is a set of connected BSSes. Access points in an ESS are connected by a distribution system. Each ESS has an ID, called the SSID, which is a 32-byte maximum character string.

4. Distribution System (DS)

A Distribution System (DS) connects access points in an extended service set. The concept of a DS can be used to increase network coverage through roaming between cells.

13.12.2 Types of Wireless LAN

There are two types of wireless LAN.

1. Peer-to-peer LAN
2. Bridges
3. Wireless distribution system

1. Peer-to-peer LAN

A peer-to-peer LAN, also called **ad-hoc wireless LAN**, is a network where stations communicate only peer to peer. There is no base and no one gives permission to talk. This is accomplished using the Independent Basic Service Set (IBSS).

This network allows wireless devices to directly communicate with each other. Wireless devices within range of each other can discover and communicate directly without involving central access points. This method is typically used by two computers so that they can connect to each other to form a network. Figure 13.20 shows a peer-to-peer LAN.

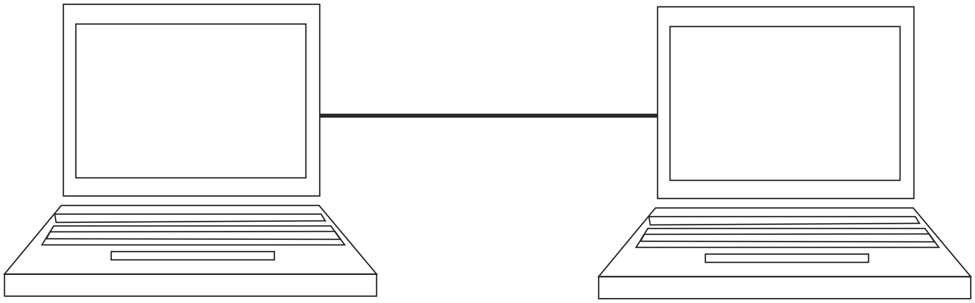


Fig. 13.20 Peer-to-peer LAN

2. Bridges

A bridge can be used to connect networks, typically of different types. A wireless Ethernet bridge allows the connection of devices on a wired Ethernet network to a wireless network. The bridge acts as the connection point to the wireless LAN.

3. Wireless Distribution System (WDS)

A Wireless Distribution System (WDS) is a system that allows a wireless network to be expanded using multiple access points without the need for a wired backbone to link them, as is traditionally required. In WDS, an access point can be either a main, relay or remote base station. A **main base station** is typically connected to the wired Ethernet. A **relay base station** relays data between remote base stations, wireless clients or other relay stations to either a main or another relay base station. A **remote base station** accepts connections from wireless clients and passes them to relay or main stations.

WDS is also referred to as repeater mode because it appears to bridge and accept wireless clients at the same time unlike traditional bridging. When it is difficult to connect all of the access points in a network by wires, it is also possible to put up access points as repeaters.

13.12.3 Roaming

There are two types of wireless LAN roaming.

1. Internal roaming
2. External roaming

Figure 13.21 shows roaming between wireless LAN.

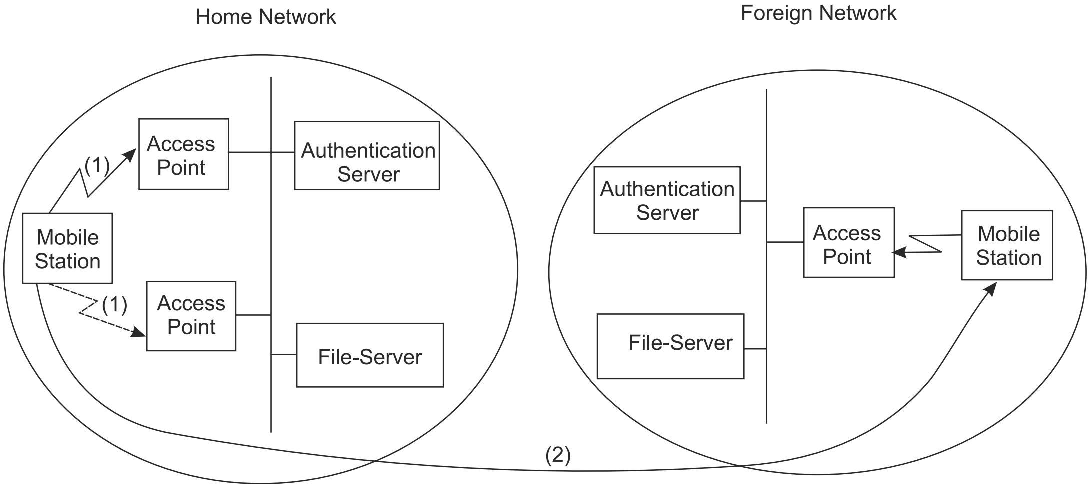


Fig. 13.21 Roaming between wireless LAN

1. Internal Roaming

The Mobile Station (MS) moves from one Access Point (AP) to another AP within a home network because the signal strength is too weak. During roaming, it often interrupts the flow of data between the mobile station and an application connected to the network. The mobile station periodically monitors the presence of alternative access points. At some point, the mobile station decides to re-associate with an access point having a stronger wireless signal. However, the mobile station may lose a connection with an access point before associating with another access point.

2. External Roaming

The MS (client) moves into a WLAN of another Wireless Internet Service Provider (WISP) and takes their services. The user can independently, of his home, network use another foreign network, if this is open for visitors. There must be special authentication and billing systems for mobile services in a foreign network.

13.12.4 WLAN Standards

The standards are designed to ensure that communications equipment from different vendors can inter-operate. The most flexible and reliable WLAN is IEEE 802.11 standard which is available in multiple forms for different needs. The following table shows the different versions of the standard.

IEEE Standard	Frequency (GHz)	Max. Data rate (Mbps)	Max. range (m)
802.11a	5	54	50
802.11b	2.4	11	100
802.11g	2.4	54	100
802.11n	2.4	600	100

IEEE 802.11b is the earliest and most widely adopted version of IEEE 802.11. It operates in 11 channels in 2.4 GHz band and the channels are spaced 5 MHz apart over the spectrum. However, each channel is 22 MHz wide so the channels overlap. Any given AP uses one of these channels.

The **IEEE 802.11a** standard was developed next and it uses a 5 GHz band. The standard provides for back-off data rates as the link conditions deteriorate due to increased range, noise or multipath interference. The major advantage of this standard is that the frequency band is much less used than the busy 2.4 GHz band which contains microwave ovens, cordless phones, Bluetooth wireless and a number of other services, all of which can cause interference at one time or another, thereby producing interference that can block communications or at least decrease the range and data rate.

IEEE 802.11g was developed to extend the data rate within the popular 2.4 GHz band. This standard provides for a maximum data rate of 54 Mbps at 100 feet indoors. This standard also accommodates the 802.11b standards and so is fully backward compatible.

IEEE 802.11n is the newest standard which was developed to further increase the data rate. A primary feature of this standard is the use of Multiple-Input-Multiple Output (MIMO) antenna systems to improve the reliability of the link. APs for 802.11n use two or more transmit antennas and three or more receive antennas. The wireless nodes use a similar arrangement. In each case, multiple transceivers are required for the AP and the node. This arrangement permits a data rate in the 100 to 600 Mbps range at a distance of up to 100 m.

13.13 WIRELESS PERSONAL AREA NETWORK (PAN)

Wireless Personal-Area Network (WPAN) is a personal-area network using wireless connections. It is used for communication among devices such as telephones, computer and its accessories, as well as personal digital assistants within a short range. The distance covered by PAN is typically within 10 metres.

WPAN includes technologies like Bluetooth, ZigBee, Ultra-wideband (UWB), IrDA, Home RF, etc. and among which, Bluetooth is the most widely used technology for WPAN communication. Each technology is optimised for specific usage, applications or domains.

Emerging WPAN standards are focused to provide the following cost-effective and smart wireless applications.

1. Wireless sensor networking
2. RFID applications
3. Mobile ad-hoc networks

WPAN typically involves only two or three nodes, but some systems permit many nodes to be connected in a small area. Although PANs can be wired, today all PANs are wireless. The most popular WPAN system is Bluetooth which is a standard developed by the cellphone company Ericcson for the purpose of cable replacement. The main objective is to provide hands-free cellphone operation by eliminating the cable connecting a cellphone to a headset.

13.14 BLUETOOTH

Bluetooth is an open-standard specification for a Radio Frequency (RF)-based, short-range connectivity technology with an aim of elimination of the need for cables. It is designed to be an inexpensive, wireless networking system for all classes of portable devices, such as laptops, PDAs (Personal Digital Assistants) and mobile phones. It will also enable wireless connections for desktop computers, making connections between monitors, printers, keyboards, and the CPU cable-free. The main purpose of Bluetooth design is to create a small, inexpensive radio chip that could be used in mobile computers, printers, mobile phones, and so on, to transmit data between these devices.

13.14.1 Components of Bluetooth

For a complete Bluetooth system, the following are the required elements.

1. An RF portion for receiving and transmitting data
2. A module with a baseband microprocessor
3. Memory
4. An interface to the host device such as a mobile phone

Among the above elements, the RF portion can be implemented as a module or as a single chip that includes a short-range radio transceiver, an external antenna and a clock reference required for synchronisation. The RF/baseband solution provides the means to communicate with the host, but there is a need to implement a connection interface, as well as any upper-layer protocols, to use applications supported by the final product.

The upper layers of the technology support, called the **Bluetooth profiles**, or a set of protocols is important since it enables interoperability among devices. Requiring a specific profile for devices that provide comparable applications ensures interoperability across a spectrum of devices.

13.14.2 Bluetooth Stack

The hardware that enables wireless communication between devices, building block of this technology, is the Bluetooth stack, that includes the hardware and software portions of the system. Figure 13.22 shows a graphic representation of the stack. This stack contains a physical-level protocol or baseband and a link-level protocol or Link Manager Protocol (LMP) with an adaptation layer enabling upper-layer protocols to interact with the lower layer.

The following are the components of the Bluetooth stack.

1. RF portion for reception and transmission
2. Baseband portion with microcontroller
3. Link-control unit
4. Link manager to support lower-layer protocols
5. Interface to the host device
6. Host processor to support upper-layer protocols
7. L2CAP to support upper-layer protocols

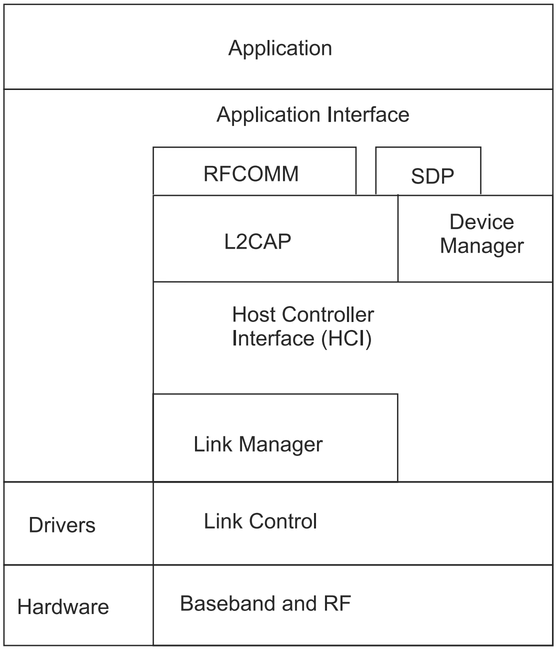


Fig. 13.22 A Bluetooth stack

The Radio Frequency (RF) portion provides the digital signal processing component of the system and the baseband processes these signals. The link controller handles all the baseband functions and supports the link manager. It sends and receives data, identifies the sending device, performs authentication and determines the type of frame to use for sending transmissions.

The link controller also directs how devices listen for transmissions from other devices and can move devices into power-saving modes. The link manager, which is located on top of the link controller, can control set-up, authentication, link configuration and other low-level protocols. A connection for the network is established by the combination of the baseband and the link manager.

The Host Controller Interface (HCI) communicates the lower-layer protocols to the host device. The host contains a processor, the L2CAP, which supports the upper-layer protocols and communicates between upper and lower layers. The upper-layer protocols consist of service-specific applications that must be integrated into the host application.

Another element in the Bluetooth stack that relates to radio communications is the RFCOMM protocol, which allows for the emulation of serial ports over the L2CAP. The Service Discovery Protocol (SDP) provides the means for Bluetooth applications to discover the services and the characteristics of the available services that are unique to Bluetooth. The Bluetooth device manager provides for device inquiry and connection-management services.

13.14.3 Links and Channels

In order to transmit the data between Bluetooth units, links and channels are used. The first step is the establishment of the links.

Bluetooth technology supports two link types:

1. Synchronous connection-oriented (SCO), and
2. Asynchronous connectionless (ACL) links.

The SCO links are used primarily for voice communications. The ACL links are used for packet data. Bluetooth devices can use either link type and can change link types during transmissions, although an ACL link must be established before an SCO link can be used.

After the link has been established, Bluetooth uses five logical channels to transfer different types of information between devices:

1. Link Control (LC) manages the flow of packets over the link interface.
2. Link Manager (LM) transports link-management information between participating stations.
3. User Asynchronous (UA) carries user data.
4. User Isochronous (UI) carries user data.
5. User Synchronous (US) carries synchronous (SCO) data.

13.14.4 Protocols

Bluetooth protocols are sets of conventions that govern the transmission of data in upper and lower layers of the system. The lower-layer protocols pertain to establishing connections and the upper layers correspond to specific types of applications.

The link-control protocol is responsible for delivery of the basic data elements. All packet information is transmitted in a specific time-slot format and specific links are designed to transport a range of data types.

The Bluetooth link-control protocol can be used to manage the associations and delivery of information between the various units within a Bluetooth network. This format is used for both synchronous (voice) and asynchronous (data) modes of operation with specific formats specified for voice transport.

The Link Manager Protocol (LMP) is a command-response system for transmitting data. It transports packets through the Bluetooth baseband link protocol, which is a time-slot-oriented mechanism. LMP packets are limited in size to ensure that they fit into a single time slot. The format of the Protocol Data Unit (PDU) is simple. Two fields are used:

1. The Op-Code identifies the type and sequence of the packet.
2. The content field contains application-specific information.

The protocol sequences are similar to client-server architectures, with the exchange of information following a similar request-response pattern. In general, a single-response PDU is sent upon receipt of the original request. Because Bluetooth is an RF broadcast technology, a set of request messages can be broadcast to all participants on a network. In this case, one request can elicit several responses.

13.14.5 Logical Link and Adaptation Protocol (L2CAP)

Logical Link and Adaptation Protocol (L2CAP) enables transmission of data between upper and lower layers of the stack. It also enables support for many third-party upper-layer protocols such as TCP/IP.

L2CAP provides group management by mapping upper-layer protocol groups to Bluetooth networks. It is also a factor in ensuring interoperability among Bluetooth units by providing application-specific protocols. The following are the other protocols interfacing to the L2CAP.

1. Service Discovery Protocol (SDP) It provides service discovery specific to Bluetooth. That is, one device can determine the services available in another connected device by implementing the SDP.
2. Radio Frequency Communication (RFCOMM) It is a transport protocol that provides serial data transfer. In other words, it enables legacy software applications to operate on a Bluetooth device.
3. Telephony Control Protocol Specification (TCS) It is for voice and data call control. It provides group-management capabilities and allows for signalling unrelated to an ongoing call.
4. Object Exchange Protocol (OBEX) It is a session protocol, and for Bluetooth devices, only connection-oriented OBEX is supported.

13.14.6 Bluetooth Networking

The Bluetooth technology provides both a point-to-point connection and a point-to-multipoint connection. In point-to-multipoint connections, the channel is shared among several Bluetooth units. In point-to-point connections, only two units share the connection.

Bluetooth protocols assume that a small number of units will participate in communications at any given time. These small groups are called **piconets**, and they consist of one master unit and up to seven active slave units. The **master** is the unit that initiates transmissions, and the **slaves** are the responding units. This type of Bluetooth network can have only one master unit. If several piconets overlap a physical area, and members of the various piconets communicate with each other, this new, larger network is known as a **scatternet**. Any unit in one piconet can communicate in a second piconet as long as it serves as master for only one piconet at a time. Figure 13.23 shows the interconnection between units in different points.

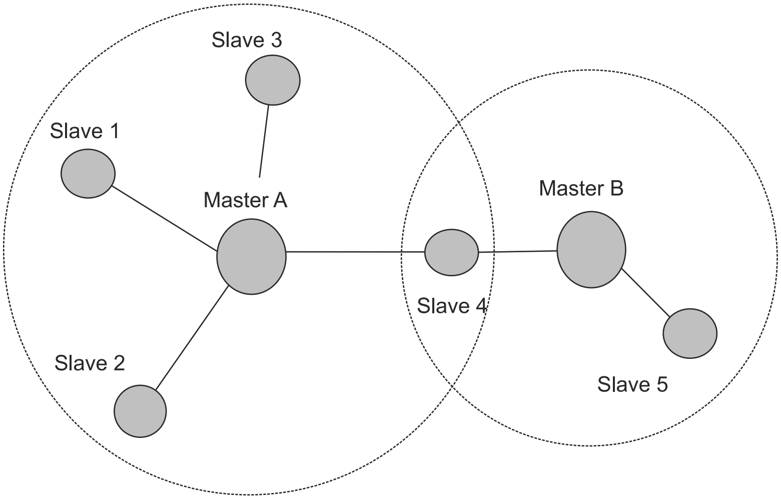


Fig. 13.23 Piconets and scatternets

13.14.7 Bluetooth Connections

Bluetooth is an emerging standard for wireless connectivity. It specifies a system that encompasses the hardware, software framework and interoperability requirements. It also primarily specifies a cable-replacement technology that targets mobile users in the global marketplace.

The major difference between Bluetooth wireless connectivity and the cellular radio architecture is that Bluetooth enables ad-hoc networking. It also relies on terminals and base stations for maintaining connections to the network via radio links, Bluetooth implements peer-to-peer connectivity in which no base stations or terminals are involved.

Using peer-to-peer connectivity, Bluetooth technology simplifies personal area wireless connections, enabling all digital devices to communicate spontaneously. Early applications are expected to include cable replacement for laptops, PDAs, mobile phones and digital cameras. Because Bluetooth supports voice transmissions, headsets are also in line to become wireless. The Bluetooth technology offers the following advantages:

1. Voice/data access points will allow mobile phone/Internet connections.
2. Cable is replaced by a Bluetooth chip that transmits information at a special radio frequency to a receiver Bluetooth chip.
3. Ad-hoc networking enables personal devices to automatically exchange information and synchronise with each other. For example, appointments made on a PDA calendar automatically appear on a desktop calendar as well.

Figure 13.24 shows the three concepts that distinguish Bluetooth technology from other wireless connections.

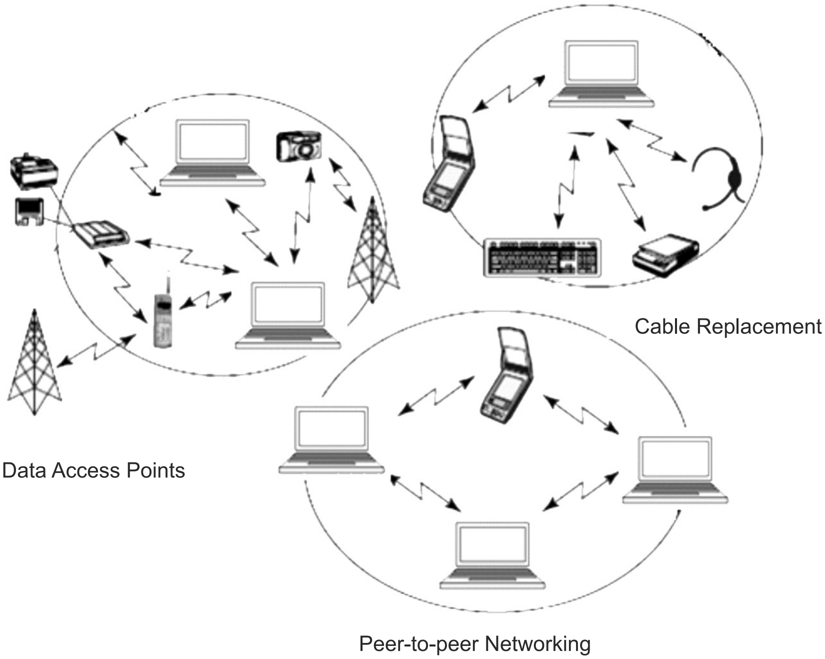


Fig. 13.24 Connecting with Bluetooth

13.14.8 Transmissions by Bluetooth

Bluetooth technology provides fast, secure voice and data transmissions. The range for connectivity is up to 10 metres and line of sight is not required. The following are the characteristics of Bluetooth transmission.

1. It functions even in noisy radio environments, ensuring audible voice transmissions in severe conditions.
2. It protects data by using error-correction methods.
3. It provides a high transmission rate.
4. It encrypts and authenticates for privacy.

13.15 ZIGBEE

ZigBee is an established set of specifications for Wireless Personal-Area Networking (WPAN), i.e. digital radio connections between computers and related devices. A network of this kind eliminates use of physical data buses like USB and Ethernet cables. The devices could include telephones, hand-held digital assistants, sensors and controls located within a few metres of each other.

In industries, ZigBee plays a major role for next-generation automated manufacturing, with small transmitters in every device on the floor, allowing for communication between devices to a central computer. This new level of communication permits finely tuned remote monitoring and manipulation. It was created to address the market need for a cost-effective, standards-based wireless networking solution that supports low data rates, low power consumption, security and reliability. It is the only standards-based technology that addresses the unique needs of most remote monitoring, and control and sensory network applications.

ZigBee targets the application domain of low power, low duty cycle and low data-rate requirement devices. Figure 13.25 shows a block diagram of a ZigBee network with five nodes.

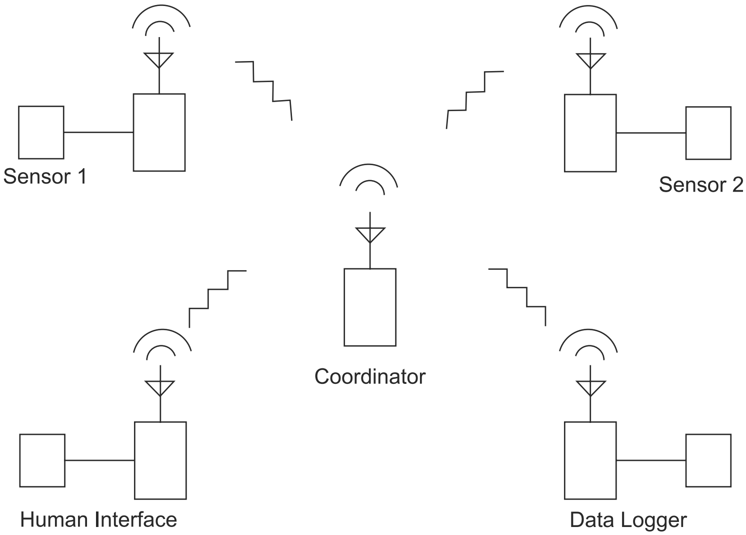


Fig. 13.25 ZigBee network with five nodes

13.15.1 ZigBee Stack

ZigBee is a product of the ZigBee Alliance. The Alliance is an association of companies working together to ensure the success of this open global standard. ZigBee is built on top of the IEEE 802.15.4 standard. ZigBee provides routing and multihop functions to the packet-based radio protocol. Figure 13.26 shows the ZigBee stack.

The ZigBee stack resides on a ZigBee logical device. There are three logical device types. They are

1. Coordinator
2. Router
3. End device

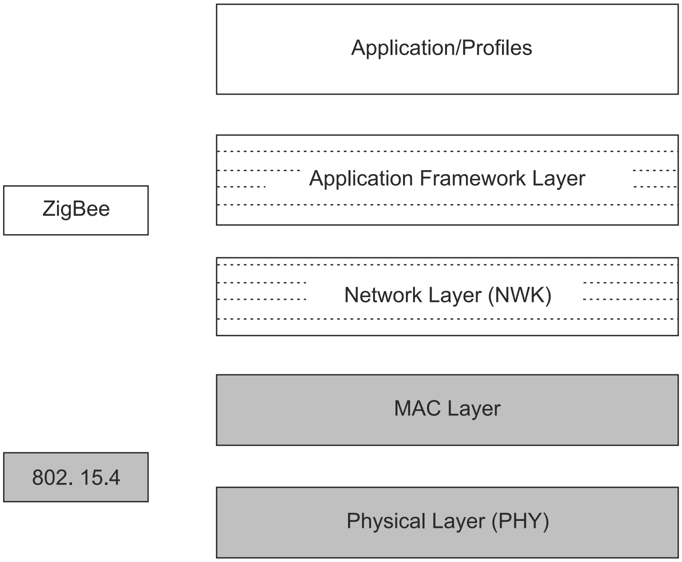


Fig. 13.26 ZigBee stack

It is at the network layer that the differences in functionality among the devices are determined. It is expected that in a ZigBee network, the coordinator and the routers will be mains-powered and that the end devices can be battery-powered.

In a ZigBee network, there is one and only one coordinator per network. The number of routers and end devices depends on the application requirements and the conditions of the physical site. Within networks that support sleeping-end devices, the coordinator or one of the routers must be designated as a Primary Discovery Cache Device. These cache devices provide server services to upload and store discovery information as well as respond to discovery requests.

13.15.2 ZigBee Stack Layers

As shown in Figure 13.25, the stack layers defined by the ZigBee specification are the network-and application-framework layers. The ZigBee stack is loosely based on the OSI 7-layer model. It implements only the functionality that is required in the intended markets.

1. Network (NWK) Layer

The network layer ensures the proper operation of the underlying MAC layer and provides an interface to the application layer. The network layer supports star, tree and mesh topologies. Among other things, this is the layer where networks are started, joined, left and discovered.

Table 13.1 shows the comparison of ZigBee devices at the network layer.

Table 13.1 Comparison of ZigBee devices at the network layer.

ZigBee Network Layer Function	Coordinator	Router	End Device
Establish a ZigBee network	-		
Permit other devices to join or leave the network	-	-	
Assign 16-bit network addresses	-	-	
Discover and record paths for efficient message delivery	-	-	
Discover and record list of one-hop neighbors	-	-	
Route network packets	-	-	
Receive or send network packets	-	-	-
Join or leave the network	-	-	-
Enter sleep mode			-

When a coordinator attempts to establish a ZigBee network, it does an energy scan to find the best RF channel for its new network. When a channel has been chosen, the coordinator assigns the logical network identifier, also known as the **PAN ID**, which will be applied to all devices that join the network.

A node can join the network either directly or through association. To join directly, the system designer must somehow add a node's extended address into the neighbour table of a device. The direct joining device will issue an orphan scan, and the node with the matching extended address (in its neighbour table) will respond, allowing the device to join. To join by association, a node sends out a beacon request on a channel, repeating the beacon request on other channels until it finds an acceptable network to join. The network layer provides security for the network, ensuring both authenticity and confidentiality of a transmission.

2. Application (APL) Layer

The APL layer is made up of several sublayers. The components of the APL layer are shown in Figure 13.27. The ovals symbolise the interface, called Service Access Points (SAP), between different sublayer entities.

(a) Application Support Sublayer (APS) The APS sublayer is responsible for

- Binding tables
- Message forwarding between bound devices
- Group address definition and management
- Address mapping from 64-bit extended addresses to 16-bit NWK addresses

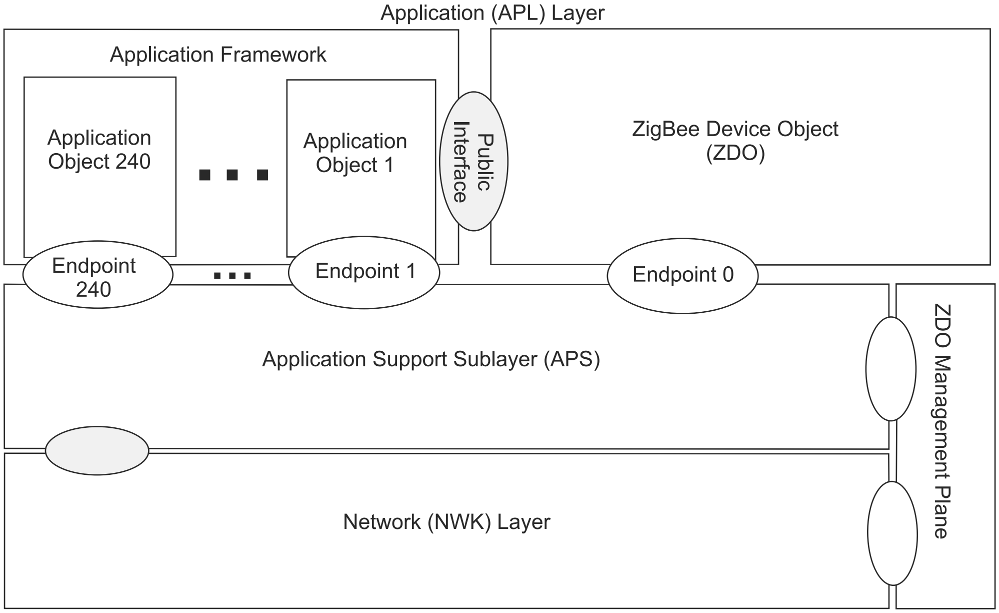


Fig. 13.27 Components of the APL layer

- Fragmentation and reassembly of packets
- Reliable data transport

The key to interfacing devices at the need/service level is the concept of binding. **Binding tables** are kept by the coordinator and all routers in the network. The binding table maps a source address and source end point to one or more destination addresses and end points. The cluster ID for a bound set of devices will be the same.

(b) Application Framework The application framework is an execution environment for application objects to send and receive data. Application objects are defined by the manufacturer of the ZigBee-enabled device. As defined by ZigBee, an application object is at the top of the application layer and is determined by the device manufacturer. An application object actually implements the application; it can be a light bulb, a light switch, an LED, an I/O line, etc. The application profile is run by the application objects.

Each application object is addressed through its corresponding end point. End point numbers range from 1 to 240. End point 0 is the address of the ZigBee Device Object (ZDO). End point 255 is the broadcast address, i.e. messages are sent to all of the end points on a particular node. End points 241 through 254 are reserved for future use.

(c) ZigBee Device Object (ZDO) The ZDO is responsible for overall device management. It is specifically responsible for:

- Initialising the APS sublayer and the NWK layer
- Defining the operating mode of the device (i.e. coordinator, router or end device)

- Device discovery and determination of which application services the device provides
- Initiating and/or responding to binding requests
- Security management

Device discovery can be initiated by any ZigBee device. In response to a device discovery inquiry, end devices send their own IEEE or NWK address. A coordinator or router will send its own IEEE or NWK address plus all of the NWK addresses of the devices associated with it.

Device discovery allows for an ad-hoc network. It also allows for a self-healing network. **Service discovery** is a process of finding out what application services are available on each node. This information is then used in binding tables to associate a device offering a service with a device that needs that service.

13.15.3 ZigBee Addressing

Before joining a ZigBee network, a device with an IEEE 802.15.4-compliant radio has a 64-bit address. When the device joins a ZigBee network, it receives a 16-bit address called the NWK address. Either of these addresses, the 64-bit extended address or the NWK address, can be used within the PAN to communicate with a device. The coordinator of a ZigBee network always has a NWK address of "0." ZigBee provides a way to address the individual components on the device of a node through the use of end-point addresses. During the process of service discovery, the node makes available its end-point numbers and the cluster IDs associated with the end-point numbers. If a cluster ID has more than one attribute, the command is used to pass the attribute identifier.

1. ZigBee Messaging

After a device has joined the ZigBee network, it can send commands to other devices on the same network. There are two ways to address a device within the ZigBee network.

(a) Direct addressing
(b) Indirect addressing

Direct addressing requires the sending device to know three kinds of information regarding the receiving device.

- Address
- End-point number
- Cluster ID

Indirect addressing requires that the above three types of information be committed to a binding table. The sending device only needs to know its own address, end-point number and cluster ID. The binding-table entry supplies the destination address(es) based on the information about the source address. The binding table can specify more than one destination address/end point for a given source address/end-point combination. When an indirect transmission occurs, the entire binding table is searched for any entries where the source address/end-point and cluster ID matches the values of the transmission. Once a matching entry is found, the packet is sent to the destination address/end point. This is repeated for each entry where the source end point/address and cluster ID match the transmission values.

2. Broadcast Addressing

There are two distinct levels of broadcast addresses used in a ZigBee network. One is a broadcast packet with a MAC layer destination address of 0xFFFF. Any transceiver that is awake will receive the packet. The packet is re-transmitted three times by each device, and thus these types of broadcasts should only be used when necessary. The other broadcast address is the use of end-point number 0xFF to send a message to all of the end points on the specified device.

3. Group Addressing

An application can assign multiple devices and specific end points on those devices to a single group address. The source node would need to provide the cluster ID, profile ID and source end point.

13.15.4 ZigBee Network Topologies

Figure 13.28 shows the ZigBee network topologies.

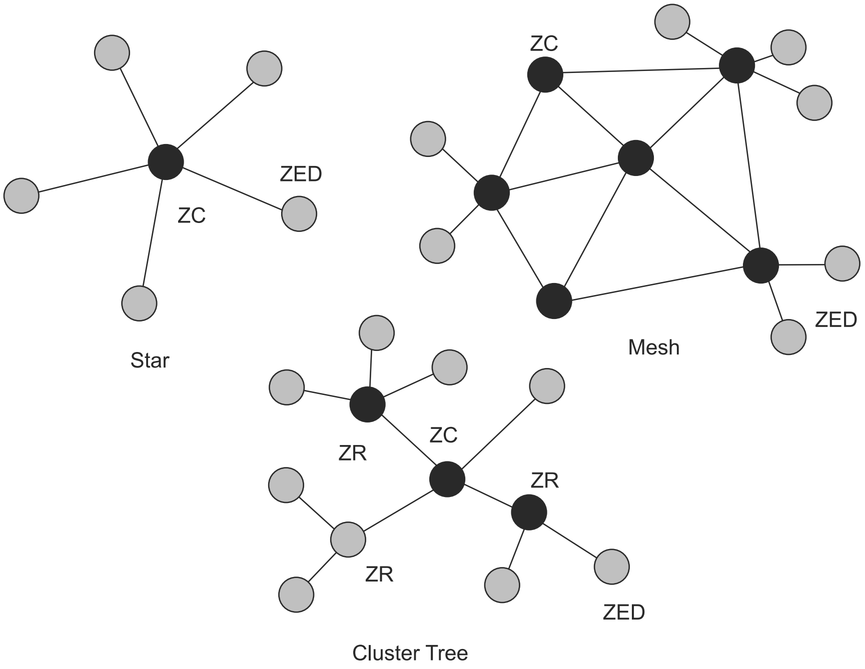


Fig. 13.28 ZigBee network topologies

There are three different network topologies supported by ZigBee.

1. Star,
2. Mesh, and
3. Cluster tree or hybrid networks.

These network topologies are made up of three types of ZigBee nodes.

1. ZigBee Coordinator (ZC)
2. ZigBee Router (ZR)
3. ZigBee End Device (ZED)

The **ZC** initiates a network formation. There is only one ZC per network. The **ZR** serves as monitor or control device that observes a sensor or initiates on/off operations on some end device. It also serves as a router as it can receive data from other nodes and retransmit it to other nodes. The **ZED** is simply an end monitor or control device that only receives data or transmits it. It does not repeat or route. The ZC and ZR nodes are called **Full Function Devices (FFD)** and the ZED is known as a **Reduced Function Device (RFD)**.

In star configuration, a centrally located ZR accepts data from or distributes control data to other ZRs or ZEDs. The central ZR then communications back to the ZC which serves as the master controller for the system.

In mesh configuration, most of the nodes are ZRs which can serve as monitor and control points but also can repeat or route data to and from other nodes. The value of the mesh topology is that it can greatly extend the range of the network. If a node lacks the power or position to reach the desired node, it can transmit its data through adjacent nodes that pass along the data until the desired location is reached. While the maximum range between nodes may be only 30 m or less, the range is multiplied by passing data from node to node over a much longer range and wider area.

Network reliability and robustness are the additional features of mesh topology. If one node is disabled, data can still be routed through other nodes over alternate paths. With redundant paths back to the ZC, a ZigBee mesh ensures the data reaches its destination regardless of unfavourable conditions.

13.15.5 Applications of ZigBee

ZigBee was designed primarily for monitoring and control.

The following are some of the ZigBee applications.

1. Home and office automation
2. Industrial automation
3. Medical monitoring
4. Low-power sensors
5. HVAC control
6. Many other control and monitoring uses

13.16 CALLER ID

Caller ID is simply caller identification or calling-line identification. It is a telephone service, available in analog and digital phone systems. This allows subscribers to screen incoming calls and decide whether the call is to be answered. Caller ID information is sent to the called party by the telephone switch as an analogue data stream using Bell 202 modulation between the first and second rings, while the telephone unit is still on hook. If the telephone call is answered too quickly after the first ring, caller ID information will not be transmitted to the recipient.

There are two types of caller ID.

1. Number-only caller ID
2. Name-and-number caller ID

Number-only caller ID is called Single Data Message Format (SDMF), which provides the caller's telephone number, the date and time of the call. **Name-and-number caller ID** is called Multiple Data Message Format (MDMF) and can provide the directory-listed name for the particular number. Caller ID readers, which are compatible with MDMF, can also read the simpler SDMF format, but an SDMF caller ID reader will not recognise an MDMF data stream, and will act as if there is no caller ID information present, e.g. as if the line is not equipped for caller ID.

Caller ID may be used to track down or limit the impact of telemarketers, prank calls and other intrusions. However, it can also impede communication by enabling users to become evasive. The concept behind caller ID is the value of informed consent. However, it also poses problems for personal privacy.

The caller ID message is a simplex transmission sent from the central office switch over the local loop to a caller ID display unit at the destination station. The caller ID information is transmitted and received using Bell 202 compatible modems. To send caller ID information, the phone company uses an FSK technique identical to a 1,200-baud modem and it sends ASCII character data to the caller ID box. In FSK, one frequency such as 1,200 hertz represents a binary 1, while another frequency such as 2,200 hertz represents a binary zero. A modem changes frequencies depending on whether it needs to send a 1 or a 0. The quickness in changing the frequencies will determine the speed, or baud rate, of the modem. The modem message is sent between the first and second ring.

So the phone rings once, and if it is listened to the phone line just after that ring, a bleep sound will be received about half a second long. If the bleep is decoded, it would find that it contains the following.

- A series of alternating 1s and 0s to help the caller ID box get the timing down
- A series of 180 1s
- A byte representing the type of message

- A byte representing the length of the message
- Month, day, hour and minute, each represented with a pair of bytes
- The 10-digit phone number in 10 bytes
- A check-sum byte

Each character is sent as a standard 8-bit ASCII character preceded by a ‘0’ start bit and followed by a ‘1’ stop bit. The caller ID box contains a modem to decode the bits, a little circuit to detect a ring signal, and a simple processor to drive the display.

To ensure the detection of caller ID signal, the telephone must ring at least twice before being answered. The callers ID signal does not begin until 500 ms after the end of the first ring and it must end 500 ms before the beginning of the second ring. Hence, the caller ID signal has a 3s window in which it must be transmitted. The format for the caller ID is shown in Figure 13.29. The 500 ms delay after the first ringing signal is immediately followed by the channel seizure field, which is a 200 ms long sequence of alternating logic 1s and 0s. A conditioning signal field immediately follows the channel-seizure field. The conditioning signal is a continuous 1200 Hz frequency lasting for 130 ms, which equates to 156 consecutive logic 1 bits.

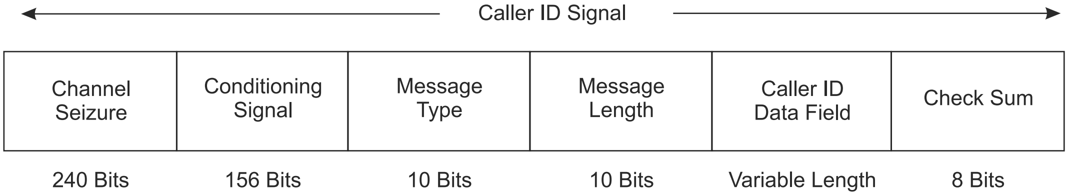


Fig. 13.29 Frame format of caller ID signal

The protocol used for the next three fields such as message type field, message length field and caller ID data field specifies asynchronous transmission of 16-bit characters formed by one start bit (logic 0) and one stop bit (logic 1) for a total of 10 bits per character. The message-type field is comprised of a 16-bit hex code, indicating the type of service and capability of the data message. There is only one message-type field currently used with caller ID. The message-type field is followed by a 16-bit message-length field, which specifies the total number of characters included in the caller ID data field. For example, a message-length code of 15 hex (0001 0101) equates to the number 21 in decimal. Therefore, a message-length code of 15 hex specifies 21 characters in the caller ID data field.

13.17 CORDLESS TELEPHONES

A cordless telephone is a telephone with a wireless handset that communicates via radio waves with a base station connected to a fixed telephone line, usually within a limited range of its base station which has the handset cradle. The base station is on the subscriber premises and attaches to the telephone network the same way a corded telephone does. In these phones, base stations are maintained by a commercial mobile network operator and users subscribe to the service.

Cordless phones first appeared around 1980 and they were quite primitive by today's standards. The earliest cordless phones operated at a frequency of 27 MHz and they had the following problems.

1. Limited range
2. Poor sound quality—noisy and ridden with static because walls and appliances interfered with the signals
3. Poor security—people could easily intercept signals from another cordless phone because of the limited number of channels

In 1986, the Federal Communications Commission (FCC) granted the frequency range of 47–49 MHz for cordless phones, which improved their interference problem and reduced the power needed to run them. However, the phones still had a limited range and poor sound quality. Because the 43–50 MHz cordless-phone frequency was becoming increasingly crowded, the FCC granted the frequency range of 900 MHz in 1990. This higher frequency allowed cordless phones to be clearer, broadcast a longer distance and choose from more channels. However, cordless phones were still quite expensive.

In 1994, digital cordless phones in the 900 MHz frequency range were introduced. Digital signals allowed the phones to be more secure and decreased eavesdropping. In 1995, Digital Spread Spectrum (DSS) was introduced for cordless phones. This technology enabled the digital information to spread in pieces over several frequencies between the receiver and the base, thereby making it almost impossible to eavesdrop on the cordless conversations. In 1998, the FCC opened up the 2.4 GHz range for cordless-phone use. This frequency has increased the distance over which a cordless phone can operate, and brought it out of the frequency range of most radio scanners, thereby further increasing security.

Unlike a corded telephone, a cordless telephone needs mains electricity to power the base station. The cordless handset is powered by a rechargeable battery, which is charged when the handset sits in its cradle. Figure 13.30 shows the components of a cordless telephone set.

A cordless telephone is basically a combination of telephone and radio transmitter/receiver. It has two main parts.

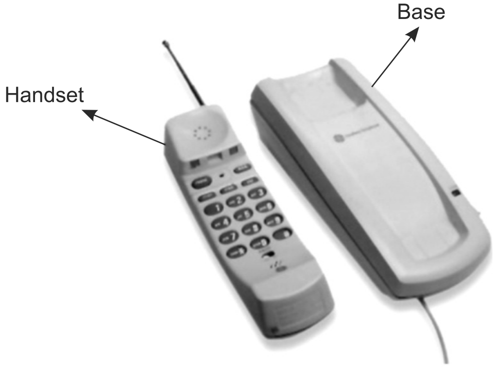


Fig. 13.30 A cordless telephone

1. Base

It is attached to the phone jack through a standard phone-wire connection, and as far as the phone system is concerned, it looks just like a normal phone. The base receives the incoming call through the phone line, converts it to an FM radio signal and then broadcasts that signal.

2. Handset

The handset receives the radio signal from the base and converts it to an electrical signal and sends that signal to the speaker. Then it is further converted into the sound which is the output from the speaker. While talking, the handset broadcasts the voice through a second FM radio signal back to the base. The base receives the voice signal and converts it to an electrical signal and sends that signal through the phone line to the other party.

The base and handset operate on a frequency pair that allows you to talk and listen at the same time, called **duplex frequency**.

13.17.1 Components of the Base

The base unit of the cordless phone is plugged into the telephone jack on the user's wall. There are several components inside the phone base. They are as follows.

1. Phone-line interface components
2. Radio components
3. Power components

1. Phone-Line Interface

The phone-line interface receives and sends telephone signals through the phone-line. Phone-line interface components do two things.

(a) First, phone-line interface components send the ringer signal to the bell or to the radio components for broadcast to the handset. This will lead the user to know that there is an incoming call.

(b) Second, they receive and send small changes in the phone line's electrical current to and from the radio components of the base. During talk, small changes are caused in the electrical current of the phone line and these changes get sent to the caller. The same happens when the caller talks to the user.

2. Radio Components

The radio components receive the electrical signals from the phone-line interface and user controls like keypads and buttons. The radio components convert the signals to radio waves and broadcast them via the antenna. Radio components use quartz crystals to set the radio frequencies for sending and receiving. There are two quartz crystals, one for sending signals and one for receiving signals. The radio components include an audio amplifier that increases the strength of the incoming electrical signals.

3. Power Components

The power components supply low voltage power to the circuits and recharge the battery of the handset. A dc power-cube transformer supplies the low voltage required by the electrical components on the circuit board. The power components on the circuit board work with the power cube to supply electrical current to re-charge the battery of the handset.

In addition to the above components, some bases also have audio amplifiers to drive speakers for speaker-phone features, keypads for dialing, Liquid Crystal Displays (LCDs) for caller ID, Light-Emitting Diodes (LEDs) for power/charging indicators and solid-state memory for answering machine or call-back features.

13.17.2 Components of the Handset

It is possible to carry the handset with the user throughout the house or outside within the range of the base transmitter. The handset has all of the equipment of a standard telephone such as speaker, microphone and dialing keypad plus the equipment of an FM radio transmitter/receiver.

1. Speaker

The speaker receives the electrical signals from the audio amplifier in the radio components and converts them into sound. In the speaker, there is a large round permanent magnet with a hole in the middle and a deep groove surrounding the hole. Within this deep groove is a coil of fine copper wire that is attached to a thin plastic membrane. The plastic membrane covers the magnet and coil.

From the radio components, the electrical signals travel in the coil of the copper wire. These electrical signals induce magnetic currents in the coil, thereby making it an electromagnet. The electromagnetic coil moves in and out of the groove within the permanent magnet. The coil moves the attached plastic membrane in and out at the same frequencies as the changes in electric currents. The movements of the membrane move air at the same frequencies, thereby creating sound waves that can be heard.

2. Microphone

The microphone changes the sound waves from the user's voice into electrical signals that are sent to the audio amplifier of the radio components. A microphone is essentially a speaker that works in reverse. When sound waves from the user's voice move the membrane, they make tiny electric currents either by moving a coil of wire within a magnet or by compressing the membrane against carbon dust.

3. Keypad

The keypad allows the user to dial a number. It transfers the pressure from the user's fingertip on the appropriate key into an electrical signal that it sends to the radio components. Below the rubber keypad is a circuit board with black conductive material under each button. The keypad works like a remote control. When a button is pressed, it makes a contact with the black material and changes its electrical conductance. The conductance sends an electrical signal to the radio components indicating that the number has been selected.

4. Ringer

When the radio components of the handset receive the ringer signal from the base, they send electrical signals to the ringer or buzzer. The buzzer changes those electrical signals into sound much like the speaker does. When the buzzer sound is heard, it is known that someone is calling. In some phones, the speaker is used to make the ringer sound and there is no need for a separate ringer.

13.17.3 Major Features of Cordless Phones

Since the cordless telephone is the combination of a telephone and a radio transmitter/receiver, the following are the issues that a standard cord phone does not have on it.

1. Range
2. Sound quality
3. Security

The **range** is the distance that the handset can be from the base. The **sound quality** can be affected by the distance, the way the information in the radio signal is transmitted and interfering structures such as walls and appliances. **Security** is an issue because the radio signals from both the handset and receiver go over open airways, where they can be picked up by other devices.

13.18 PAGER

A pager is also called a **beeper**, a small telecommunications device that receives and in some cases transmits the alert signals and/or short messages. This type of device is convenient for people expecting telephone calls, but who are not near a telephone set to make or return calls immediately.

A typical one-way pager fits easily in a shirt pocket and some are as small as a wristwatch. A miniature size, short-range wireless receiver captures a message, usually accompanied by a beep. The simplest one-way pagers display the return-call telephone number of the person who sent the message. Alternatively, a code can be displayed that indicates which of the designated party is requesting a return phone call.

Recently, pagers were designed as receive-only devices. A one-way numeric pager can only receive a message consisting of a few digits and typically a phone number that the user is then requested to call. Two-way numeric pagers have the ability to send and receive email, numeric pages and SMS messages.

In general, all pagers are given unique phone numbers while the latest alphanumeric pagers, which are two-way pagers, have an email address usually consisting of the phone number. When calling a phone number assigned to a pager, the calling party reaches a recorded greeting asking the caller to enter a numeric message and sometimes giving the caller an option to leave a voice-mail message. The numeric message given is usually a phone number. Generally, the paged person will receive an alert from the pager with the phone number and/or a pager code within a few minutes. In case of email paging, the text will be displayed.

13.18.1 Types of Pagers

There are several types of pagers. They are as follows.

1. Beepers
2. Voice/Tone pagers
3. Numeric pagers
4. One-way alphanumeric pagers
5. Two-way alphanumeric pagers

1. Beepers

Beepers are the first pagers and they are the simplest form of paging. They are called beepers because they originally made a beeping noise, but current pagers in this category use other forms of alert as well. Some of them use audio signals, some will light up and some will vibrate. The majority of restaurant pagers fall into the vibrant category.

2. Voice/Tone Pagers

Voice/Tone pagers provide the ability to listen to a recorded voice message when an alert is received.

3. Numeric Pagers

Numeric pagers are devices offering only a numeric display of the phone number to be called and pager codes, an informal language wherein sets of numbers symbolise preset messages. Figure 13.31 shows a Motorola numeric pager.

Fig. 13.31 Motorola numeric pager (Courtesy: pager service)

4. Alphanumeric Pagers

Alphanumeric pagers are essentially modified versions of numeric pagers with sophisticated display to accommodate text. These devices are usually given an email address to receive text messages.

5. Two-way Alphanumeric Pagers

Two-way alphanumeric pagers are capable of both sending and receiving text messages and email. To do this, the units either have a small built-in keypad that allows the user to input messages or the message can be typed from a wireless keyboard and is received by the pager. Figure 13.32 shows the earlier and current versions of the numeric pager.

Fig. 13.32 Earlier and current versions of the numeric pager (Courtesy: Unication Paging Company)

6 . Modern Paging Systems

Some of the pager models rely on existing message templates that the user can choose to send back. This has the advantage of increasing speed of a message reply and reducing the chance of miscommunication. In addition, two-way pagers are also offered with Global Positioning System (GPS). GPS allows field agent’s location information to be sent back to a control centre that can use the information to send only location-relevant information and to improve response times by designating jobs or activities only to the closest field personnel.

Most modern paging systems use flexible message delivery by satellite-controlled networks. This type of distributed system makes them inherently more reliable than terrestrial-based cellular networks for message delivery. Many paging transmitters may overlap a coverage area, while cellular systems are built to fill holes in existing networks. When terrestrial networks go down in an emergency, satellite systems continue to perform. Because of superior building penetration and availability of service in disaster situations, pagers are often used by first responders in emergencies.

13.18.2 Paging Operation

A person with a touchtone phone, cellphone or Internet access can send a message to a person with a pager. The page can consist of a numeric message or an alphanumeric message, depending on the type of pager the recipient owns. If the person who owns the pager has an alphanumeric pager, he/she can receive emails and text messages. In order to send an email to an alphanumeric pager, simply log into your email account and address the email to the email address that corresponds with the pager. Type your message and click "Send." To send a text message, use a cellphone or two-way pager.

Many paging network operators now allow numeric and textual pages to be submitted to the paging networks via email. This is convenient for many users, due to the widespread adoption of email; but email-based message-submission methods do not usually provide any way to ensure that messages have been received by the paging network. This can result in pager messages being delayed or lost. Older forms of message submission using the alphanumeric input protocol involve modem connections directly to a paging network, and are less subject to these delays. For this reason, older forms of message submission retain their usefulness for disseminating highly important alerts to users such as emergency services personnel.

13.19 FASCIMILE (FAX)

In addition to basic signals consisting of speech, music or telegraph codes, a telecommunication system is often required to transmit signals of visual nature by a system like Facsimile which is simply called Fax.

Facsimile means an exact reproduction and in Facsimile transmission, an exact reproduction of a document or picture is provided at the receiving end.

13.19.1 Applications of FAX

There are several applications of FAX. Some of them are listed below.

1. Transmission of photographs for the press
2. Transmission of documents
3. Transmission of weather maps
4. Transmission of language texts for which a teleprinter is not suitable

13.19.2 Facsimile Transmitter

The message to be scanned may take any one of the three formats.

1. A single page, which may be a photograph, which is usually wrapped around a cylindrical drum in the sender to permit scanning to take place,
2. Narrow continuous tape
3. Continuous sheet paper, which may be thought of as broad tape

There are two methods of scanning in use:

1. **Optical scanning** in which a light spot traverses the message
2. **Resistance scanning** in which the characters of the message offer varying resistance and these are brought into circular mode by means of a stylus touching and moving over them.

1. Cylindrical Scanning

The message is first fixed around the drum by means of clips. The drum is simultaneously rotated about its axis and made to traverse along it under a fixed scanning spot. The light reflected from the scanning area is focused into a photocell, the electrical output of which represents the signal. The layout of cylindrical scanning is shown in Figure 13.33.

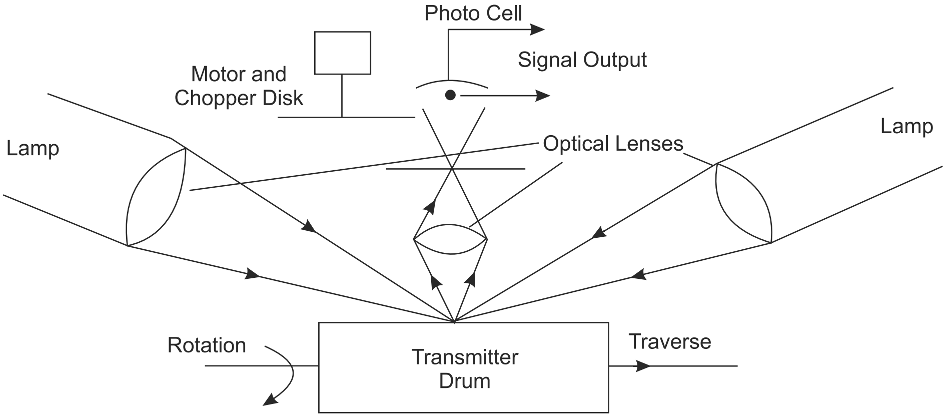


Fig. 13.33 Layout of cylindrical scanning

The chopper disk found in earlier equipment converts the signal into a modulated wave, the carrier frequency being determined by the speed of the disk. The modulated signal is easier to amplify than a direct signal from the photocell. In newer equipment, solid-state amplifiers capable of amplifying the photocell outputs directly are used which eliminates the need of a chopper.

In the usual scanning arrangement, the scanning spot follows a spiral path around the drum. An alternative arrangement is to scan in a series of closed rings, the spot moving from one ring to the next as the fixed clips pass under it.

Typical scanning values are a traversing speed of 1/100 inch per second and a speed of rotation of 60 revolutions per minute. This means that there will be 100 scanning lines on each 1 inch width of picture.

2. Tape Scanning

In this system, the message is taken directly off a printed tape. The scanning beam is arranged to travel across the tape at right angles to the direction of tape travel. The layout of tape scanning is shown in Figure 13.34.

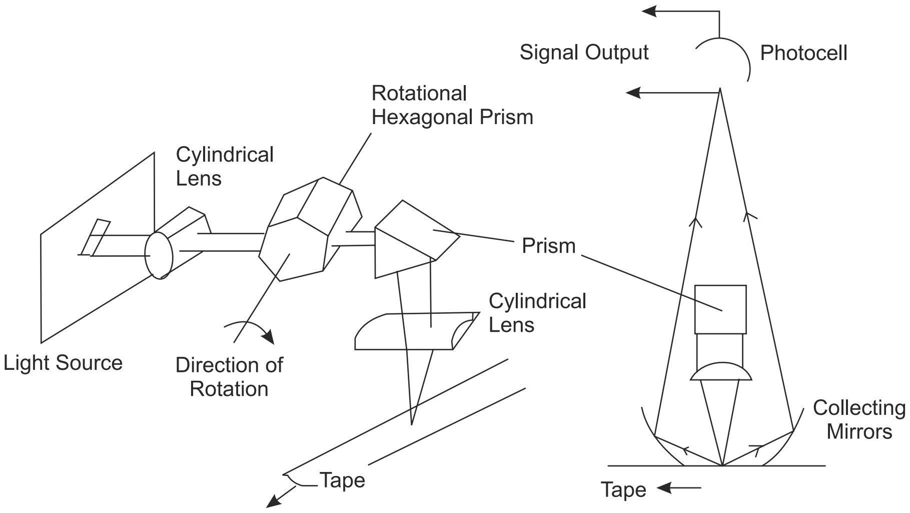
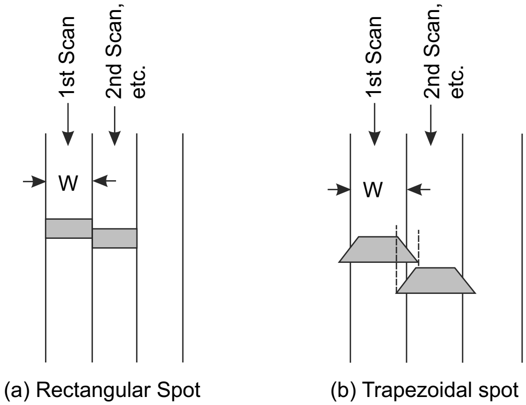


Fig. 13.34 Tape-scanning layout

The light beam will leave the hexagonal prism parallel to the incident beam. But its position will be deflected as the prism rotates. The spot thus travels across the tape and starts a new scan each time a new face of the prism intercepts the incident beam. Thus, the scanning of wide tape has been carried out in practice, but the difficulties encountered have limited its application.

3. Scanning Spot

The shape of the scanning spot is important as it determines the wave shape of the signal output. The rectangular-shaped scanning spot is preferred. This is arranged so that there is no overlap or gap at the sides as shown in Figure 13.35 (a). Less frequently, a trapezoidal-shaped spot, as shown in Figure 13.35 (b), is used and this is arranged such that the average length of the top and bottom sides is equal to the scanning width.



(a) Rectangular Spot (b) Trapezoidal spot

Fig. 13.35 Scanning spot shapes

13.19.3 Facsimile Receiver

The mechanical aspects of scanning in the receiver are similar to those in the sender and very often identical equipment is used at both ends. Scanning in the receiver must produce an optical output from the electrical input signal, the reverse of what happens in the transmitter. In order to have the correct relationship to the transmitted signal, it is necessary for the signal to be synchronised, to be phased correctly and to have same height/width ratio.

1. Synchronisation

Where the message is documentary, it is sufficient to use synchronous motors for both sender and receiver and operate frequency-controlled supply mains. On the other hand, where picture transmission is involved, a synchronising signal must be sent and this by international agreement has a frequency signal of 1020 Hz. The sender speed bears a known relationship to this and the receiver speed is adjusted by means of a stroboscope to correspond to the relationship. An accuracy of about 1 in 10^5 can be obtained in this way.

With carrier transmission, it is necessary to send the carrier along with the side band transmitted. The carrier being present enables the exact 1020 Hz synchronising signal to be recovered. This is an added requirement of Facsimile transmission, as with normal telegraph signals it is not necessary to send the carrier, a local oscillator at the receiver being adequate for recovery of the signal. The effect of a constant speed error is shown in Figure 13.36.

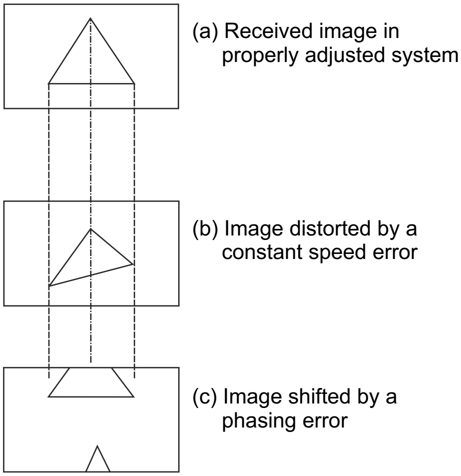

(a) Received image in properly adjusted system
(b) Image distorted by a constant speed error
(c) Image shifted by a phasing error

Fig. 13.36 Facsimile distortion

2. Phase

Correct phasing is necessary to ensure that the image of the clips holding the paper to the drum does not intersect the transmitted picture. Phasing adjustment need only be made once for each picture transmitted and is carried out as follows.

The operator at the receiver first adjusts the speed to correct value by means of the synchronising signal and then sets the drum in the correct start position. This is held in position by means of a switch. At the sender, a pulsed signal of 1020 Hz is sent to indicate the start of the transmission and the pulse releases the switch holding the receiver drum. The effect of incorrect phasing is shown in Figure 13.36 (c).

3. Index of Cooperation

It is defined as the ratio of the diameter of the sending drum to the scanning pitch of the senders and is normally abbreviated as IOC. The height/width ratio must be the same for both the transmitted and received pictures and this in turn depends on the scanning pitch and the diametres of the drums used in the sender and the receiver.

The width of the transmitted picture is *nP* where *P* is the scanning pitch of the sender and *n* is the number of lines scanned as shown in Figure 13.37.

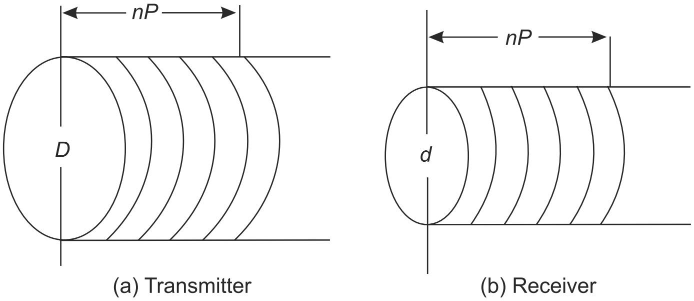


Fig. 13.37 Index of cooperation

The height of the transmitted picture is proportional to the diameter *D* of the sending drum and that of the received picture to the diameter *d* of the receiver by the same constant. Therefore, for maintaining correct height/width ratio,

$$\frac{D}{nP} = \frac{d}{nP} \tag{13.6}$$

where *p* is the scanning pitch of the receiver.

$$\frac{D}{P} = \frac{d}{P} \tag{13.7}$$

Thus, the ratio of the diameter to the scanning pitch is the same for the sender and the receiver.

$$\therefore \qquad \text{IOC} = \frac{D}{P} \tag{13.8}$$

The effect of having different indexes of cooperation at the receiver and the transmitter are illustrated in Figure 13.38.

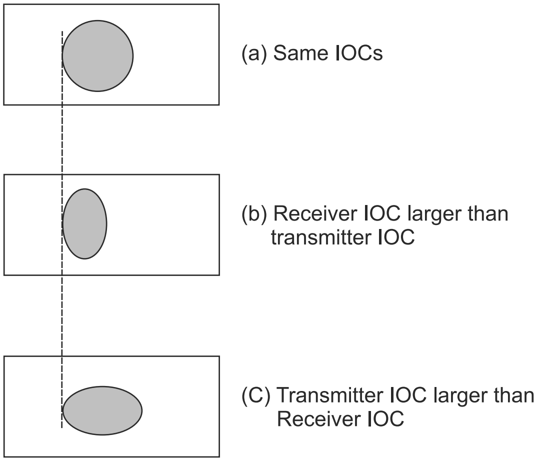

(a) Same IOCs
(b) Receiver IOC larger than transmitter IOC
(C) Transmitter IOC larger than Receiver IOC

Fig. 13.38 Effect of different indexes of cooperation

13.19.4 Conversion of Electrical Signals to an Optical Image

When the electrical signal is received, it must be converted to an optical image and this may be achieved either through photographic reception or through direct record reception.

1. Photographic Reception

In this method, the received signal is used to control the intensity of a light beam, which in turn scans the photographic material. Figure 13.39 shows the essential features of a **Duddell mirror oscillograph**, which may be used for the photographic reception.

A small coil of the fine wire is suspended in a strong magnetic field. Mounted on the coil is a very small mirror, the dimensions of which are about 0.033 inches × 0.06 inches. The loop and the mirror, which form the movement of the system, offer negligible inertia. The basic signal representing the message from the sender is passed through the loop and causes it to deflect, the angle turned through being proportional to the signal current. A beam of light focused onto the mirror is reflected through the various apertures and lenses on to the photographic paper or film on the drum. It can then be arranged that maximum deflection of the mirror corresponds to maximum light through the aperture if photographic paper is to be used and, therefore, lesser degrees of deflection put lesser amounts of light onto the photographic paper. Alternatively, if film is to be used to produce negatives, minimum deflection of the mirror can be arranged to give maximum light through the aperture.

Photographic positives are usually made, since these can be processed and finished much faster and speed is usually the important factor. Negatives have some advantages in that more copies can be obtained from the original and are more sensitive. Hence, they can be produced with lower power of the lamp. Negatives also permit retouching, thus improving the quality of the final picture.

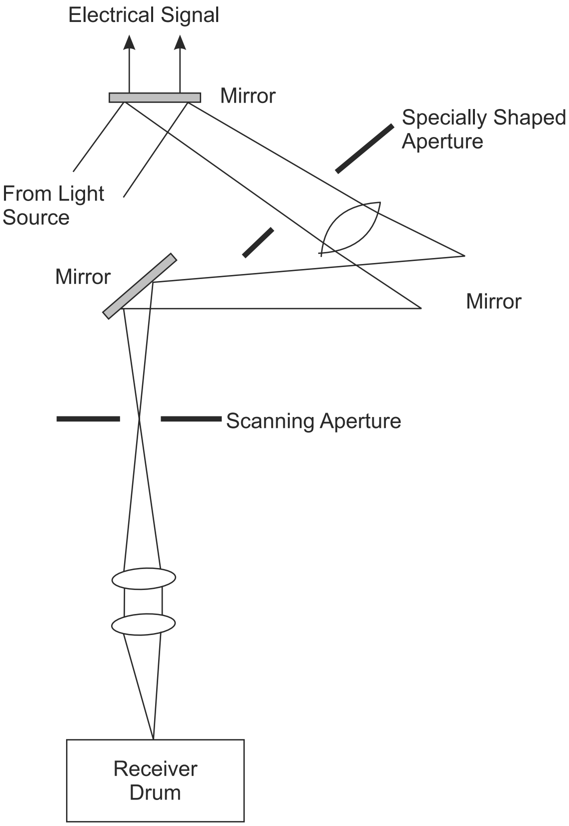


Fig. 13.39 Photographic reception

2. Direct Recording Reception

There are two methods of direct recording reception. In one method, a highly absorbent, chemically treated paper is used in which the electrolyte held by the paper dissociates when a voltage is applied across it. The signal voltage is applied to the paper via a metal stylus, producing dissociates, one of which is a metallic salt. This, in turn, reacts with a colour chemical in the paper, which produces a mark on the paper. The intensity of the mark depends on the signal voltage. A steel stylus is often used, as this produces a very intense black colouration.

The paper used is damp and must be kept in sealed containers. It has a lifetime of about one month after opening. It is reasonably cheap, but the total range is much less than that obtained with photographic methods. It is usually found to be adequate for low-grade work.

A second form of the method employs a resistance paper known commercially as **telephone paper**. This contains a metallised baking on which is deposited a substance similar to carbon black and on top of this, there is a very thin layer of insulation. A stylus exerts a study pressure on the paper and when the signal voltage is applied, burning occurs, which causes blackening of the paper. The total range is similar to the previous method, but definition is not so good and the paper is fairly expensive.

13.19.5 Transmission of Facsimile Telegraph Signals

The bandwidth requirements for the transmission of Facsimile telegraph signals are as given in Figure 13.40. The bandwidth requirements may be found by a method similar to that used for telegraph signals. The worst condition will be assumed that in which alternate black and white squares occur, each square being the width of the scanning pitch. Each of these squares is a minimum-sized picture element or **pixel**. It follows that the vertical resolution needs be no better that the horizontal resolution as shown in Figure 13.40 (a). The ideal output wave would be produced using a scanning slit as shown in Figure 13.40 (b). However, the available output from a slit is very small and a compromise must be reached between the output and desired waveform. A square scanning spot would produce a triangular waveform as shown in Figure 13.40 (c) which would give maximum output.

The triangular waveform is not the best and a compromise is reached using a rectangular scanning spot shown in Figure 13.40 (b) in which the height is about 0.8 times the width. The number of pixels along one circumference of the drum is

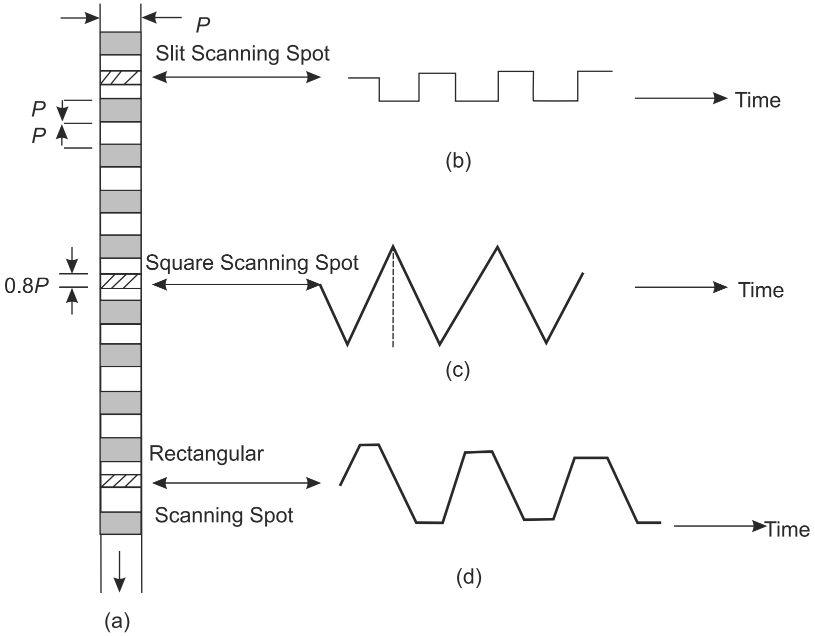


Fig. 13.40 (a) A one-line scan with equal horizontal and vertical resolution (b) Signal produced by slit scanning, (c) Signal produced by square scanning, (d) Signal produced by rectangular scanning midway between (b) and (c)

$$\text{Pixels per scan} = \frac{\pi D}{P} \tag{13.9}$$

and, therefore, this is also the number of pixels scanned in 1 second.

$$\text{Therefore, Pixel rate} = \frac{\pi D n}{P} \tag{13.10}$$

Every two pixels form one cycle of output and, therefore, the output frequency is expressed as follows.

$$f = \frac{\pi D n}{2P}\ \text{Hz} \tag{13.11}$$

This only gives the fundamental frequency of the trapezoidal waveform shown in Figure 13.41 (d).

The bandwidth is determined by the range of frequencies to be transmitted in an actual signal. The equation for *f* gives the highest frequency and as the lowest frequency is very close to zero, the bandwidth is approximately *f* Hz.

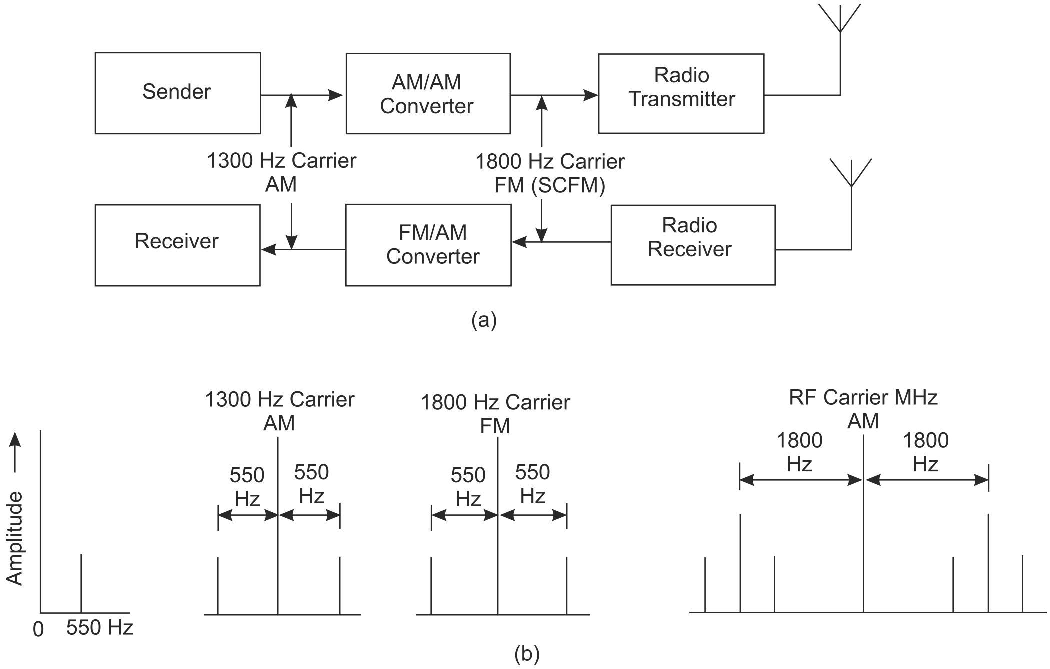


Fig. 13.41 (a) Subcarrier frequency modulation system for transmitting Facsimile signals (b) Frequency spectra of signal at various points in the system

1. Line Transmission

The basic signal, as obtained from the information scanned, is not suitable for direct transmission since it is difficult to amplify the low frequencies involved. AM or FM modulation is, therefore, employed. This also allows FDM to be used and two Facsimile channels can be fitted in the normal 300 to 3400 Hz telephone channels. The carrier frequencies for AM are 1300 Hz and 1900 Hz. All forms of distortion and interference must be kept at a very low level and SNR of at least 35 dB is recommended.

Echo signals must also be avoided and because of echo signals, long lines are not suitable for phototelegraphy. The gain stability of amplifiers must also be high. Level changes and impulsive-type noise affect AM more than FM, and the latter method, known as **subcarrier FM** is preferred. However, it must be limited to narrowband FM.

2. Radio Transmission

The main difficulty with radio transmission is fading which can occur over the radio path and can completely destroy the pictorial transmission. Although special methods are in use to compensate for fading in normal telegraph and telephone radio services, these are not satisfactory for picture telegraphy. As a result, Sub Carrier FM (SCFM) is used over some part of the transmission, as it is less affected by amplitude fading. A simplified block diagram is shown in Figure 13.41 (a).

The frequency spectra of the various sections in the transmission chain are shown in Figure 13.41 (b). Any fading over the radio path, provided that it is not frequency selective, will not affect the FM. The transmitter/receiver equipment is shown as producing AM and in the case of AM/FM, FM/AM converters are required.

Summary

Wireless communication is the transfer of information over a distance without the use of wires. The distances involved may be from a few metres to thousands or millions of kilometres. Wireless operations permit services, such as long-range communications, that are impossible to implement with the use of wires.

Wireless communication can be performed by any of the following.

- Radio Frequency (RF) communication
- Microwave communication, for example long-range line-of-sight via highly directional antennas, or short-range communication
- Infrared (IR) short-range communication

The cellular concept offers very high capacity in a limited spectrum allocation without any major technological changes. The cellular concept has the following system-level ideas.

- Replacing a single, high-power transmitter with many low-power transmitters, each providing coverage to only a small area.
- Neighbouring cells are assigned different groups of channels in order to minimise interference.

- The same set of channels is then reused at different geographical locations.

A cell is the basic geographical unit of a cellular system. The actual radio coverage of a cell is known as the cell footprint The coverage area is divided into honeycomb-shaped areas from which the term 'cellular' is derived. Cells are base stations transmitting over geographic areas represented as hexagons. Each cell size varies depending on the landscape. Any group of cells in smaller areas is known as a cluster. No channels are reused within a cluster.

The concept of frequency reuse is based on assigning to each cell a group of radio channels used within a small geographical area. The purpose of cell splitting is to increase the channel capacity and improve the availability and reliability of a cellular telephone network.

Cell sectoring is another means of increasing the channel capacity of a cellular telephone system to decrease the co-channel reuse ratio while maintaining the same cell radius. Capacity improvement can be achieved by reducing the number of cells in a cluster, thus increasing the frequency reuse.

Handover, or handoff, is a problem in cellular systems and it occurs as a mobile moves into a different cell during an existing call, or when going from one cellular system into another. In Mobile Assisted Handover (MAHO), the mobile station measures the received power from surrounding channels and continually reports the results of these measurements to the serving channel.

Advanced Mobile Phone Service (AMPS) is released using 800 MHz to 900 MHz frequency band and 30 kHz bandwidth for each channel as a fully automated mobile telephone service. It is the first standardised cellular service in the world and is currently the most widely used standard for cellular communications.

Multiple-access-scheme techniques are used to allow many mobile users to share simultaneously a common bandwidth. The techniques include

- Frequency Division Multiple Access (FDMA)
- Time-Division Multiple Access (TDMA)
- Code-Division Multiple Access (CDMA)
- Spatial-Division Multiple Access (SDMA)

Wireless LAN is one in which a mobile user can connect to a Local Area Network (LAN) through a wireless (radio) connection. This technology allows the users to connect several computers together wirelessly, without an entire jungle of cables running everywhere.

Wireless Personal-Area Network (WPAN) is a personal-area network using wireless connections. It is used for communication among devices such as telephones, computer and its accessories, as well as personal digital assistants within a short range. The distance covered by PAN is typically within 10 metres.

Bluetooth is an open-standard specification for a Radio Frequency (RF)-based, short-range connectivity technology with an aim of elimination of the need for cables. The main purpose of Bluetooth design is to create a small, inexpensive radio chip that could be used in mobile computers, printers, mobile phones, and so on, to transmit data between these devices.

ZigBee is an established set of specifications for wireless personal-area networking, i.e. digital radio connections between computers and related devices. A network of this kind eliminates use of physical data buses like USB and Ethernet cables. The devices could include telephones, hand-held digital assistants, sensors and controls located within a few metres of each other.

REVIEW QUESTIONS

PART-A

1. State the purpose of wireless communication.
2. Mention some examples of wireless equipment.
3. What are the transmission media used by wireless communication?
4. List the applications of wireless communication.
5. State the principle of mobile communication.
6. Define the cellular concept.
7. What are cells and cell clusters?
8. What do you mean by frequency reuse?
9. Define cell splitting.
10. What is cell sectoring?
11. What is handoff?
12. List the types of multiple-access techniques.
13. State the principle of frequency-division multiple-access method.
14. What is TDMA?
15. State the principle of code-division multiple-access method.
16. What is SDMA?
17. State the significance of wireless LAN.
18. What are the hardware elements of wireless LAN architecture?
19. What are the two types of wireless LAN roaming?
20. List out the important WLAN standards.
21. What is WPAN?
22. Give the applications of WPAN.
23. What is Bluetooth?
24. What are the components of Bluetooth?

25. State the advantages of Bluetooth technology.
26. What are the characteristics of Bluetooth transmission?
27. What is ZigBee?
28. Give the logical device types of ZigBee.
29. What are the network topologies supported by ZigBee?
30. List the applications of ZigBee.

PART-B

1. Explain the concept and principle of operation of mobile communication in detail.
2. With neat sketches, explain the concept of cellular architecture.
3. Describe the principle of cell splitting and cell sectoring in detail with neat sketches.
4. With a neat block diagram of AMPS, explain the functioning of each block.
5. Compare the digital cellular system with the analog cellular system.
6. Differentiate between FDMA and TDMA.
7. Explain the principle of CDMA and SDMA.
8. With a neat sketch, explain the WLAN architecture.
9. What are the different WLAN standards? Give details.
10. What is Bluetooth? How does it work?
11. What is ZigBee? Explain its functions and ZigBee stack.
12. Explain the various ZigBee topologies and their functioning.
13. What is the significance of caller ID. Explain its functioning.
14. How does a cordless phone work? Explain.
15. With neat sketches, explain the functioning of a Facsimile system.
16. How does a pager work? Explain its functioning in detail.

14

TRANSMISSION LINES

Objectives

- To know about the purpose and different types of transmission lines in detail
- To model an electrical transmission line as a two-port network
- To provide the details about the elements of a transmission line like inductance, capacitance and resistance
- To discuss the details about characteristic impedance and various line losses for a transmission line.
- To know the details about transmission media and its two major categories such as guided transmission media and unguided transmission media.

14.1 INTRODUCTION

A transmission line is a device designed to guide electrical energy from one point to another. It is mainly used to transfer the output RF energy of a transmitter to an antenna. In other words, it is a material medium or structure that forms a path for directing the transmission of energy from one place to another, such as electromagnetic waves as well as electric-power transmission.

In the communication field, transmission lines are specialised cables designed to carry alternating current and electromagnetic waves of high frequency. Ordinary electrical cables can carry low-frequency ac, such as mains power. But they cannot carry currents in the RF range or higher which causes power losses. Transmission lines use specialised construction to transmit electromagnetic signals with minimal reflections and power losses.

14.2 TYPES OF TRANSMISSION LINES

There are several types of transmission lines. They are

1. Ladder line

2. Coaxial cable
3. Dielectric slabs
4. Strip line
5. Optical fibre
6. Waveguides

For higher frequency ranges, the waves are shorter in a transmission medium. Transmission lines must be used when the frequency is high enough that the wavelength of the waves begin to approach the length of the cable used. To conduct energy at frequencies above the radio range, the waves become much smaller than the dimensions of the structures used to guide them, so transmission-line techniques become inadequate. For such cases, the optical methods are used.

14.3 A MODEL OF ELECTRICAL TRANSMISSION LINE

An electrical transmission line can be modelled as a two-port network, which is also called a **quadruple network**. Figure 14.1 shows a two-port network.

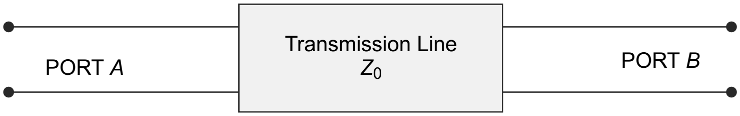


Fig. 14.1 A two-port network

The network is assumed to be linear and the two ports are assumed to be interchangeable. If the transmission line is uniform along its length then its behaviour is largely described by a single parameter called the **characteristic impedance** (Z_0). It is the ratio of the complex voltage of a given wave to the complex current of the same wave at any point on the line.

Typical values of Z_0 are

1. 50 or 75 Ω for a coaxial cable,
2. about 100 Ω for a twisted pair of wires and
3. about 300 Ω for a untwisted pair used in radio transmission.

When sending power down a transmission line, maximum amount of power will be absorbed by the load and minimum amount of power will be reflected back to the source. This can be ensured by making the load impedance equal to Z_0 and the transmission line is said to be matched. Some of the power that is fed into a transmission line is lost due to its resistance. This effect is called **ohmic loss**. At high frequencies, dielectric loss is caused when the insulating material inside the transmission line absorbs energy from the alternating electric field and converts it to heat. The total loss of power in a transmission line is often specified in decibels per metre (dB/m), and usually depends on the frequency of the signal. High-frequency transmission lines are designed to carry electromagnetic waves whose wavelengths are shorter than the length of the line.

14.4 TRANSMISSION-LINE THEORY

The construction of the transmission line determines the electrical characteristics. The two-wire line acts like a long capacitor. When the frequency applied to the transmission line changes, its capacitive reactance will also be changed. Since long conductors have a magnetic field about them when electrical energy is being passed through them, they exhibit the properties of inductance.

The values of inductance and capacitance presented depend on the various physical factors like the type of line used, the dielectric in the line and the length of the line. The effects of the inductive and capacitive reactance of the line depend on the frequency applied. Since no dielectric is perfect, electrons manage to move from one conductor to the other through the dielectric. Each type of two-wire transmission line also has a conductance value. This conductance value represents the value of the current flow that may be expected through the insulation.

14.4.1 Lumped Constants

A transmission line has the properties of inductance, capacitance and resistance similar to conventional circuits. However, the constants in conventional circuits are lumped into a single device or component. For example, a coil of wire has the property of inductance. When a certain amount of inductance is needed in a circuit, a coil of the proper dimensions is inserted. The inductance of the circuit is lumped into the one component. Two metal plates separated by a small space can be used to supply the required capacitance for a circuit. In such a case, most of the capacitance of the circuit is lumped into this one component. Similarly, a fixed resistor can be used to supply a certain value of circuit resistance as a lumped sum. Ideally, a transmission line would also have its constants of inductance, capacitance and resistance lumped together, as shown in Figure 14.2.

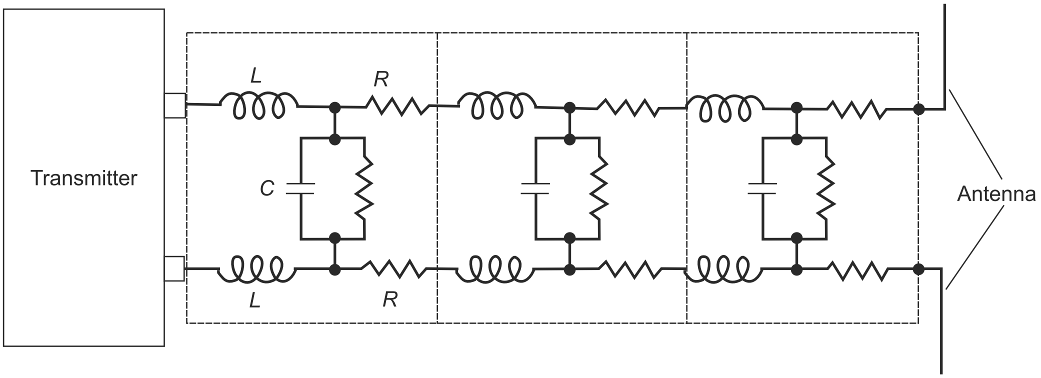


Fig. 14.2 Two-wire transmission line

14.4.2 Distributed Constants

Transmission-line constants, called distributed constants, are spread along the entire length of the transmission line and cannot be distinguished separately. The amount of inductance, capacitance and resistance depends on the length of the line, the size of the conducting wires, the spacing between the wires and the dielectric between the wires.

1. Inductance of a Transmission Line

When current flows through a wire, magnetic lines of force are set up around the wire. As the current increases and decreases in amplitude, the field around the wire expands and collapses accordingly. The energy produced by the magnetic lines of force collapsing back into the wire tends to keep the current flowing in the same direction. This represents a certain amount of inductance, which is expressed in μH per unit length. Figure 14.3 illustrates the inductance and magnetic fields of a transmission line.

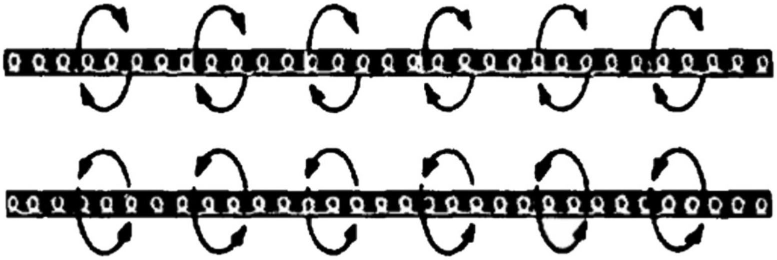


Fig. 14.3 Distributed inductance

2. Capacitance of a Transmission Line

Capacitance also exists between the transmission line wires, as illustrated in Figure 14.4.

Notice that the two parallel wires act as plates of a capacitor and that the air between them acts as a dielectric. The capacitance between the wires is usually expressed in pF per unit length. This electric field between the wires is similar to the field that exists between the two plates of a capacitor.

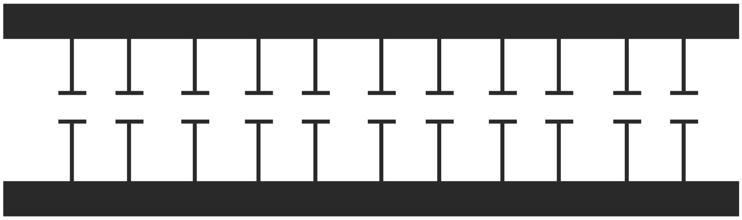


Fig. 14.4 Distributed capacitance

3. Resistance of a Transmission Line

The transmission line shown in Figure 14.5 has electrical resistance along its length.

This resistance is usually expressed in ohms per unit length and is shown as existing continuously from one end of the line to the other.

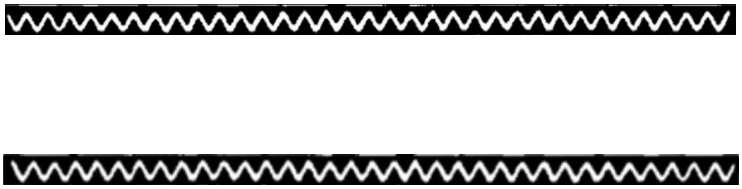


Fig. 14.5 Distributed resistance

4 Leakage Current

Since any dielectric is not a perfect insulator, a small current known as 'leakage current' flows between the two wires. In effect, the insulator acts as a resistor, permitting current to pass between the two wires. Figure 14.6 shows this leakage path as resistors in parallel connected between the two lines. This property is called **conductance** (G) and is the opposite of resistance. Conductance in transmission lines is expressed as the reciprocal of resistance and is usually given in μΩ per unit length.

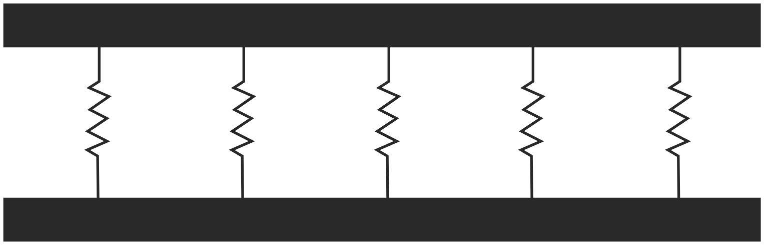


Fig. 14.6 Leakage in a transmission line

14.5 CHARACTERISTIC IMPEDANCE

The characteristic impedance, also called **surge impedance**, of a uniform transmission line, usually written Z_0, is the ratio of the amplitudes of a single pair of voltage and current waves propagating along the line in the absence of reflections. The unit of characteristic impedance is ohm. The characteristic impedance of a lossless transmission line is purely real, that is there is no imaginary component ($Z_0 = | Z_0 | + j0$).

Characteristic impedance appears like a resistance in this case, such that power generated by a source on one end of an infinitely long lossless transmission line is transmitted through the line but is not 'dissipated in' the line itself. A transmission line of finite length that is terminated at one end with a resistor equal to the characteristic impedance ($Z_L = Z_0$) appears to the source like an infinitely long transmission line.

A transmission line will only exhibit its characteristic impedance when it is terminated in its characteristic impedance. A 50 Ω cable is 50 Ω when it is connected to a load consisting of

50 Ω pure resistance. If the transmission line is terminated in a load not equal to its characteristic impedance then the impedance on that line will vary from one point to the next along its length due to the presence of reflected waves.

14.5.1 Factors that Determine Characteristic Impedance

The characteristic impedance of any transmission line is a function of the size and spacing of the conductors and the type of insulating material between them.

If the distributed inductance and capacitance per unit length of a line is known then the characteristic impedance can be found from

$$Z_0 = \sqrt{\frac{L}{C}} \tag{14.1}$$

Using the above equation, with the known values of distributed inductance and capacitance, Z_0 will be calculated. Figure 14.7 illustrates the calculation of characteristic impedance for co-axial cable and parallel pair.

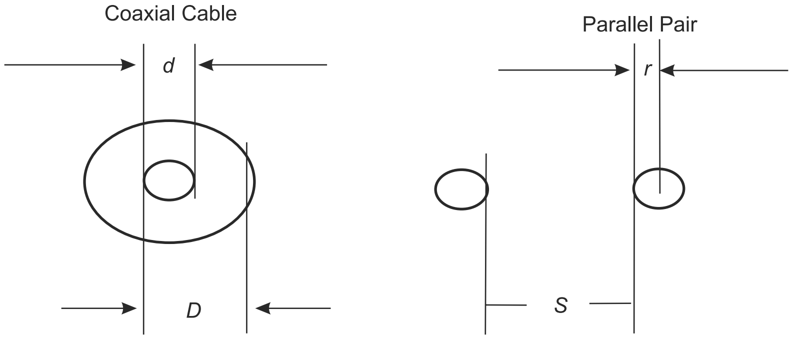


Fig. 14.7 Characteristic impedance for co-axial cable and parallel pair

The characteristic impedances for a co-axial cable and parallel pair will be expressed as follows.

$$Z_0 = \frac{138}{\sqrt{e}} \log D/d \quad \text{for co-axial cable} \tag{14.2}$$

$$Z_0 = \frac{138}{\sqrt{e}} \log S/r \quad \text{for a parallel pair} \tag{14.3}$$

14.5.2 Ranges of Z_0

In the design of transmission lines, there are certain constraints which restrict the range of practical impedances that can be achieved. For two-wire parallel lines, the Z_0 is usually restricted to a range of 100 to 600 Ω, while for co-axial cables, the practical range of characteristic impedance is typically 30 to 100 Ω.

14.6 LINE LOSSES

Practically, it is difficult to have a transmission line without transmission losses. Actually, some losses occur in all lines. Line losses may be any of three types.

1. Copper losses
2. Dielectric losses
3. Radiation and induction losses

14.6.1 Copper Losses

The first type of copper loss is $\boldsymbol{I^2R}$ **loss**. In transmission lines, the resistance of the conductors is never equal to zero. Whenever current flows through one of these conductors, some energy is dissipated in the form of heat. This heat loss is called **power loss**. With a copper braid, which has a resistance higher than solid tubing, this power loss is higher.

The second type of copper loss is due to **skin effect**. When dc current flows through a conductor, the movement of electrons through the conductor's cross section is found to be uniform. The situation is somewhat different when ac current is applied. The expanding and collapsing fields about each electron encircle other electrons. This phenomenon, called **self-induction**, retards the movement of the encircled electrons. The flux density at the centre is so great that electron movement at this point is reduced.

As frequency is increased, the opposition to the flow of current in the centre of the wire increases. Current in the centre of the wire becomes smaller and most of the electron flow is on the wire surface. When the frequency applied is 100 MHz or higher, the electron movement in the centre is so small that the centre of the wire could be removed without any noticeable effect on current. Since resistance is inversely proportional to the cross-sectional area, the resistance will increase as the frequency is increased. Also, since power loss increases as resistance increases, power losses increase with an increase in frequency because of skin effect.

Copper losses can be minimised and conductivity increased in a transmission line by plating the line with silver. Since silver is a better conductor than copper, most of the current will flow through the silver layer. The tubing then serves primarily as a mechanical support.

14.6.2 Dielectric Losses

Due to the heating effect on the dielectric material between the conductors, dielectric losses will appear. Power from the source is used in heating the dielectric. The heat produced is dissipated into the surrounding medium. When there is no potential difference between two conductors, the atoms in the dielectric material between them are normal and the orbits of the electrons are circular. When there is a potential difference between two conductors, the orbits of the electrons change. The excessive negative charge on one conductor repels electrons on the dielectric toward the positive conductor and thus distorts the orbits of the electrons. A change in the path of electrons requires more energy, introducing a power loss. The atomic structure of rubber is more difficult to distort than the structure of some other dielectric materials. The atoms of materials, such as polyethylene, distort easily. Therefore, polyethylene is often used as a dielectric because less power is consumed when its electron orbits are distorted.

14.6.3 Radiation and Induction Losses

Radiations and induction losses are similar in that both are caused by the fields surrounding the conductors. Induction losses occur when the electromagnetic field about a conductor cuts through any nearby metallic object and a current is induced in that object. As a result, power is dissipated in the object and is lost. Radiation losses occur because some magnetic lines of force about a conductor do not return to the conductor when the cycle alternates. These lines of force are projected into space as radiation and this results in power losses.

14.7 TRANSMISSION OF ENERGY

When the load is connected directly to the source of energy, or when the transmission line is short, problems concerning current and voltage can be solved by applying Ohm's law. When the transmission line becomes long enough so the time difference between a change occurring at the generator and a change appearing at the load becomes appreciable, it is important to have the analysis of the transmission line.

14.8 STANDING-WAVE RATIO

The measurement of standing waves on a transmission line yields information about equipment-operating conditions. Maximum power is absorbed by the load when $Z_L = Z_0$. If a line has no standing waves, the termination for that line is correct and maximum power transfer takes place. A wide variation in voltage along the length means a termination far from Z_0. A small variation means termination near Z_0. Therefore, the ratio of the maximum to the minimum is a measure of the perfection of the termination of a line. This ratio is called the Standing-Wave Ratio (SWR) and is always expressed in whole numbers. For example, a ratio of 1:1 describes a line terminated in its characteristic impedance (Z_0).

14.8.1 Voltage Standing-Wave Ratio

The ratio of maximum voltage to minimum voltage on a line is called the Voltage Standing-Wave Ratio (VSWR). It is expressed as follows.

$$\text{VSWR} = \left|\frac{E_{\max}}{E_{\min}}\right| \tag{14.4}$$

The vertical lines in the formula indicate that the enclosed quantities are absolute and that the two values are taken without regard to polarity. Depending on the nature of the standing waves, the numerical value of VSWR ranges from a value of 1 to an infinite value for theoretically complete reflection. Since there is always a small loss on a line, the minimum voltage is never zero and the VSWR is always some finite value. However, if the VSWR is to be a useful quantity, the power losses along the line must be small in comparison to the transmitted power.

14.8.2 Power Standing-Wave Ratio

The square of the VSWR is called the Power Standing-Wave Ratio (PSWR). It is expressed as follows.

$$\text{PSWR} = \frac{P_{\max}}{P_{\min}} \tag{14.5}$$

This ratio is useful because the instruments used to detect standing waves react to the square of the voltage. Since power is proportional to the square of the voltage, the ratio of the square of the maximum and minimum voltages is called the power standing-wave ratio.

14.8.3 Current Standing-Wave Ratio

The ratio of maximum to minimum current along a transmission line is called Current Standing-Wave Ratio (ISWR). It is expressed as follows.

$$\text{ISWR} = \left|\frac{I_{\max}}{I_{\min}}\right| \tag{14.6}$$

This ratio is the same as that for voltages. It can be used where measurements are made with loops that sample the magnetic field along a line. It gives the same results as VSWR measurements.

14.9 TRANSMISSION MEDIA

There are two basic categories of transmission media. They are as follows.

1. Guided transmission media
2. Unguided transmission media

14.9.1 Guided Transmission Media

Guided transmission media uses a cabling system that guides the data signals along a specific path. The data signals are bound by the cabling system. Guided media is also known as **bound media**. Cabling is meant in a generic sense in the previous sentences and is not meant to be interpreted as copper-wire cabling only.

There four basic types of Guided Media. They are listed as follows.

1. Open wire type
2. Twisted pair
3. Coaxial cable
4. Optical fibre

1. Open-wire Type

Open wire is traditionally used to describe the electrical wire strung along power poles. There is a single wire strung between poles. No shielding or protection from noise interference is used. We are going to extend the traditional definition of open wire to include any data-signal path without shielding or protection from noise interference. This can include multiconductor cables or single wires. This media is susceptible to a large degree of noise and interference and consequently not acceptable for data transmission except for short distances under 20 feet. Figure 14.8 shows open-wire transmission media.

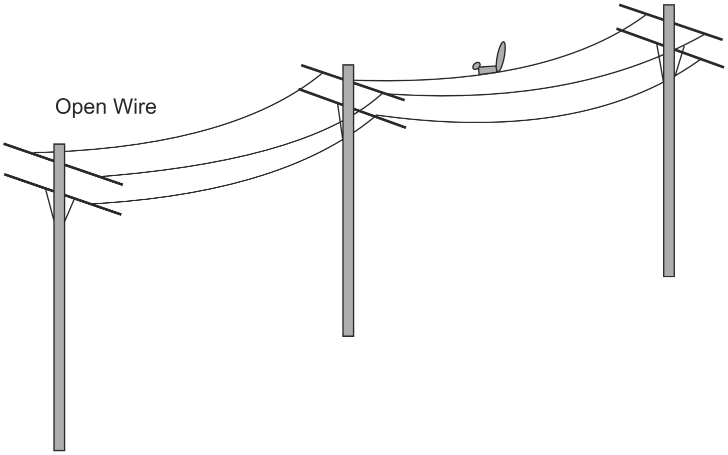


Fig. 14.8 Open-wire transmission media

2. Twisted Pair

Since the wires are twisted together in pairs, it is named as twisted-pair cabling. Each pair would consist of a wire used for the positive data signal and a wire used for the negative data signal. Any noise that appears on one wire of the pair would occur on the other wire. Because

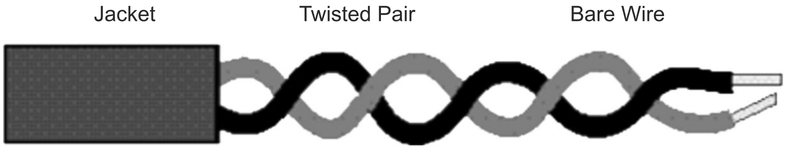


Fig. 14.9 Twisted-pair cable

the wires have opposite polarities, they are 180° out of phase. When the noise appears on both wires, it cancels or nulls itself out at the receiving end. Figure 14.9 shows a twisted-pair cable.

The degree of reduction in noise interference is determined specifically by the number of turns per foot. Increasing the number of turns per foot reduces the noise interference. To further improve noise rejection, a foil or wire-braid shield is woven around the twisted pairs. This 'shield' can be woven around individual pairs or around a multipair conductor. Figure 14.10 shows shielded twisted-pair cable.

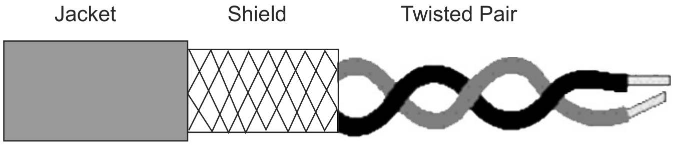


Fig. 14.10 Shielded twisted-pair cable

3. Co-axial Cable

A co-axial cable consists of two conductors. The inner conductor is held inside an insulator with the other conductor woven around it providing a shield. An insulating protective coating called a jacket covers the outer conductor. Figure 14.11 shows a co-axial cable.

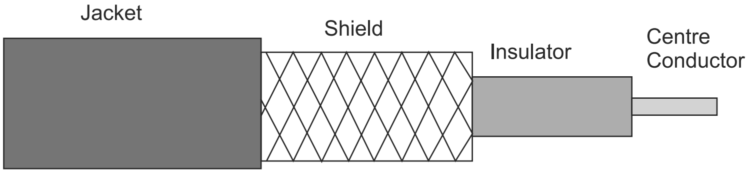


Fig. 14.11 Co-axial cable

The outer shield protects the inner conductor from outside electrical signals. The distance between the outer conductor (shield) and inner conductor plus the type of material used for insulating the inner conductor determines the cable properties or impedance. The excellent control of the impedance characteristics of the cable allow higher data rates to be transferred than in the twisted-pair cable.

4. Optical Fibre

Optical fibres consist of thin glass fibres that can carry information at different frequencies in the visible light spectrum and beyond. A typical optical fibre consists of a very narrow strand of glass called **core**. A concentric layer of glass around the core is called **Cladding**. The diameter of a typical core is 62.5 microns and for a typical cladding is 125 microns. A protective coating covering the cladding is called **jacket** which is made by plastics. Side view and end view of a typical fibre is shown in Figure 14.12. More details about fibre-optic transmission are given in Chapter 10.

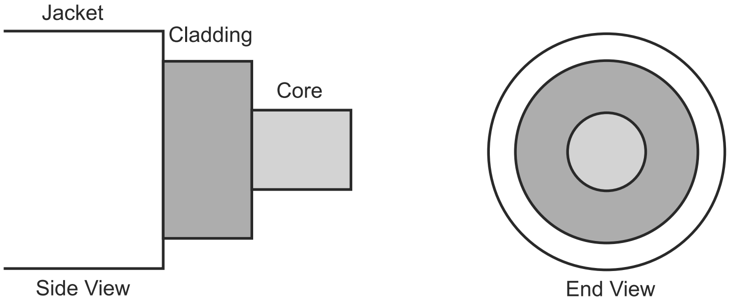


Fig. 14.12 An optical fibre

5. Bandwidth and Different Transmission Media

Table 14.1 compares the usable bandwidth between the different guided transmission media.

Table 14.1 Comparison of usable bandwidth between the different guided transmission media

Cable Type	Bandwidth
Open cable	0–5 MHz
Twisted pair	0–100 MHz
Coaxial cable	0–600 MHz
Optical fibre	0–1 GHz

14.9.2 Unguided Transmission Media

Unguided transmission media consists of a means for the data signals to travel but nothing to guide them along a specific path. The data signals are not bound to a cabling media and as such are often called **unbound media**. They are classified by the type of wave propagation.

They are listed as follows.

1. Radio-frequency propagation
2. Microwave propagation
3. Satellite propagation

1. Radio-Frequency Propagation

Radio frequencies are in the range of 300 kHz to 10 GHz. An emerging technology called wireless LANs uses radio frequencies. Among them, some use radio frequencies to connect the workstations together and some use infrared technology.

There are three major types of RF propagation. They are listed as follows.

(a) Ground-wave propagation
(b) Ionospheric propagation
(c) Line-of-sight (LOS) propagation

(a) Ground-Wave Propagation Ground-wave propagation follows the curvature of the earth. Ground waves have carrier frequencies up to 2 MHz. AM radio is an example of ground-wave propagation. Figure 14.13 shows ground-wave propagation.

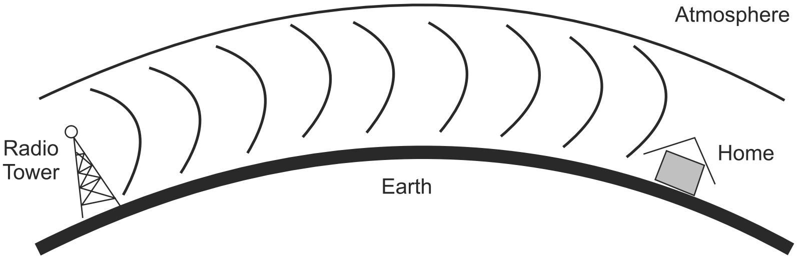


Fig. 14.13 Ground-wave propagation

(b) Ionospheric Propagation Ionospheric propagation bounces off the earth's ionospheric layer in the upper atmosphere. It is sometimes called **double-hop propagation**. It operates in the frequency range of 30–85 MHz. Because it depends on the earth's ionosphere, it changes with weather and time of day. The signal bounces off the ionosphere and back to the earth. It is illustrated in Figure 14.14.

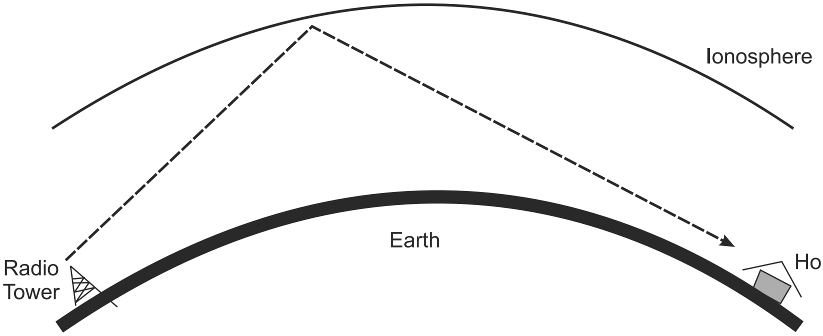


Fig. 14.14 Ionospheric propagation

(c) Line-of-Sight (LOS) Propagation Line-of-sight propagation transmits exactly in the line of sight. The receive station must be in the view of the transmit station. It is sometimes called **space waves** or **tropospheric propagation**. It is limited by the curvature of the earth for ground-based stations. Examples of line-of-sight propagation are FM radio, microwave and satellite. Figure 14.15 illustrates line-of-sight propagation.

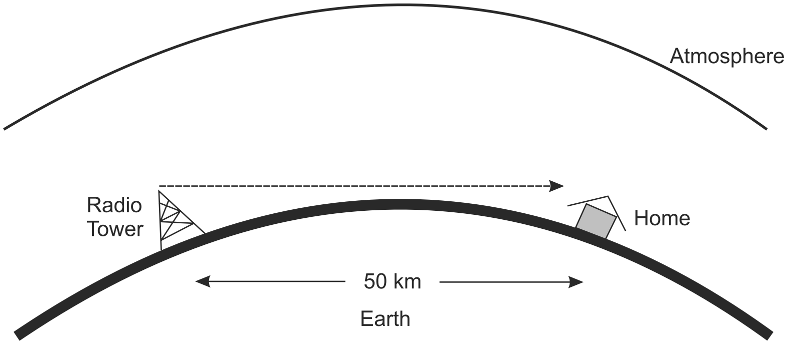


Fig. 14.15 Line-of-sight propagation

2. Microwave Propagation

Microwave transmission is line-of-sight transmission. The transmit station must be in visible contact with the receive station. This sets a limit on the distance between stations depending on the local geography. Typically, the line of sight due to the earth's curvature is only 50 km to the horizon. Figure 14.16 illustrates microwave transmission.

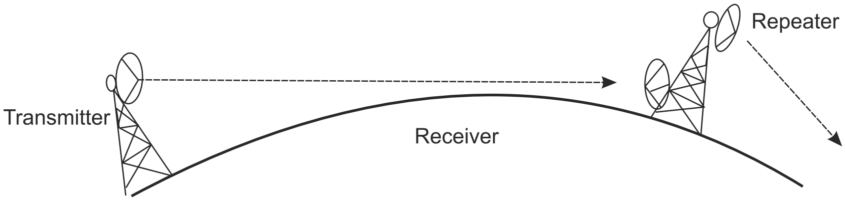


Fig. 14.16 Microwave transmission

Microwaves operate at high operating frequencies of 3 to 10 GHz. This allows them to carry large quantities of data due to the large bandwidth.

3. Satellite Propagation

Satellites are transponders that are set in a geostationary orbit directly over the equator. A transponder is a unit that receives on one frequency and retransmits on another. The geostationary orbit is 36,000 km from the earth's surface. At this point, the gravitational pull

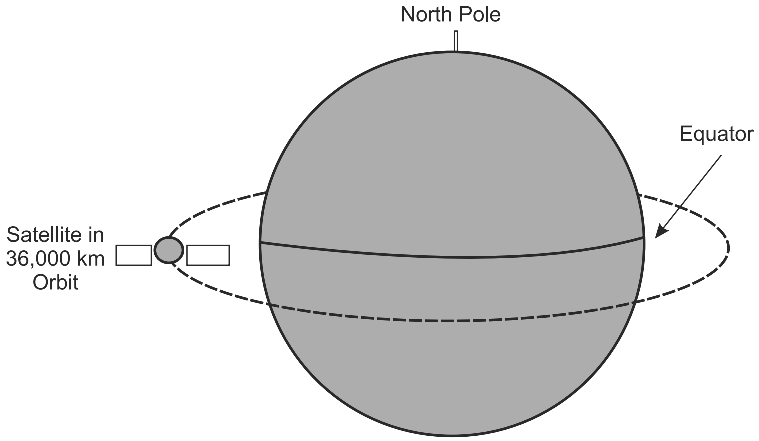


Fig. 14.17 Satellite propagation

of the earth and the centrifugal force of the earth's rotation are balanced and cancel each other out. Centrifugal force is the rotational force placed on the satellite that wants to fling it out to space. Figure 14.17 illustrates satellite propagation.

Summary

A transmission line is a material medium or structure that forms a path for directing the transmission of energy from one place to another, such as electromagnetic waves as well as electric power transmission. They are specialised cables designed to carry alternating current and electromagnetic waves of high frequency. Transmission lines must be used when the frequency is high enough so that the wavelength of the waves begins to approach the length of the cable used.

The characteristic impedance is the ratio of the amplitudes of a single pair of voltage and current waves propagating along the line in the absence of reflections. The unit of characteristic impedance is ohm. A transmission line will only exhibit its characteristic impedance when it is terminated in its characteristic impedance. A 50 Ω cable is 50 Ω when it is connected to a load consisting of 50 Ω pure resistance. If the transmission line is terminated in a load not equal to its characteristic impedance then the impedance on that line will vary from one point to the next along its length due to the presence of reflected waves. The characteristic impedance of any transmission line is a function of the size and spacing of the conductors and the type of insulating material between them.

Practically, it is difficult to have a transmission line without transmission losses. Actually, some losses occur in all lines. Line losses may be any of three types.

- Copper losses
- Dielectric losses and

- Radiation and induction losses

The measurement of standing waves on a transmission line yields information about equipment-operating conditions. If a line has no standing waves, the termination for that line is correct and maximum power transfer takes place.

The ratio of maximum voltage to minimum voltage on a line is called the Voltage Standing-Wave Ratio (VSWR), and the square of the VSWR is called the Power Standing-Wave Ratio (PSWR). The ratio of maximum to minimum current along a transmission line is called Current Standing-Wave Ratio (ISWR).

There are two basic categories of transmission media. They are as follows.

- Guided transmission media
- Unguided transmission media

Guided transmission media uses a cabling system that guides the data signals along a specific path. The data signals are bound by the cabling system. There are four basic types of guided media. They are listed as follows.

- Open-wire type
- Twisted pair
- Coaxial cable
- Optical fibre

REVIEW QUESTIONS

PART-A

1. What is a transmission line? Give its significance.
2. What are the types of transmission lines?
3. Define characteristic impedance.
4. What are lumped constants?
5. What are distributed constants?
6. Mention the factors to be considered for the determination of characteristic impedance.
7. List the different line losses.
8. What is meant by copper loss?
9. Give the significance of dielectric losses.
10. What do you mean by radiation losses?
11. Define and give the expression of voltage standing-wave ratio.

12. Define and give the expression of power standing-wave ratio.
13. What is meant by ISWR?
14. List the types of transmission media.
15. Define and give the types of guided transmission media.
16. What is meant by unguided transmission media? Give its types.
17. What are the types of types of RF propagation?
18. What do you mean by double-hop propagation?
19. What is line-of-sight propagation?
20. Define satellite propagation.

PART-B

1. How will you model an electrical transmission line? Explain.
2. Explain various transmission line constants in detail.
3. Briefly explain the various line losses in a transmission line.
4. Explain the various types of guided transmission media with suitable diagrams.
5. With neat sketches, describe the types of unguided transmission media.

Appendix

Chapter 1

Problems

1. For the following continuous time signals, determine whether they are periodic or aperiodic?

   (i) $x(t) = \sin\sqrt{3}t + \cos 2t$

   (ii) $x(t) = 2 \cos 200\, \pi t + 7 \sin 60t$

2. Find out the even and odd components for the given signals.

   (i) $x(t) = \cos^2\left(\dfrac{3\pi t}{2}\right)$

   (ii) $x(t) = \begin{cases} t & 0 \le t \le 1 \\ 4 - t & 1 \le t \le 2 \end{cases}$

3. Evaluate the even and odd parts of the following rectangular pulse signal denoted as

   $x(t) = A \text{ rec}\left(\dfrac{t}{T} - \dfrac{1}{4}\right)$

4. Check whether the following signals are energy signals or power signals.
   (i) $x(t) = \cos^2 (2\omega_0 t)$
   (ii) $x(t) = A \sin (\omega t + \theta)$

5. Determine the Fourier transform of the following signal:

   $x(t) = \dfrac{2}{\pi t}\left[\sin^2(\pi\omega t)\right]$

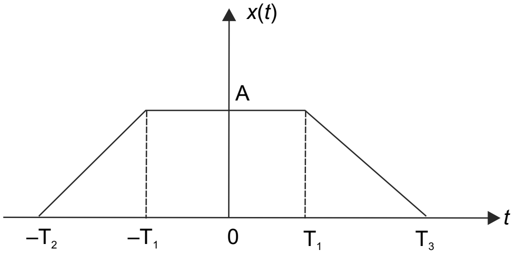
6. Determine the Fourier transforms of a single-sided exponential signal and a double -sided signal.

7. Evaluate the Fourier transforms of the following signals.

   (i) $x(t) = \cos \omega_0 t$

   (ii) $x(t) = \sin \omega_0 t$

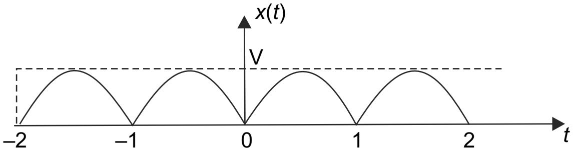
8. For the following trapezoidal function shown below, determine the Fourier transform.



9. Determine the Fourier transform of the following function represented as

   $x(t) = \exp(-t) \sin(2\pi f_c t)\, u(t)$

10. For the following full wave rectifier output function represented below, determine the Fourier series.



Chapter 3

Solved Problems

3.11 For the series diode detector, the R_1C_1 load consists of a 100 pF capacitor in parallel with a 10 kΩ resistance. Calculate the maximum modulation index that can be considered for sinusoidal modulation at a frequency of 10 kHz if the diagonal peak clipping is to be avoided.

Solution

Admittance for the load of the diode detector is

$$Y = \frac{1}{R} + j2\pi f_m C$$

$$= \frac{1}{10\times 10^3} + j2\pi \times 10\times 10^3 \times 100\times 10^{-12}$$

$$= 10^{-4} + j6.283\times 10^{-6}$$

$$|Y| = \sqrt{\left(10^{-4}\right)^2 + \left(6.283\times 10^{-6}\right)^2}$$

$$= 1.002 \times 10^{-4}$$

$$|Z| = \frac{1}{|Y|} = \frac{1}{1.002\times 10^{-4}}$$

$$= 9980\ \Omega$$

The maximum modulation index that can be handled is given as

$$m = \frac{|Z|}{R}$$

$$= \frac{9980}{10\times 10^3} = 0.998$$

3.12 Calculate the *Q* of the *LC* tuned circuit having a resonant frequency of 1.5 MHz, internal resistance of coil as 65 Ω and a capacitor of 125 μF.

Solution:

Q of the tuned circuit $= \dfrac{X_L}{R}$

$$=\frac{2\pi f_r L}{R}$$

$$=\frac{2\pi\times 1.5\times 10^{6}\times 125\times 10^{-6}}{65}\ = 19.63$$

3.13 Calculate the maximum allowable modulation index which may be applied to the practical diode detector with a DC load for diode as 300 kΩ and the impedance value of 230 kΩ.

Solution:

Maximum allowable modulation index will be

$$m=\frac{Z_m}{R_L}$$

$$=\frac{230\times 10^{3}}{300\times 10^{3}}\ = 0.77$$

3.14 For a superheterodyne receiver, the Q factor of the antenna section is 125. If the intermediate frequency is 465 kHz, determine the image frequency. Also its image frequency rejection ratio is at 1000 kHz.

Solution:

It is known that,

$$IFRR=\sqrt{(1+Q^2\rho^2}\ \text{ and }\ \rho=\left({}^{f_{\text{im}}}\!/\!_{f_{\text{RF}}}\right)-\left({}^{f_{\text{RF}}}\!/\!_{f_{\text{im}}}\right)$$

f_{im} = 1000 + 2 × 465 = 1930 kHz

$$\rho=\left({}^{f_{im}}\!/\!_{f_{RF}}\right)-\left({}^{f_{RF}}\!/\!_{f_{im}}\right)$$

$$\therefore \rho=\left(\frac{1930}{1000}\right)-\left(\frac{1000}{1930}\right)$$

= 1.93 – 0.52 = 1.41

$$\therefore IFRR=\sqrt{(1+\left(125^2\times 1.41^2\right)}\ = 176.25$$

3.15 For Problem 3.12, determine the IFRR at 30 MHz and check whether it is sufficient or not for a practical receiver.

Solution:

$$f_{\text{im}} = 30 + 2 \times 0.465 = 30.93 \text{ MHz}$$

$$\rho = \left({}^{f_{\text{im}}}\!/\!_{f_{\text{RF}}} \right) - \left({}^{f_{\text{RF}}}\!/\!_{f_{\text{im}}} \right)$$

$$\therefore \rho = \left(\frac{1930}{1000} \right) - \left(\frac{1000}{1930} \right)$$

$$= 1.03 - 0.97 = 0.06$$

$$\therefore \text{IFRR} = \sqrt{(1 + \left(125^2 \times 0.06^2\right)} = 7.57$$

The resultant IFRR is not sufficient for a practical receiver.

3.16 How will you make the IFRR as a sufficient ratio at the same 30 MHz? Find the new loaded *Q*-factor for the practical receiver.

Solution:

$$\text{The IFRR'} = \frac{176.25}{7.57} = 23.3$$

$$\text{IFRR}' = 23.3 = \sqrt{(1 + Q'^2 \times 0.06^2}$$

$$Q'^2 = \frac{23.3^2 - 1}{0.06}$$

$$\therefore \text{new loaded } Q' = \frac{\sqrt{541.9}}{0.06} = 387.97$$

3.17 Determine the image rejection of a receiver which is tuned to 555 kHz. Its local oscillator provides the mixer with an input at 1010 kHz and the *Q* factor of the tuned circuit is 40.

Solution:

$$f_{\text{im}} = 1010 + 2 \times 555 = 2120 \text{ kHz}$$

$$\rho = \left(f_{im} \big/ f_{RF} \right) - \left(f_{RF} \big/ f_{im} \right)$$

$$\therefore \rho = \left(\frac{2120}{1010} \right) - \left(\frac{1010}{2120} \right)$$

$$= 2.1 - 0.48 = 1.62$$

$$\therefore \text{IFRR} = \sqrt{(1 + \left(40^2 \times 1.62^2\right)} \quad = 64.\ 81$$

Problems

1. An AM receiver is tuned to a station whose carrier frequency is 750 kHz. What frequency should the local oscillator be set to in order to provide an intermediate frequency of 475 kHz if the local oscillator tracks below the received frequency?
2. Repeat Problem 1 for a station having a carrier frequency of 1300 kHz.
3. Calculate the image rejection ratio for a receiver having an RF amplifier and an IF amplifier of 500 kHz, if the Q-factors of relevant coils are 65 and at an incoming frequency of 30 MHz.
4. A superheterodyne receiver having an RF amplifier and an IF of 460 kHz is tuned to 20 MHz. Determine Q-factor of the coupled tuned circuits, both being the same, if the receiver's image rejection is to be 125.
5. An AM commercial broadcast receiver operates in a frequency range of (550 to 1600) kHz with an input filter factor of 45. Determine the bandwidth at the low and high ends of RF spectrum.
6. The load on an AM diode detector which consists of a resistance of 50 kΩ in parallel with a capacitor of 0.005 F. Calculate the maximum modulation index that the detector can handle without distortion when modulating frequency is 3 kHz.
7. A diode detector uses a load consisting of a capacitor value of 0.01μF in parallel with a resistor value of 6 kHz. Calculate the maximum depth of modulation that the diode can detect without diagonal clipping when the modulating frequency is 1.5 kHz.
8. What is the value of Q-factor if the resonant frequency of an RF amplifier of a receiver is 1.5 MHz and its bandwidth is 20 kHz.

Chapter 4

Problems

1. A 94.2 MHz is frequency modulated by a 5 kHz sine wave and the frequency deviation of the resultant FM signal is 40 kHz.

   (a) Determine the carrier swing of the FM signal.

   (b) Determine the highest and lowest frequencies attained by the modulated signal.

   (c) Also calculate the modulation index of FM wave.
2. Find the upper and lower frequencies that are reached by an FM wave that has a rest frequency of 104.02 MHz and a frequency deviation of 60 kHz. What is the carrier swing of the modulated signal?
3. The carrier swing of a frequency-modulated signal IS 125 kHz. The modulating signal is a 5 kHz sine wave. Determine the modulation index of the FM signal.
4. An FM signal for broadcast in the 85-105 MHz range has a frequency deviation of 15 kHz. Find the percent modulation of this signal. If this signal were prepared for broadcast, what would the percent modulation be?
5. An FM signal to be broadcast in the 85-105 MHz broadcast band is to be modulated at 75%. Determine the carrier swing and the frequency deviation.
6. An FM signal has a bandwidth of 120 kHz when its frequency deviation is 35 kHz. Find the frequency of the modulating signal.
7. A 20 kHz sine wave is frequency modulating a 105.500 MHz carrier. Find the modulation index of the FM signal and determine the bandwidth of the FM signal if the carrier swing is 120 kHz.
8. Calculate the bandwidth of a narrowband FM signal which is generated by a 6 kHz audio signal modulating a 116 MHz carrier.
9. A 50.001 MHz carrier is to be frequency modulated by a 4 kHz audio tone resulting in a narrowband FM signal. Determine the bandwidth of the FM signal.
10. Determine the bandwidth of a signal generated by a 3.1 kHz audio tone frequency modulating a 96.003 MHz carrier resulting in a frequency deviation of 3.5 kHz.

Chapter 5

Solved Problems

5.4 Find the bandwidth when the modulating frequency is three times the modulating frequency for an FM broadcast system in which the maximum deviation allowed is 50 kHz and the modulating signal is of 15 kHz.

Solution:

Given that

$\Delta f = 50$ kHz and $f_m = 15$ kHz

$$\text{BW} = 2\,(\Delta f + f_m)$$
$$= 2(50 + 15) = 130 \text{ kHz}$$

When the modulating frequency is three times the modulating frequency,

$$\text{Bandwidth BW} = 2\,(\Delta f + 3\,f_m)$$
$$= 2\,(50 + 45) = 190 \text{ kHz}$$

5.5 Consider an angle-modulated signal with carrier $V_c = 7\cos\,(2\pi(112 \text{ MHz})t)$ with frequency deviation of 75 kHz and single frequency interfering signal $V_n = 0.35\cos\,(2\pi(111.985 \text{ MHz})t)$. Determine the frequency of the demodulated signal.

Solution:

From $V_c = 7\cos\,(2\pi\,(112 \text{ MHz})t), f_c = 112$ MHz

From $V_n = 0.35\cos\,(2\pi\,(111.985 \text{ MHz})t), f_n = 111.985$ MHz

$\therefore f_c - f_n = 112 \text{ MHz} - 111.985 \text{ MHz} = 15 \text{ kHz}$

5.6 For the above Problem 5.5, find the peak phase deviation and peak frequency deviation. Also determine the voltage signal-to-noise ratio at the output of the demodulator.

Solution:

Peak phase deviation can be expressed as

$$\Delta\theta_{\text{peak}} = \frac{V_n}{V_c}$$

From the above problem, $V_c = 7$ and $V_n = 0.35$

$$\therefore \Delta\theta_{\text{peak}} = \frac{V_n}{V_c} = \frac{0.35}{7}$$

$$= 0.05 \text{ rad}$$

Peak frequency deviation can be expressed as

$$\Delta f_{\text{peak}} = \frac{V_n}{V_c} \cdot f_n$$

$$\therefore \Delta f_{\text{peak}} = \frac{0.35}{7} \cdot (15 \times 10^3)$$

$$= 750 \text{ Hz}$$

Voltage signal to noise $= \dfrac{V_c}{V_n} = \dfrac{7}{0.35} = 20$

Voltage signal to noise after demodulation is calculated as

$$\left(\frac{S}{N}\right) = \frac{75\,\text{kHz}}{750\,\text{Hz}} = 100$$

There is an improvement in voltage signal to noise of $\dfrac{100}{20} = 5$

20 log 5 = 14 dB

5.7 Design a de-emphasis circuit for a cut-off frequency of 3 kHz.

Solution:

The cutoff frequency for a de-emphasis circuit is expressed as

$$f_c = \frac{1}{2\pi RC}$$

It is given that cut-off frequency = 3 jHz

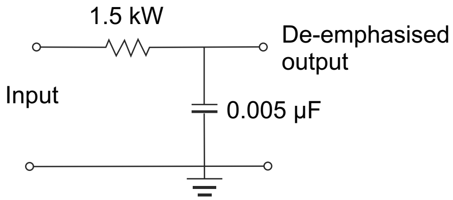
Assume $C = 0.01$ μF

$$3 \times 10^3 = \frac{1}{2\pi R \times 0.01 \times 10^{-6}}$$

$$\therefore R = \frac{1}{2\pi fC}$$

$$= \frac{1}{2\pi \times 3 \times 10^3 \times 0.01 \times 10^{-6}}$$

5.8 Determine the cut-off frequency for the de-emphasis circuit shown below.



Solution:

$R = 1.5$ kΩ $C = 0.005$ μF

The cutoff frequency for the de-emphasis circuit is expressed as

$$f_c = \frac{1}{2\pi RC}$$

It is given that $R = 1.5$ kΩ and $C = 0.005$ μF

$$3 \times 10^3 = \frac{1}{2\pi R \times 0.01 \times 10^{-6}}$$

$$= 21.22 \text{ kHz}$$

Problems

1. For a de-emphasis network used with an FM receiver, the time constant of the *RC* circuit is 80 μs. Compute the cut-off frequency of the circuit.
2. For an angle modulated signal with carrier V_c = 5 cos (2π(250 MHz)*t* with frequency deviation of 75 kHz and single frequency interfering signal V_c = 0.25 cos (2π(250.015 MHz)*t*, determine the frequency of the demodulated signal.
3. Design a de-emphasis circuit for a cut-off frequency of 4.7 kHz.

Chapter 6

Solved Problems

6.11 For a PAM transmission of voice signal, determine the bandwidth if f_s=7 kHz and τ = 0.1T_s. If the rise time is 15 of the width of the pulse, determine the minimum transmission bandwidth required for PDM.

Solution:

The sampling period is expressed as

$$T_s = \frac{1}{f_s}$$

$$f_s = 7 \text{ kHz}$$

$$\therefore T_s = \frac{1}{f_s} = \frac{1}{7\times 10^3} = 143\,\mu\text{s}$$

It is given that

$$\tau = 0.1\ T_S$$

$$\tau = 0.1 \times 143\ \mu\text{s} = 14.3\ \mu\text{s}$$

$$\text{Rise time } t_r = \tau \times 0.10$$

$$= 14.3 \times 10^{-6} \times 0.10 = 1.43 \times 10^{-7} \text{ second}$$

Bandwidth for PAM signal is

$$\text{BW} \geq \frac{1}{2\tau}$$

$$\therefore \text{BW} \geq \frac{1}{2\times 1.43 \times 10^{-7}}$$

$$\geq 3.5 \text{ MHz}$$

6.12 A PCM system uses a uniform quantiser followed by a 7-bit binary encoder. The bit rate of the system is 60 Mb/s. What is the maximum message bandwidth for desired performance?

Solution:

Bit rate > nf_s

$$f_s \le \frac{\text{Bit rate}}{n} \le \frac{60}{7} = 8.57$$

$$2\,f_M \le f_s$$

$$f_M \le \frac{f_s}{2}$$

$$f_M = \frac{8.57}{2} = 4.3\text{ MHz}$$

6.13 Determine the signalling rate needed to achieve $(S/N) \geq 45$ dB or number of binary bits as 7 for a binary PCM transmission of a video signal with $f_s = 10$ MHz.

Solution

$(S/N) \geq 45$ dB

Number of binary bits $n \geq 7$ dB

Signalling rate = nf_s

$= 7 \times 10 \times 10^6 = 70$ MHz

6.14 Consider the bandwidth of a TV radio plus audio signal at 5 MHz which is converted to PCM with 1024 quantizing levels. Determine the bit rate of the resulting PCM signal. Assume that the signal is sampled at a rate 20% above the Nyquist rate.

Solution

$n = \log_2 (1024) = 10$

Nyquist rate = $2f_m$

$= 2 \times 5$ MHz = 10 MHz

Sampling rate $f_s = 10 \times 1.2$ MHz = 12.0 MHz

Bit rate = nf_s

$= 10 \times 12$ MHz = 120 Mb/s

6.15 Bandwidth of a PCM system is 4 kHz and the input varies from –3.8 to 3.8 V. The average power of the PCM system is 30 mW and the required signal to quantisation power is 20 dB. The output of the modulator is in binary. Determine the number of bits required per sample.

Solution:

$$\left(\frac{S}{N_q}\right) = 20\text{ dB}$$

$$10\log\left(\frac{S}{N_q}\right)_0 = 20$$

$$\left(\frac{S}{N_q}\right)_0 = 100$$

Quantisation step size $\Delta = \dfrac{2A}{L}$

$L = 2^n$

The average quantising power is

$$N_q = \frac{\Delta^2}{12} = \frac{A^2}{3L^2}$$

$$100 = \frac{30\times10^{-3}}{A^2/3L^2}$$

$$3L^2 = \frac{30\times10^{-3}}{100\times(3.8)^2}$$

$$L = \sqrt{\frac{30\times10^{-3}}{3\times100\times(3.8)^2}}$$

$L = 126.67$

$2^n = 128 \Rightarrow n = 7$

6.16 Consider an 8-bit PCM for speech signal ranging up to 1.5V. Calculate the resolution and quantisation error.

Solution

Resolution can be expressed as

$$\text{Resolution } = \frac{V_{\max}}{(2^n - 1)}$$

$$= \frac{1.5}{(2^8 - 1)} = 5.88 \text{ mV}$$

Resolution = Quantisation step q = 5.88 mV

$$\text{Quantisation error} = \frac{q}{2} = \frac{5.88\text{mV}}{2} = 2.94 \text{ mV}$$

6.17 For Problem 6.16, find out the coding efficiency for a resolution of 0.01V.

Solution

Dynamic Ratio (DR) is the largest possible magnitude to the smallest possible magnitude. For 0.01V resolution, Dynamic Ratio (DR) will be

$$20\log\left(\frac{1.5}{0.01}\right) = 43.5 \text{ dB OR } 150$$

Minimum number of bits required to get the dynamic range is

$$n = \frac{\log(DR+1)}{\log 2} = 7.24$$

$$\text{Coding efficiency} = \frac{\text{Maximum no. of bits}}{\text{Actual no. ofbits}} \times 100$$

$$= \frac{7.24}{8} \times 100 = 90.5\%$$

6.18 Calculate the minimum line speed in bits per second to transmit speech signal as a 10-bit PCM.

Solution

It is well known that the frequency range of a speech signal will be 300 Hz to 3300 Hz.

Lowest frequency =300 Hz

Highest frequency = 3300 Hz

For minimum line speed, the condition of Nyquist rate is to be satisfied which is twice the maximum frequency component.

Assuming the maximum frequency component of speech signal as 4 kHz, the sampling rate for speech signal will be 8 K samples/second.

For 7-bit PCM, minimum line speed will be

$$= \frac{8K\text{ samples}}{\text{Second}} \times \frac{10\text{ bits}}{\text{Sample}} = 80 \text{ kbp}$$

Problems

1. Calculate the transmission bandwidth for a PAM system in which the transmission of a signal has a maximum frequency of 3 kHz with sampling frequency of 8 kHz and the pulse duration of 0.1 T_s.
2. Determine the bandwidth of a PAM system in which a voice signal is being transmitted with modulating frequency of 6.3 kHz, if the sampling frequency is 12 kHz and sampling period is 0.01 $T_{s.}$
3. What is the code-word length for a TV signal having a bandwidth of 6 MHz is transmitted using PCM and that the number of quantisation levels is 1024?
4. For Problem- 3, find the bandwidth and (*S*/*N*) in decibels.
5. An analog voltage signal has a bandwidth of 150 Hz and amplitude range of -6.5 to + 6.5 V is transmitted over a PCM system with an accuracy of ±0.2% (full scale). Find the required sampling rate and number of bits in each PCM word.
6. A sinusoidal signal is to be transmitted utilising PCM so that the output signal-to-quantising noise ratio is 48.9 dB. Determine the minimum number of representation level *L* and binary code-word length '*n*' to achieve the required response.
7. A sinusoidal modulating signal of amplitude A_m uses all the representation levels provided for quantisation in case of full load condition. Determine the $(S/N)_{dB}$. Assume the number of quantisation levels to be 512.
8. The input voltage of a PCM compander with a minimum voltage range of 1 V and a μ of 285 is 0.35. Calculate the output voltage and gain of μ-law and A-law compander circuits.
9. Determine the minimum line speed in bps to transmit speech signal as an 8-bit PCM signal.
10. Calculate the resolution and quantisation error for a musical signal sent as an audio signal ranging from 20 Hz to 20 kHz as a 9-bit PCM if the musical signal swing is upto 1.7 volts

Chapter 7

Solved Problems

7.9 There are 12 message signals, each having a bandwidth of 15 kHz, to be multiplexed and transmitted. Calculate the minimum bandwidth required for a TDM system.

Solution

Number of channels = 12

Bandwidth of each channel = 15 kHz

Minimum bandwidth required is calculated as

$$f_c = Nf_m$$
$$= 12 \times 15 = 180 \text{ kHz}$$

7.10 There are 24 channels of voice signals are sampled uniformly and then time division multiplexed. The highest frequency component of each signal is 3.5 kHz. Calculate the minimum channel bandwidth required and the sampling rate in samples per second for an 8-bit encoder provided the signalling rate of the overall system as 1.5×10^6 bps.

Solution

Minimum channel bandwidth required = No. of channels × Max. frequency

$$= 24 \times 3.5 = 84 \text{ kHz}$$

Signaling rate for the overall system = 1.5×10^6 bps

For an individual channel, the bit rate will be

$$= \frac{1.5 \times 10^6 \text{ bps}}{24} = 62{,}500 \text{ bps}$$

Each sample is encoded using 8 bits, the samples per second will be

7.11 Calculate the minimum bandwidth for a binary PSK modulator with a carrier frequency of 45 MHz and an input bit rate of 550 kbps.

Solution

The input bit rate gives the maximum frequency of the baseband signal.

$$\therefore f_{\text{baseband}} = 550 \text{ kbps} = 550 \text{ kHz}$$

Bandwidth of BPSK signal is calculated as

Bandwidth = $2f_{\text{baseband}}$

$\therefore$ bandwidth = 2 × 550 = 1.1 MHz

7.12 In an FSK system, the amplitude of the received signal is 1 μV and the transmitted binary data rate is 3 Mbps. Calculate the transmitted power and the data rate for the signal transmission.

Solution

From $A = \sqrt{2P}$

Transmitted power is calculated as

$$P = \frac{A^2}{2}$$

$$P = \frac{(1\times10^{-6})^2}{2} = \frac{(1\times10^{-6})^2}{2} = 5 \times 10^{-13} \text{ Watts}$$

The bit duration is expressed as

$$T = \frac{1}{\text{Data rate}}$$

$$T = \frac{1}{3\times10^6} = 3.33\times10^{-7}\,\text{second}$$

7.13 Consider an FSK signal with a MARK frequency of 50 kHz, a SPACE frequency of 52 kHz and an input bit rate of 2 kbps. Calculate the peak frequency deviation, minimum bandwidth required and baud rate of the FSK signal.

Solution:

Peak frequency deviation $\Delta f = \frac{|50-52|}{2}$ = 1 kHz

Minimum bandwidth required

= 2 (1 kHz + 2 kbps) = 6 kHz

Baud rate $= \frac{\text{Bit rate}}{N}$

For FSK, $N = 1$

$$\therefore \text{baud rate} = \frac{2000}{1} = 2000$$

7.14 Consider a BPSK modulator with a fundamental frequency of 10 MHz and a carrier frequency of 75 MHz and an input bit rate of 12 Mbps. Calculate the upper and lower side band frequencies and draw its frequency spectrum.

Solution:

$$\text{BPSK output} = \sin(2\pi f_m t) \times \sin(2\pi f_c t)$$

$$= \sin(2\pi(10\text{ MHz})t) \times \sin(2\pi(75\text{ MHz})t)$$

$$= \frac{1}{2}\cos(2\pi(75\text{ MHz} - 10\text{ MHz})t) - \frac{1}{2}\cos(2\pi(75\text{ MHz} - 10\text{ MHz})t)$$

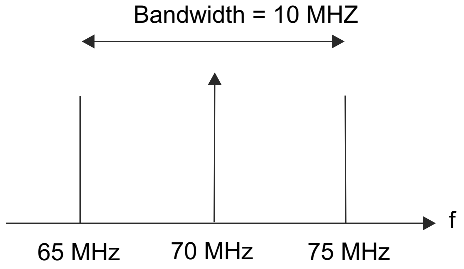
Lower side-band frequency (LSB) = 75 MHz – 10 MHz

= 65 MHz

Upper side-band frequency (USB) = 75 MHz + 10 MHz

= 85 MHz

Frequency spectrum is drawn as follows.



7.15 Consider a quadrature PSK modulator with a carrier frequency of 75 MHz and an input bit rate of 12 Mbps. Calculate the double-sided bandwidth.

Solution

For both *I* and *Q* channels, the bit rate is half of the transmission bit rate.

$$f_{bI} = f_{bQ} = \frac{1}{2} f_b = \frac{12\,\text{Mbps}}{2}$$

$$= 6\text{ Mbps}$$

The highest fundamental frequency to the channels will be

$$f_{\text{m}} = \frac{f_{\text{bI}}}{2} \text{ or } \frac{f_{\text{bQ}}}{2}$$

$$= \frac{6 \text{ Mbps}}{2} = 3 \text{ Mbps}$$

$$\text{QPSK output} = \sin(2\pi f_m t) \times \sin(2\pi f_c t)$$

$$= \sin(2\pi (10 \text{ MHz})t) \times \sin(2\pi (75 \text{ MHz})t)$$

$$= \frac{1}{2}\cos(2\pi (75 \text{ MHz} - 3 \text{ MHz})t) - \frac{1}{2}\cos(2\pi (75 \text{ MHz} + 3 \text{ MHz})t)$$

Lower side band frequency (LSB) = 75 MHz – 3 MHz

= 72 MHz

Upper side band frequency (USB) = 75 MHz + 3 MHz

= 78 MHz

Double sided bandwidth = 78 MHz + 72 MHz = 6 MHz

7.16 Consider an 8-PSK system which is operating with a bit rate of 24 kbps. Calculate the baud rate, minimum bandwidth and bandwidth efficiency.

Solution

$$\text{Baud rate} = \frac{\text{Bit rate}}{N}$$

$$= \frac{24000}{3} = 8000$$

$$\text{Minimum bandwidth required} = \frac{\text{Bit rate}}{N}$$

$$= \frac{24000}{3} = 8000$$

$$\text{Bandwidth efficiency} = \frac{24000}{8000}$$

= 3 bps/cycle of bandwidth

7.16 Determine the maximum baud rate for a 16-PSK transmission system with a 15 kHz bandwidth.

Solution

Maximum baud rate = Bandwidth × Bandwidth efficiency

Bandwidth efficiency of 16 PSK = 4

$\therefore$ maximum baud rate = 4 × 15000 = 60,000

Problems

1. For an FSK signal with a MARK frequency of 35 kHz, a SPACE frequency of 25 kHz and a bit rate of 5 kbps, calculate the baud rate and the bandwidth required.
2. For an FSK signal with a MARK frequency of 135 kHz, a SPACE frequency of 137 kHz and an available bandwidth of 10 kHz, calculate the maximum bit rate.
3. For a 16 PSK modulator with an input bit rate of 25 Mbps and a carrier frequency of 130 MHz, determine the double-sided bandwidth and the baud rate. Also sketch the frequency spectrum.
4. Determine the bandwidth efficiency for an 8 PSK with an input baud rate of 20 Mbps.
5. Calculate the minimum bandwidth required and the baud rate for a binary PSK modulator with a carrier frequency of 85 MHz and an input bit rate of 1 Mbps. Also sketch its output spectrum.
6. Consider an 8 PSK modulator with an input bit rate of 12 Mbps and a carrier frequency of 75 MHz. Determine the bandwidth and also plot its frequency spectrum.
7. For the DBPSK modulator, determine the output phase sequence for the following input bit sequence 00110011010101. Assume that the reference bit = 1.
8. Calculate the minimum channel bandwidth required for a system consisting of 36 channels of voice signals which are sampled uniformly and then time-division multiplexed. The highest frequency component of each signal is 4.5 kHz.
9. For Problem 8, calculate the sampling rate in samples per second for an 8-bit encoder provided the signalling rate of the overall system is 1.8×10^6 bps.
10. Draw the MSK modulated signals for a modulating signal sequence of 011010.

Chapter 8

Problems

1. Consider a source X that produces five symbols with probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$ and $\frac{1}{16}$. Determine the source entropy $H(X)$.
2. By assuming all 26 characters in the alphabet occur with equal probability, calculate the average information content in the English language.
3. A channel is described by the following channel matrix.

   $$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

   Find the channel diagram and the channel capacity.
4. A system X consists of five symbols x_1, x_2, x_3, x_4 and x_5 with respective probabilities 0.2, 0.15, 0.05, 0.1 and 0.5. Construct a Shannon-Fano code for X and calculate the code efficiency.
5. Construct a Huffman code and calculate the code efficiency for a system X consisting of five symbols x_1, x_2, x_3, x_4 and x_5 with respective probabilities 0.2, 0.15, 0.05, 0.1 and 0.5.
6. A source emits one of four symbols s_0, s_1, s_2 and s_3 with probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}$ and $\frac{1}{8}$ respectively. The successive symbols emitted by the source are statistically independent. Calculate the entropy of the source.
7. Consider a sequence of letters of the English alphabet with their probabilities of occurrence as given below.

Letter	a	i	l	m	n	o	p	y
Probability	0.1	0.1	0.2	0.1	0.1	0.2	0.1	0.1

   Construct a Huffman code and calculate the average codeword length and entropy.
8. A discrete memoryless source has an alphabet of seven symbols whose probabilities of occurrence are as follows.

Symbol	S_0	S_1	S_2	S_3	S_4	S_5	S_6
Probability	0.25	0.25	0.125	0.125	0.125	0.0625	0.0625

Compute the Huffman code for the above data and show the computed source has an efficiency of 100%

9. For the following ASCII coded message 'THE CAT', determine the VRC and LRC. Use odd parity for the VRC and even parity for the LRC.

10. Determine the BCS for the following data and CRC generating polynomials:

    $G(x) = x^9 + x^7 + x^5 + x^4 + x^2 + x^1 + x^0 = 110110111$

    CRC $P(x) = x^5 + x^4 + x^1 + x^0 = 110011$

Chapter 9

Solved Problems

9.8 Determine the noise margins for an RS-232 interface with driver signal voltages of ± 7 V.

Solution

Noise Margin (NM) is the difference between the driver signal voltage and the terminator receive voltage.

Noise margin = 7 – 3 = 4V or

Noise margin = 25 – 7 = 18V

Minimum noise margin is 4 V.

9.9 Consider a message given as 'KONGU.43'. Represent the message by ASCII code. Use even parity for 8th bit in MSB.

Solution

Characters in message	Parity bit 8	ASCII code						
		7	6	5	4	3	2	1
K	0	1	0	0	1	0	1	1
O	1	1	0	0	1	1	1	1
N	0	1	0	0	1	1	1	0
G	0	1	0	0	0	1	1	1
U	0	1	0	1	0	1	0	1
.	0	0	1	0	1	1	1	0
4	1	0	1	1	0	1	0	0
3	0	0	1	1	0	0	1	1

9.10 Represent the following message by ASCII code. Use odd parity for 8th bit in MSB. Message: SHARMIE.16

Solution

Characters in message	Parity bit 8	ASCII code						
		7	6	5	4	3	2	1
S	1	1	0	1	0	0	1	1
H	1	1	0	0	1	0	0	0
A	1	1	0	0	0	0	0	1
R	0	1	0	1	0	0	1	0
M	1	1	0	0	1	1	0	1
I	0	1	0	0	1	0	0	1
E	0	1	0	0	0	1	0	1
.	1	0	1	0	1	1	1	0
1	0	0	1	1	0	0	0	1
6	1	0	1	1	0	1	1	0

9.11 Obtain odd parity bits for EBCDIC characters S, v, C and 3. The hex codes for EBCDIC S, v, C and 3 characters are E2, A5, C3 and F3.

Solution:

Hex code for ‘S’ is E2 (i.e. 11100010)

Odd parity in the position of MSB is ‘1’.

Thus, ‘S’ with added parity is 111100010.

Hex code for ‘v’ is A5 (i.e. 10100101)

Odd parity in the position of MSB is ‘1’.

Thus, ‘v’ with added parity is 110100101.

Hex code for ‘C’ is C3 (i.e. 11000011)

Odd parity in the position of MSB is ‘1’.

Thus, ‘C’ with added parity is 111000011.

Hex code for ‘3’ is F3 (i.e. 11110011)

Odd parity in the position of MSB is ‘1’.

Thus, ‘3’ with added parity is 111110011.

9.12 Obtain even-parity bits for EBCDIC characters P, m, s and 4. The hex codes for EBCDIC P, m, s and 4 characters are D7, 94, A2 and F4.

Solution:

Hex code for ‘P’ is D7 (i.e. 11010111)

Even parity in the position of MSB is ‘0’.

Thus, ‘P’ with added parity is 011010111

Hex code for ‘m’ is 94 (i.e. 10010100)

Even parity in the position of MSB is ‘1’.

Thus, ‘m’ with added parity is 110010100.

Hex code for‘s’ is A2 (i.e. 10100010)

Even parity in the position of MSB is ‘1’.

Thus,‘s’ with added parity is 110100010.

Hex code for‘4’ is F4 (i.e. 11110100)

Even parity in the position of MSB is ‘1’.

Thus,‘4’ with added parity is 111110100.

Problems

1. Determine the speed of transmission in bits/second if a block of 256 sequential 10-bit data words is transmitted serially in 0.015 s. Also determine the time duration of one word.
2. During serial transmission, a group of 1024 sequential 12-bit data words is transmitted in 0.012 s. Find the speed of transmission in bps.
3. Calculate the maximum data rate if the bandwidth of the transmission channel is 3000 Hz and SNR is 1000 dB.
4. If the number of links required is 36, how many nodes can be connected together in a data communication network?
5. If a 1500-byte group of data is transmitted on a 16 Mbps token ring packet, find the speed of transmission.
6. Consider a message given as ‘NILA.97’. Represent the message by ASCII code. Use even parity for 8th bit in MSB.
7. Represent the following message by ASCII code. Use odd parity for 8th bit in MSB. Message: BALU.2013
8. Determine the noise margins for an RS-232 interface with driver output signal voltages of ± 10 V.
9. Obtain odd parity bits for EBCDIC characters s, p, M and 4. The hex codes for EBCDIC s, p, M and 4 characters are A2, 97, D4 and F4.
10. Obtain even parity bits for EBCDIC characters R, a, J and 5. The hex codes for EBCDIC R, a, J and 5 characters are D9, 81, D1 and F5.

Chapter 10

Problems

1. The optical fibre is made up of a glass with the refractive index of 1.45. The clad surrounding the fibre has a refractive index of 1.52. The launching of the light in the fibre is done through air. Calculate the numerical aperture of the fibre and acceptance angle.
2. Consider a fibre which has a core diameter of 50 μm, used at light of wavelength of 0.8 μm and the numerical aperture is 0.35. Determine V number and the number of modes supported.
3. A multimode step index fibre with a core refractive index of 1.49, core diameter of 65 μm and normalised frequency $V = 12$ at wavelength of 1.2 μm. Determine the numerical aperture, relative index difference and critical angle at the core-cladding interface.
4. Calculate the intermodal dispersion for a fibre cable of 11.5 km having n_1=1.45 and n_2 =1.51.
5. Find out the acceptance-cone half angle for an optical fibre with core and cladding refractive index of n_1=1.48 and n_2 =1.43.
6. A multimode graded index fibre has an acceptance angle of 7° in air. Find out the relative refractive index difference between the core axis and the cladding when the refractive index at the core axis is 1.51.
7. A step index fibre has n_1=1.43 and n_2=1.42 respectively. Calculate the acceptance angle in air for skew rays which change direction by 145° at each reflection.
8. Find the core radius necessary for single-mode operation at 750 nm of the step index fibre with refractive index of core as 1.51 and refractive index of cladding as 1.48.
9. Calculate the number of propagating modes at an operating wavelength of 900 nm for a step index fibre with a core diameter of 175 μm and NA of 0.28.
10. Determine the cut-off parameter and the number of modes supported by a fibre with refractive index of core as 1.51 and refractive index of cladding as 1.48. Core radius is 27 μm and operating wavelength is 1250 nm.

Chapter 11

Problems

1. A satellite moving in an elliptical eccentric orbit has the semi-major axis of the orbit equal to 15000 km. If the difference between the apogee and the perigee is 28000 km, determine the orbit eccentricity.
2. Determine the relationship between their orbital periods if Satellite-1 in an elliptical orbit has the orbit semi-major axis equal to 17500 km and Satellite-2 in an elliptical orbit has the semi-major axis equal to 24750 km.
3. Determine the semi-major axis of the elliptical orbit for a satellite in an elliptical orbit haring an apogee of 27,500 km and a perigee of 1500 km.
4. The farthest and the closest points in a satellite's elliptical eccentric orbit from earth's surface are 32000 km and 320 km respectively. Determine the apogee, the perigee and the orbit eccentricity. Assume radius of earth to be 6370 km.
5. Calculate the escape velocity for an object to be launched from the surface of earth from a point where earth's radius is 6240 km. Gravitation constant is 6.67×10^{-11} Nm^2/kg^2 and mass of earth is 5.98×1024 kg.
6. The distance from the centre of an ellipse to the centre of the earth is 26500 km. Gravitation constant is 6.67×10^{-11} Nm^2/kg^2 and mass of the earth is 5.98×1024 kg. Find out the orbital period of a satellite in an eccentric elliptical orbit.
7. Determine the orbital time period and the velocity at the apogee and perigee points if a satellite moving in a highly eccentric orbit having the farthest and closest points as 36,250 km and 450 km respectively from the surface of earth. Assume the earth's radius is 6360 km.
8. The sum of apogee and perigee distances of a certain elliptical satellite orbit is 48000 km and the difference of apogee and perigee distances is 28000 km. Determine the target eccentricity.
9. Determine the apogee and perigee distances if the semi-major axis and the semi-minor axis of an elliptical satellite orbit are given as 24,000 km and 18500 km respectively.
10. A geosynchronous satellite orbiting at 43,064 km from the earth's centre has a circular equatorial orbit. The orbit gets inclined due to some reason and it is observed that the maximum displacement due to latitude deviation is 480 km. Determine the angle of inclination between the new orbital plane and the equatorial plane.

Chapter 12

Problems

1. Consider that a pulsed radar with a Pulse Repetition Frequency (PRF) of 1.2 kHz receives an echo pulse exactly 0.20 ms after it transmits. Determine the maximum unambiguous range of the radar and also find the target range in km.
2. Find out the centre of the frequency spectrum, interline spacing of the spectrum and the matched bandwidth of a pulse that is 11.5 μs wide and has an RF signal frequency of 12 GHz. Also determine the frequencies of the spectral lines closest to the centre if the PRF is 1.0 kHz.
3. The transmitted pulse has a 4.3 μs wide envelope. The frequency across this width is swept linearly from 485 to 495 MHz. Calculate the centre of spectrum, matched bandwidth and the compressed pulse width.
4. With a CW transmit frequency of 5 GHz, determine the Doppler frequency seen by a stationary radar when the target radial velocity is 125 km/hour.
5. A vehicle is moving towards stationary CW Doppler radar transmitting at 12.3 GHz along the axis of the radar with a speed of 110 km/hour. If the vehicle is moving away from the radar along the same axis, what would be the received signal frequency? Also determine the Doppler shift and the frequency of the received signal.
6. Determine the Doppler shift if a pulse Doppler radar emitting at 12.5 GHz has PRF of 2.4 kHz and also the radar is capable of measuring the radial velocity of 170 m/s of a closing target without any ambiguity. What would be the desired pulse-repetition rate to achieve that?
7. Calculate the approximate bandwidth of the spectrum of the waveform and the range resolution of the radar if a CW radar waveform is modulated by a 105 Hz sinusoidal signal and the amplitude of the modulating signal is such that it causes a maximum frequency deviation of 575 Hz.
8. Determine the radial-velocity component for an MTI radar system operating at 15 GHz and a repetition rate of 1200 Hz receiving echoes from an aircraft approaching the radar with a radial-velocity component of 1.5 km/s.
9. Find the lowest blind speed if a 4.2 cm MTI is operating at a PRF of 2180 Hz.
10. Consider a vehicle is moving towards stationary CW Doppler radar transmitting at 11.5 GHz along the axis of the radar with a speed of 120 km/hour. Determine the Doppler shift and the frequency of the received signal.

Chapter 13

Solved Problems

13.1 Consider a cellular telephone area comprised of 15 clusters with seven cells in each cluster and 12 channels in each cell. Find out the number of channels per cluster and the total channel capacity.

Solution

Total number of cellular channels in a cluster can be expressed as

F = Number of channels in a cell × Number of cells in a cluster

$\therefore\ F = 15 \times 7 = 105$ channels/cluster

Total channel capacity can be expressed as

C = Number of 6 channels in a cluster × Number of cluster in a given area

$C = F \times m = 105 \times 12$

$= 1260$ channels

13.2 A cellular telephone area is comprised of 7 macro cells with 15 channels per cell. If each macro cell is further divided into four mini cells, find out the channel capacity.

Solution:

Channel capacity is calculated as follows.

$$\frac{15\,\text{Channels}}{\text{Cell}} \times \frac{7\text{ Cells}}{\text{Area}} = 105\text{ channels/area}$$

13.3 A cellular telephone area is comprised of 7 macro cells with 15 channels per cell. Find out the channel capacity if each mini cell is further divided into four micro cells.

Solution

The total number of cells in the area will be increased if each macro cell is split into 4 mini cells and which will be 4 × 7 = 28.

$$\therefore \frac{15\,\text{Channels}}{\text{Cell}} \times \frac{28\text{ Cells}}{\text{Area}} = 420\text{ channels/area}$$

13.4 If the channel capacity is 140 channels per area for a cellular telephone area which is comprised of seven macro cells, find out the number of channels occupied per cell.

Solution:

$$\frac{\text{X Channels}}{\text{Cell}} \times \frac{\text{7 Cells}}{\text{Area}} = 140 \text{ channels/area}$$

Number of channels occupied per cell = 20

13.5 Find out the total channel capacity for a cellular telephone area comprised of 12 clusters with seven cells in each cluster and 10 channels in each cell.

Solution:

Total number of cellular channels in a cluster can be calculated as

$\therefore F = 12 \times 7 = 84$ channels/cluster

Total channel capacity can be expressed as

C = Number of channels in a cluster × Number of clusters in a given area

$C = F \times m = 84 \times 10$

$= 840$ channels

Problems

1. Calculate the number of channels per cluster and the total channel capacity for a cellular telephone area comprised of 12 clusters with 7 cells in each cluster and 16 channels in each cell.
2. If the channel capacity is 210 channels per area for a cellular telephone area which is comprised of seven macro cells, find out the number of channels occupied per cell.
3. A cellular company has 150 full-duplex channels for a given service area. It is decided to divide the working area into 15 clusters and use a seven-cell reuse pattern and use the same number of channels in each cell. Calculate the total number of channels available for its subscribers at any time.
4. Determine the channel capacity for a cellular telephone area comprised of seven macro cells with 15 channels per cell. Also find out the channel capacity if each macro cell is divided into four mini cells and the channel capacity of each mini cell is further divided into four micro cells.
5. Find out the distance from the nearest co-channel for a cell radius of 0.3 miles and a co-channel reuse factor of 10.

Chapter 14

Solved Problems

14.1 Find out the characteristic impedance for an air dielectric two-wire parallel transmission line with *D*/*r* ratio of 12.34.

Solution:

The characteristic impedance of a two-wire parallel transmission line with an air dielectric can be determined by

$$Z_0 = 276 \log \frac{D}{r}$$

$$\therefore Z_0 = 276 \log (12.34) = 301.2\ \Omega$$

14.2 Calculate the characteristic impedance for a concentric co-axial cable with $d = 0.025$ inches, $D = 0.12$ inches and $e = 2.25$.

Solution:

The characteristic impedance of a concentric co-axial cable can be calculated by

$$Z_0 = \frac{138}{\sqrt{e}}\left(\log \frac{D}{d}\right)$$

$$= \frac{138}{\sqrt{2.25}}\left(\log \frac{0.12}{0.025}\right) = 62.67\ \Omega$$

14.3 For a concentric co-axial cable with $L = 0.12$ μH/ft and $C = 18$ pF/ft, determine the characteristic impedance.

Solution

$$Z_0 = \sqrt{\frac{L}{C}}$$

$$\therefore Z_0 = \sqrt{\frac{0.12 \times 10^{-6}}{18 \times 10^{-12}}} = 82\ \Omega$$

14.4 For a transmission line with an incident voltage $E_i = 6$V and reflected voltage $E_r = 3$V, determine the reflection coefficient and Standing Wave Ratio (SWR).

Solution:

$$\text{Reflection coefficient } = \frac{E_r}{E_i} = \frac{3}{6} = 0.5$$

$$\text{Standing Wave Ratio (SWR) } = \frac{E_i + E_r}{E_i - E_r} = \frac{6+3}{6-3}$$

$$= \frac{9}{3} = 3$$

14.5 How far down the cable is the impairment if a pulse down a cable has a velocity of propagation of 0.9 C and the reflected signal is received 1μs later.

Solution:

The distance between the impairment and the source can be determined by

$$d = \frac{v \times t}{2}$$

$$\therefore d = \frac{0.9(3 \times 10^8) \times 1 \times 10^{-6}}{2}$$

= 135 m

14.6 Determine the characteristic impedance of a transmission line which has a capacitance of 30 pF/ft and an inductance of 0.20 μH/ft.

Solution:

$$Z_0 = \sqrt{\frac{L}{C}}$$

$$\therefore Z_0 = \sqrt{\frac{0.20 \times 10^{-6}}{30 \times 10^{-12}}} = 81.6\ \Omega$$

14.7 A particular cable has a capacitance of 40 pF/ft and a characteristic impedance of 70 Ω. Find out the inductance of this cable.

Solution:

It is known that,

$$Z_0 = \sqrt{\frac{L}{C}}$$

By substituting the values of Z_0 and C,

$$70 = \sqrt{\frac{L}{40 \times 10^{-12}}}$$

$$70^2 = \frac{L}{40 \times 10^{-12}}$$

$L = 70^2 \times 40 \times 10^{-12}$

$\therefore L = 0.196$ μH/ft

14.8 Voltage and current readings are taken on a transmission line at different points. The maximum voltage reading is 60 $E_{\max}$ and the minimum voltage reading is 20 $E_{\min}$. Calculate VSWR and ISWR. If the maximum current reading on this line is 3 A, what would be the lowest current reading be?

Solution:

It is known that

$$\text{VSWR} = \left|\frac{E_{\max}}{E_{\min}}\right|$$

$$\therefore \text{VSWR} = \left|\frac{60}{20}\right| = 3$$

It is also known that

$$\text{ISWR} = \left|\frac{I_{\max}}{I_{\min}}\right|$$

Since VSWR = ISWR = 3

To find the lowest current reading,

$$\text{ISWR} = \left|\frac{I_{\max}}{I_{\min}}\right|$$

$$3 = \frac{3\,\text{A}}{I_{\min}}$$

$$\therefore I_{\min} = \frac{3\,\text{A}}{3} = 1\,\text{A}$$

14.9 A transmission line having characteristic impedance of 80 Ω is delivering power to a 160 Ω load. Determine SWR on this line and also find the minimum voltage reading on this line if the maximum voltage is 25 V.

Solution:

SWR can be expressed as

$$\text{SWR} = \frac{Z_L}{Z_0}$$

$$\therefore \text{SWR} = \frac{160}{80} = 2$$

To find the minimum voltage reading,

$$\text{VSWR} = \left| \frac{E_{\text{max}}}{E_{\text{min}}} \right|$$

By substituting the given values,

$$2 = \frac{25\,\text{V}}{V_{\text{min}}}$$

$$\therefore V_{\text{min}} = \frac{25}{2} = 12.5\,\text{V}$$

14.10 A 50 Ω load is being applied from a 70 Ω transmission line. (i) Find the SWR resulting from this mismatch. (ii) Determine the reflection coefficient resulting from this mismatch. (iii) What percentage of the incident power is reflected from the load? (iv) What percentage of the incident power is absorbed by the load?

Solution:

i) To find SWR:

$$\text{SWR} = \frac{Z_L}{Z_0}$$

$$\text{SWR} = \frac{70}{50} = 1.4$$

ii) To find the reflection coefficient:

$$K_r = \left| \frac{Z_L - Z_0}{Z_L - Z_0} \right|$$

$$= \left|\frac{70-50}{70+50}\right| = \left|\frac{20}{120}\right| = 0.17$$

iii) To find the percentage of the incident power reflected from the load

$$K_r^{\,2} = \frac{P_{\text{ref}}}{P_{\text{inc}}}$$

$$0.17^2 = \frac{P_{\text{ref}}}{P_{\text{inc}}}$$

$$0.0289 = \frac{P_{\text{ref}}}{P_{\text{inc}}}$$

$$\%\ P_{\text{ref}} = 0.0289 \times 100 = 2.89\%$$

iv) To find the percentage of the incident power absorbed by the load

$$\%\ P_{\text{abs}} = 100 - 2.89 = 97.11\%$$

Problems

1. A cable has a capacitance of 10 pF/ft and an inductance of 0.03 μH/ft. What is the characteristic impedance of the cable?
2. Find the characteristic impedance of the cable having a capacitance of 25 pF/ft and an inductance of 0.18 μH/ft.
3. Calculate the capacitance of a transmission cable that has a characteristic impedance of 65 Ω and an inductance of 0.29 μH/ft.
4. A cable has a capacitance of 45 pF/ft and a characteristic impedance of 350 Ω. What is the inductance of this cable?
5. Find the inductance of a transmission cable that has a characteristic impedance of 250 Ω and a capacitance of 35 pF/ft.
6. Determine the SWR on a transmission line on which the maximum voltage and minimum voltage are 80 V and 25 V respectively.
7. Calculate SWR on a transmission line which has the maximum voltage of 120 V and a minimum voltage of 50 V.
8. Find SWR on a transmission line which has the maximum current of 1.25 A and a minimum current of 0.75 A.

9. From a transmission line, voltage and current readings are taken at different places. The maximum voltage and minimum voltage are 140 V and 65 V respectively. The maximum current reading on the line is 4.5 A. Find (i) VSWR, (ii) ISWR, and (iii) Determine the lowest current reading on the line.
10. The maximum current reading along the transmission line is 8.2 A while the minimum current reading on the line is 2.5 A. (i) What is the ISWR? (ii) If the maximum voltage on the line is 170 V, what is the minimum voltage on the line?

INDEX

A

Absorption loss, 425
Acceptance angle, 400
Adaptive delta modulation, 219
Additive noise, 38
Aliasing, 272
AM detectors, 104
AM receivers, 113
AM superheterodyne receivers, 117
  Advantages, 128
  Applications, 128
AM transmitters, 93
Amplitude Modulation AM, 52
  Definition, 53
  Modulation index, 54
  Frequency spectrum, 54
  Graphical representation, 55
  Phasor representation, 58
  Degrees of freedom, 58
  Power relation, 60
  Current relation, 61
  Drawbacks, 69
  Modulation by several sine waves, 70
  Applications, 82
  Generation, 82
Amplitude Shift Keying (ASK), 52, 238, 240
  Generation, 241
  ASK generation using 555 timer, 242
  ASK detection, 243
Analog pulse modulation, 187

Definition, 188
Angle of incidence, 397
Angle of refraction, 397
Antenna directivity, 508
Antenna radiation pattern lobes, 510
Antenna scanning pattern, 534
Anti aliasing filter, 273
Aperiodic signal, 9
Aperture, 537
Armstrong method, 157
ASCII Code, 330
Asynchronous modem, 269
Asynchronous transmission, 337
Atmospheric noise, 34
Automatic Frequency Control (AFC), 180
Automatic Gain Control (AGC), 128
AGC circuit, 129
AGC circuit with ϖ-filter, 130
Delayed AGC, 130
Characteristics, 132
Amplified AGC, 131
Avalanche Photo Diode (APD), 445
Azimuth angle, 489

B

Back propagation loss, 428
Back scatter, 519
Balanced modulators, 88
Balanced slope detector, 166
Bandlimited signals, 27
Band pass signals, 28
Bar code, 333
Baseband signal, 49
Basic radar system, 518
Classification, 526
Antennas, 532
Basic service set, 587
Baudot teletype code, 331
Bauds, 340
BCDIC code, 333
Beam width, 536
Bending loss, 427
Binary Erasure Channel (BEC), 289
Binary Phase Shift Keying (BPSK), 252

Implementation, 252
Error rate calculation, 252
Generation, 253
Detection, 253
Binary Symmetric Channel (BSC), 289
Bit Error Rate (BER), 308, 404
Bit rate, 336
Blind dopplers, 549
Blind speed, 548
Block Check Sequence, 312
Bluetooth, 591
Components, 592
Stack, 593
Links, 594
Protocols, 594
L2CAP, 594
Networking, 595
Connection, 596
Transmission, 597
Bridges, 589

C

Caller ID, 605
Cassegrain antenna, 495
Cell footprint, 570
Cell splitting, 573
Cellular architecture, 569
Cellular concept, 569
Channel capacity, 286
Properties, 287
Examples, 288
Character Error Rate, 309
Characteristic impedance, 630
Check sum error detection, 310
Circular polar orbit, 474
Cladding, 407
Classification of signals, 7
Cluster, 570
Coarse WDM, 454
Coherent detector, 108, 243, 248, 253
Coherent PSK, 251
Common-Base detector, 107
Common-Emitter detector, 106
Communication, 1

History, 3
Components, 3
Types, 5
Companding in PCM, 215
Comparison between TDM and FDM, 237
Complex signal, 8
Conditional entropy, 284
Conical scanning, 557
Continuous Wave (CW) radar, 527,552
Advantages, 553
Drawbacks, 553
Continuous-Time signal, 8
Convolutional code, 322
Cordless Telephones, 606
Base, 608
Handset, 609
Features, 610
Core of an optical fibre, 401
Critical angle, 399
CSMA/CD- IEEE 802.3, 379
Cyclic Redundancy Checking (CRC), 312
Cylindrical scanning, 614

D

Data communication codes, 329
Data representation, 330
Hardware, 341
Data Communication Equipment (DCE), 341
Data communication networks, 347
Applications, 348
Components, 349
Functions, 350
Data signal, 7
Data Terminal Equipment (DTE), 341
Data transmission, 334
Parallel, 334
Serial, 334
Modes, 337
De- Emphasis, 181
Dedicated client/server network, 355
Delta modulation, 218
Demodulation of PM, 176
Dense WDM, 454
Deterministic signal, 9

Differential PCM (DPCM), 210
Differential PSK, 261
  Generation, 261
  Detection, 262
Digital cellular system, 581
Digital modulation, 237
  Types, 238
  Principles, 238
  I/Q formats, 239
  Error rate, 240
  Constellation diagram, 240
Digital pulse modulation, 205
Digital signals, 6
Directional capabilities, 340
Directive gain, 522
Directivity, 522
Discrete memoryless source, 279
Discrete-Time signal, 8
Distributed constants, 628
Doppler effect, 543
Double Side Band Suppressed Carrier AM (DSB-SC-AM), 71
  Phasor representation, 72
  Power calculation, 73
  Advantages, 74
  Disadvantages, 75
  Applications, 75
  Generation, 87
Downlink system, 494
Duplex WDM, 453

E

Earth beam, 511
Earth coverage angle, 490
Earth segment, 464, 469
EBCDIC code, 333
Edge emitting diode, 442
Electrical signals, 5
Electromagnetic spectrum, 41
  Range, 41
  Types of radiation, 43
Electronic communication, 2
Elevation angle, 489

Elliptical orbit, 475
Energy signal, 12
Entropy, 282
Error control, 306
Error correction, 308
  Methods, 317
Error detection, 307
  Methods, 309
Even signal, 11
Expendable rocket, 498
Extended service set, 588
Eye pattern, 304
  Interpretation, 305

F

Facsimile (FAX), 613
  Applications, 613
  Transmitter, 613
  Receiver, 616
  Transmission, 620
Fan beam, 509
Filter method, 90
FM detectors, 165
FM receivers, 177
  Characteristics, 178
FM superheterodyne receiver, 177
FM transmitters, 158
  Directly modulated, 158
  Indirectly modulated, 159
Footprints, 510
Forward error correction, 317
Foster-Seeley discriminator, 167
  Advantages, 173
  Disadvantages, 173
Fourier transform, 25
Frame error rate, 309
Frequency demodulation, 166
Frequency discriminators, 165
Frequency Division Multiplexing (FDM), 232
  Illustration, 233
  Operation, 234
  Frequency translation, 235
Frequency modulated CW radar, 553
Frequency modulation, 136

Definition, 136
Frequency spectrum, 137
Bessel function, 139
Bandwidth, 144
Power calculation, 147
Vector representation, 148
Comparison with AM, 148
Conversion to PM, 152
Generation, 153
Advantages, 162
Disadvantages, 162
Frequency reuse factor, 572
Frequency reuse, 571
Frequency Shift Keying (FSK), 52, 246
Generation, 247
FSK detection, 248
Frequency synthesiser, 578
Full duplex modem, 266
Fundamental period, 9

G

GaAs Laser diode, 438
Gas lasers, 437
Geostationary orbit, 475
Geosynchronous orbit, 473
Glass fibres, 416
Global Positioning System (GPS), 565
Go-Back–N, 324
Graded-index fibres, 413
Ground range, 520
Ground wave transmission, 499

H

Half duplex modem, 266
Hamming code, 320
Hamming distance, 307
Hemispherical beam, 511
Heterojunction laser, 438
Hierarchical communication, 365
Homojunction laser, 438
Huffman coding, 298
Humanmade noise, 34

I

Image frequency Rejection Ratio, 122
Image frequency, 121
Index of cooperation, 617
Index profile, 425
Information rate, 284
Information theory, 278
Insertion loss, 428
Integrated Services Digital Network (ISDN), 371
  Principle, 371
  Subscriber's view, 371
  Objectives, 371
  Architecture, 372
  Broadband ISDN, 374
Inter Symbol Interference (ISI), 301
  Causes, 303
Interpolation, 270
Isosynchronous transmission, 339

K

Kepler Laws, 477

L

LAMDANET Star network, 456
LAN protocol architecture, 377
LASER, 433
  Operation, 433
  Types, 436
  Uses, 440
Latitude drift, 488
Launch phase, 497
Launch vehicles, 497
Layer entity, 364
Leakage current, 630
Light Emitting Diode (LED), 442
  Requirements, 442
  Structure, 442
  Properties, 443
Light guide, 397
Limitations of optical fibre cables, 418, 447
Line losses, 632

Copper losses, 632
Dielectric losses, 632
Radiation losses, 633
Induction losses, 633
Liquid Lasers, 437
Lobe switching, 556
Local operating system, 352
Longitude drift, 488
Longitude Redundancy Checking (LRC), 311
Low-earth orbit, 472
Lumped constants, 628

M

Mathematical models, 38
Additive noise channel, 39
Linear filter channel, 39
Linear time-variant filter channel, 40
Maximum Shift Keying (MSK), 250
Medium-earth orbit, 472
Mobile communication, 567
Modems, 263, 340
Connections, 264
Blocks, 264
Features, 265
Types, 265
4-wire modem, 267
2-wire modem, 267
Bell system modem, 269
Modulated signal, 49
Modulation, 49
Need for modulation, 50
Types, 51
Monopulse tracking, 558
Moving Target Indicator (MTI), 543
MTI principle, 546
MTI radar, 546
Multi fibre cable system, 419
Multi mode fibres, 409
Multi mode graded- index fibre, 424
Multi mode step-index fibre, 423
Multipath propagation, 303
Multiple Access (MA) techniques, 582
Frequency division MA, 582
Time division MA, 583

Code division MA, 585
Spatial division MA, 585
Multiplex WDM, 453
Multiplexing, 224
Need, 224
Concept, 225
Multivibrator modulator, 155

N

Network classification, 362
Local Area Network (LAN), 363, 375
Metropolitan Area Network (MAN), 363
Wide Area Network (WAN), 363
Global Area Network (GAN), 364
Network Interface Card (NIC),352
Network models, 353
Network operating system, 353
Network topologies, 356
Point-to-point line configuration, 356
Multipoint configuration, 357
Mesh, 357
Star, 358
Tree, 359
Bus, 360
Ring, 361
Hybrid, 362
Newton’s Laws, 478
Noise figure-37
Noise in communication system, 33
External noise, 33
Internal noise, 35
Noise power, 38
Noise temperature, 38
Noiseless binary channel, 288
Noisy channel, 288
Non-coherent detection, 244,248
Advantages, 245
Drawbacks, 245
Non linear AM modulators, 82
Normalized frequency parameter, 409
Numerical Aperture,(NA), 402
Nyquist frequency, 272
Nyquist rate, 27

O

Odd signal, 11
Omnidirectional beam, 508
Open System Interconnection (OSI), 367
Optical connectors, 429
  Structure, 429
  Interface, 430
  Losses, 431
  Applications, 431
  Features, 431
Optical detectors, 443
  Requirements, 443
Optical Fibre Communication, 388
  Generations, 389
  Spectra, 389
  Advantages, 391
  Disadvantages, 392
Optical fibre configurations, 421
Optical fibre splicing, 432
Optical fibres, 406, 636
  Structure, 407
  Types, 408, 416
  Losses, 425
  Applications, 447
Optical sources, 432
  Requirements, 433
Orbit-injection phase, 497
Orbital adjustments, 476
Orbital parameter, 479
OSI model, 367

P

Pager, 610
  Types, 611
  Operation, 612
Parabolic dish antenna, 532
Parallel interface, 346
Parity checking, 310
Partition noise, 36
Peer-to-peer client/server network, 354
Peer-to-peer communication, 366
Peer-to-peer LAN, 588

Pencil beam, 509
Periodic signal, 9
Phase modulation, 149
  Definition, 149
  Modulation index, 150
  Conversion to FM, 151
  Comparison with FM, 153
Phase Shift Keying (PSK), 250
  Representation, 251
Phase shift method, 91
Phased array antenna, 534
p-i-n photo diodes, 444
Plan Position Indicator (PPI), 519, 541
Planck's law, 394
Plastic clad silica fibre, 417
Plastic fibres, 417
PN Junction Laser diode, 438
Polarisation, 394, 508
Population inversion, 435
Power signal, 12
Pre-emphasis, 181
Propagation of light waves, 393, 399
Protocol, 366
Pulse Amplitude Modulation (PAM), 188, 190
  Definition, 190
  Mathematical analysis, 190
  Frequency spectrum, 191
  Bandwidth, 192
  Generation, 194
  Sampling considerations, 195
  Demodulation, 195
  Transmission, 196
  Limitations, 196
Pulse Code Modulation (PCM), 206
  Binary coding, 207
  Bandwidth, 208
  Detection, 209
  Transmission system, 210
  Reception system, 210
  Differential PCM, 210
  SNR, 211
  Companding, 215
Pulse Duration Modulation (PDM), 197
  Definition, 197
  Graphical representation, 197

Mathematical analysis, 198
Bandwidth, 199
Generation, 200
PDM transmission, 202
Demodulation, 202
Pulse modulation, 186
Definition, 186
Types, 187
Characteristics, 189
Pulse Position Modulation (PPM), 203
Definition, 203
Definition from PDM, 203
Definition from PAM, 204
Comparison with PDM, 204
Pulse radar system, 529
Pulse radar, 528
Pulse repetition time, 519

Q

Quadrature Phase Shift Keying (QPSK), 255
Implementation, 255
Error rate, 255
Generation, 256
Detection, 257
Timing diagram, 258
Offset QPSK, 258
ϖ/4 QPSK, 259
Quadrature PLL detector, 110
Quantisation process, 270
Characteristics, 274
Noise, 274

R

Radar displays, 538
A-scope, 539
B-scope, 540
F-scope, 541
PPI, 541
Radar measurement, 517
Radar signal, 7
Radar, 517
Frequencies, 523
Radiation pattern, 506

Parameters, 507
Types, 508
Categories, 510
Radio broadcast signals, 6
Radio receiver, 113
Classification, 113
Features, 113
Radio wave transmission, 499
Random signal, 9
Range tracking, 558
Ratio detector, 174
Amplitude limiting, 175
Rayleigh scatter loss, 427
Real signal, 8
Reception of data bits, 336
Redundancy checking, 310
Refraction, 395
Refractive index, 396
Residual Error Rate, 309
Return loss, 428
Reverse error correction, 323
Ring modulators, 88
Roaming, 589
Internal, 590
External, 590
Round trip time, 521
RS-232 Interface, 342
RS-422 Interface, 344
RS-485 Interface, 345

S

Sampling 270
Sampling theorem, 272
Satellite antennas, 502
Dipole, 503
Folded dipole, 504
Directional, 504
Parasitic arrays, 504
Horn, 505
Helical, 506
Satellite communication, 461
Features, 461
Advantages, 463
Frequency bands, 463

Elements, 464
Applications, 464
Block diagram, 466
Theory, 471
Satellite earth station subsystem, 493
Power supply subsystem, 494
Transmitting sub system, 494
Receiving sub system, 494
Antenna sub system, 495
Satellite launching, 497
Satellite orbits, 473
Satellite transponder subsystem, 496
Satellite, 461, 464
Scattering, 519
Search radar, 543
Selective retransmission, 324
Semiconductor Lasers, 437
Semiconductor reactance modulator, 154
Sequential lobing, 557
Serial interfaces, 341
Series diode detector, 105
Shannon-Fano coding, 293
Shaped beam, 509
Shot noise, 36
Shunt diode detector, 106
Signal representation, 19
Time domain, 19
Frequency domain, 20
Signal routing, 519
Signal- to-Noise ratio (SNR), 180
Single mode dual step-index fibre, 422
Single mode step-index fibre, 421
Single Side Band Suppressed Carrier AM (SSB-SC-AM), 75
Frequency spectrum, 76
Graphical representation, 76
Phasor representation, 77
Power calculation, 77
SSB over AM with carrier, 79
Advantages, 79
Disadvantages, 80
Applications, 80
Generation, 89
Single tuned discriminator, 166
Single-mode fibres, 409
Skew ray propagation, 405

Sky wave transmission, 500
Slant angle, 490
Slant range, 520
Slope detector, 166
Snell's law, 396
Solid lasers, 437
Source coding, 290
  Efficiency, 291
  Theorem, 292
Space shuttle, 499
Space wave transmission, 501
Spectral slicing, 457
Spontaneous emission, 434
Spot beam, 511
Square law modulators, 82
SSB AM transmitters, 99
Stagger tuned discriminator, 166
Standing Wave Ratio (SWR), 633
  Voltage SWR, 633
  Current SWR, 634
  Power SWR, 634
Step-index fibres, 408
Stimulated absorption, 433
Stop and wait, 323
Surface emitting diode, 442
Surge impedance, 630
Synchronous modem, 269
Synchronous transmission, 338

T

Tape scanning, 614
Telegraph signals, 6
Telephone signals, 5
Thermal agitation noise, 35
Time Division Multiplexing (TDM), 225
  Slow-speed TDM, 226
  High-speed TDM, 227
  PAM/TDM system, 227
  Implementation, 229
  Synchronous TDM, 229
  Asynchronous TDM, 230
  High order multiplexing, 231
Timing errors, 336
Token Bus-IEEE 802.4, 378
Token passing IEEE 802.5, 380

Toroidal radiation pattern, 506
Total Internal Reflection (TIR), 397
Track modes, 555
Tracking radar, 543,555
  Block diagram, 556
  Types, 556
Transmission channel, 339
Transmission lines, 626
  Types, 626
  Model, 626
  Theory, 627
Transmission media, 634
  Guided, 635
  Unguided, 637
Transponders, 468, 496
Tropospheric scattering, 501
Tuned Radio Frequency (TRF) receiver, 114
TV picture signals, 6
Types of modulation, 52
Types of satellites, 470

U

Uplink system, 494
Ultraviolet rays, 45

V

Varactor FM modulator, 156
Velocity tracking, 559
Vertical Redundancy Checking (VRC), 311
Vestigial Side Band (VSB AM), 80
  Block diagram, 81
  Frequency spectrum, 81
  Generation, 92
VSB demodulation, 111

W

Wavelength Division Multiplexing (WDM), 453
  Architecture, 454
  Types, 454
  Advantages, 457

Wireless communication, 564
- Needs, 564
- Examples, 565
- Transmission media, 565
- Applications, 565

Wireless distribution system, 589

Wireless LAN, 586
- Types, 588

Wireless Personal Area Network (WPAN), 591

WLAN standards, 590

Z

Zigbee, 597
- Stack, 598
- Layers, 599
- Addressing, 602
- Topologies, 603
- Applications, 604

Zonal beam, 511