MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

Department of Electronics & Communication Engineering

QUESTION BANK

EC8751 - OPTICAL COMMUNICATION

VII SEMESTER
Regulation – 2017

EC8751 OPTICAL COMMUNICATION L T P C 3 0 0 3

OBJECTIVES:

☐ To study about the various optical fiber modes, configuration and transmission characteristics of optical fibers

☐ To learn about the various optical sources, detectors and transmission techniques

☐ To explore various idea about optical fiber measurements and various coupling techniques

☐ To enrich the knowledge about optical communication systems and networks

UNIT I INTRODUCTION TO OPTICAL FIBERS 9

Introduction-general optical fiber communication system- basic optical laws and definitionsoptical modes and configurations -mode analysis for optical propagation through fibersmodes in planar wave guide-modes in cylindrical optical fiber-transverse electric and transverse magnetic modes- fiber materials-fiber fabrication techniques-fiber optic cables classification of optical fiber-single mode fiber-graded index fiber.

UNIT II TRANSMISSION CHARACTERISTIC OF OPTICAL FIBER 9

Attenuation-absorption --scattering losses-bending losses-core and cladding losses-signal dispersion –inter symbol interference and bandwidth-intra model dispersion-material dispersion- waveguide dispersion-polarization mode dispersion-intermodal dispersion-dispersion optimization of single mode fiber-characteristics of single mode fiber-R-I Profile cutoff wave length-dispersion calculation-mode field diameter.

UNIT III OPTICAL SOURCES AND DETECTORS 9

Sources: Intrinsic and extrinsic material-direct and indirect band gaps-LED-LED structuressurface emitting LED-Edge emitting LED-quantum efficiency and LED power-light source materials-modulation of LED-LASER diodes-modes and threshold conditions- Rate equations-external quantum efficiency-resonant frequencies-structures and radiation patterns-single mode laser-external modulation-temperature effort.
Detectors: PIN photo detector-Avalanche photo diodes-Photo detector noise-noise sources-SNR-detector response time-Avalanche multiplication noise-temperature effectscomparisons of photo detectors.

UNIT IV OPTICAL RECEIVER, MEASUREMENTS AND COUPLING 9

Fundamental receiver operation-preamplifiers-digital signal transmission-error sources-Front end amplifiers-digital receiver performance-probability of error-receiver sensitivity-quantum limit.
Optical power measurement-attenuation measurement-dispersion measurement- Fiber Numerical Aperture Measurements- Fiber cut- off Wave length Measurements- Fiber diameter measurements-Source to Fiber Power Launching-Lensing Schemes for Coupling Management-Fiber to Fiber Joints-LED Coupling to Single Mode Fibers-Fiber Splicing- Optical Fiber connectors.

UNIT V OPTICAL COMMUNICATION SYSTEMS AND NETWORKS 9

System design consideration Point – to –Point link design –Link power budget –rise time budget, WDM –Passive DWDM Components-Elements of optical networks-SONET/SDH Optical
Interfaces-SONET/SDH Rings and Networks-High speed light wave Links-OADM configuration-Optical ETHERNET-Soliton.

TOTAL: 45 PERIODS

OUTCOMES:

At the end of the course, the student should be able to:

☐ Realize basic elements in optical fibers, different modes and configurations.

☐ Analyze the transmission characteristics associated with dispersion and polarization tech.

☐ Design optical sources and detectors with their use in optical communication system.

☐ Construct fiber optic receiver systems, measurements and coupling techniques.

☐ Design optical communication systems and its networks.

TEXT BOOKS:

1. P Chakrabarti, "Optical Fiber Communication", McGraw Hill Education (India) Pvt. Ltd. 2016 (UNIT I, II, III)
2. Gred Keiser,"Optical Fiber Communication", McGraw Hill Education (India) Pvt. Ltd. Fifth Edition, Reprint 2013. (UNIT I, IV, V)

REFERENCES:

1. John M.Senior, "Optical fiber communication", Pearson Education, second edition.2007.
2. Rajiv Ramaswami, "Optical Networks " , Second Edition, Elsevier , 2004.
3. J.Gower, "Optical Communication System", Prentice Hall of India, 2001.
4. Govind P. Agrawal, "Fiber-optic communication systems", third edition, John Wiley & sons, 2004.

UNIT -1 INTRODUCTION

PART A

1. **why do we prefer step index single mode fiber for long distance communication?[APR/MAY 2019]**
   Step index single mode fiber has (1) low attenuation due to smaller core diameter, (2) Higher bandwidth, (3)Low dispersion.
2. **What is the necessity of cladding for an optical fiber? [APR/MAY 2019]**
   a) To provide proper light guidance inside the core b) To avoid leakage of light from the fiber c) To avoid mechanical strength for the fiber d) To protect the core from scratches and other mechanical damages
3. **Distinguish between meridional rays from skew Rays. [NOV/DEC 2018]**
   A **skew ray** is a **ray** that travels in a non planar zig zag path and never crosses the axis of an optical fibre! A **meridional ray** is a **ray** that passes through the axis of an optical fiber.
4. **Manufacturing engineer wants to to make an optical fibre that has core intex of 1.40 and cladding intex of 1.47 8 what should be the core size for single[NOV/DEC 2018]**

$$V = \frac{2\pi a}{\lambda} NA$$

5. **A silica optical fiber with a large core diameter has a core refractive index of 1.5 and a cladding refractive index of 1.47.Determine the acceptance angle in air for thefiber. [April 2017, APR 2018]**

Given data:

$n_1 = 1.5$

$n_2 = 1.47$

Formula:

$$\theta_a = \sin^{-1}\sqrt{n_1^2 - n_2^2}$$

Solution:

$$\theta_a = \sin^{-1}\sqrt{1.5^2 - 1.47^2}$$

$$\theta_a = 17.36^{\circ}$$

6. Write short notes on ray optics theory.

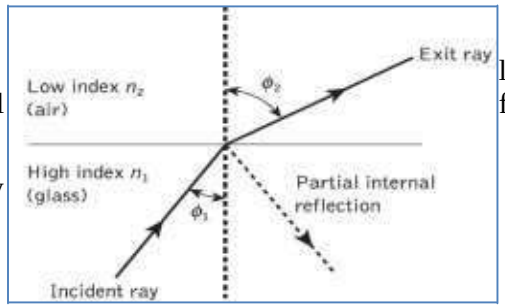
Laws governing the nature of light are called as ray optics. These laws are stated as:

1. Light rays in homogenous media travel in straight lines.
2. Laws of reflection: Angle of reflection θ_r equals angle of incidence θ_i
3. Snell's Law: The angle of refraction θ_t is related to angle of incidence θ_i by

$n1 \sin \theta_i = n2 \sin \theta_t$

7. What are the advantages and disadvantages of the ray optics? [April 2017]

The advantages of ray optics are:



a) Ray optics is used to develop some of the fundamental parameters like acceptance angle, numerical aperture that are associated with optical fiber transmission.

b) It provides an excellent approximation, when the wavelength is very small compared with the size of structures, with which the light interacts.

The disadvantages of the ray optics are:

a) Ray optics fails to account for optical effects such as diffraction and interference.

8. Define Phase and group velocity. (Nov-Dec 2015)

The group velocity of a wave is the velocity with which the overall shape of the waves amplitude known as modulation or envelope of wave propagates through space.

$$v_g = \frac{d\omega}{d\beta}$$

The Phase velocity of a wave is the rate at which the phase of the wave propagates in space. This is the velocity at which the phase of any one frequency component of wave travel.

$$v_p = \frac{\omega}{\beta}$$

9. **Define – Critical Angle [DEC 2016]**

The critical angle is defined as the minimum angle of incidence ($\varphi 1$) at which the ray strikes the interface of the two medium and causes an angle of refraction ($\varphi 2$) equal to 90^o.

10. Assume that there is a glass rod of refractive index 1.5, surrounded by air. Find the critical incidence angle.[MAY 2016]

Given data:

$n_1 = 1.5$

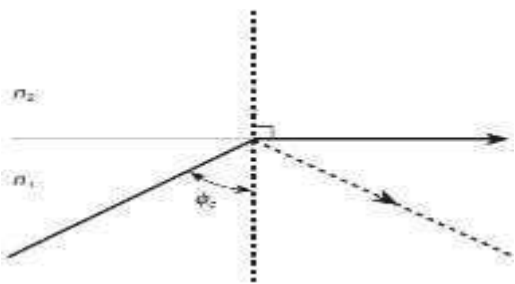
$n_2 = 1$

Formula: $\phi_c = \sin^{-1} \frac{n2}{n1}$

Solution:

$$\phi_c = \sin^{-1} \frac{1}{1.5} = 41.81°$$

11. State Snell's law.[DEC 2013, MAY 2016]



The Snell's law is an expression that describes the relationship between the angles of incidence θ_1 and refraction θ_2 and to the refractive indices of the dielectrics, when referring to waves passing through a boundary between two isotropic medium.

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

where n_1 is the refractive index of the core and n_2 is the refractive index of the cladding.

12. What are the conditions for total internal reflection? [NOV 2015]

The conditions for total internal reflection are:

a) The ray should travel from denser to rarer medium i.e. from core to clad region of the optical fiber.
b) The angle of incidence in the denser medium should be greater than the critical angle of that medium.

13. Calculate the critical angle of incidence between two substances with different refractive indices, where n1 =1.5 and n2 =1.46. (Apr-May 2015)

$n_1 = 1.5$

$n_2 = 1.46$

$\emptyset = \sin^{-1}\frac{n_2}{n_1}$

$= \sin^{-1}\frac{1.46}{1.5}$

$\emptyset = 76.74°$

14. List any two advantages of single mode fibers.(Nov-Dec 2014)

Single mode fiber has only one ray passes through fiber. Ray passes along the axis-axial ray. Core diameter is small (typically 10 - 12 µm). Intermodal dispersion is not present. Coupling efficiency is less.

15. **Define - Numerical Aperture [NOV2014]**

Numerical Aperture (NA) of the fiber is the light collecting efficiency of the fiber and is the measure of the amount of light rays that can be accepted by the fiber. It is equal to the sine of acceptance angle θ_a

$$\mathbf{NA = \sin\theta_a = (n_1^2 - n_2^2)^{1/2}}$$

where n_1 and n_2 are the refractive indices of core and cladding respectively.

16. For n1 = 1.55 and n2 =1.52 , Calculate the critical angle and numerical aperture.(May-June 2013)

Critical angle $\emptyset_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) = \sin^{-1}\left(\frac{1.52}{1.55}\right) = 78.7°$ Numerical aperture NA = $\sqrt{n_1^2 - n_2^2}$ = 0.3

17. Define – Relative Refractive Index Difference

The relative refractive index difference is the ratio of the refractive index difference between core and cladding and refractive index of core.

Where, $\Delta = \frac{n_1 - n_2}{2n_1}$

Δ is the relative refractive index
n_1 is the numerical aperture of the core
n_2 is the numerical aperture of the cladding

18. What is the energy of the single photon of the light whose λ = 1550nm in eV? (N/D2011)

The energy of the single photon of the light is given by the equation

$E = h \times f$

Sub $f = \frac{c}{\lambda}$ in the above equation

$E = h \times \frac{c}{\lambda}$

Given data:

$h = 6.625 \times 10^{-34}$

$c = 3 \times 10^{8}$ m/sec

$\lambda = 1550 \times 10^{-9}$ m

$$E = \frac{6.625 \times 10^{-34} \times 3 \times 10^{8}}{1550 \times 10^{-9}}$$

$$= 0.0128 \times 10^{-17} \text{ J}$$

19. step index fiber has the normalized frequency of 26.6 at 1300nm. If the core radius is 25μm, find the numerical aperture.

Given data: V= 26.6 , $\lambda = 1300 \times 10^{-9}$m a = 25×10^{-6}m

Formula:

Normalized frequency V is given by $V = 2\pi a (NA) / \lambda$

$NA = \lambda V / 2\pi a$

Solution:

Numerical Aperture = $\lambda V / 2\pi a$

$$NA = \frac{1300 \times 10^{-9} \times 26.6}{2 \times 3.14 \times 25 \times 10^{-6}}$$

NA = **0.22**

20. What is meant by refractive index of the material?

The refractive index (or index of refraction) 'n' is defined as the ratio of the velocity of light in vacuum to the velocity of light in the medium.

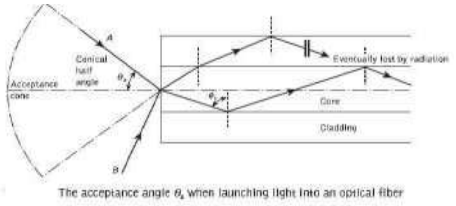
$$n = \frac{c}{v}$$

c = speed of light in free space

v = speed of light in a given material

21. Define - Acceptance angle

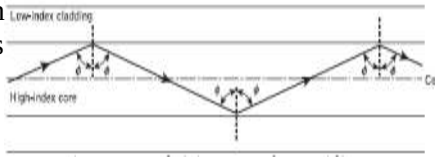
The maximum angle 'θ_a' with which a ray of light can enter through the fiber and still be totally internally reflected is called acceptance angle of the fiber.



The acceptance angle θ_a when launching light into an optical fiber

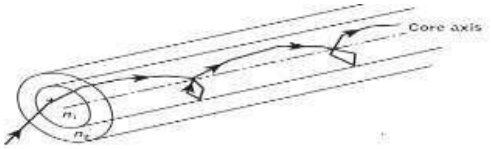
22. What are meridional rays?

Meridional rays are the rays following zig- zag path when they travel through fiber and for every reflection it will cross the fiber axis. The figure below shows the meridional rays.



23. What are skew rays?

Skew rays are the rays following the helical path around the fiber axis when they travel through the fiber and they would not cross the fiber axis at any time. The figure below shows the propagation of skew rays.



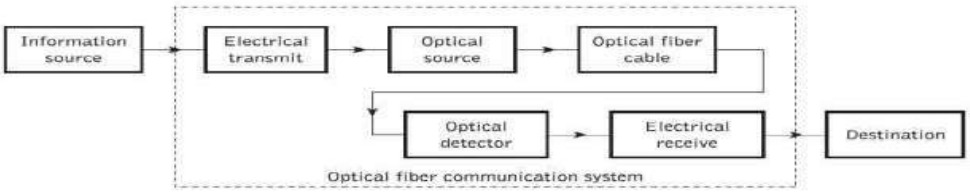
24. Write the acceptance angle condition for the skew rays.

The acceptance conditions for skew rays is given by the equation

$$\theta_{as} = \sin^{-1} \frac{NA}{\cos\gamma}$$

where NA is the numerical aperture and γ is the angle between the projection of the ray in two dimensions and the radius of the fiber core at the point of reflection.

25. Draw the block diagram of an optical communication system.



The block diagram of an optical communication system is represented above:

26. The relative refractive index difference (Δ) for an optical fiber is 1%. Determine the critical angle at the core cladding interface if the core refractive index is 1.46.

$$\Delta = \frac{n1 - n2}{n1} \quad (1)$$

$$\phi_c = \sin^{-1} \frac{n2}{n1} \quad (2)$$

Find n_2 from equation (1)

$$n_2 = n1(1 - \Delta) \quad (3)$$

Solution:

$$n_2 = 1.46(1 - 0.01)$$
$$= 1.4454$$
$$\phi_c = \sin^{-1} \frac{1.4454}{1.46}$$
$$\phi_c = 81.19^{\circ}$$

27. Which photodiode is used for a low power optical signal and Why?

Avalanche Photo Diode (APD) is used for a low power optical signal because it has a greater sensitivity due to an inherent internal gain mechanism produced by avalanche effect.

28. What is V number of a fiber?

Normalized frequency or V number is a dimensionless parameter and represents the relationship among three design variables of the fiber i.e. core radius **a**, relative refractive index Δ and the operating wavelength **λ**. It is expressed as V= 2πa (NA)/λ.

29. What are guided modes?

Guided modes are a pattern of electric and magnetic field distributions that is repeated along the fiber at equal intervals.

30. Define – Phase Velocity

As a monochromatic light wave propagates along a waveguide in the *z* direction the points of constant phase travel at a phase velocity V_p given by

$$V_p = \frac{\omega}{\beta}$$

where ω is the angular frequency and β is the propagation constant

31. Define – Group Velocity

Group of waves with closely similar frequencies propagate so that their resultant forms packet of waves. This wave packet does not travel at the phase velocity of individual but it moves with the group velocity Vg given by

$V_p = \frac{\omega}{\beta}$ where ω is the angular frequency and β is the propagation constant

32. What is meant by mode coupling? What causes it?

The effect of coupling energy from one mode to another mode is known as mode coupling. The cause of mode coupling is due to waveguide perturbations such as deviations of the fiber axis from straightness, variations in the core diameter, irregularities at the core -cladding interface and refractive index variations.

33. What are the uses of optical fibers?

The uses of optical fiber are

- To transmit analog and digital information.
- To transmit the optical images.(Endoscopy Images)
- To act as a light source at the inaccessible places.
- To act as sensors for mechanical, electrical and magnetic measurements.

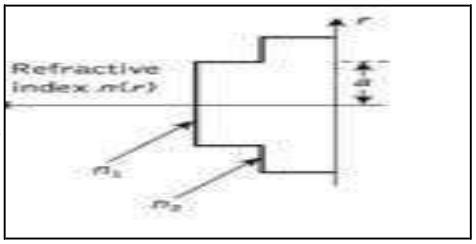
34. What is the necessity of cladding for an optical fiber?

The necessity of cladding for an optical fiber is:

- To provide proper light guidance inside the core.
- To avoid leakage of light from the fiber.
- To provide mechanical strength for the fiber.
- To protect the core from scratches and other mechanical damages

35. What is step index fiber?

Step index fiber is a cylindrical waveguide that has the central core with uniform refractive index of n1, surrounded by outer cladding with refractive index of n2.The refractive index of the core is constant and is larger than the refractive index of the cladding. It makes a step change at the core cladding interface as indicated in the figure,



36. Write the refractive index expression for step index fiber.

In step index fiber, the refractive index of a core is constant and is larger than the refractive index of the cladding. The refractive index profile is defined as

n(r) = n1; r < a (core)

n2; r ≥ a (cladding)

37. What are the advantages of Graded Index Fiber?

The advantages of Graded Index Fiber are

- It exhibits **less intermodal dispersion** because the different group velocities of the modes tend to be normalized by the index grading.
- It provides **higher bandwidth**

38. Write the refractive index expression for graded index fiber.

Graded index fibers does not have a constant refractive index in the core but a gradually decreasing core index $n(r)$ with radial distance from a maximum value of $n1$ at the axis to a constant value $n2$ beyond the core radius 'a' in the cladding. This index variation may be represented as:

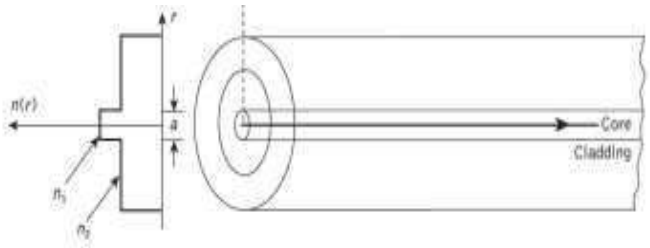
$n(r) = \{ n_1 = (1 - 2\Delta(r/a)^{\alpha})^{1/2}$; **r < a (core)}**

$\{n_1 (1 - 2\Delta)^{1/2} = n_2$; **r ≥ a (cladding)}**

Where, n_1 is the refractive index of the core and n_2 is the refractive index of the cladding Δ is the index difference, α is the index profile

39. Write a short note on single mode fiber.

For single-mode operation, only one mode (the fundamental LP01) can exist and it does not suffer from mode delay. The core diameter is small so that there is only one path for light ray to propagate inside the core. Typical core sizes are 2μm to 5μm. It provides larger bandwidth and less coupling efficiency. It is used for long haul transmission.



40. List out the advantages of multimode fiber over single mode fibers. (A/M2008) The advantages of multimode fiber are:[DEC 2016]

- The larger core radii of multimode fibers make it easier to launch optical power into the fiber. Connecting together of similar fibers is easy.
- Light can be launched into a multimode fiber using an LED source, whereas single-mode fibers with LASER diodes. LED's are easier to make, less expensive, less complex circuitry and have longer life times.

41. List the advantages and disadvantages of monomode fiber.

The advantages of single mode fiber are:

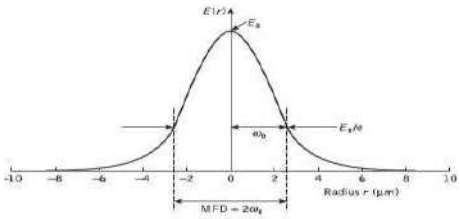
- No intermodal dispersion
- Information capacity of single mode fiber is large

The disadvantages of single mode fiber are:

- Launching of light into single mode and joining of two fibers are very difficult
- Fabrication is very difficult and so that fiber is so costly

42. Define – Mode Field Diameter

Mode-Field Diameter is an important parameter for characterizing single mode fiber properties that accounts the wavelength dependent field penetration into the fiber cladding. This can be determined from the mode field distribution of the fundamental LP01 mode. The MFD equals $2\omega_0$ where ω_0 is the nominal half width of the input excitation.



43. Why is step index single mode fiber preferred for long distance communication?

The step index single mode fiber is preferred for long distance communication because,

- They exhibit higher transmission bandwidth because of low fiber losses.
- They have superior transmission quality because of the absence of modal noise.
- The installation of single mode fiber is easy and will not require any fiber replacement over twenty plus years.

44. Define – Birefringence

Manufactured optical fibers have imperfections, such as asymmetrical lateral stresses, non circular cores, and variations in refractive index profiles. These imperfections break the circular symmetry of the ideal fiber and lift the degeneracy of the two modes. These modes propagate with different phase velocity and it is called as fiber birefringence. Birefringence is expressed as

$B_f = \beta_x - \beta_y / \frac{2\pi}{\lambda}$ where β is the propagation constant.

45. **State the reasons to opt for optical fiber communication.**

- **Broad bandwidth**: A single optical fiber can carry over 3,000,000 full-duplex voice calls or 90,000 TV channels.
- **Immunity to electromagnetic interference:** Light transmission through optical fibers is unaffected by otherelectromagnetic radiation nearby. The optical fiber is electrically non-conductive, so it does not act as an antenna to pick up electromagnetic signals. Information traveling inside the optical fiber is immune to electromagnetic interference, evenelectromagnetic pulses generated by nuclear devices.
- **Low attenuation loss over long distances:** Attenuation loss can be as low as 0.2 dB/km in optical fiber cables, allowing transmission over long distances without the need for repeaters.
- **Electrical insulator:** Optical fibers do not conduct electricity, preventing problems with ground loops and conduction oflightning. Optical fibers can be strung on poles alongside high voltage power cables.
- **Material cost and theft prevention:** Conventional cable systems use large amounts of copper. Global copper prices experienced a boom in the 2000s, and copper has been a target of metal theft.
- **Security of information passed down the cable:** Copper can be tapped with very little chance of detection.

45. Differentiate between Mono Mode Fiber and Multimode Fiber.

S. No	Mono Mode Fiber	Multi Mode Fiber
1	Only one ray passes	More than one ray passes
2	Ray passes along the axis-axial ray	MMSI – Meridional and Skew MMGI – Paraxial
3	Core diameter is small typically10 - 12µm	Core diameter is large typically 50 - 200µm
4	Intermodal dispersion is not present	Intermodal dispersion is present
5	Fabricating single mode fiber is difficult	Fabricating multimode fiber is easy
6	Coupling efficiency is less	Coupling efficiency is large
7	LED is not suitable source for single mode	LED is suitable for multimode

46. Point out the limitations of Optical Fiber Communication system?

- Optical fiber is made up of glass because of the impurities present within the fiber result in absorption leads to loss of light in the Optical fiber.
- Maximum limitation of the bandwidth of the signals can be carried by the fiber due to spreading of pulse.
- It is costly.
- Optical fiber has limited band radius (□ 10mm)
-

47. Distinguish between Step Index fiber and Graded Index fiber.

S. No	Step Index Fiber	Graded Index Fiber

1	The core has uniform refractive index but step change in core-cladding interface.	The core has high refractive index along the axis which gradually decreases towards the clad-core interface (radially decreases)
2	Axial ray – SMSI, Meridonial rays & Skew - MMSI	Paraxial rays – MMGI
3	Intermodal dispersion is present in MMSI	Intermodal dispersion is reduced in MMGI
4	Numerical Aperture is constant	Numerical Aperture is a function of radius
5	Step index profile	Graded index profile □ □profile Factor
6	No of modes, $m \square v^2 / 2$. Step index supports twice the number of modes than GI	No of modes, $m \square v^2 / 4$
7	Fabrication is easy	Fabrication is difficult

48. Compare Ray Optics with Wave Optics.

S. No	Ray Optics	Wave Optics
1	It is used to represent the direction of light propagation	It is used to analyze mode theory
2	It is used to study reflection and refraction of light	It is used to analyze diffraction and Interference of light waves

49. Define Mode.

Mode is the pattern of distribution of electric and magnetic fields

- Transfers Electric Mode TE_{02}
- Transfers Magnetic Mode TM_{02}

50. List out the ways to minimize leaky modes.

A mode remains guided as long as β satisfies the condition $n_2K < \beta n_1K$.

$n_1, n_2 \rightarrow$ Refractive index of core are cladding $K = 2\pi/\lambda$

$\beta \geq n_2K$ = To prevent power leaks out of the core.

51. What are the three windows of Optical Communication?

The three wave lengths 850nm, 1300nm and 1500nm are three optical windows of optical communication system. Since only at this wavelength silica fiber loss is minimum.

PART B & C

1. **a. Explain a neat block diagram of fundamentals of optical fibre communication.[8]**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. NO.8**[DEC 2016, APR 2018]**
   b. Discuss the mode theory of circular waveguides. [Nov 2008]
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 43

2. **For multi-mode step-index fibre with glass core (n1 =1.5) and a fused quartz cladding (n2 =1.46), determine the acceptance angle $\theta_{in})$ and numerical aperture. The source to fibre medium is air. (Apr-May 2015, NOV/DEC 2018)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 37
3. **Explain the ray propagation into and down an optical fibre cable. Also derive the expression for acceptance angle. (Apr-May 2015 , 2019)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 37
4. **Contrast the advantages and disadvantages of step-index, graded- index, single-mode propagation and multi-mode propagation. (Apr-May 2015)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 37
5. **Classify fibers and explain them. (Nov-Dec 2015)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 37
6. **Describe and derive the modes in planar guide. (Nov-Dec 2015)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 43
7. **Define the normalized frequency for an optical fiber and explain its use.(Nov-Dec 2014)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 25
8. **Explain the features of multimode and single mode step index fiber and compose them. (Nov-Dec 2014)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 37
9. **A Single mode step index fiber has a core diameter of 7μm and a core refractive index of 1.49. Estimate the shortest wavelength of light which allows single mode operation when the relative refractive index difference for fiber is 1%. (Nov-Dec 2014)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 37
10. **a. Discuss briefly about linearly polarized modes. [6][APR/MAY 2019]**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 56
    b. Draw the structure of single and multi mode step index fibres and graded index fibres with typical dimensions. [6] [Apr 2018]
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 37
    c. Mention the advantage of optical fibre system [4]. [Nov 2008]
11. **Explain mode propagation in circular waveguides. Obtain its wave equation and modal equations for step index fibers. [Nov-2009]**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 43
12. **A typical refractive index difference for an optical fiber designated for long distance transmission is 1% .Determine the NA and the solid acceptance angle in air for the fiber when the core index is 1.46.calculate the critical angle at the core-cladding interface with in the fiber [Nov2009]**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 23
13. **Draw and explain the working principle of single mode and multimode fiber(Nov 2010)**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 37
14. **Distinguish step index and graded index fiber [DEC2016] (Nov 2011)**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 37
15. **Derive the mode equations for a circular fibre using maxwell's equations (May 2012)**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 43

UNIT- II
TRANSMISSION CHARACTERISTICS OF OPTICAL FIBER

1. **Distinguish between intra model and inter modal dispersion[NOV/DEC 2018]**
   Intramodal dispersion:
   Pulse broadening within a single mode is called as intramodal dispersion or chromatic dispersion
   Intermodal dispersion:
   • Dispersion caused by multipath propagation of light energy is referred to as intermodal dispersion.
2. **Give the measure of information capacity in optical wave guide.[Apr /May 2019]**
   It is usually specified by bandwidth distance product in MHz.For a step index fiber the various distortion effects tend to limit the bandwidth distance product to 20MHz.
3. **What is elastic and inelastic scattering? Give examples. [Apr 2018]**

   Purely elastic scattering means all the pre-collision kinetic energy of the colliding objects goes into kinetic energy of the post-collision objects. A collision between two hard things, like billiard balls, is a good example of a collision that's *mostly* elastic.
   Inelastic scattering means that at least *some* of the pre-collision kinetic energy ends up somewhere else, besides post-collision kinetic energy. For example, the pre-collision kinetic energy can be used to cause an internal state change in one of the colliding objects.
4. **What is meant by attenuation coefficient of a fiber? (N/D2011)[DEC2016]**
   Attenuation coefficient is defined as the ratio of the input optical power P_i launched into the to the output optical power P_o from the fiber.

$$\propto_{dB} = \frac{10}{L} \log_{10} \frac{P_i}{P_o}$$

   where $\propto_{dB}$ is the attenuation coefficient in decibels per kilometer.

5. **A 30 km long optical fiber has an attenuation of 0.8 dB/km. If 7 dBm of optical power is launched into the fiber, determine the output optical power in dBm. (M/J 2012)**

Given Data: $P_i = 7$ dBm
$L = 30$ km $= 3 \times 10^4$ m;

$\alpha_{dB} = 0.8$dB/km $= 0.8 \times 10^{-3}$ dB/m

Solution:

$$\alpha_{dB} = \frac{10}{L} \log_{10} \frac{Pi}{Po}$$

$$= \frac{10}{3 \times 10^4 \log \frac{7}{P_o}}$$

$\log (7/P_0) = 2.4$;

$$P_0 = \left(\frac{7}{e^{2.4}}\right) = 0.63\text{dBm}$$

6. **What are the types of material absorption losses in silica glass fibers? [DEC 2016]**
   The types of material absorption losses in silica fiber are:
   - Absorption by atomic defects in the glass composition
   - Extrinsic absorption by impurity atoms in the glass material

     Intrinsic absorption by the basic constituent atoms in the glass material

7. **Compare Rayleigh scattering and Mie scattering.**

S.No	Rayleigh Scattering	Mie Scattering
1	Caused due to refractive index variation in the core glass.	Caused by fiber imperfections such as irregularities in the core –cladding interface, core- cladding refractive index difference along the fiber length, diameter fluctuations, strains and bubbles.
2.	When the inhomegenetics size is smaller than the wavelength of light Rayleigh scattering occurs.	When the in homegenetics size is greater than the Wavelength of light, Mie scattering occurs
3	Scattering occurs both is forward and backward direction.	Scattering is mainly in the forward direction.
4.	Rayleigh scattering can be reduced by minimizing the compositional fluctuations by using best manufacturer methods.	Mie scattering can be reduced by Removing imperfections due to the glass manufacturing process. Carefully controlled extrusion and coating of the fiber Increasing the fiber guidance by increasing the relative refractive index difference.

8. Compare Linear scattering and Non- Linear scattering.

S.No	Linear Scattering	Non-Linear Scattering
1.	Linear scatterings are observed only at low optical power densities below the threshold power levels.	Non-Linear scattering are only observed at high optical power densities above the threshold power levels in long single mode fibers.
2.	There are two types are Linear Scattering namely, Rayleigh Scattering Mie Scattering	There are two types of Non-Linear scattering namely, Stimulated Brillouin Scattering (SBS) Stimulated Raman Scattering (SRS)
3.	The Incident light frequency and scattered light frequency is the same. There is no frequency shift during scattering.	The Incident light frequency and scattered light frequency are different. There is a frequency shift during scattering.

9. Compare SRS and SBS.

S.No	SRS	SBS
1.	SRS can occur both in forward and backward direction.	It is mainly backward process.
2.	The threshold power level of SRS is three times higher than SBS threshold in a particular fiber.	The SBS threshold power level is less.
3.	The scattering process produces high frequency optical phonon.	The scattering process produces acoustic phonon as well as a scattered photon.

10. What is meant by intrinsic absorption in optical fibers?

The absorption caused by the interaction of one or more of the major components of the glass is known as intrinsic absorption.

11. What is meant by extrinsic absorption in optical fibers?

The absorption caused by the impurities within the glass is known as extrinsic absorption.

12. Differentiate linear scattering from nonlinear scattering.

- Linear scattering mechanisms transfers linearly some or all of the optical power contained within one propagating mode to a different mode.
- Non-linear scattering causes the optical power from one mode to be transferred in either the forward or backward direction to the same or other modes at different frequencies.

13. What are the types of linear scattering losses? [MAY 2016]

Linear scattering is of two types. They are:

- Rayleigh scattering
- Mie scattering

14.What are the types of nonlinear scattering losses?

Non-linear scattering is of two types. They are

- Stimulated Brillouin Scattering (SBS)
- Stimulated Raman Scattering (SRS)

15. What is meant by Fresnel Reflection? (N/D 2011)

When the two joined fiber ends are smooth and perpendicular to the axes, and the two fiber axes are perfectly aligned, small proportion of the light may be reflected back into the transmitting fiber causing attenuation at joint. This is known as Fresnel reflection.

16. What is meant by linear scattering?

Linear scattering mechanisms transfers linearly some or all of the optical power contained within one propagating mode to a different mode.

17. What are the factors that cause Rayleigh scattering in optical fibers? (M/J 2012)

The inhomogeneties of a random nature occurring on a small scale compared with the wavelength of the light in optical fiber causes Rayleigh scattering. These inhomogeneities manifest themselves as refractive index fluctuations and arise from density and compositional variations that are frozen into the glass lattice on cooling.

18. What are the factors that cause Mie scattering in optical fibers?

The factors that cause Mie scattering in optical fibers are:

- Fiber imperfections such as irregularities in the core – cladding interface
- Core – cladding refractive index differences along the fiber length, diameter fluctuations

19. What are the ways to reduce macro bending losses? (N/D 2009) (N/D 2010)

The ways to reduce macro bending losses are

- Designing fibers with large relative refractive index differences
- Operating at the shortest wavelength possible.

20. What is meant by dispersion in optical fiber? (A/M 2008)

Different spectral components of the optical pulse travel at slightly different group velocities and cause pulse broadening within the fiber. This phenomenon is referred as dispersion.

21. What are the different types of dispersion? (N/D 2008)

There are two types of dispersion. They are

- Intramodal Dispersion:
  - Material Dispersion
  - Waveguide Dispersion
- Intermodal Dispersion:
  - Multimode step index
  - Multimode graded index

22. **What is meant by intermodal dispersion? (A/M 2010 A/M 2008)**
    Pulse broadening due to propagation delay differences between modes within a multimode fiber is known as intermodal dispersion.

23. **Define – Group Velocity Dispersion (GVD) (A/M 2011), (N/D 2010)**
    Intra-modal dispersion is pulse spreading that occurs within a single mode. The spreading arises from the finite spectral emission width of an optical source. This phenomenon is known as Group Velocity Dispersion (GVD).

24. **What is meant by modal noise? (A/M 2011)**
    The speckle patterns are observed in multimode fiber as fluctuations which have characteristic times longer than the resolution time of the detector. This is known as modal or speckle noise.

25. **What is meant by chromatic dispersion? (N/D2011)**
    The dispersion due to the variation of the refractive index of the core material as a function of wavelength is known chromatic dispersion. This causes a wavelength dependence of the group velocity of any given mode. Pulse spreading occurs even when different wavelengths follow the same path.

26. **What is meant by polarization mode dispersion? (N/D 2007) [Apr 2018]**
    Polarization refers to the electric - field orientation of a light signal, which can vary significantly along the length of the fiber.

27. **Distinguish between dispersion shifted and dispersion flattened fibers. (N/D 2007)**
    Reduction in the fiber core diameter with an increase in the relative or fractional index difference to create dispersion is known a dispersion shifted fiber. Fibers which relax the spectral requirements for optical sources and allow flexible wavelength division multiplexing are known as dispersion flattened fibers.

28. **What are the two types of fiber joints? The two types of fiber joints are:**
    (i) Fiber splices: These are semi permanent or permanent joints.
    (ii) Demountable fiber connectors or simple connectors: These are removable joints.

29. **What is meant by fiber splicing?**
    A permanent joint formed between two individual optical fibers in the field or factory is known as fiber splice.

30. **What are the techniques used in splicing?**
    Generally used splicing techniques are:
    - Fusion splice
    - V-groove mecha/ nical splice
    - Elastic tube splice

31. **List the types of mechanical misalignments that occur between two joined fibers.**
    There are three types of mechanical misalignments:
    - Lateral/radial/axial misalignment
    - Longitudinal misalignment
    - Angular misalignment

32. What are the causes of absorption?

- Absorption by atomic defects in glass composition.
- Extrinsic absorption by impurity atoms in the glass materials.
- Intrinsic absorption by basic constituent atoms.

33. What is polarization mode dispersion?

The difference in propagation times between the two orthogonal polarization modes will result pulse spreading. This is called as polarization mode dispersion. (PMD)

34. Define signal attenuation.

If P(0) is the optical power in a fiber at the origin (at Z = 0), then the power P(Z) at a distance z further down the fiber is

$P(z) = P(0)\, e^{-\alpha_{pz}}$.

The above equation can be rewritten as

$\alpha_p = (1 / z) \{ P(0) / P(z) \}$. Where α_p is the fiber attenuation coefficient given in units of km^{-1}

35. What are bending losses? Name any two types,

Micro bend losses

Macro bend losses

36. What is meant by Polarization of light?

The polarization of light describes by a specifying the orientation of the waves electric field at a point in space over one period of the oscillation. When light travels in free space, it propagates as a transverse wave, i.e. the polarization is perpendicular to the wave‟s direction of travel.

37. What is fiber Bi - refrigence?

Fiber bi-refrigence is the optical property of a material having a refractive index that depends on the polarization and propagation direction of light. These optically anisotropic materials are said to be bi-refringence. The bi-refringence is often quantified by the maximum difference in the refractive index within the material.

38. Define Beat length?

Beat length is defined as the period of interference effects in a bi- refringence medium. When two waves with different linear polarization states propagate in a bi-refringence medium, their phases will evolve differently. It is assumed that the polarization of each wave is along the principle directions of the medium (x – axis (or) y – axis), so that this polarization will be preserved during propagation. This means that the phase relation between both waves is restored after integer multiples called the polarization beat length.

39. Define PMF (Polarization Mode Fiber)?

PMF is an optical fiber in which the polarization of linearly polarized light waves launched into the fiber is maintained during propagation, with less or no cross-coupling of optical power between the polarization modes. Such fiber is used in special application where processing the polarization is essential.

40. What are Dispersion Flattened Fibers (DFF)?

DFF is a type of glass optical fiber that provides low pulse Dispersion over a broad portion of the light spectrum and as a result can operate at 1300 nm and 1550 nm wavelength simultaneously.

41. What are Dispersion Shifted Fibers (DSF)?

DSF is a type of optical fiber made to optimize both low dispersion and low attenuation. DSF is a type of single mode optical fiber with a core-clad index profile tailored to shift the zero-dispersion wavelength from the natural 1300 nm in silica-glass fibers to the minimum loss at 1550 nm.

42. What is meant by Fresnel reflection in Fiber cable?

Fresnel reflection at the air-glass interfaces at the entrance and exit of an optical fiber.

43. List out the advantages of elastic tube splicing?

The advantages of elastic tube splicing are,

a) This type of splicing allows accurate and automatic alignment of axes of the two fibers to be joined.
b) In this method the fibers to be splices do not have to be equal in diameter.

44. List out the advantages of V-groove splicing?

a) There is no thermal stress.

b) No change in refraction index of the two fibers. c)

45. What are bending losses? Name any two types. (Apr-May 2015)

(i) Micro bending losses - The light power is dissipated through the micro bends because if the respective coupling of energy between guided modes and leaky modes.

(ii) Macro bending losses - Macrobending losses occur when fibres are physically bent beyond the point at which the critical angle is exceeded.

46. What are the types of fiber losses which are given per unit distance?(Nov-Dec 2014)

(i)Absorption

(ii) Scattering

(iii)Bending Loss

47. List the factors that cause intrinsic joint losses in a fiber. (N0v-Dec 2014)

(i) Different core and / or cladding diameters

(ii) Different numerical apertures and / or relative refractive index differences.

(iii) Different refractive index profiles.

(iv) Fiber faults.

48. Define dispersion in multimode fibers. What is its effect? (Nov-Dec 2013) (R)

In multimode fiber many modes are propagating along the fiber at a time. Different modes are taking different ray path and they reach at different time at the output end of the fiber. So a time delay is experienced between modes. This is called intermodal delay and pulse broadening occurs due to intermodal delay is called intermodal dispersion

Effect:

1. It restricts bandwidth of the optical fiber cable.
2. The intermodal dispersion causes the light rays to spread out through the fiber.
3. It accounts for a significant loss occurring in the fiber.

49. What are the two reasons for Chromatic Dispersion? (Nov-Dec 2012)

- Dispersive Properties of the waveguide material – **Material Dispersion**
- Guidance effects within the fiber structure– **Waveguide Dispersion**

50. What are the most important non-linear effects of optical fibre communication? (Nov-Dec 2012)

Non linear effects of Scattering are:

- Stimulated Brillouin Scattering (SBS)
- Stimulated Raman Scattering (SRS)

51. A fiber has an attenuation of 0.5dB/Km at 1500nm. If 0.5mW of optical power is initially launched into the fiber estimate the power level after 25km.

$$P_{out}\ (dBm) = P_{in}(dBm) - \alpha \left(\frac{dB}{Km}\right) \times l$$

$$P_{in}(dBm) = 10^{\log_{10} \frac{P_{in}(dBm)}{1mW}}$$

$$P_{out}\ (dBm) = - = 10^{-(0.5 \times 25)}$$

$$P_{out}\ (dBm) = -$$

PART B & C

1. **Discuss about the design optimization of single mode fiber.(Nov-Dec 2016)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 115·
2. **What is waveguide dispersion? Derive and expression for time delay produced due to waveguide dispersion.(Nov-Dec 2016,APR/MAY 2019)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 109·
3. **With necessary diagrams, explain the causes and types of fiber attenuation loss. (Nov-Dec 2015)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 94·
4. **With diagram, derive the expression for intra modal dispersion. (Nov-Dec 2015)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 109·
5. **What are the loss or signal attenuation mechanism in a fibre? Explain.(Apr-May 2015, NOV/DEC 2018)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 94·
6. **Discuss the pulse broadening in graded index fiber (April 2005,Nov 08)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 105
7. **Derive the expression for pulse broadening due to material dispersion.**
   Refer Book: Optical fiber Communications - Gerd Kaiser -Pg. No. 108
8. **What is meant by waveguide dispersion ? Derive the expression for the same (Nov 2006)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 109
9. **Explain the attenuation mechanisms in Optical fibers. (Dec 2007)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 94
10. **Explain bending losses and type of dispersion (Nov 2010)**
    Refer Book: Optical fiber Communications - Gerd Kaiser -Pg. No. 108
11. **Describe the linear and non-linear scattering losses in optical fibers (Nov 2012)**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 94
12. **Discuss the attenuation encountered in optical fiber communication due to:**
    1. Bending 2. Scattering 3. Absorption.
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 94
13. **Explain the attenuation and losses in fiber. (May 2014) [DEC2016]**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 94
14. **With diagram, explain intra and inter modal dispersion. (May 2014, AAPR/MAY2019)**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 115
15. **Explain signal distortion in single mode fibers.**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 115·

UNIT-III

SOURCES AND DETECTORS

1. **What is meant by heterojunction? List out the advantages of heterojunction.(A/M 2011) (N/D 2007)**
   A heterojunction is an interface between two adjoining single crystal semiconductors with different bandgap energies. Devices that are fabricated with heterojunction are said to have hetrostructure.
   Advantages of heterojunction are:
   - Carrier and optical confinement
   - High output power
   - High coherence and stability

2. **Distinguish between direct and indirect band gap materials. (N/D2010), (N/D 2008)**

Direct bandgap materials	**Indirect bandgap material**
The electron and hole have the same momentum value	The conduction band minimum and the valence band maximum energy level occur at different values of momentum.
Direct transition is possible from valence band to conduction	Direct transition is not possible from valence band to conduction

3. **Why is silicon not used to fabricate LED or Laser diode? (N/D2011, 2018)[DEC2016]**
   Silicon is not used to fabricate LED or Laser diode because
   - It is an indirect bandgap semiconductor
   - It has Eg level of 1.1eV, the radiated emission corresponds to infrared but not the visible light.
4. **What are the advantages of LED?(M/J2012)**
   The advantages of LED are:
   - Less expensive
   - Less complex
   - Long life time
   - Used for short distance communication
5. **When an LED has 2V applied to its terminals, it draws 100mA and produces 2mW of optical power. Determine conversion efficiency of the LED from electrical to optical power. (N/D2008)**

   Given Data: $V_{in} = 2$ V, $I_{in} = 100\times10^{-3}$ A, $P_{out} = 2\times10^{-3}$

   Formula: LED conversion efficiency = $\frac{P_{out}}{P_{in}}$

   Solution:

   $P_{in} = V_{in} \times I_{in} = 2 \times 100\times10^{-3}$

   Conversion Efficiency = $\frac{2\times10^{-3}}{2\times100\times10^{-3}} = \mathbf{0.01}$
6. **What are the advantages and disadvantages of LED?**
   Advantages:
   - Small size and light weight;
   - High Speed;
   - Low operating temperature;
   - Longer life;
   - No complex driver capacity required.

   Disadvantages:
   - Quantum efficiency is low;
   - Damages due to over voltage and over current;

- Temperature dependent.

7. What are the three requirements of Laser action? [DEC2016]

The three requirements of Laser action are

- Absorption
- Spontaneous emission
- Stimulated emission

8. What is the principle of operation of LASER? (N/D2008)

The principle of operation of LASER is population inversion, the most photons incident on the system. The population of the upper energy level is greater than lower energy level i.e. $N_2 > N_1$. This condition is known as population inversion.

9. Write the three modes of the cavity of LASER diode. (N/D2009)

The three modes of the cavity of LASER are:

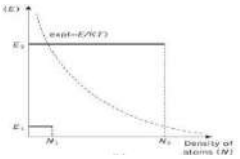
- Longitudinal modes, related to the length L of the cavity
- Lateral Modes lie in the plane of the P-N junction. These modes depend upon the side wall preparation and width of the cavity.
- Transverse modes are associated with the Electro Magnetic Field and beam profile in the direction perpendicular to the plane of the PN junction. These modes determine the radiation pattern of the LASER.

10. What is a DFB Laser? Differentiate DFB LASER from other types of LASER.(N/D2009)

In DFB Laser, the lasing action is obtained by periodic variations of refractive index, which are incorporated into multilayer structure along the length of the diode. DFB LASER does not require optical feedback unlike the other LASERs.

11. What is population inversion? (A/M 2008)

Under thermal equilibrium, the lower energy level E_1 of the two level atomic system contains more atoms than upper energy level E_2.To achieve optical amplification it is necessary to create non-equilibrium distributions of atoms such that population of the upper energy level is greater than lower energy level i.e. $N_2 > N_1$ as shown in the figure. This condition is known as population inversion.



12.Compare LED and ILD sources.[Apr 2017] (A/M 2008)

Sl.N	LED	ILD
1.	Incoherent	Coherent
2.	For multimode fibers only	For multi and single mode fibers
3.	Large beam divergence due to s	Low beam divergence due to s

13. Write the three key processes of laser action.(A/M 2008)

The three key processes of laser actions are:

- The atomic system must have population inversion. This means the number of atoms in the excited state should be more than that of ground state
- There should be photons with proper energy to start the stimulated emission
- There should be an arrangement for multiple reflections to increase the intensity of LASER beam

14. What are the advantages of Quantum Well Lasers?(N/D2009)

The advantages of Quantum Well Lasers are:

- High threshold current density
- High modulation speed
- High line width of the device

15. Define – Internal Quantum Efficiency[NOV /DEC 2018]

Internal Quantum Efficiency is defined as the ratio of radiative recombination rate to the total recombination rate.

$$\eta_{in} = \frac{R_r}{R_r + R_{nr}}$$

where R_r is radiative recombination rate, R_{nr} is the non-radiative recombination rate.

16. Define – External Quantum Efficiency

The external quantum efficiency is defined as the ratio of photons emitted from LED to the number of photons generated internally.

17. **Define – Quantum efficiency of a photo detector(A/M2008,10) (M/J2009)(N/D2011)**

Quantum efficiency is defined as the number of the electron-hole carrier pairs generated per incident photon of energy hv, is given bynumber of electron−hole pairs generated

number of incident photons

$$\eta = \frac{I_P/q}{P_o/hv}$$

where I_p is the photon current
q is the charge of the electron
p_o is the optical output power
h is the Planck's constant
v is the frequency of the optical signal

18. **An LED has radiative and nonradiative recombination times of 30 and 100 ns respectively. Determine the internal quantum efficiency. (N/D 2007) (N/D 2010)**

Given data: $\tau = 30\times 10^{-9} sec$, $\tau_{nr} = 100\times 10^{-9} sec$

Formula: $\tau = \frac{\tau_r \times \tau_{nr}}{\tau_r + \tau_{nr}} = \frac{30\times10^{-9}\times100\times10^{-9}}{130\times10^{-9}} = 23.1ns$

Solution: $\eta_{int} = \frac{\tau}{\tau_r} = \frac{23.1\ ns}{30ns} = 0.77 = 77\%$

19. **Calculate the external differential quantum efficiency of a laser diode operating at 1.33μm.The slope of the straight line portion of the emitted optical power P versus drive current I is given by 15 mW/mA. (N/D2011)**

Given data: $\lambda = 1.33\times10^{-6}$

$\frac{q}{Eg} = 0.8065\,\lambda = 0.8065\times1.33\times10^{-6}$ $\quad M = (q)/Eg \times \frac{dP}{dI} = 0.8065\times1.33\times10^{-6}\times 15\times10^{-3}$

$\frac{dP}{dI} = 15\times10^{-3}$ $\quad = 16.089\%$

20. **What are the necessary features of a photo detector? (N/D2007)**

The necessary features of a photo detector are:

- High Quantum efficiency
- Low rise time or fast response
- Low dark current

21. **Define – Responsivity of a photodetector (N/D2008, 2018),(N/D 2010)**

Responsitivity is defined as the ratio of output photo current to the incident optical power.

R= P_O= $\frac{I_p}{P_o}$ $\frac{\eta q}{hv}$

where, R=Responsivity.

I_P=Output photo current P_O=Incident optical power

22. **Compare the performance of APD with PIN diode. (N/D2008)**

APD	PIN
No internal gain	Internal gain is high
Thermal current noise dominates photo detector noise current	Photo detector noise current dominates thermal noise current
Low responsivity	High responsivity
Low dark current	High dark current
Suitable for high intensity application	Suitable for low intensity application

Required low reverse bias voltage	Required high reverse bias voltage

23. List out the operating wavelengths and responsivities of Si, Ge, and InGaAs photodiodes. (N/D2009)

The Operating Wavelengths and Responsivities of Si, Ge, and InGaAs photodiodes are:

Silicon (Si) :

- Operating wavelength range λ = 400 - 1100 nm
- Responsivity R = 0.4-0.6

Germanium (Ge) :

- Operating wavelength range λ = 800 - 1650 nm
- Responsivity R = 0.4 - 0.5

Indium Gallium Arsenide (InGaAs):

- Operating wavelength range λ = 1100 - 1700 nm
- Responsivity R = 0.75 - 0.95

24. List the benefits and drawbacks of avalanche photodiodes.

Benefits of APD are:

- Carrier multiplication takes place.
- Sharp threshold

Drawbacks of APD are:

- High biasing voltage.
- Noisy

25. Photons of energy 1.53×10^{-19} J are incident on a photodiode that has the responsivity of 0.65Amps/W. If the optical power level is 10μW, find the photo current generated.(M/J 2012)

Given data : E = 1.53×10^{-19} J, R = 0.65Amps/W, P_0 = 10×10^{-6} W

Formula : $I_p = R\times P_0$

Solution : $I_p = 0.65\times 10\times10^{-6}$ = **6.5 μA**

26. GaAs has band gap energy of 1.43ev at 300k. Determine the wavelength above which an intrinsic photo detector fabricated from this material will cease to operate.(A/M 2008)

Given data: Eg(eV) = 1.43ev

Formula: $\lambda(\mu m)$ = 1.24/Eg(eV)

Solution: $\lambda(\mu m)$ =1.24/1.43

$\lambda(\mu m)$ =**0.86 μm**.

27. Define – Photocurrent

The high electric field present in the depletion region causes the carriers to separate and be collected across the reverse- biased junction. This gives to a current flow in the external circuit, with one electron flowing for every carrier pair generated. This current flow is known as photocurrent.

28. Define – Impact Ionization

In order for carrier multiplication to take place, the photo-generated carriers must traverse a region where a very high electric field is present. In this high field region, a photo generated electron or hole can gain energy so that it ionizes bound electrons in the valence band upon colliding with them. This current multiplication mechanism is known as impact ionization.

29. Define – Avalanche Effect

The newly created carriers are accelerated by the high electric field, thus gaining enough energy to cause further impact ionization. This phenomenon is called avalanche effect.

30. llustrate the factor that determine the response time of the photodiode.

The resistance and capacitance of the photodiode and the external circuitry give rise to another response time known as RC time constant{\display style \tau =RC}. This combination of R and C integrates the photo response over time and thus lengthens the impulse response of the photodiode. When used in an optical communication system, the response time determines the bandwidth available for signal modulation and thus data transmission.

31. Define dark current?

The photo diode dark current is the current that continues to flow through the bias circuit of the device when no light is incident on the photo diode.

32. Define Johnson or thermal noise?

When current is flowing continuously across the load resistor, heat will be dissipated. This is called thermal noise.

33. What is known as detector response time? (May 2012, NOV/DEC2018)

It is defined as the time taken for the photo detector to respond to an optical input pulse. The response time determines the bandwidth available for signal modulation and data transmission.

34. What are the factors that limit the response time of the photo detectors?

- Transit time of photo carriers within the depletionregion.
- Diffusion time of photo carriers outside the depletion region.
- RC time constant of the photo diode and its associated circuit.

35. What are inherent connection problems when joining fibers?

The inherent connection problems when jointing fibers are,

- Different core and/or cladding diameters.
- Different numerical apertures and/or relative refractive index differences.
- Different refractive index profiles.
- Fiber faults(core elliptically, core concentricity etc

36. Compare PIN and APD?

S.No	PIN	APD
1	No internal gain.	Internal gain.
2	Thermal noise current dominates photo detector noise current.	Photo detector noise current dominates thermal noise current.
3	Low responsivity.	High responsivity.
4	Low dark current.	High dark current.
5	Suitable for high intensity application.	Suitable for low intensity application.
6	Required low reverse bias voltage.	Required high reverse biasvoltage.

37. List out the different types of mechanical misalignments during fiber connection?

The three possible types of misalignment which may occur when joining compatible optical fibers are,

- Longitudinal misalignment
- Lateral misalignment
- Angular misalignment

38. What is fiber splicing?

Fiber splicing is the process of joining two fibers by melting the fiber ends.

39. Compare splices and connectors.

S.No	Splices	Connectors
1	Permanent or semi permanent joints	Temporary joint
2	Splice loss is low	Connector loss is high

40. Define power- bandwidth product.(Apr-May 2015)

High output power and high bandwidth are two important parameters in the design of photo-detector. The Product of photodetector bandwidth and power at which bandwidth ismeasured.

41.Contrast the advantages of PIN diode with APD diode. (Apr-May 2015)

- Low dark current

- It is affected but only thermal noise
- No speed limitation due to capacitance effect

42. What is meant by Mechanical splicing? (May-June 2013)

Mechanical splicing, in which the fibers are held in alignment by some mechanical means, may be achieved by including the use of V-groove into which the butted fibers are placed (or) the use of tubes around the fiber ends.

43. Calculate the Band gap energy for an LED to emit $850nm$? (May- June 2013)

Solution $\lambda = 850nm = 0.85\mu m$

$$E_g = \frac{hc}{\lambda} = \frac{6.625 \times 10^{-34} \times 3 \times 10^8}{0.85 \times 10^{-6}} = 2.33 \times 10^{-19} = 1.45\ eV$$

$$E_g = 1.45\ eV$$

44. Define external quantum efficiency.(Nov-Dec 2016).

The external quantum efficiency is defined as the ratio of the photons emitted from the LED to the number of internally generated photons.

45. Write two difference between a Laser diode and a LED. (Nov-Dec 2013)

S.no	Laser Diode	LED
1.	Coherent radiation Takes place.	In coherent radiation takes place.
2.	Narrow spectral width	Wider Spectral width

46. Why silicon is preferred for fabrication of photo receiver?

- Silica is used for fabrication photo receiver, because it has larger band gap, it generates low noise and it supports multiple channels as it has larger bandwidth.
- Silicon is available plenty in nature.

47. Why are semiconductor based photo detectors preferred to other types of photo detectors?

Semiconductor laser diode generates low noise and they support multiple channels as they have larger bandwidth.

48. What are the requirements of photo detector?

- The photo detector must have high quantum efficiency to generate a large signal power.
- The photo detector and amplifier noises should be kept as low as possible.

49. What is the significance of intrinsic layer in PIN diode.[APR/MAY 2019]

The intrinsic region in the diode is in contrast to a PN junction diode. This region makes the PIN diode an lower rectifier, but it makes it appropriate for fast switches, attenuators, photo detectors and applications of high voltage power electronics.

50. What are the factors that limit the response time of the photo detectors?

- Transit time of photo carriers within the depletion region.
- Diffusion time of photo carriers outside the depletion region.
- RC time constant of the photo diode and its associated circuit.

51.What are inherent connection problems when joining fibers?

The inherent connection problems when jointing fibers are,

- Different core and/or cladding diameters.
- Different numerical apertures and/or relative refractive index differences.
- Different refractive index profiles.
- Fiber faults(core elliptically, core concentricity etc)

PART B & C

1. Discuss the LASER diode principle, modes and threshold conditions. (Jun 2007)

Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 163 **(Apr./May 2008)**

2. **a. Derive the threshold condition for lasing. [DEC2016]**
   b. Explain in detail the fabry perot resonator cavity Laser diode.
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 163 **(Apr./May 2008)**

3. **a. Explain the various lensing schemes. (Nov 2004,Dec 05,07,April 09, Apr 2017. Apr 2018)**
   b. Explain the various splicing techniques.
4. **Discuss about modulation of LED & Quantum LASER (Nov 2010)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 152
5. **With neat diagram explain the working of surface emitting LED (Nov 2011, 2018, may 2012, Apr 2017, 2019)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 152
6. **Explain the structure of silicon ADP**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 249
7. **Explain any two injection laser structure with neat diagram (May 2012)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 163 **(Apr./May 2008)**
8. **Derive laser diode rate equations. (16 marks) [APR/MAY 2019] (Nov 2012)[DEC2016]**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 163 **(Apr./May 2008, Apr 2018)**
9. **(i) Explain the working of n hetero structure LED. (Nov2013, NOV 2018)**
   (ii) Define internal quantum efficiency of a LED. Deduce the expression for the same.
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 152
10. **Explain expanded beam connectors with neat diagram (Nov 2011)**
    Refer Book: Optical fiber Communications –John M.Senior - Pg. No. 238-244.
11. **Write brief note on fiber alignment and joint loss (May 2012)**
    Refer Book: Optical fiber Communications –John M.Senior - Pg. No. 227-234.
12. **Describe about connectors, splices and couplers. (Nov-Dec 2015)**
    Refer Book: Optical fiber Communications –John M.Senior - Pg. No. 238-244.
13. **A Photodiode is constructed of GaAs which has a band gap energy of 1.43eV at 300K. Find the long wavelength cut-off. (Apr-May 2015)**
    Refer Book: Optical fiber Communications –John M.Senior - Pg. No. 238-244.
14. **What do you understand by optical-wave confinement and current confinement in LASER diode? Explain with suitable structures. (Nov-Dec 2013)**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 163
15. **Discuss about optical detection noise. (Nov-Dec 2015)**
    Refer Book: Optical fiber Communications –John M.Senior - Pg. No. 238-244.
16. **Explain gain guided and Index guided laser diodes.**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 163

UNIT-IV

FIBER OPTIC RECEIVER AND MEASUREMENTS

1. What is bit rate?

The transmitted signal is two level binary data stream consisting of either 0 or 1 in a time slot of duration T. this time slot is referred to a bit period.

2. What are the error sources of receiver? (M / J 2013)

The error sources of receiver are

- Thermal noise
- Dark current noise
- Quantum noise

3. A digital fiber optic link operating at 1310 nm, requires a maximum BER of 10^{-8}. Calculate the required average photons per pulse. (N / D 2013)

The probability error $P_r(o) = e^{-N} = 10^{-8}$

Solving for $N = 8 \log_e 10 = 18.42$

An average of 18 photons per pulse is required for this BER.

4. **Define: Probability of error**.

Probability of error means that a transmitted' l' is misinterpreted as a '0' or transmitted '0' is misinterpreted as a '1' by the receiver.

5. **Define – Quantum Limit (M/J2012), (N/D2007)9DEC2016)**

The minimum received power level required to maintain a specific Bit – Error - Rate (BER) of an optical receiver is known as the quantum limit.

6. **What is meant by (1/f) noise corner frequency? (N/D2009)**

The (1/f) noise corner frequency is defined as the frequency at which (1/f) noise, which dominates the FET noise at low frequencies and has (1/f) power spectrum.

7. **Why silicon is preferred to make fiber optical receivers? (N/D2010), (A/M2011)**

Silicon is preferred to make fiber optical receivers because

It has high sensitivity over the 0.8–0.9 μm wavelength band with adequate speed

It provides negligible shunt conductance, low dark current and long-term stability

8. **Define – Modal noise and Mode Partition Noise.(A/M2011)(M/J2009)(A/M2010)**

Disturbances along the fibre such as vibrations, discontinuities, connectors, splices and source/detector coupling may cause fluctuations in the speckle patterns. It is known as modal noise.

Phenomenon that occurs in multimode semiconductor lasers when the modes are not well stabilized is known as mode partition noise.

9. **Mention the error sources in fiber optical receiver. (N/D2011)**

There are three main error sources in fiber optical receiver. They are:

- Thermal noise
- Dark current noise
- Quantum noise

10. Define – Bit Error Rate(DEC2016)

Bit Error Rate (BER) is defined as the ratio of the number of errors occurred over a certain time interval't' to the number of pulses transmitted during this interval.

11. **How does dark current arise?**

When there is no optical power incident on the photo detector a small reverse leakage current flows from the

device terminals known as dark current. Dark current contributes to the total system noise and gives random fluctuations about the average particle flow of the photocurrent

12. **What is Inter Symbol Interference?**

Each pulse broadens and overlaps with its neighbors, eventually becoming indistinguishable at the receiver input. This effect is known as Inter Symbol Interference.

13. **Define – Extinction ratio**

The extinction ratio is usually defined as the ratio of the optical energy emitted in the '0' bit period to that emitted during the '1' bit period.

14. **What are the requirements of an optical receiver**?**(Nov. / Dec. 2006)**

- It should have fast response;
- High sensitivity;
- Tolerable SNR;
- Should reproduce the signal without any distortion.

15. Define – Minimum Detectable Optical Power

It is defined as the optical power necessary to produce a photocurrent of the same magnitude as the root mean square of the total current.

16. What are the noise effects on system performance?

The main penalties are modal noise, wavelength chirp, spectral broadening, mode-partition noise.

17. Why the attenuation limit curve slopes towards to the right?

As the minimum optical power required at the receiver for a given BER becomes higher for increasing data rates, the attenuation limit curve slopes downward to the right.

18. What do you mean thermal noise?

Thermal noise is due to the random motion of electrons in a conductor. Thermal noise arising from the detector load resistor and from the amplifier electronics tend to dominate in applications with low signal to noise ratio.

19. What is meant by excess noise factor?

The ratio of the actual noise generated in an avalanche photodiode to the noise that would exist if all carrier pairs were multiplied by exactly m is called the excess noise factor. (F).

20. What are the system requirements?

The key system requirements are as follows

- The desired or possible transmission distance
- The data rate or channel bandwidth
- Bit error rate (BER)

21. Give the two analyses that are used to ensure system performance.

The two analyses that are used to ensure system performance are

- Link power budget analysis
- Rise time budget analysis.

22. What are the requirements of preamplifier?

It should have low noise level, high bandwidth, high dynamic range, and high sensitivity to avoid non linearity and high gain.

23. What are the types of pre - amplifiers?

The types of pre-amplifier are

- Low- impedance preamplifier
- High – impedance preamplifier

Transimpedance preamplifier

24. List the advantages of preamplifiers.

The advantages of pre amplifiers are

- Low noise level
- High Bandwidth
- High dynamic range

- High Sensitivity
- High gain

25. What are the standard fiber measurement techniques?

The standard fiber measurement techniques are

- Fiber attenuation measurement
- Fiber dispersion measurement
- Fiber refractive index profile measurement
- Fiber cutoff wavelength measurement
- Fiber numerical aperture measurement
- Fiber diameter measurement

26. Define – Bend Attenuation

A peak in the wavelength region where the radiation losses resulting from the small loop are much higher than the fundamental mode is known as bend attenuation.

27. **What is the technique used for measuring the total fiber attenuation?**

Total fiber attenuation per unit length can be determined using cut-back method. Taking a set of optical output power measurements over the required spectrum using a long length of fiber usually at least a kilometre is known as cut back technique. The fiber is then cut back to a point 2 meters from the input end and maintaining the same launch conditions, another set of power output measurements are taken.

Relationship for the optical attenuation per unit length α_{db} for the fiber may be obtained from,

$$\alpha_{db} = \frac{10}{(L_1 - L_2) log_{10} \frac{P_{02}}{P_{01}}}$$

where,

L_1, L_2 - original and cut-back fiber length respectively

P_{02}, P_{01} -output optical powers at a specific wavelength from the original and cut back fiber lengths.

28. What are the factors that produce dispersion in optical fibers?

The factors that produce dispersion in optical fibers are:

- Propagation delay difference between the different spectral components of the transmitted signal.
- Variation in group velocity with wavelength

29. What are the methods used to measure fiber dispersion?

The methods used to measure fiber dispersion are:

- Time domain measurement
- Frequency domain measurement

30. What are the methods used to measure fiber refractive index profile? (M/J2012)

The methods used to measure fiber refractive index profile are

- Interferometric method
- Near infra scanning method
- Refracted near field method

31. List the process associated with fiber optic receiver section.

Although the photo-detector is the major element in the fibre optic receiver, the other elements to the whole unit. Once the light has been received by the fibre optic receiver and converted into electronic pulses, the signals are processed by the electronics in the receiver. Typically these will include various forms of amplification including a limiting amplifier. These serve to generate a suitable square wave that can then be processed in any logic circuitry that may be required.

32. What is Mode Coupling and what are its causes?

It is another type of pulse distortion which is common in optical links.

The pulse distortion will be increased less rapidly after a certain initial length of fiber, due to this mode coupling and differential mode losses occur.

33. Define Quantum limit (Q). (May-June 2013)

The minimum received power level required for a specific BER of digital system is known as Quantum limit.

34. List out the methods used to measure fiber refractive index profile.

1. Inter-ferometric method
2. Near field scanning method
3. End field scanning method

35. What are the error sources in fiber optic receiver? (May-June 2013, Nov-Dec 2012)

The error sources in fiber optic receiver are

- Shot Noise
- DarkCurrent

  Bulk Dark Current SurfaceDarkCurrent
- Thermal Noise.
- Amplifier noise

36. What are the different techniques for determining attenuation in optical fiber?

The different techniques for determining attenuation are

i) Cut-back ii) Insertion-loss

37. Write the expression to measure attenuation using cut back method.

$$dB = 0 \frac{1}{L_1 - L_2} \log_{10} \frac{V_1}{V_2}$$

Where

38. List any two advantages of trans-impedance amplifiers.(Apr-May 2015)

(i) Reduces thermal noise

(ii) Provide wide bandwidth

39. State the significance of maintaining the fiber outer diameter constant.(Nov-Dec 2014)

It is essential during the fiber manufacturing process (at the drawing stage) that the fiber outer diameter (cladding diameter) is maintained constant to within 1%. Any diameter variations may cause excessive radiation losses and make accurate fiber – fiber connection difficult.

40. Mention few fiber diameter measurement techniques. (Nov-Dec 2015)

There are two very broad classifications of diameter measurements techniques

(i) Contacting or destructive methods

(ii) Non-contacting and nondestructive methods

41. What is dark current?(Nov-Dec 2012)

The photo diode dark current is the current that continues to flow through the bias circuit of the device when no light is incident on the photo diode.

42. A digital fiber optic link operating at 1310 nm, requires a maximum BER of 10-8. Calculate the required average photons per pulse. (Nov- Dec 2013)

Solution:

Given

Probability error $(_r)P0 = \square N \square 10^{\square 8}$

L_1 = original fiber length

L_2 = Cut-back fiber lengthV1, V2are the output voltages

$N \square 8\log_e 10 \square 18.42 \square 18$

An average of 18 photons per pulse ie required for this BER.

43. A trignometrical measurement is performed in order to determine the numerical aperture of a step index fiber. The screen is positioned 10.0cm from the fiber end face. When illuminated from a wide angled visible source the measurement output pattern size is6.2 cm. Calculate the Numerical Aperture of the fiber.

NA = A / (A^2 +4D^2) = 6.2 / (38.44 + 400) = 0.30

44. In a fiber optic system operating at 1.3 μm, the transmitter power = -3 dB m. Fiber cable loss = 0. 2 dB/km. Total connector loss at the transmitter and receiver = - 2dB. PINreceiver sensitivity when operating at 400 Mbps = - 44 dB m. Safety margin = 6 dB. Calculate the maximum transmission distance without repeaters (i) when there is no dispersion equalization penalty and (ii) when there is a dispersion equalization penalty o f1.5 dB.

(i) P Tr- P Rr= a fL + a c+ P SM= - 3 dB - (- 44 dB)

=0.2 x L + 2 + 641-8=0.2L

L= 33/0.2 = 165 km

(ii) P Tr- P Rr= a fL + a c+ P SM+ D L

41 =0.2xL+2+6+1.5

L = 31.5/0.2 = 157.5 km

45. State detector response time

This is the measure of the photodiode response speed to a stepped light input signal. It is the time required for the photodiode to increase its output from 10% to 90% of final output level. is the rise timeof the device.

46. Give the 2 analysis that are used to ensure system performance?

The 2 analysis that are used to ensure system performance are: • link power budget analysis • rise time budget analysis

47. Explain briefly about link power budget analysis?

In the optical power loss model for a pt-to-pt link, the optical power rxed at the photo detector depends on the amount of light coupled into the fiber & losses occurring in the fiber at the connectors & splices. The link loss budget is derived from the sequential loss contribution of each element in the link. Loss=10 log (Pout) (Pin) The total optical power loss is, PT = PS - PR

48. Give the range of system margin in link power budget? The system margin is usually (6-8) dB. A positive system margin ensures proper operation of the circuit. A negative value indicates that insufficient power will reach the detector to achieve the required bit error rate, BER.

49. The specifications of the light sources are converted to equivalent rise time in rise time budget. Why? A rise time budget is a convenient method to determine the dispersion limitation of an optical link. This is particularly useful for digital systems. For this purpose, the specifications of the light sources (both the fiber & the photo detector) are converted to equivalent rise time. The overall system rise time is given in terms of the light source rise time, fiber dispersion time & the photo detector rise time.

50. What are the system components of system rise time? The 4 basic system components that contribute to the system rise time are: • transmitter (source) rise time • receiver rise time • material dispersion time of the fiber • modal dispersion time of the fiber link All these 4 basic elements may significantly limit system speed.

51. Why the attenuation limit curve slopes downwards to the right? As the minimum optical power required at the rxer for a given BER becomes higher for increasing data rates, the attenuation limit

curve slopes downward to the right.

PART B & C

1. **Draw and explain the high impedence of high impedence pre-amplifier designed based on BJT and FET (8)**
   b) Write a brief note on trans impedence amplifier. [8] (Nov/Dec 2008, 2018)
   Refer Book: Optical fiber Communications - John M.Senior - Pg. No . 377.
2. **Explain the operation of preamplifier built using FET (Nov 2011)**
   Refer Book: Optical fiber Communications - John M.Senior - Pg. No . 377.
3. **Explain measurement technique used in the case of fiber diameter, fiber cut off length, refractive index profile, Numerical aperture (Nov 2011, 2018)**
   Refer Book: Optical fiber Communications - John M.Senior - Pg. No . 779 -781.
4. **Draw and explain the operation of high impedance FET and BJT preamplifiers(May 2012)**
   Refer Book: Optical fiber Communications - John M.Senior - Pg. No . 377.
5. **Explain a) Attenuation measurement using cut back technique**
   b)Frequency domain measurement of fiber dispersion (May 2012) (DEC2016)
   Refer Book: Optical fiber Communications - John M.Senior - Pg. No . 782 -783.
6. **Considering the probability distributions for received logic 0 and 1 signal pulses.**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 282**(Nov. / Dec. 2007)**
7. **Derive the expressions for BER and error function. (Nov 2012)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 282**(Nov. / Dec. 2007)**
8. **Explain the types of preamplifiers used in a receiver.**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 282**(Nov. / Dec. 2007)**
9. **Define the terms- Quantum limit and probability of Error with respect to a receiver with typical values. (Nov 2013, 2018)**
   Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 282**(Nov. / Dec. 2007)**
10. **With suitable diagram, explain optical receiver operation and its performance.(May 2014, May 2015)**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 279
11. **Describe the dispersion and numerical aperture measurements of fiber. (May 2014)**
    Refer Book: Optical fiber Communications - John M.Senior - Pg. No . 779 -781.
12. **With schematic diagram, explain the blocks and their functions of an optical receiver. (Apr-May 2015, Nov-Dec 2014, APR/MAY 2019)**
    Refer Book: Optical fiber Communications - John M.Senior - Pg. No . 779 -781.
13. **A digital fibre optic link operationg at 850nm requires a maximum BER of 10^{-9}. Find the quantum limit in terms of the quantum efficiency of the detector and the energy of the incident photon. (Apr-May 2015)**
    Refer Book: Optical fiber Communications - John M.Senior - Pg. No . 779 -781.
14. **What are the performance measures of a digital receiver? Derive an expression for bit error rate of a digital receiver. (Nov-Dec 2016,Nov-Dec 2015, 2018)**

Refer Book: Optical fiber Communications - John M.Senior - Pg. No . 779 -781.

15. **Draw the block diagram of OTDR. Explain the measurements of any two fiber optic measurements with this. (Nov-Dec 2014)**
Refer Book: Optical fiber Communications - John M.Senior - Pg. No . 779 -781.

UNIT-V OPTICAL NETWORKS

1. **Define power penalty [NOV/DEC 2018]**
The power penalty may become significant if the semiconductor laser is biased above the threshold. For lasers biased below threshold, the extinction ratio is typically 0.05 and the power penalty is less than 0.4 dB. Timing Jitter In a digital system the signal is generally sampled at the center of the pulse.
2. **Distinguish between fundamental and higher order solitons.[APR/MAY 2019]**
A fundamental soliton is an optical pulse which can propagate in a dispersive medium (e.g. an optical fiber) with a constant shape of the temporal intensity profile, i.e., without any temporal broadening as is usually caused by dispersion.
A higher-order soliton is a soliton pulse the energy of which is higher than that of a fundamental soliton by a factor which is the square of an integer number
3. **Mention the drawbacks of Broadcast and select networks for wide area networkapplications.[Apr 2018]**
The drawbacks of broadcast and select networks for wide area network applications are:
More wavelengths are needed as the number of nodes in the network grows
Without the use of optical booster amplifiers splitting losses occurs
4. **What are the three topologies used for fiber optical network? (N/D 2011)**
The three topologies used for fiber optical network are:
Bus
Ring
Star
5. **Calculate the number of independent signals that can be sent on a single fiber in the 1525-1565 nm bands. Assume the spectral spacing as per ITU-T recommendation G.692. (A/M 2011)**
Given data: Mean frequency spacing as per ITU-T is 0.8 nm
Wavelength = 1565 nm -1525 nm = 40 nm
Solution:
Number of independent channel = (40 nm/0.8 nm) = 50 Channels
6. **What is meant by power penalty?**
When nonlinear effects contribute to signal impairment, an additional amount of power will be needed at the receiver to maintain the same BER. This additional power(dB) is known as the power penalty.
7. **Define – Network**
Network is defined as to establish connections between these stations; one interconnects them by transmission paths to form a network.
8. **What is meant by topology?**
The topology is the logical manner in which nodes are linked together by information transmission channels to form a network.
9. **What are the drawbacks of broadcast and select networks for wide area network applications?**

(M/J 2012)

The drawbacks of broadcast and select networks for wide area network applications are:

More wavelengths are needed as the number of nodes in the network grows

Without the use of optical booster amplifiers splitting losses occurs

10. Define – WDM (A/M2011)

In fiber-optic communications, wavelength-division multiplexing (WDM) is a technology which multiplexes a number of optical carrier signals onto a single optical fiber by using different wavelengths (i.e. colors) of laser light. This technique enables bidirectional communications over one strand of fiber, as well as multiplication of capacity.

11. **What are the advantages of WDM? (N/D2007)**

The advantages of WDM are

- Various optical channels can support different transmission formats
- Increase in the capacity of optical fiber compared to point-to-point link

12. What is the purpose of rise-time budget analysis? (A/M2008)

Rise-time budget ensures that the link is able to operate for a given data rate at specified BER. All the components in the link must operate fast enough to meet the band width or rise time requirements.

13. The specifications of the light sources are converted to equivalent rise time in rise time budget. Why?

A rise time budget is a convenient method to determine the dispersion limitation of an optical link. This is particularly useful for digital systems. For this purpose, the specifications of the light sources (both the fiber and the photo detector) are converted to equivalent rise time. The overall system rise time is given in terms of the light source rise time, fiber dispersion time and the photo detector rise time.

14. What is EDFA?(A/M2008, 2019), (M/J2012)NOV /DEC2018

An erbium-doped fiber amplifier (EDFA) is a device that amplifies an optical fiber signal. A trace impurity in the form of a trivalent erbium ion is inserted into the optical fiber's silica core to alter its optical properties and permit signal amplification.

15. What are the two different types of WDM? (DEC2016)

The two different types of WDM are

a. Unidirectional WDM

b. Bidirectional WDM

16. Define – Crosstalk

Crosstalk is defined as the feed through one of the channel signals into another channel.

17. Give the important features of time-slotted optical TDM network.

The important features of time slotted optical TDM network are

c. To provide backbone to interconnect high speed networks
d. To transfer quickly very large data blocks
e. To switch large aggregations of traffic
f. To provide both high- rate.

18. How the speckle pattern can form?

The speckle patterns are formed by the interference of the modes from a coherent source when the coherence time of the source is greater than the intermodal dispersion time within the fiber.

19. Define – Full- Width Half- Maximum(FWHM)

The FWHM is a pulse defined as the full width at its half-maximum power level.

20. **What are the advantages of using soliton signals through fiber? (M/J2009)**

The advantages of using soliton signals through fiber are, it is very narrow, high-intensity optical pulses that retain their shape through the interaction of balancing pulse dispersion with the nonlinear properties of an optical fiber

21. **What is chirping? (N/D2009)**

The d.c. modulation of a single longitudinal mode semiconductor laser can cause a dynamic shift of the peak wavelength emitted from the device This phenomenon, which results in dynamic line width broadening under the direct modulation of the injection current, is referred to as frequency chirping.

22. What is the best way to minimize chirping?

It is to choose the LASER emission wavelength close to the zero-dispersion of the wavelength of the fiber.

23. What do you mean by bidirectional WDM?

A single WDM which operates as both multiplexing and demultiplexing device is said to be bidirectional WDM.

24. What are the basic performances of the WDM?

The basic performances of WDM are

- ✓ Insertion loss
- ✓ Channel width
- ✓ Cross talk

25. Distinguish between fundamental and higher order soliton. (N/D2007)

The optical pulse that does not change in shape is called fundamental solitons.

The pulses that undergo periodic shape changes are called higher order solitons.

26. What are the two different types of WDM? (MAY 2016)

The two different types of WDM are

g. Unidirectional WDM

h. Bidirectional WDM

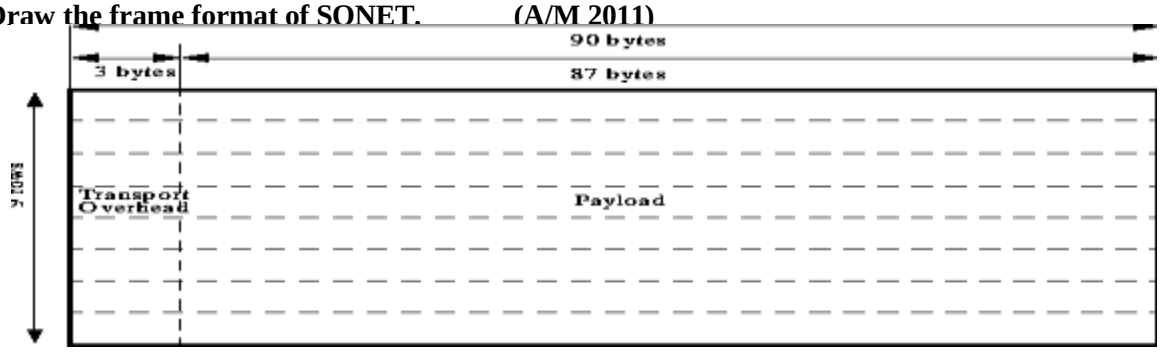
27. What is DWDM?

Dense Wavelength Division Multiplexing (DWDM) is an optical technology used to increased bandwidth over existing fiber-optic bones. It works by combining and transmitting multiple signals simultaneously at different wavelengths on the same fibers.

28. What is SONET/SDH?[Apr 2018]

Synchronous Optical NETworking (SONET) or Synchronous Digital Hierarchy (SDH) is a standardized protocol that transfers multiple digital bit streams over optical fiber using lasers or highly coherent light from light emitting diodes. At low transmission rates data can also be transferred via an electrical interface.

29. **Draw the frame format of SONET. (A/M 2011)**



30. **What are the drawbacks of broadcast and select network for wavelength multiplexing?**

The problems that arise in broadcast and select networks are:

- More wavelengths are needed as the number of nodes in the network grows.
- Without the width spread use of optical booster amplifiers, due to this splitting loss is high.

31. **What is optical CDMA? (Nov-Dec 2015)**

Optical CDMA is a multiple access technique in which each user is assigned an unique optical code. When a receiver is placed anywhere on the network with a bar code that matches a transmitter that signal line is decoded and extracted from the network.The codes are orthogonal to each other.

32. **Distinguish SONET and SDH. (Nov-Dec 2015)**

ONET	**DH**
t means synchronous optical network developed by ANSI.	t means synchronous digital hierarchy developed by ITU
asic signaling unit is OC-I (51.84Mbps)	asic signaling unit is STM-1 (155.52 Mbps)
ONET uses the term section, line and path.	DH uses the term path, multiplex section and egenerator section.

33. **Name two popular architectures of SONET/SDH network.(Nov-Dec 2016) (R)**

The two popular architectures of SONET/SDH networks are:

- UPSR - Unidirectional Path Switched Ring, two-fiber.
- BLSR – Bidirectional Line Switched Ring, two-fiber or four-fiber.

34. **Obtain the transmission bit rate of the basic SONET frame in Mbps.(Nov-Dec 2013) (E)**

STS-1 frame rate = (810 bytes/frame)*(8000 frames/sec)
= 51.840 Mbps.

35. **Illustrate inter-channel cross talk that occurs in a WDM system.(Nov- Dec 2013) (A)**

Inter-channel crosstalk arises when an interfacing signal comes from a neighboring channel that operates at a different wavelength. This nominally occurs when a wavelength selecting device imperfectly rejects or isolates the signals from other near-by wavelength channels.

36. **What is a broadcast and select network?(May-June 2013) (R)**

In broadcast and select networks, a node sends its transmission to the star coupler on the available wavelength using a laser which produces an optical information stream. The information stream from

multiple sources is optically combined by the star and the signal and the signal power of each stream is equally spilt and forwarded to all the nodes on their receiver fiber.

37. What is SONET?(Apr-May 2015) (R)

SONET means synchronous optical network which is developed by ANSI, standardized protocol that transfer multiple digital bit stream synchronously over optical fiber using laser.

38. What were the problems associated with PDH networks?(Nov-Dec 2012) (AZ)

1. PDH- Plesiochronous Digital Hierarchy
2. It is difficult to "pick out" (drop) a low bit rate stream out of a high bit rate stream it is completely demultiplexing stream**.**
3. Expensive and compromises network reliability.

39. Enumerate the various SONET/SDH layers. ?(Nov-Dec 2012)

The various SONET/SDH layers are,

- Photonic layer
- Section layer
- Line layer
- Path layer.

40. What is DWDM?

Dense Wavelength Division Multiplexing (DWDM) is an optical technology used to increased bandwidth over existing fiber-optic bones. It works by combining and transmitting multiple signals simultaneously at different wavelengths on the same fibers.

41.What are solitons? (N/D2010) (DEC2016)

Solitons are nonlinear optical pulses that have the potential to support very high optical transmission rates of many terabits per second over long distances.

42. Give the important features of time-slotted optical TDM network.

The important features of time slotted optical TDM network are

- ✓ To provide backbone to interconnect high speed networks
- ✓ To transfer quickly very large data blocks
- ✓ To switch large aggregations of traffic
- ✓ To provide both high- rate.

43. How the speckle pattern can form?

The speckle patterns are formed by the interference of the modes from a coherent source when the coherence time of the source is greater than the intermodal dispersion time within the fiber.

44.What do you mean by bidirectional WDM?

A single WDM which operates as both multiplexing and demultiplexing device is said to be bidirectional WDM.

45. Define – Full- Width Half- Maximum(FWHM)

The FWHM is a pulse defined as the full width at its half-maximum power level.

46. What are the types of broadcast and select network?

The types of broadcast and select network are

- ✓ Single – hop networks
- ✓ Multi – hop networks

47. What is meant by cross- phase modulation (XPM)?

Cross- phase modulation, which converts power fluctuations in particular wavelength channel to phase fluctuations in the copropating channels.

48. Define self-healing rings.

The SONET/SDH rings are called **self-healing rings,** since the traffic flowing along a certain path can automatically be switched to an alternate path.

49. Mention the architectures for SONET/SDH networks.

- ✓ Two fiber unidirectional path switched ring.
- ✓ Two or four fiber bidirectional path switched ring.

50. Define single –hop network.

Single –hop network refers to networks where information transmitted in the form of light reaches its destination without being converted to an electrical form at any intermediate point.

og

PART B & C

1. With neat diagram, explain the elements of SONET infrastructure. (16) (May 2007) (MAY 2016)

Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. NO.472

2. Explain the principle and operation of Erbium doped fiber amplifiers with neat diagrams. (10) (Jan 2010)

Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 514- 516.

3. Describe the principle and performance of DT-WDMA protocol. (8)

Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 477- 482

4. Explain the architecture of SONET and discuss the nonlinear effects on network performance (Nov 2011)

Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. NO.472

5. Explain the principle of solitons and discuss the soliton parameters with necessary expressions (May 2012)

Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No . 506.

6. Write short notes on optical CDMA,WDM and EDFA performance (May 2012)

Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 426.

7. Describe the non-linear effects on network performance in detail.(8) (Nov 2012, 2018) [DEC 2016] MAY 2019

Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 514- 516.

8. Explain the basics of optical CDMA systems. (8) (Nov 2012)

Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 514- 516.

9. (i) What is a four fiber BLSR ring in a SONET? Explain the reconfiguration of the same during node or fiber failure

(ii) What is broadcast and select multi hop network? Explain. (Nov 2013)

Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. NO.4 72

10. (i) Explain the following requirements for the design of an optically amplified WDM link: 1) Link band width

2) Optical power requirements for a specific BER.

Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 477- 482

(ii) Write a note on solitons. (Nov 2013)

11. **Expalin SONET layers and frame structure with diagram (May 2014) [DEC 2016,2018]**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. NO.4 72
12. **Discuss in detail about the effect of noise on system performance.(Nov-Dec 2016)**

13. **Discuss the performance improvement of WDM and EDFA systems.(Nov- Dec 2015,Apr-May 2015, 2019, Nov-Dec 2014, NOV/DEC2018)**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 426.

14. **Discuss the non-linear effects on optical network performance.(Apr-May 2015, 2019, Nov-Dec 2012)**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 514- 516.
15. **Explain i)Optical CDMA ii)Optical Wavelength Routing Network.(Nov-Dec 2012)**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 514- 516.
16. **Discuss about Ultra High Capacity Networks. (Apr-May 2015,Nov-Dec 2014)**
    Refer Book: Optical fiber Communications - Gerd Kaiser - Pg. No. 514- 516.

UNIT –I

Introduction to Optical Fibers:

Historical Development

- Fiber optics deals with study of propagation of light through transparent dielectric wageguides. The fiber optics are used for transmission of data from point to point location. Fiber optic systems currently used most extensively as the transmission line between terrestrial hardwired systems.

- The carrier frequencies used in conventional systems had the limitations in handlinmg the volume and rate of the data transmission. The greater the carrier frequency larger the available bandwith and information carrying capacity.

First generation

- The first generation of lightwave systems uses GaAs semiconductor laser and operating region was near 0.8 μm. Other specifications of this generation are as under:

  i) Bit rate : 45 Mb/s

  ii) Repeater spacing : 10 km

Second generation

i) Bit rate : 100 Mb/s to 1.7 Gb/s

ii) Repeater spacing : 50 km

iii) Operation wavelength : 1.3 μm

iv) Semiconductor : In GaAsP

Third generation

i) Bit rate : 10 Gb/s

ii) Repeater spacing : 100 km

iii) Operating wavelength : 1.55 μm

Fourth generation

Fourth generation uses WDM technique.

Bit rate : 10 Tb/s

Repeater spacing : > 10,000 km

Operating wavelength : 1.45 to 1.62 μm

Fifth generation

Fifth generation uses Roman amplification technique and optical solitiors.

Bit rate : 40 - 160 Gb/s

Repeater spacing : 24000 km - 35000 km

Operating wavelength : 1.53 to 1.57 µm

Need of fiber optic communication

- Fiber optic communication system has emerged as most important communication system. Compared to traditional system because of following requirements :

  1. In long haul transmission system there is need of low loss transmission medium
  2. There is need of compact and least weight transmitters and receivers.
  3. There is need of increase dspan of transmission.
  4. There is need of increased bit rate-distrance product.

- A fiber optic communication system fulfills these requirements, hence most widely acception.

General Optical Fiber Communication System

- Basic block diagram of optical fiber communication system consists of following important blocks.

  1. Transmitter
  2. Information channel
  3. Receiver.

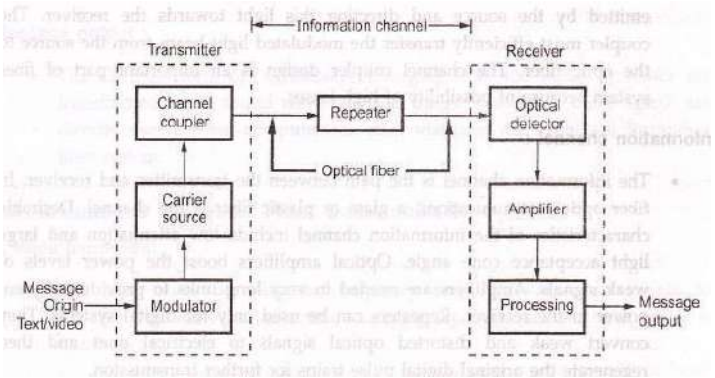


Fig. 1.2.1 shows block diagram of OFC system.

Message origin :

- Generally message origin is from a transducer that converts a non-electrical message into an electrical signal. Common examples include microphones

for converting sound waves into currents and video (TV) cameras for converting images into current. For data transfer between computers, the message is already in electrical form.

Modulator :

- The modulator has two main functions.

  1) It converts the electrical message into the proper format.

  2) It impresses this signal onto the wave generated by the carrier source.

  Two distinct categories of modulation are used i.e. analog modulation and digital modulation.

Carrier source :

- Carrier source generates the wave on which the information is transmitted. This wave is called the carrier. For fiber optic system, a laser diode (LD) or a light emitting diode (LED) is used. They can be called as optic oscillators, they provide stable, single frequency waves with sufficient power for long distance propagation.

Channel coupler :

- Coupler feeds the power into the information channel. For an atmospheric optic system, the channel coupler is a lens used for collimating the light emitted by the source and directing this light towards the receiver. The coupler must efficiently transfer the modulated light beam from the source to the optic fiber. The channel coupler design is an important part of fiber system because of possibility of high losses.

Information channel :

- The information channel is the path between the transmitter and receiver. In fiber optic communications, a glass or plastic fiber is the channel. Desirable characteristics of the information channel include low attenuation and large light acceptance cone angle. Optical amplifiers boost the power levels of weak signals. Amplifiers are needed in very long links to provide sufficient power to the receiver. Repeaters can be used only for digital systems. They convert weak and distorted optical signals to electrical ones and then regenerate the original disgital pulse trains for further transmission.

- Another important property of the information channel is the propagation time of the waes travelling along it. A signal propagating along a fiber normally contains a range of optic frequencies and divides its power along several ray paths. This results in a distortion of the propagating signal. In a digital system, this distortion appears as a spreading and deforming of the pulses. The spreading is so great that adjacent pulses begin to overlap and become unrecognizable as separate bits of information.

Optical detector :

- The information being transmitted is detector. In the fiber system the optic wave is converted into an electric current by a photodetector. The current developed by the

- detector is proportional to the power in the incident optic wave. Detector output current contains the transmitted information. This detector output is then filtered to remove the constant bias and thn amplified.

- The important properties of photodetectors are small size, economy, long life, low power consumption, high sensitivity to optic signals and fast response to quick variations in the optic power.

Signal processing :

- Signal processing includes filtering, amplification. Proper filtering maximizes the ratio of signal to unwanted power. For a digital system decision circuit is an additional block. The bit error rate (BER) should be very small for quality communications.

Message output :

- The electrical form of the message emerging from the signal processor are transformed into a soud wave or visual image. Sometimes these signals are directly usable when computers or other machines are connected through a fiber system.

Advantages of Optical Fiber Communications

1. Wide bandwidth

- The light wave occupies the frequency range between 2 x 1012 Hz to 3.7 x 1012 Hz. Thus the information carrying capability of fiber optic cables is much higher.

2. Low losses

- Fiber optic cables offers bery less signal attenuation over long distances. Typically it is less than 1 dB/km. This enables longer distance between repeaters.

3. Immune to cross talk

- Fiber optic cables has very high immunity to electricaland magnetic field. Since fiber optic cables are non-conductors of electricity hence they do not produce magnetic field. Thus fiber optic cables are immune to cross talk between cables cause dby magnetic induction.

4. Interference immune

- Fiber optic cable sar eimmune to conductive and radiative interferences caused by electrical noise sources such as lighting, electric motors, fluorescent lights.

5. Light weight

- As fiber cables are made of silica glass or plastic which is much lighter than copper or aluminium cables. Light weight fiber cables are cheaper to transport.

6. Small size

- The diameter of fiber is much smaller compared to other cables, therefore fiber calbe is small in size, requires less storage space.

7. More strength

- Fiber cables are stronger and rugged hence can support more weight.

8. Security

- Fiber cables are more secure than other cables. It is almost impossible to tap into a fiber cable as they do n ot radiate signals.

No ground loops exist between optical fibers hence they are more secure.

9. Long distance transmission

- Becaujse of less attenuation transmission at a longer distance is possible.

10. Environment immune

- Fiber calbes are more immune to environmental extremes. They can operate over a large temperature varioations. Also they are not affected by corrosive liquids and gases.

11. Sage and easy installation

- Fiber cables are safer and easier to install and maintain. They are non-conductors hence there is no shock hazards as no current or voltage is associated with them. Their small size and light weight feature makes installation easier.

12. Less cost

- Cost of fiber optic system is less compated to any other system.

Disadvantages of Optical Fiber Communicaitons

1. High initial cost

- The intial cost of installation or setting up cost is very high compared to all other system.

2. Maintenance and repaiding cost

- The maintenance and repaiding of fiber optic systems is not only difficult but expensive also.

3. Jointing and test procedures

- Since optical fibers are of very small size. The fiber joining process is very constly and requires skilled manpower.

4. Tensile stress

- Optical fibers are more susceptible to buckling, bending and tensile stress than copper cables. This leades to restricted practice to use optical fiber technology to premises and floor backbones with a few interfaces to the copper cables.

5. Short links

- Eventhough optical fiber calbes are inexpensive, it is still not cost effective to replace every small conventional connector (e.g. between computers and peripherals), as the price of optoelectronic transducers are very high.

6. Fiber losses

- The amount of optical fiber available to the photodetector at the end of fiber length depends on various fiber losses such as scattering, dispersion, attenuation and reflection.

Applications of Optical Fiber Communicaitons

- Applications of optical fiber communications include telecommunications, data communications, video control and protection switching, sensors and power applications.

1. Telephone networks

- Optical waveguide has low attenuation, high transmission bandwidth compated to copper lines, therefore numbers of long haul co-axial trunks l;links between telephone exchanges are being replaced by optical fiber links.

2. Urban broadband service networks

- Optical waveguide provides much larger bandwidth than co-axial calbe, also the number of repeaters required is reduced considerably.

- Modern suburban communications involves videotext, videoconferencing videotelephony, switched broadband communication network. All these can

be supplied over a single fiber optic link. Fiber optic calbes is the solution to many of today's high speed, high bandwidth data communication problems and will continue to play a large role in future telecom and data-com networks.

Optical Fiber Waveguides

- In free space light ravels as its maximum possible speed i.e. 3 x 10^8 m/s or 186 x 10^3 miles/sec. When light travels through a material it exnibits certain behavior explaned by laws of reflection, refraction.

Electromagnetic Spectrum

- The radio waves and light are electromagnetic waves. The rate at which they alternate in

  polarity is called their frequency (f) measured in hertz (Hz). The speed of electromagnetic wave (c) in free space is approximately 3 x 10^8 m/sec. The distance travelled during each cycle is called as wavelength (λ)

$$\text{Wavelength } (\lambda) = \frac{\text{Speed of light}}{\text{Frequency}} = \frac{c}{f}$$

- In fiber optics, it is more convenient to use the wavelength of light instead of the frequency with light frequencies, wavlengfth is often stated in microns or nanometers.

  1 micron (μ) = 1

  Micrometre (1 x 10^{-6}) 1

  nano (n) = 10^{-9} metre

- Fiber optics uses visible and infrared light. Infrared light covers a fairly wide range of wavelengths and is generally used for all fiber optic communications. Visible light is normally used for very short range transmission using a plastic fiber.

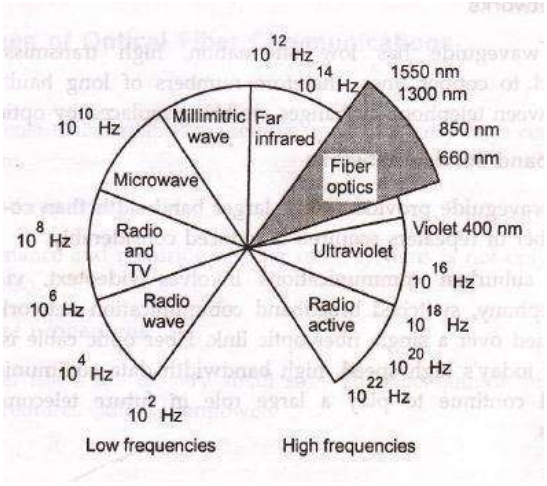


Fig. 1.6.1 shows electromagnetic frequency spectrum

Ray Transmission Theory

- Before studying how the light actually propagates through the fiber, laws governing the nature of light m ust be studied. These was called as **laws of optics (Ray theory).** There is conception that light always travels at the same speed. This fact is simply not true. The speed of light depends upon the material or medium through which it is moving. In free space light travels at its maximum possible speed i.e. 3 x 108 m/s or 186 x 103 miles/sec. When light travels through a material it exhibits certain behavior explained by laws of reflection, refraction.

Reflection

- The law of reflection states that, when a light ray is incident upon a reflective surface at some incident angle ϕ_1 from imaginary perpendicular normal, the ray will be reflected from the surface at some angle ϕ_2 from normal which is equal to the angle of incidence.

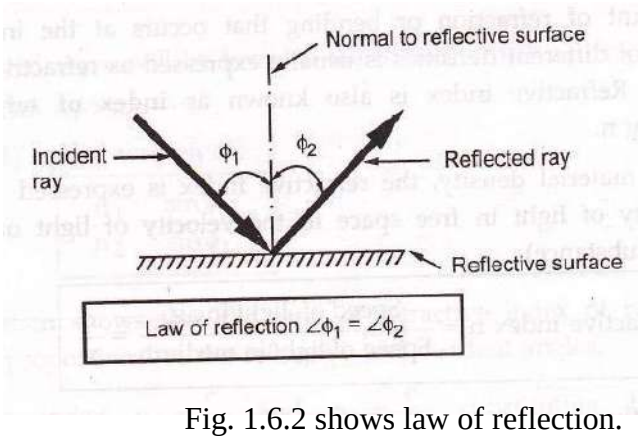


Fig. 1.6.2 shows law of reflection.

Refraction

- Refraction occurs when light ray passes from one medium to another i.e. the light ray changes its direction at interface. Refraction occurs whenever density of medium changes. E.g. refraction occurs at air and water interface, the straw in a glass of water will appear as it is bent.

The refraction can also observed at air and glass interface.

- When wave passes through less dense medium to more dense medium, the wave is refracted (bent) towards the normal. Fig. 1.6.3 shows the refraction phenomena.

- The refraction (bending) takes place because light travels at different speed in different mediums. The speed of light in free space is higher than in water or glass.

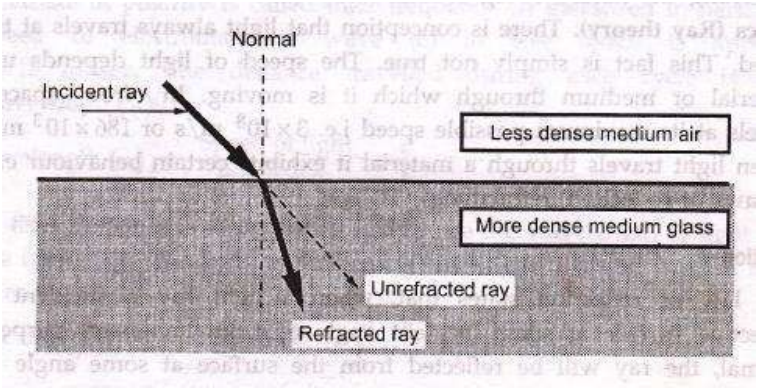


Fig.1.6.3 Refraction

Refractive Index

- The amount of refraction or bending that occurs at the interface of two materials of different densities is usually expressed as refractive index of two materials. Refractive index is also known as **index of refraction** and is denoted by n.

- Based on material density, the refractive index is expressed as the ratio of the velocity of light in free space to the velocity of light of the dielectric material (substance).

$$\text{Refractive index } n = \frac{\text{Speed of light in air}}{\text{Speed of light in medium}} = \frac{c}{v}$$

The refractive index for vacuum and air os 1.0 for water it is 1.3 and for glass refractive index is 1.5.

Snell's Law

- Snell's law states how light ray reacts when it meets the interface of two media having different indexes of refraction.

- Let the two medias have refractive indexes n1 and n_2 where $n_1 > n_2$.

$\phi 1$ and $\phi 2$ be the angles of incidence and angle of refraction respectively. Then according to Snell's law, a relationship exists between the refractive index of both materials given by

$$n1 \sin\phi 1 = n2 \sin\phi 2 \qquad \ldots (1.6.1)$$

- A refractive index model for Snell's law is shown in Fig. 1.6.4.

- The refracted wave will be towards the normal when $n_1 < n_2$ and will away from it when $n_1 > n_2$.

Equation (1.6.1) can be written as,

$$\frac{n_1}{n_2} = \frac{\sin \phi_2}{\sin \phi_1}$$

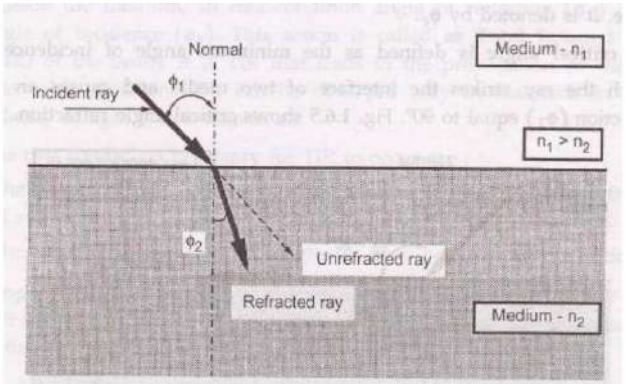


Fig 1.6.4 Refractive model for Snells Law

- This equation shows that the ratio of refractive index of two mediums is inversely proportional to the refractive and incident angles.

As refractive index $n_1 = \frac{c}{v_1}$ and $n_2 = \frac{c}{v_2}$ substituting these values in equation (1.6.2)

$$\frac{c/v_1}{c/v_2} = \frac{\sin \phi_2}{\sin \phi_1}$$

$$\frac{v_2}{v_1} = \frac{\sin \phi_2}{\sin \phi_1}$$

Critical Angle

- When the angle of incidence (ϕ_1) is profressively increased, there will be progressive increase of refractive angle (ϕ_2). At some condition (ϕ_1) the refractive angle (ϕ_2) becomes
  90o to the normal. When this happens the refracted light ray travels along the interface. The angle of incidence (ϕ_1) at the point at which the refractive angle (ϕ_1) becomes 90^o is called the critical angle. It is denoted by ϕ_c.
- The **critical angle** is defined as the minimum angle of incidence (ϕ_1) at which the ray strikes the interface of two media and causes an agnle of refraction (ϕ_2) equal to 90^o. Fig 1.6.5 shows critical angle refraction.

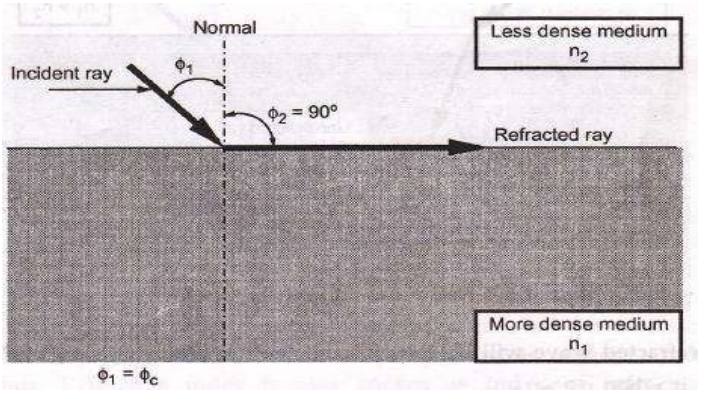


Fig.1.6.5 Critical Angle

Hence at critical angle $\phi_1 = \phi_C$ and $\phi_2 = 90^o$

Using Snell's law : $n_1 \sin \phi_1 = n_2 \sin \phi_2$

$$\sin \phi_c = \frac{n_2}{n_1} \sin 90°$$

$\because$ $\sin 90° = 1$

Therefore, $\sin \phi_c = \frac{n_2}{n_1}$

$$\text{Critical angle } \phi_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) \quad \ldots (1.6.3)$$

- The actual value of critical angle is dependent upon combination of materials present on each side of boundary.

Total Internal Refleciton (TIR)

- When the incident angle is increase dbeyond the critical angle, the light ray does not pass through the interface into the other medium. This gives the effect of mirror exist at the
  interface with no possibility of light escaping outside the medium. In this condition angle of reflection (ϕ_2) is equal to angle of incidence (ϕ_1). This action is called as **Total Internal Reflection (TIR)** of the beam. It is TIR that leads to the propagation of waves within fiber-cable medium. TIR can be observed only in materials in which the velocity of light is less than in air.

The refractive index of first medium must be greater than the refractive index of second one.

1. The angle of incidence must be greater than (or equal to) the critical angle.

Example 1.6.1 : A light ray is incident from medium-1 to medium-2. If the refractive indices of medium-1 and medium-2 are 1.5 and 1.36 respectively then determine the angle of refraction for an angle of incidence of 30^o.

Solution : Medium-1 $n1 = 1.5$

Medium-2 $n2 = 1.36$

Angle of incidence $\phi_1 = 30^o$.

Angle of incident $\phi_2 = ?$

$$\text{Snell's law : } n_1 \sin\phi_1 = n_2 \sin\phi_2$$

$$1.5 \sin 30° = 1.36 \sin\phi_2$$

$$\sin\phi_2 = \frac{1.5}{1.36} \sin 30°$$

$$\sin\phi_2 = 0.55147$$

$$\therefore \quad \phi_2 = 33.46°$$

Angle of refraction 33.46^o from normal. **… Ans.**

Example 1.6.2 : A light ray is incident from glass to air. Calculate the critical angle (ϕ_C).

Solution : Refractive index of glass $n_1 = 1.50$

Refrative indes of air $n_2 = 1.00$

$$\text{Snell's law : } n_1 \sin\phi_1 = n_2 \sin\phi_2$$

$$\sin\phi_1 = \frac{n_2}{n_1} \sin\phi_2$$

$$\therefore \quad \sin\phi_1 = \frac{n_2}{n_1} \sin 90°$$

Example 1.6.3 : Calculate the NA, acceptance angle and critical angle of the fiber having n_1 (Core refractive index) = 1.50 and refractive index of cladding = 1.45.

Soluiton : $n_1 = 1.50$, $n_2 = 1.45$

$$\Delta = \frac{(n_1 - n_2)}{(n_1)} = \frac{1.50 - 1.45}{1.50} = 0.033$$

$$\text{Numerical aperture, } NA = n_1 \sqrt{2\Delta}$$

$$NA = 1.50 \sqrt{2 \times 0.033}$$

$$NA = 0.387$$

$$\text{Acceptance angle } \phi_0 = \sin^{-1} NA \qquad \phi_0 = \sin^{-1} 0.387$$

$$\text{Critical angle } \phi_c = \sin^{-1} \frac{n_2}{n_1} \qquad \phi_c = \sin^{-1} \frac{1.45}{1.50}$$

$$\phi_0 = 22.78°$$

Optical Fiver as Waveguide

- An optical fiber is a cylindrical dielectric waveguide capable of conveying electromagnetic waves at optical frequencies. The electromagnetic energy is in the form of the light and propagates along the axis of the fiber. The structural of the fiver determines the transmission characteristics.
- The propagation of light along the waveguide is decided by the modes of the waveguides, here mode means path. Each mode has distict pattern of electric and magnetic field distributions along the fiber length. Only few modes can satisfy the homogeneous wave

  equation in the fiver also the boundary condition a waveguide surfaces. When there is only one path for light to follow then it is called as single mode propagation. When there is more than one path then it is called as multimode propagation.

Single fiber structure

- A single fiber structure is shown in Fig. 1.6.6. It consists of a solid dielectric cylinder with radius 'a'. This cylinder is called as **core** of fiber. The core is surrounded by dielectric, called **cladding**. The index of refraction of core (glass fiber) is slightly greater than the index of refraction of cladding.

If refractive index of core (glass fiver) = n_1

and refractive index of cladding = n_2

then $n_1 > n_2$.

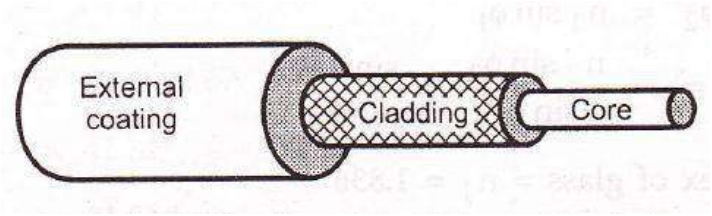


Fig.1.6.6. Single optical Fibre Structure

Propagation in Optical Fiber

- To understand the general nature of light wave propagation in optical fiber. We first consider the construction of optical fiber. The innermost is the glass core of very thin diameter with a slight lower refractive index n2. The light wave can propagate along such a optical fiber. A single mode propagation is illustrated in Fig. 1.6.7 along with standard size of fiber.

Single mode fibers are capable of carrying only one signal of a specific wavelength.

- In multimode propagation the light propagates along the fiber in zigzag fashion, provided it can undergo total internal reflection (TIR) at the core cladding boundaries.
- Total internal reflection at the fiber wall can occur only if two conditions are satisfied.

Condition 1:

The index of refraction of glass fiber must be slightly greater than the index of refraction of material surrounding the fiber (cladding).

If refractive index of glass fiber = n_1

and refractive index of cladding = n_2

then $n_1 > n_2$.

Condition 2 :

The angle of incidence (ϕ_1) of light ray must be greater than critical angle (ϕ_c).

- A light beam is focused at one end of cable. The light enters the fibers at different angles.
  Fig. 1.6.8 shows the conditions exist at the launching end of optic fiber. The light source is surrounded by air and the refractive index of air is $n_0 = 1$. Let the incident ray makes an angle ϕ_0 with fiber axis. The ray enters into glass fiber at point P making refracted angle ϕ_1 to the fiber axis, the ray is then propagated diagonally down the core and reflect from the core wall at point Q. When the light ray reflects off the inner surface, the angle of incidence is equal to the angle of reflection, which is greater than critical angle.
- In order for a ray of light to propagate down the cable, it must strike the core cladding interface at an angle that is greater than critical angle (ϕ_c).

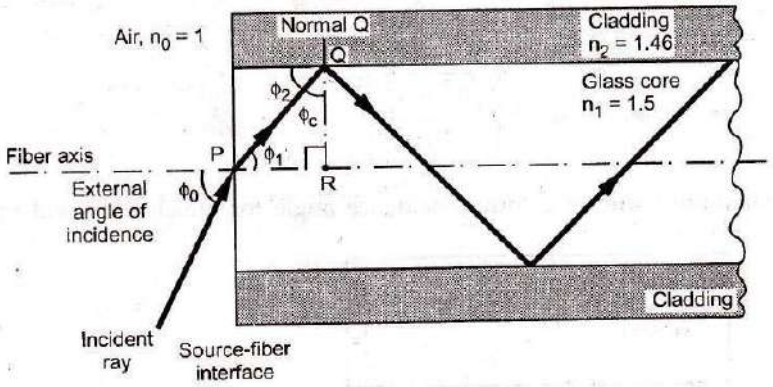


Fig. 1.6.8 Ray propagation by TIR

Acceptance Angle

Applying Snell's law to external incidence angle.

$$n_0 \sin \phi_0 = n_1 \sin \phi_1$$

But $\phi_1 = (90 - \phi_c)$

$$\sin \phi_1 = \text{sing } (90 - \phi_c) = \cos \phi_c$$

Substituting $\sin \phi_1$ in above equation.

$$n_0 \sin \phi_0 = n_1 \cos \phi_c$$

$$\sin \phi_c = \frac{n_i}{n_0} \cos \phi_c$$

Applying Pythagorean theorem to ΔPQR.

$$\cos \phi = \frac{\sqrt{n_1^2 - n_2^2}}{}$$

The maximum value of external incidence angle for which light will propagate in the fiber.

$$\phi_{0(max)} = \sin^{-1}\left[\frac{\sqrt{n_1^2 - n_2^2}}{n_0}\right]$$

When the light rays enters the fivers from an air medium $n_0 = 1$. Then above equation reduces to,

$$\phi_{0(max)} = \sin^{-1}\left(\sqrt{n_1^2 - n_2^2}\right)$$

The angle ϕ_0 is called as **acceptance angle** and $\phi_{0(max)}$ defines the maximum angle in which the light ray may incident on fiber to propagate down the fiber.

Acceptance Cone

- Rotating the acceptance angle $\phi_{0(max)}$ around the fiber axis, a cone shaped pattern is obtained, it is called as **acceptance cone** of the fiber input. Fig 1.6.10 shows formation of acceptance cone of a fiber cable.

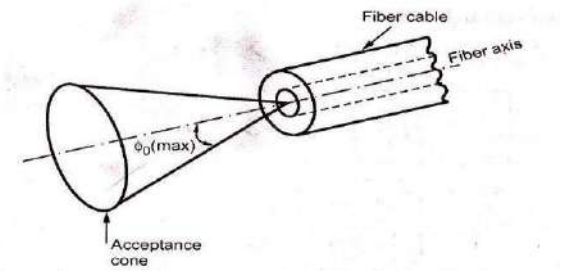


FIG: 1.6.10 shows formation of acceptance cone of a fiber cable.

- The Cone of acceptance is the angle within which the light is accepted into the core and is able to travel along the fiber. The launching of light wave becomes easier for large acceptance come.
- The angle is measured from the axis of the positive cone so the total angle of convergence is actually twice the stated value.

Numerical Aperture (NA)

- The **numerical aperture** (NA) of a fiber is a figure of merit which represents its light gathering capability. Larger the numerical aperture, the greater the

amount of light accepted by fiber. The acceptance angle also determines how much light is able to be enter the fiber and hence there is relation between the numerical aperture and the cone of acceptance.

Numerical aperture (NA) = $\sin \phi_{0(max)}$

$$NA = \frac{\sqrt{n_1^2 - n_2^2}}{n_0}$$

For air $n_o = 1$

$\therefore$ $$NA = \sqrt{n_1^2 - n_2^2}$$

$$\boxed{NA = \sqrt{n_{core}^2 - n_{cladding}^2}} \quad \ldots (1.6.4)$$

Hence acceptance angle = $\sin^{-1}$ NA

By the formula of NA note that the numerical aperture is effectively dependent only on refractive indices of core and cladding material. NA is not a function of fiber dimension.

- The index difference (Δ) and the numerical aperture (NA) are related to the core and cladding indices:

$$\Delta = \frac{(n_1 - n_2)}{n_1} \qquad \boxed{\Delta = \frac{NA^2}{2n_1^2}}$$

Also $$NA = \sqrt{n_1^2 - n_2^2}$$

Example 1.6.5 : Calculate the numerical aperture and acceptance angle for a fiber cable of which $n_{core} = 1.5$ and $n_{cladding} = 1.48$. The launching takes place from air.

Solution $NA = (n_1^2 - n_2^2)^{1/2}$

$$NA = \sqrt{n_{core}^2 - n_{cladding}^2}$$

$$\boxed{NA = n_1 (2\Delta)^{1/2}}$$

$$NA = \sqrt{1.5^2 - 1.48^2}$$

NA = **0.244** **…Ans.**

Acceptance angle =

$$\boxed{\sin^{-1} \sqrt{n_{core}^2 - n_{cladding}^2} = \sin^{-1} NA}$$

Acceptance angle = $\sin^{-1}$ 0.244

$\phi_0 = \mathbf{14.12^o}$ **…Ans.**

Types of Rays

- If the rays are launched within core of acceptance can be successfully propagated along the fiber. But the exact path of the ray is determined by the position and angle of ray at which it strikes the core.
  There exists three different types of rays.
  i) Skew rays ii) Meridional rays iii) Axial rays.

- **The skew rays** does not pass through the center, as show in Fig. 1.6.11 (a). The skew rays reflects off from the core cladding boundaries and again bounces around the outside of the core. It takes somewhat similar shape of spiral of helical path.

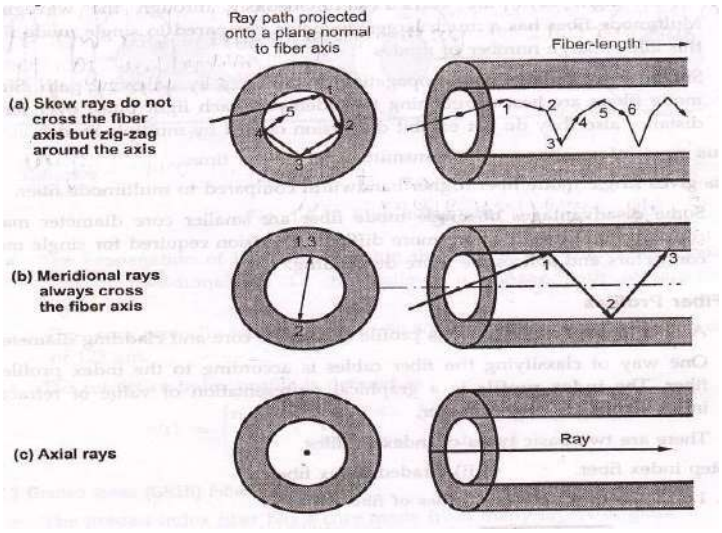


Fig:1.6.11 Different Ray Propagation

- The **meridional** ray enters the core and passes through its axis. When the core surface is parallel, it will always be reflected to pass through the enter. The meridional ray is shown in fig. 1.6.11 (b).
- The **axial ray** travels along the axis of the fiber and stays at the axis all the time. It is shown in fig. 1.6.11 (c).

Modes of Fiber

- Fiber cables cal also be classified as per their mode. Light rays propagate as an electromagnetic wave along the fiber. The two components, the electric field and the magnetic field form patterns across the fiber. These patterns are called **modes** of transmission. The **mode** of a fiber refers to the number of paths for the light rays within the cable. According to modes optic fibers can be classified into two types.
  i) Single mode fiber ii) Multimode fiber.

- Multimode fiber was the first fiber type to be manufactured and commercialized. The term multimode simply refers to the fact that numerous modes (light rays) are carried simultaneously through the waveguide. Multimode fiber has a much larger diameter, compared to single mode fiber, this allows large number of modes.
- Single mode fiber allows propagation to light ray by only one path. Single mode fibers are best at retaining the fidelity of each light pulse over longer distance also they do not exhibit dispersion caused by multiple modes.

Thus more information can be transmitted per unit of time.

This gives single mode fiber higher bandwidth compared to multimode fiber.

- Some disadvantages of single mode fiber are smaller core diameter makes coupling light into the core more difficult. Precision required for single mode connectors and splices are more demanding.

Fiber Profiles

- A fiber is characterized by its profile and by its core and cladding diameters.
- One way of classifying the fiber cables is according to the index profile at fiber. The **index profile** is a graphical representation of value of refractive index across the core diameter.
- There are two basic types of index profiles.

i) Step index fiber. ii) Graded index fiber.

Fig. 1.6.12 shows the index profiles of fibers.

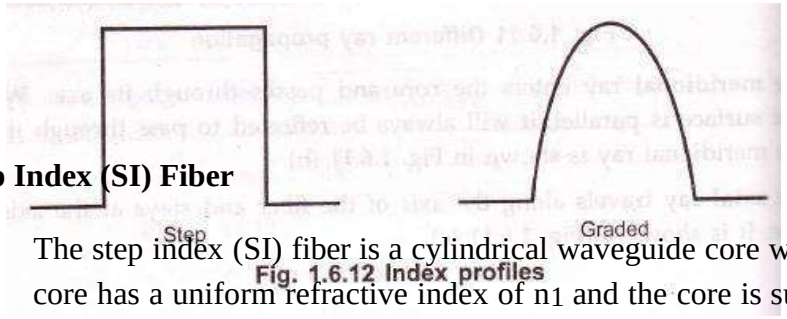


Fig. 1.6.12 Index profiles

Step Index (SI) Fiber

- The step index (SI) fiber is a cylindrical waveguide core with central or inner core has a uniform refractive index of n_1 and the core is surrounded by outer cladding with uniform refractive index of n_2. The cladding refractive index (n_2) is less than the core refractive index (n_1). But there is an abrupt change in the refractive index at the core cladding interface. Refractive index profile of step indexed optical fiber is shown in Fig. 1.6.13. The refractive index is plotted on horizontal axis and radial distance from the core is plotted on vertical axis.

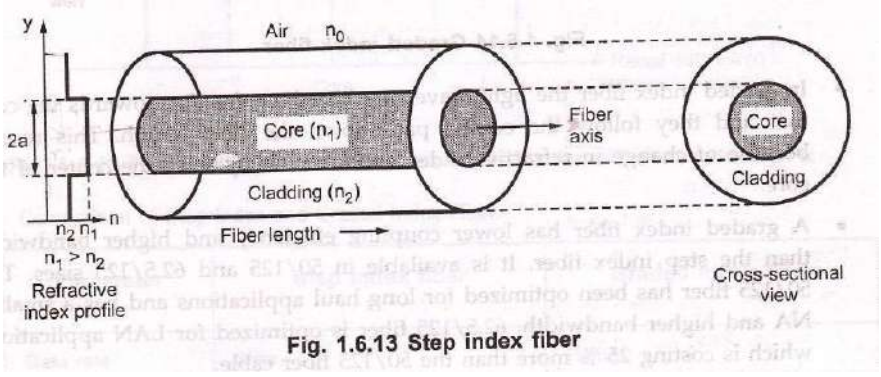


Fig. 1.6.13 Step index fiber

- The propagation of light wave within the core of step index fiber takes the path of meridional ray i.e. ray follows a zig-zag path of straight line segments.
  The core typically has diameter of 50-80 μm and the cladding has a diameter of 125 μm.
- The refractive index profile is defined as –

$$n(r) = \begin{cases} n_1 & \text{when } r < a \text{ (core)} \\ n_2 & \text{when } r \geq a \text{ (cladding)} \end{cases}$$

Graded Index (GRIN) Fiber

- The graded index fiber has a core made from many layers of glass.
- In the **graded index (GRIN)** fiber the refractive index is not uniform within the core, it is highest at the center and decreases smoothly and continuously with distance towards the cladding. The refractive index profile across the core takes the parabolic nature. Fig. 1.6.14 shows refractive index profile of graded index fiber.

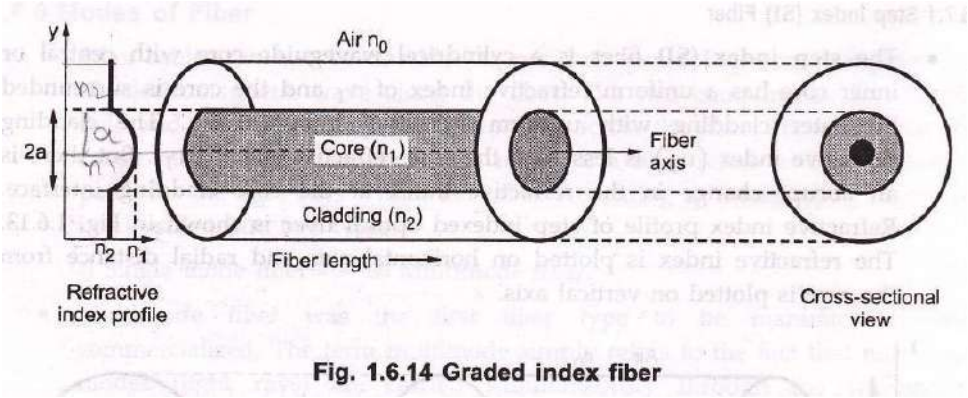


Fig. 1.6.14 Graded index fiber

- In graded index fiber the light waves are bent by refraction towards the core axis and they follow the curved path down the fiber length. This results because of change in refractive index as moved away from the center of the core.
- A graded index fiber has lower coupling efficiency and higher bandwidth than the step index fiber. It is available in 50/125 and 62.5/125 sizes. The 50/125 fiber has been optimized for long haul applications and has a smaller NA and higher bandwidth. 62.5/125 fiber is optimized for LAN applications which is costing 25% more than the 50/125 fiber cable.
- The refractive index variation in the core is giver by relationship

$$n(r) = \begin{cases} n_1\left(1 - 2\Delta\left(\frac{r}{a}\right)^{\alpha}\right) & \text{when } r < a \text{ (core)} \\ n_1(1-2\Delta)^{\frac{1}{2}} \approx n_2 & \text{when } r \geq a \text{ (cladding)} \end{cases}$$

where,

r = Radial distance from fiber axis

a = Core radius

n_1= Refractive index of core

n_2 = Refractive index of cladding

α = Shape of index profile.

- Profile parameter α determines the characteristic refractive index profile of fiber core. The range of refractive index as variation of α is shown in Fig. 1.6.1

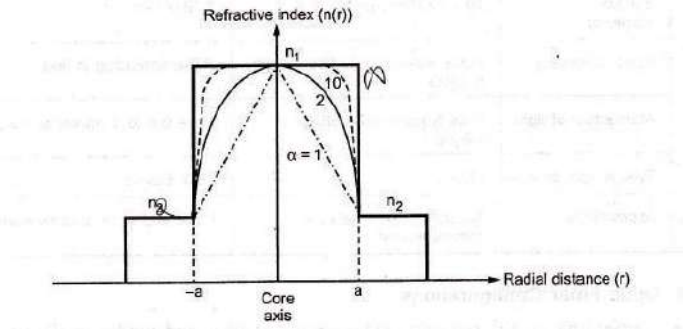


Fig. 1.6.15 Possible fiber refractive index profiles for different values of α

Comparison of Step Index and Graded Index Fiber

Sr. No.	Parameter	Step index fiber	Graded index fiber
1.	Data rate	Slow.	Higher
2.	Coupling efficiency	Coupling efficiency with fiber is higher.	Lower coupling efficiency.
3.	Ray path	By total internal reflection.	Light travelled oscillatory fashion.
4.	Index variation		
5.	Numerical aperture	NA remains same.	Changes continuously distance from fiber axis.
6.	Material used	Normally plastic or glass is preferred.	Only glass is preferred.
7.	Bandwidth efficiency	10 – 20 MHz/km	1 GHz/km
8.	Pulse spreading	Pulse spreading by fiber length is more.	Pulse spreading is less
9.	Attenuation	Less typically 0.34	More 0.6 to 1 dB/km at 1.3

	of light	dB/km at 1.3 μm.	μm.
10	Typical light source	LED.	LED, Lasers.
	Applications	Subscriber local network communicat ion.	networks.

Optic Fiber Configurations

- Depending on the refractive index profile of fiber and modes of fiber there exist three types of optical fiber configurations. These optic-fiber configurations are -

  i) Single mode step index fiber.
  ii) Multimode step index fiber.
  iii) Multimode graded index fiber.

Single mode Step index Fiber

- In single mode step index fiber has a central core that is sufficiently small so that there is essentially only one path for light ray through the cable. The light ray is propagated in the fiber through reflection. Typical core sizes are 2 to 15 μm. Single mode fiber is also known as fundamental or monomode fiber.

Fig. 1.6.16 shows single mode fiber.

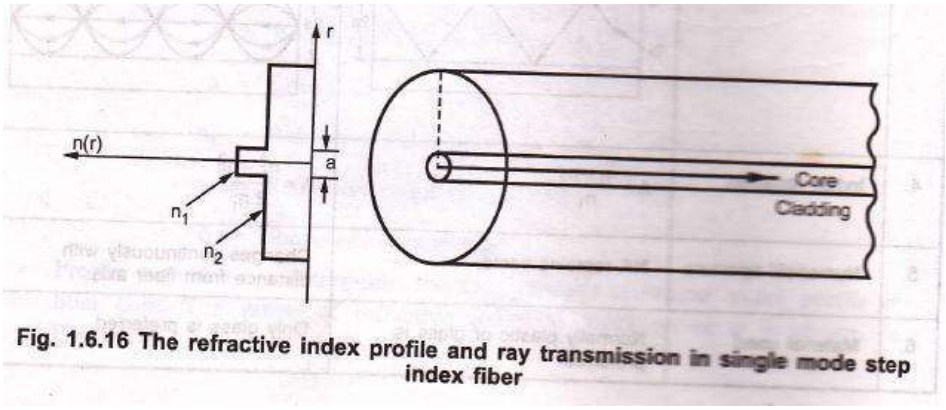


Fig. 1.6.16 The refractive index profile and ray transmission in single mode step index fiber

- Single mode fiber will permit only one mode to propagate and does not suffer from mode delay differences. These are primarily developed for the 1300 nm window but they can be also be used effectively with time division multiplex (TDM) and wavelength division multiplex (WDM) systems operating in 1550 nm wavelength region.
- The core fiber of a single mode fiber is very narrow compared to the wavelength of light being used. Therefore, only a single path exists through the cable core through which light can travel. Usually, 20 percent of the light in a single mode cable actually

travels down the cladding and the effective diameter of the cable is a blend of single mode core and degree to which the cladding carries light. This is referred to as the 'mode field diameter', which is larger than physical diameter of the core depending on the refractive indices of the core and cladding.

- The disadvantage of this type of cable is that because of extremely small size interconnection of cables and interfacing with source is difficult. Another disadvantage of single mode fibers is that as the refractive index of glass decreases with optical wavelength, the light velocity will also be wavelength dependent. Thus the light from an optical transmitter will have definite spectral width.

Multimode step Index Fiber

- **Multimode step index fiber** is more widely used type. It is easy to manufacture. Its core diameter is 50 to 1000 μm i.e. large aperture and allows more light to enter the cable. The light rays are propagated down the core in zig-zag manner. There are many many paths that a light ray may follow during the propagation.
- The light ray is propagated using the principle of total internal reflection (TIR). Since the core index of refraction is higher than the cladding index of refraction, the light enters at less than critical angle is guided along the fiber.

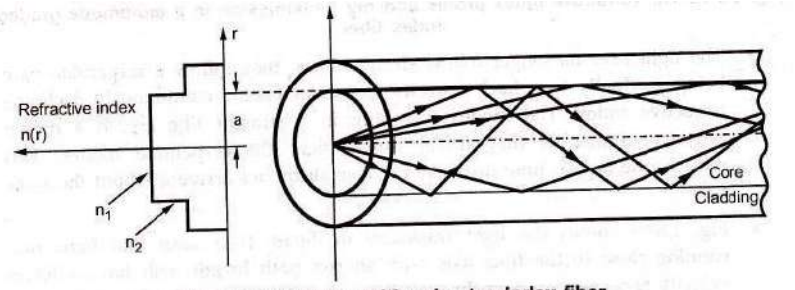


Fig. 1.6.17 TIR in multimode step index fiber

- Light rays passing through the fiber are continuously reflected off the glass cladding towards the centre of the core at different angles and lengths, limiting overall bandwidth.
- The disadvantage of multimode step index fibers is that the different optical lengths caused by various angles at which light is propagated relative to the core, causes the

transmission bandwidth to be fairly small. Because of these limitations, multimode step index fiber is typically only used in applications requiring distances of less than 1 km.

Multimode Graded Index Fiber

- The core size of **multimode graded index fiber** cable is varying from 50 to 100 µm range. The light ray is propagated through the refraction. The light ray enters the fiber at

  many different angles. As the light propagates across the core toward the center it is intersecting a less dense to more dense medium. Therefore the light rays are being constantly being refracted and ray is bending continuously. This cable is mostly used for long distance communication.

Fig 1.6.18 shows multimode graded index fiber.

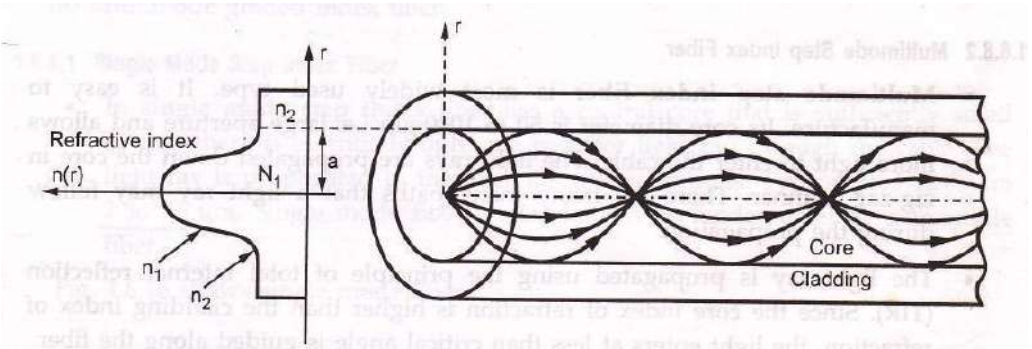


Fig. 1.6.18 The refractive index profile and ray transmission in a multimode graded index fiber

- The light rays no longer follow straight lines, they follow a serpentine path being gradually bent back towards the center by the continuously declining refractive index. The modes travelling in a straight line are in a higher refractive index so they travel slower than the serpentine modes. This reduces the arrival time disparity because all modes arrive at about the same time.
- Fig 1.6.19 shows the light trajectory in detail. It is seen that light rays running close to the fiber axis with shorter path length, will have a lower velocity because they pass through a region with a high refractive index.

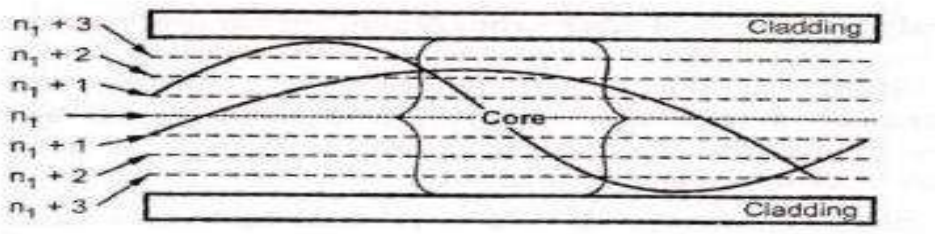


Fig. 1.6.19 Light trajectories in a graded index fiber

Rays on core edges offers reduced refractive index, hence travel more faster than axial rays and cause the light components to take same amount of time to travel the length of fiber, thus minimizing dispersion losses. Each path at a different angle is termed as

'transmission mode' and the NA of graded index fiber is defined as the maximum value of acceptance angle at the fiber axis.

- Typical attenuation coefficients of graded index fibers at 850 nm are 2.5 to 3 dB/km, while at 1300 nm they are 1.0 to 1.5 dB/km.
- The main advantages of graded index fiber are:

1. Reduced refractive index at the centre of core.
2. Comparatively cheap to produce.

Standard fibers

Sr. No.	Fiber type	Cladding Diameter (μm)	Core diameter (μm)		Applications
1.	Single mode (8/125)	125	8	0.1% to 0.2%	1. Long distance 2. High data rate
2.	Multimode (50/125)	125	50	1% to 2%	1. Short distance 2. Low data rate
3.	Multimode (62.5/125)	125	62.5	1% to 2%	LAN
4.	Multimode (100/140)	140	100	1% to 2%	LAN

Mode Theory for Cylindrical Waveguide

- To analyze the optical fiber propagation mechanism within a fiber, Maxwell equations are to solve subject to the cylindrical boundary conditions at core-cladding interface. The core-cladding boundary conditions lead to coupling of electric and magnetic field components resulting in hybrid modes. Hence the analysis of optical waveguide is more complex than metallic hollow waveguide analysis.
- Depending on the large E-field, the hybrid modes are HE or EH modes. The two lowest order does are HE_{11} and TE_{01}.

Overview of Modes

- The order states the number of field zeros across the guide. The electric fields are not completely confined within the core i.e. they do not go to zero at core-cladding interface and extends into the cladding. The low order mode confines the electric field near the axis of the fiber core and there is less penetration into the cladding. While the high order mode distribute the field towards the edge of the core fiber and penetrations into the cladding. Therefore cladding modes also appear resulting in power loss.
- In leaky modes the fields are confined partially in the fiber core attenuated as they propagate along the fiber length due to radiation and tunnel effect.

- Therefore in order to mode remain guided, the propagation factor β must satisfy the condition

$$n_2k < \beta < n_1k$$

where, n_1 = Refractive index of fiber core

n_2 = Refractive index of cladding

k = Propagation constant = $2\pi / \lambda$

- The cladding is used to prevent scattering loss that results from core material discontinuities. Cladding also improves the mechanical strength of fiber core and reduces surface contamination. Plastic cladding is commonly used. Materials used for fabrication of optical fibers are silicon dioxide (SiO_2), boric oxide-silica.

Summary of Key Modal Concepts

- Normalized frequency variable, V is defined as

$$V = \frac{2\pi a(n_1^2 - n_2^2)^{1/2}}{\lambda} \quad \ldots (1.7.1)$$

where, a = Core radius

λ = Free space wavelength

$$V = \frac{2\pi a}{\lambda} NA$$

Since $(n_1^2 - n_2^2)^{1/2}$ = NA … (1.7.2)

- The total number of modes in a multimode fiber is given by

$$M = \frac{1}{2}\left(\frac{2\pi a}{\lambda}\right)^2 (n_1^2 - n_2^2)$$

$$M = \frac{1}{2}\left[\frac{2\pi a}{\lambda} NA\right]^2 = \frac{[V]^2}{2}$$

$$M = \frac{1}{2}\left[\frac{\pi d}{\lambda}.NA\right]^2$$

Example 1.7.1 : Calculate the number of modes of an optical fiber having diameter of 50 μm, n_1 = 1.48, n_2 = 1.46 and λ = 0.82 μm.

Solution : d = 50 μm

n_1 = 1.48

n_2 = 1.46

$\lambda = 0.82$ µm

$$NA = (n_1^2 - n_2^2)^{1/2}$$

$$NA = (1.48^2 - 1.46^2)^{1/2}$$

$$NA = 0.243$$

Number of modes are given by,

$$M = \frac{1}{2}\left[\frac{\pi d}{\lambda}.NA\right]^2$$

$$M = \frac{1}{2}\left[\frac{\pi (50 \text{ X } 10^{-6})}{0.82 \text{ X } 10^{-6}} \text{X } 0.243\right]^2$$

$$M = 1083 \quad \text{…Ans.}$$

Example 1.7.2 : A fiber has normalized frequency V = 26.6 and the operating wavelength is 1300nm. If the radius of the fiber core is 25 µm. Compute the numerical aperture.

Solution : $V = 26.6$

$\lambda = 1300 \text{ nm} = 1300 \text{ X } 10^{-9} \text{ m}$

$a = 25 \text{ µm} = 25 \text{ X } 10^{-6} \text{ m}$

$$\boxed{V = \frac{2\pi a}{\lambda} NA}$$

$$NA = V.\frac{\lambda}{2\pi a}$$

$$NA = 26.6 \frac{1300 \text{ X } 10^{-9}}{2\pi \text{ X } 25 \text{ X } 10^{-6}}$$

$$NA = \mathbf{0.220} \quad \textbf{… Ans.}$$

Example 1.7.3 : A multimode step index fiber with a core diameter of 80 µm and a relative index difference of 1.5 % is operating at a wavelength of 0.85 µm. If the core refractive index is 1.48, estimate the normalized frequency for the fiber and number of guided modes.

[July/Aug.-2008, 6 Marks]

Solution : Given : MM step index fiber, 2 a = 80 µm

∴ Core radians a = 40 µm

Relative index difference, = 1.5% = 0.015

Wavelength, λ = 0.85µm

Core refractive index, n1 = 1.48

Normalized frequency, V = ?

Number of modes, M = ?

Numerical aperture

$$NA = n_1 (2\Delta)^{1/2}$$

$= 1.48 (2 \text{ X } 0.015)^{1/2}$

$= 0.2563$

Normalized frequency is given by,

$$V = \frac{2\pi a}{\lambda} NA$$

$$V = \frac{2\pi \text{ X } 40}{0.85} \text{ x } 0.2563$$

V = 75.78 **… Ans.**

Number of modes is given by,

$$M = \frac{V^2}{2}$$

$$M = \frac{(75.78)^2}{2} = \mathbf{2871.50}$$ **Ans**

Example 1.7.4 : A step index multimode fiber with a numerical aperture of a 0.20 supports approximately 1000 modes at an 850 nm wavelength.

i) What is the diameter of its core?
ii) How many modes does the fiber support at 1320 nm?
iii) How many modes does the fiber support at 1550 nm? **[Jan./Feb.-2007, 10 Marks]**

Solution : i) Number of modes is given by,

$$M = \frac{1}{2}\left[\frac{\pi\, a}{\lambda}.NA\right]^2$$

a = **60.49 µm … Ans.**

ii)

$$M = \frac{1}{2}\left[\frac{\pi \text{ X } 60.49 \text{ X } 10^{-6}}{1320 \text{ X } 10^{-9}} \text{ X } 0.20\right]^2$$

$M = (14.39)^2 =$ **207.07**

iii)

$$M = \frac{1}{2}\left[\frac{\pi \text{ X } 6.49 \text{ X } 10^{-6}}{1320 \text{ X } 10^{-9}} \text{ X } 0.20\right]^2$$

M = **300.63**

Wave Propagation

Maxwell's Equations

Maxwell's equation for non-conducting medium:

∇ X E = - ∂B /

∇ X H = - ∂D /

∇. D = 0

∇. B 0

where,

E and H are electric and magnetic field vectors.

- The relation between flux densities and filed vectors:

  $D = \varepsilon_0 E + P$

  $B = \mu_0 H + M$

where,

ε_0 is vacuum permittivity.

μ_0 is vacuum permeability.

P is induced electric polarization.

M is induced magnetic polarization (M = 0, for non-magnetic silica glass)

- P and E are related by:

$$P(r,t) = \varepsilon_0 \int_{-\infty}^{\infty} \chi(r, t - t') E(r,t') dt'$$

Where,

X is linear susceptibility.

- Wave equation:

$$\nabla \times \nabla \times E = \frac{-1}{c^2} \frac{\partial^2 E}{\partial E^2} - \mu_0 \frac{\partial^2 P}{\partial t^2}$$

Fourier transform of E (r, t)

$$\tilde{E}(r,\omega) = \int_{-\infty}^{\infty} E(r,t) e^{i\omega t} dt$$

$$\nabla \times \nabla \times \tilde{E} = -\varepsilon(r,\omega) \frac{\omega^2}{c^2} \tilde{E}$$

where,

$$\varepsilon = \left(n + \frac{i\alpha c}{2\omega}\right)^2$$

n is refractive index.

α is absorption coefficient.

$$n = \sqrt{(1 + R_e \chi)}$$

$$\alpha = \left(\frac{\omega}{nc}\right) I_m \chi$$

- Both n and α are frequency dependent. The frequency dependence of n is called as chromatic dispersion or material dispersion.
- For step index fiber,

$$\nabla \times \nabla \times \tilde{E} = \nabla(\nabla . \tilde{E}) - \nabla^2 . \tilde{E} = -\nabla^2 \tilde{E}$$

Fiber Modes

Optical mode : An optical mode is a specific solution of the wave equation that satisfies boundary conditions. There are three types of fiber modes.

a) Guided modes
b) Leaky modes
c) Radiation modes

- For fiber optic communication system guided mode is sued for signal transmission.

Considering a step index fiber with core radius 'a'.

The cylindrical co-ordinates ρ, ϕ and can be used to represent boundary conditions.

$$\frac{\partial^2 E_z}{\partial \rho^2} + \frac{1}{\rho} \cdot \frac{\partial E_z}{\partial \rho} + \frac{1}{\rho^2} \cdot \frac{\partial^2 E_z}{\partial \phi^2} + \frac{\partial^2 E_z}{\partial z^2} + n^2 k_0^2 E_z = 0$$

- The refractive index 'n' has values

$$n = \begin{cases} n_1; & \rho \leq a \\ n_2; & \rho > a \end{cases}$$

- The general solutions for boundary condition of optical field under guided mode is

infinite at $\rho = 0$ and decay to zero at $\rho = \infty$. Using Maxwell's equation in the core region.

$$E_\rho = \frac{i}{p^2}\left(\beta \frac{\partial E_z}{\partial \rho} + \mu_0 \frac{\omega}{\rho} \cdot \frac{\partial H_z}{\partial \phi}\right)$$

- The **cut-off condition** is defined as –

$$V = k_0 a \sqrt{(n_1^2 - n_2^2)}$$

$$V = \left(\frac{2\pi}{\lambda}\right) a\, n_1 \sqrt{2\Delta}$$

It is also called as **normalized frequency**.

Graded Index Fiber Structure

- The Refractive index of graded index fiber decreases continuously towards its radius from the fiber axis and that for cladding is constant.
- The refractive index variation in the core is usually designed by using power law relationship.

$$n(r) = \begin{cases} n_1 \left[1 - 2\Delta \left(\frac{r}{a}\right)^{\propto}\right]^{\frac{1}{2}}, & when\ 0 \leq r \leq a \\ n_1 (1 - 2\Delta)^{\frac{1}{2}} \approx n_1 (1 - \Delta) = n_2, & when\ r \geq a \end{cases}$$

Where, r = Radial distance from fiber axis

a = Core radius

n_1 = Refractive index core

n_2 Refractive index of cladding and

α = The shape of the index profile

- For graded index fiber, the index difference is given by,
  - In graded index fiber the incident light will propagate when local numerical aperture at distance r from axis, NA is axial numerical aperture NA(0). The local numerical aperture is given as,

$$NA(r) = \begin{cases} [n^2(r) - n_2^2]^{\frac{1}{2}} \approx NA(0)\sqrt{1-\left(\frac{r}{a}\right)^{\alpha}}, & for\ r \le a \\ 0, & for\ r > a \end{cases}$$

  - The axial numerical aperture NA(0) is given as,

$$NA(0) = [n^2(0) - n_2^2]^{1/2}$$

$$NA(0) = [n_1^2 - n_2^2]^{1/2}$$

$$NA(0) = n_1\sqrt{2\Delta} \approx n_1(2\Delta)^{1/2}$$

Hence Na for graded index decreases to zero as it moves from fiber axis to core-cladding boundary.

- The variation of NA for different values of α is shown in Fig. 1.7.1.

- The number of modes for graded index fiber in given as,

$$M = \frac{\alpha}{\alpha+2} a^2 k^2 n_1^2 \Delta$$ …

Single Mode Fibers

- Propagation in single mode fiber is advantageous because signal dispersion due to delay differences amongst various modes in multimode is avoided. Multimode step index fibers cannot be used for single mode propagation due to difficulties in maintaining single mode operation. Therefore for the

transmission of single mode the fiber is designed to allow propagation in one mode only, while all other modes are attenuated by leakage or absorption.

- For single mode operation, only fundamental LP_{01} mode many exist. The single mode propagation of LP_{01} mode in step index fibers is possible over the range.

$$0 \leq V < 2405$$

- The normalized frequency for the fiber can be adjusted within the range by reducing core radius and refractive index difference < 1%. In order to obtain single mode operation with maximum V number (2.4), the single mode fiber must have smaller core diameter than the equivalent multimode step index fiber. But smaller core diameter has problem of launching light into the fiber, jointing fibers and reduced relative index difference.

- Graded index fibers can also be sued for single mode operation with some special fiber design. The cut-off value of normalized frequency Vc in single mode operation for a graded index fiber is given by,

$$V_c = 2.405\left(1+\frac{2}{\alpha}\right)^{\frac{1}{2}}$$

Example 1.8.1 : A multimode step index optical fiber with relative refractive index difference 1.5% and core refractive index 1.48 is to be used for single mode operation. If the operating wavelength is 0.85μm calculate the maximum core diameter.

Solution : Given :

$n_1 = 1.48$
$\Delta = 1.5\ \% = 0.015$
$\lambda = 0.85\ \mu m = 0.85 \times 10^{-6}\ m$

Maximum V value for a fiber which gives single mode operations is 2.4.

Normalized frequency (V number) and core diameter is related by expression,

$$V = \frac{2\pi}{\lambda} a\ (NA)$$

$$\boxed{V = \frac{2\pi}{\lambda} a\ n_1 (2\Delta)^{\frac{1}{2}}}$$

$$a = \frac{V\lambda}{2\pi\ n_1 (2\Delta)^{\frac{1}{2}}}$$

$$a = \frac{2.4 \times (0.85 \times 10^{-6})}{2\pi \times (1.48) \times (0.03)^{\frac{1}{2}}}$$

a = **1.3 μm** **… Ans.**

Maximum core diameter for single mode operation is 2.6 μm.

Example 1.8.2 : A GRIN fiber with parabolic refractive index profile core has a refractive index at the core axis of 1.5 and relative index difference at 1%. Calculate maximum possible core diameter that allows single mode operations at λ = 1.3 μm.

Solution : Given :

$$n_1 = 1.5$$

$$\Delta = 1\ \% = 0.01$$

$$\lambda = 1.3\ \mu m = 1.3 \times 10^{-6} \text{m}$$

for a GRIN

Maximum value of normalized frequency for single mode operation is given by,

$$V = 2.4\left(1 + \frac{2}{\alpha}\right)^{\frac{1}{2}}$$

Maximum core radius is given by expression,

$$a = \frac{V\lambda}{2\pi\, n_1 (2\Delta)^{\frac{1}{2}}}$$

$$a = \frac{24\sqrt{2} \times 1.3 \times 10^{-6}}{2\pi \times 1.5 \times (0.02)^{\frac{1}{2}}}$$

a = 3.3 μm **… Ans.**

∴ Maximum core diameter which allows single mode operation is 6.6 μm.

Cut-off Wavelength

- One important transmission parameter for single mode fiber us cut-off wavelength for the first higher order mode as it distinguishes the single mode and multim0de regions.
- The effective cut-off wavelength λ_C is defined as the largest wavelength at which higher order $(L_{P_{11}})$ mode power relative to the fundamental mode $(L_{P_{01}})$ power is reduced to 0.1 dB. The range of cut-off wavelength recommended to avoid modal noise and dispersion problems is : 1100 to 1280 nm (1.1 to 1.28μm) for single mode fiber at 1.3 μm.
- The cut-off wavelength λ_C can be computed from expression of normalized frequency.

$$V = \frac{2\pi}{\lambda} a\, n_1 (2\Delta)^{\frac{1}{2}} \Rightarrow \lambda = \frac{2\pi a n_1}{V} (2\Delta)^{\frac{1}{2}} \quad \text{…. (1.8.1)}$$

∴ .... (1.8.2)

$$\lambda = \frac{2\pi a n_1}{V}(2\Delta)^{\frac{1}{2}}$$

where,

V_C is cut-off normalized frequency.

- λ_C is the wavelength above which a particular fiber becomes single moded. For same fiber dividing λ_C by λ we get the relation as:

$$\frac{\lambda_c}{\lambda} = \frac{V}{V_C}$$

$$\lambda = \frac{V\lambda}{V_C} \quad \ldots (1.8.3)$$

But for step index fiver $V_C = 2.405$ then

$$\lambda_c = \frac{V\lambda}{2.405}$$

Example 1.8.3 : Estimate cut-off wavelength for step index fiber in single mode operation. The core refractive index is 1.46 and core radius is 4.5 μm. The relative index difference is 0.25 %.

Solutions : Given :

$n_1 = 1.46$

$a = 4.5\ \mu m$

$\Delta = 0.25\ \% = 0.0025$

Cut-off wavelength is given by,

$$\lambda_c = \frac{2\pi a n_1 (2\Delta)^{\frac{1}{2}}}{V_c}$$

For cut-off wavelength, $V_C = 2.405$

$$\lambda_c = \frac{2\pi \times 4.5 \times 1.46\ (0.005)^{\frac{1}{2}}}{2.405}$$

$$\lambda_c = 1.214\ \mu m$$

Mode Field Diameter and Spot Size

- The mode filed diameter is fundamental parameter of a single mode fiber. This parameter is determined from mode field distributions of fundamental LP_{01} mode.
- In step index and graded single mode fibers, the field amplitude distribution is

approximated by Gaussian distribution. The **mode Field diameter** (MFD) is distance between opposite 1/e – 0.37 times the near field strength)amplitude) and power is $1/e^2 = 0.135$ times.

- In single mode fiber for fundamental mode, on field amplitude distribution the mode filed diameter is shown in fig. 1.8.1.

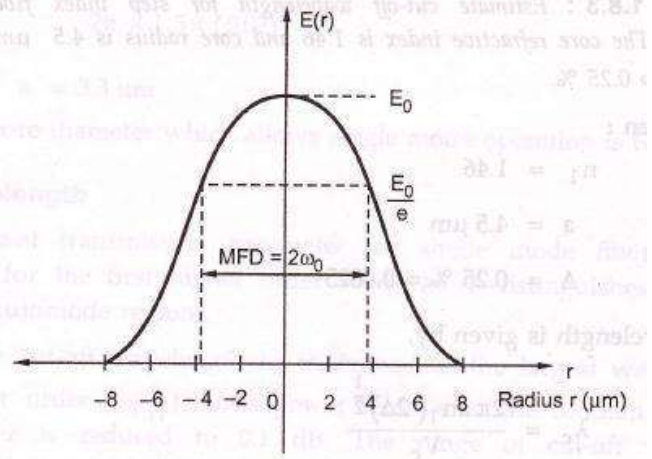


Fig. 1.8.1 Mode field diameter

- The spot size ω_0 is gives as –

$$\omega_0 = \frac{MFD}{2}$$

$$MFD = 2\ \omega_0$$

The parameter takes into account the wavelength dependent filed penetration into the cladding. Fig. 1.8.2 shows mode field diameters variation with λ.

QUESTIONS

1. State and explain the advantages and disadvantages of fiber optic communication systems.
2. State and explain in brief the principle of light propagation.
3. Define following terms with respect to optical laws –
   A) Reflection
   B) Refraction
   C) Refractive index
   D) Snell's law
   E) Critical angle
   F) Total internal reflection (TIR)
4. Explain the important conditions for TIR to exit in fiber.
5. Derive an expression for maximum acceptance angle of a fiber.
6. Explain the acceptance come of a fiber.
7. Define numerical aperture and state its significance also.
8. Explain the different types of rays in fiber optic.
9. Explain the
   A) Step index fiber
   B) Graded index fiber
10. What is mean by mode of a fiber?
11. Write short notes on following –
    A) Single mode step index fiber
    B) Multimode step index fiber
    C) Multimode graded index fiber.

UNIT - 2

SIGNAL DEGRADATION OPTICAL FIBERS.

Introduction

1. One of the important property of optical fiber is signal attenuation. It is also known as fiber loss or signal loss. The signal attenuation of fiber determines the maximum distance between transmitter and receiver. The attenuation also determines the number of repeaters required, maintaining repeater is a costly affair.
2. Another important property of optical fiber is distortion mechanism. As the signal pulse travels along the fiber length it becomes more broader. After sufficient length the broad pulses starts overlapping with adjacent pulses. This creates error in the receiver. Hence the distortion limits the information carrying capacity of fiber.

Attenuation

- Attenuation is a measure of decay of signal strength or loss of light power that occurs as light pulses propagate through the length of the fiber.
- In optical fibers the attenuation is mainly caused by two physical factors absorption and scattering losses. Absorption is because of fiber material and scattering due to structural imperfection within the fiber. Nearly 90 % of total attenuation is caused by Rayleigh scattering only. Microbending of optical fiber also contributes to the attenuation ofsignal.
- The rate at which light is absorbed is dependent on the wavelength of the light and the characteristics of particular glass. Glass is a silicon compound, by adding different additional chemicals to the basic silicon dioxide the optical properties of the glass can be changed.
- The Rayleigh scattering is wavelength dependent and reduces rapidly as the wavelength of the incident radiation increases.
- The attenuation of fiber is governed by the materials from which it is fabricated, the manufacturing process and the refractive index profile chosen. Attenuation loss is measured in dB/km.

Attenuation Units

- As attenuation leads to a loss of power along the fiber, the output power is significantly less than the couples power. Let the couples optical power is p(0) i.e. at origin (z = 0).

Then the power at distance z is given by,

$$P(z) = P(0)e^{-\alpha_p z} \quad \ldots (2.1.1)$$

where, α_p is fiber attenuation constant (per km).

$$\alpha_p = \frac{1}{z} \ln \left[\frac{P(0)}{P(z)}\right]$$

$$\boxed{\alpha_{dB/km} = 10.\frac{1}{z}\log\left[\frac{P(0)}{P(z)}\right]}$$

$$\alpha_{dB/km} = 4.343\ \alpha_p \text{ per km}$$

This parameter is known as fiber loss or fiber attenuation.

- Attenuation is also a function of wavelength. Optical fiber wavelength as a function of wavelength is shown in Fig. 2.1.1.

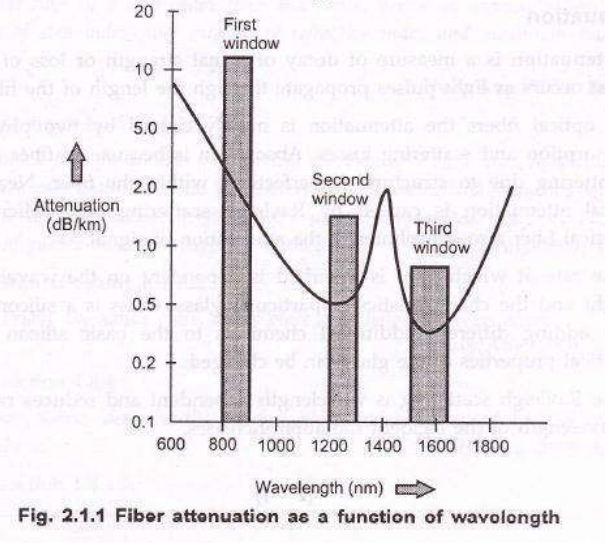


Fig. 2.1.1 Fiber attenuation as a function of wavelength

Example 2.1.1 : A low loss fiber has average loss of 3 dB/km at 900 nm. Compute the length over which –

a) Power decreases by 50 % b) Power decreases by 75 %.

Solution : $\alpha = 3$ dB/km

a) Power decreases by 50 %.

$\Rightarrow$ $\frac{p(0)}{p(z)} = 50\ \% = 0.5$

 is given by,

$$\alpha = 10 . \frac{1}{z} \log \left[\frac{P(0)}{P(z)}\right]$$

$$3 = 10 . \frac{1}{z} \log [0.5]$$

∴

z = **1 km** **… Ans.**

b) $\frac{P(0)}{P(z)} = 25\ \% = 0.25$

Since power decrease by 75 %.

$$3 = 10 \text{ x } \frac{1}{z} \log[0.25]$$

∴ z = **2 km** **… Ans.**

Example 2.1.2 : For a 30 km long fiber attenuation 0.8 dB/km at 1300nm. If a 200 µwatt power is launched into the fiber, find the output power.

Solution : z = 30 km

□ = 0.8 dB/km

P(0) = 200 µW

Attenuation in optical fiber is given by,

$$\alpha = 10 \text{ x } \frac{1}{z} \log \left[\frac{P(0)}{P(z)}\right]$$

∴ $P(z) = \left[\frac{200\ \mu W}{P(z)}\right] = 0.7962\ \mu W$

$$0.8 = 10 \text{ x } \frac{1}{30} \log \left[\frac{200\ \mu W}{P(z)}\right]$$

$$2.4 = 10 \text{ x } \log \left[\frac{200\ \mu W}{P(z)}\right]$$

$$\left[\frac{200\ \mu W}{P(z)}\right] = 10^{2.4}$$

Example 2.1.3 : When mean optical power launched into an 8 km length of fiber is 12 μW, the mean optical power at the fiber output is 3 μW.

Determine –

- Overall signal attenuation in dB.
- The overall signal attenuation for a 10 km optical link using the same fiber with splices at 1 km intervals, each giving an attenuation of 1 dB.

Solution : **Given** : z = 8 km

P(0) = 120 μW

P(z) = 3 μW

1) Overall attenuation is given by,

$$\alpha = 10 \cdot \log \left[\frac{P(0)}{P(z)}\right]$$

$$\alpha = 10 \cdot \log \left[\frac{120}{3}\right]$$

$$\alpha = 16.02 \text{ dB}$$

2) Overall attenuation for 10 km,

Attenuation per km $\alpha_{dB} = \frac{16.02}{z} = \frac{16.02}{8} = 2.00 \text{ dB/km}$

Attenuation in 10 km link = 2.00 x 10 = 20 dB

In 10 km link there will be 9 splices at 1 km interval. Each splice introducing attenuation of 1 dB.

Total attenuation = 20 dB + 9 dB = **29 dB**

Example 2.1.4 : A continuous 12 km long optical fiber link has a loss of 1.5 dB/km.

- What is the minimum optical power level that must be launched into the fiber to maintain as optical power level of 0.3 μW at the receiving end?
- What is the required input power if the fiber has a loss of 2.5 dB/km?

[July/Aug.-2007, 6 Marks]

Solution : Given data : z = 12 km

= 1.5 dB/km

P(0) = 0.3 μW

- Attenuation in optical fiber is given by,

$$\alpha = 10 \times \frac{1}{z} \log\left(\frac{P(0)}{P(z)}\right)$$

$$1.5 = 10 \times \frac{1}{12} \log\left(\frac{0.3\ \mu W}{P(z)}\right)$$

$$\log\left(\frac{0.3\ \mu W}{P(z)}\right) = \frac{1.5}{0.833}$$

$$= 1.80$$

$$\left(\frac{0.3\ \mu W}{P(z)}\right) = 10^{1.8}$$

$$P(z) = \left(\frac{0.3\ \mu W}{10^{1.8}}\right) = \frac{0.3}{63.0}$$

$$P(z) = 4.76 \times 10^{-9} W$$

Optical power output = **4.76×10^{-9} W** **… Ans.**

ii) Input power = ? P(0)

When α = 2.5 dB/km

$$\boxed{\alpha = 10 \times \frac{1}{z} \log\left(\frac{P(0)}{P(z)}\right)}$$

$$2.5 = 10 \times \frac{1}{z} \log\left(\frac{P(0)}{4.76 \times 10^{-9}}\right)$$

$$\log\left(\frac{P(0)}{4.76 \times 10^{-9}}\right) = \frac{2.5}{0.833} = 3$$

$$\frac{P(0)}{4.76 \times 10^{-9}} = 10^3 = 1000$$

$\therefore$ $P(0) = 4.76\ \mu W$

Input power= **4.76 µW** **… Ans.**

Example 2.1.5 : Optical power launched into fiber at transmitter end is 150 µW. The power at the end of 10 km length of the link working in first windows is – 38.2 dBm. Another system of same length working in second window is 47.5 µW. Same length system working in third window has 50 % launched power. Calculate fiber attenuation for each case and mention wavelength of operation. **[Jan./Feb.-2009, 4 Marks]**

Solution : Given data:

$P(0) = 150\ \mu W$

$z = 10$ km

$$P(z) = -38.2\ \text{dBm} \Rightarrow \begin{cases} -38.2 = 10 \log \dfrac{P(z)}{1\ \text{mW}} \\ P(z) = 0.151\ \mu W \end{cases}$$

$z = 10$ km

$$\boxed{\alpha = 10 \times \frac{1}{z} \log \left[\frac{P(0)}{P(z)}\right]}$$

Attenuation in 1st window:

$$\alpha_1 = 10 \times \frac{1}{10} \log \left[\frac{150}{0.151}\right]$$

$$\alpha_1 = \mathbf{2.99\ dB/km}$$ **… Ans.**

Attenuation in 2nd window:

$$\alpha_2 = 10 \times \frac{1}{10} \log \left[\frac{150}{47.5}\right]$$

$$\alpha_2 = \mathbf{0.49\ dB/km}$$ **… Ans.**

Attenuation in 3rd window:

$$\alpha_3 = 10 \times \frac{1}{10} \log \left[\frac{150}{75}\right]$$

$\alpha_3 = 0.30$ dB/km … **Ans.**

Wavelength in 1st window is 850 nm.

Wavelength in 2nd window is 1300 nm.

Wavelength in 3rd window is 1550 nm.

Example 2.1.6 : The input power to an optical fiber is 2 mW while the power measured at the output end is 2 µW. If the fiber attenuation is 0.5 dB/km, calculate the length of the fiber.

[July/Aug.-2006, 6 Marks]

Solution : Given : P(0) = 2 mwatt = 2 x 10^{-3} watt

P(z) = 2 µwatt = 2 x 10^{-6} watt

α = 0.5 dB/km

$$\alpha = 10 \times \frac{1}{z}\left[\frac{P(0)}{P(z)}\right]$$

$$0.5 = 10 \times \frac{1}{z}\log\left[\frac{2\times10^{-3}}{2\times10^{-6}}\right]$$

$z = 60$ km … **Ans.**

$$0.5 = \frac{1}{z} \times 3$$

$$z = \frac{3}{0.05}$$

Absorption

- Absorption loss is related to the material composition and fabrication process of fiber. Absorption loss results in dissipation of some optical power as hear in the fiber cable. Although glass fibers are extremely pure, some impurities still remain as residue after purification. The amount of absorption by these impurities depends on their concentration and light wavelength.

- Absorption is caused by three different mechanisms.
  Absorption by atomtic defects in glass composition.
  Extrinsic absorption by impurity atoms in glass matts.
  Intrinsic absorption by basic constituent atom of fiber.

Absorption by Atomic Defects

- Atomic defects are imperfections in the atomic structure of the fiber materials such as missing molecules, high density clusters of atom groups. These absorption losses are negligible compared with intrinsic and extrinsic losses.
- The absorption effect is most significant when fiber is exposed to ionizing radiation in nuclear reactor, medical therapies, space missions etc. The radiation dames the internal structure of fiber. The damages are proportional to the intensity of ionizing particles. This results in increasing attenuation due to atomic defects and absorbing optical energy. The total dose a material receives is expressed in rad (Si), this is the unit for measuring radiation absorbed in bulk silicon.

  1 rad (Si) = 0.01 J.kg

The higher the radiation intensity more the attenuation as shown in Fig 2.2.1.

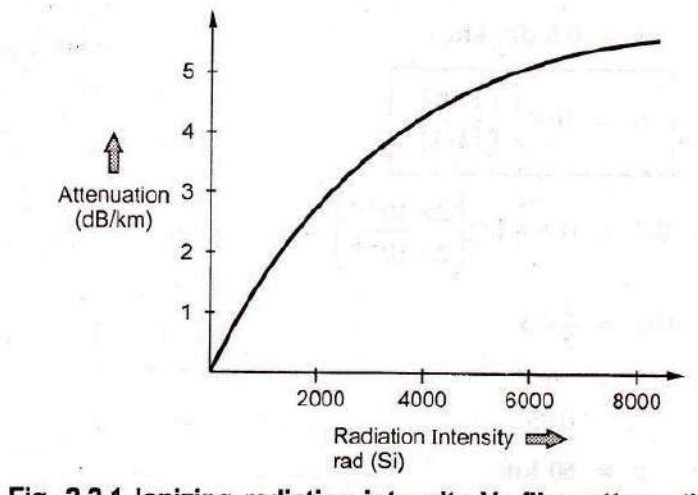


Fig. 2.2.1 Ionizing radiation intensity Vs fiber attenuation

Extrinsic Absorption

- Extrinsic absorption occurs due to electronic transitions between the energy level and because of charge transitions from one ion to another. A major source of attenuation is

  from transition of metal impurity ions such as iron, chromium, cobalt and copper. These losses can be upto 1 to 10 dB/km. The effect of metallic impurities can be reduced by glass refining techniques.
- Another major extrinsic loss is caused by absorption due to **OH (Hydroxil)** ions impurities dissolved in glass. Vibrations occur at wavelengths between 2.7 and 4.2 μm.

The absorption peaks occurs at 1400, 950 and 750 nm. These are first, second and third overtones respectively.

- Fig. 2.2.2 shows absorption spectrum for OH group in silica. Between these absorption peaks there are regions of low attenuation.

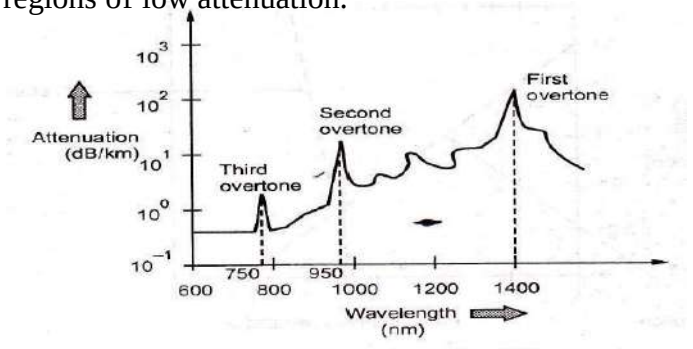


Fig. 2.2.2 Absorption spectra for OH group

Intrinsic Absorption

- Intrinsic absorption occurs when material is in absolutely pure state, no density variation and inhomogenities. Thus intrinsic absorption sets the fundamental lower limit on absorption for any particular material.
- Intrinsic absorption results from electronic absorption bands in UV region and from atomic vibration bands in the near infrared region.
- The electronic absorption bands are associated with the band gaps of amorphous glass materials. Absorption occurs when a photon interacts with an electron in the valene band and excites it to a higher energy level. UV absorption decays exponentially with increasing wavelength (λ).
- In the IR (infrared) region above 1.2 μm the optical waveguide loss is determined by presence of the OH ions and inherent IR absorption of the constituent materials. The inherent IR absorption is due to interaction between the vibrating band and the electromagnetic field of optical signal this results in transfer of energy from field to the band, thereby giving rise to absorption, this absorption is strong because of many bonds present in the fiber.

6. The ultraviolet loss at any wavelength is expressed as,

$$\alpha_{uv} = \frac{154.2}{46.6\,x+60} \times 10^{-2} \times e^{\left(\frac{4.63}{\lambda}\right)} \quad \ldots (2.2.1)$$

where, x is mole fraction of GeO_2.

λ is operating wavelength.

α_{uv} is in dB/km.

9. The loss in infrared (IR) region (above 1.2 μm) is given by expression :

$$\alpha_{IR} = 7.81 \times 10^{11} \times e^{\left(\frac{-48.48}{\lambda}\right)} \quad \ldots (2.2.2)$$

The expression is derived for GeO_2-SiO_2 glass fiber.

Rayleigh Scattering Losses

13. Scattering losses exists in optical fibers because of microscopic variations in the material density and composition. As glass is composed by randomly connected network of molecules and several oxides (e.g. SiO_2, GeO_2 and P_2O_5), these are the major cause of compositional structure fluctuation. These two effects results to variation in refractive index and Rayleigh type scattering of light.
14. **Rayleigh scattering** of light is due to small localized changes in the refractive index of the core and cladding material. There are two causes during the manufacturing of fiber.

---

3. The first is due to slight fluctuation in mixing of ingredients. The random changes because of this are impossible to eliminate completely.
4. The other cause is slight change in density as the silica cools and solidifies. When light ray strikes such zones it gets scattered in all directions. The amount of scatter depends on the size of the discontinuity compared with the wavelength of the light so the shortest wavelength (highest frequency) suffers most scattering. Fig. 2.3.1 shows graphically the relationship between wavelength and Rayleigh scattering loss.

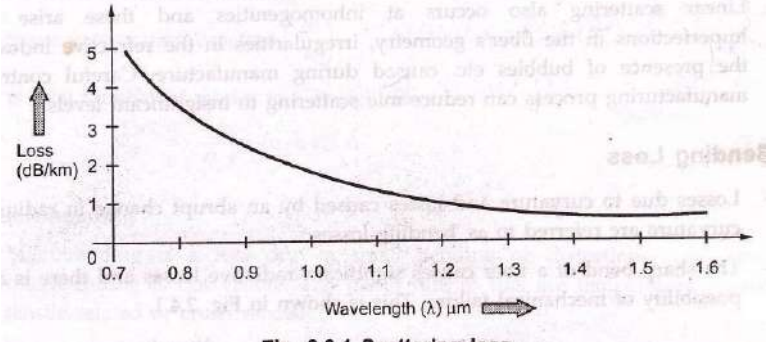


Fig. 2.3.1 Scattering loss

7. Scattering loss for single component glass is given by,

$$\alpha_{scat} = \frac{8\pi^3}{3\lambda^4} (n^2 - 1)^2 \, k_B \, T_f \, \beta_T \text{ nepers} \quad \ldots (2.3.1)$$

where, n = Refractive index

k_B = Boltzmann's constant

β_T = Isothermal compressibility of material

T_f = Temperature at which density fluctuations are frozen into the glass as it solidifies (fictive temperature)

Another form of equation is

$$\alpha_{scat} = \frac{8\pi^3}{3\lambda^4} n^8 \, p^2 \, k_B \, T_f \, \beta_T \text{ neper} \quad \alpha_{scat} = \frac{8\pi^3}{3\lambda^4} (\delta_n^2)^2 \, \delta v \quad \ldots (2.3.2)$$

where, P = Photoelastic coefficient

where, δ_n^2 = Mean square refractive index fluctuation

δv = Volume of fiber

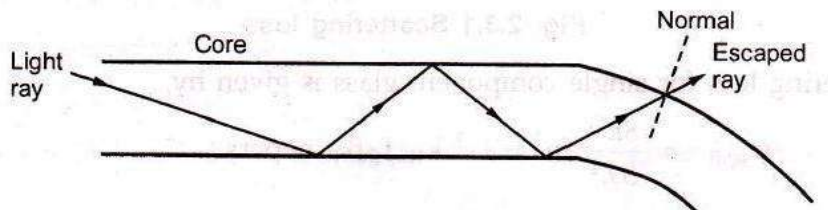
- Multimode fibers have higher dopant concentrations and greater compositional fluctuations. The overall losses in this fibers are more as compared to single mode fibers.

Mie Scattering :

- Linear scattering also occurs at inhomogenities and these arise from imperfections in the fiber's geometry, irregularities in the refractive index and the presence of bubbles etc. caused during manufacture. Careful control of manufacturing process can reduce mie scattering to insignificant levels.

Bending Loss

- Losses due to curvature and losses caused by an abrupt change in radius of curvature are referred to as 'bending losses.'
- The sharp bend of a fiber causes significant radiative losses and there is also possibility of mechanical failure. This is shown in Fig. 2.4.1.



2 As the core bends the normal will follow it and the ray will now find itself on the wrong side of critical angle and will escape. The sharp bends are therefore avoided.

3 The radiation loss from a bent fiber depends on –

Field strength of certain critical distance x_C from fiber axis where power is lost through radiation.

The radius of curvature R.

4 The higher order modes are less tightly bound to the fiber core, the higher order modes radiate out of fiber firstly.

5 For multimode fiber, the effective number of modes that can be guided by curved fiber is where, α is graded index profile.

 is core – cladding index difference.

n_2 is refractive index of cladding.k is

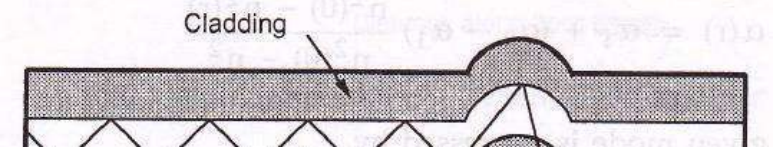
wave propagation constant $\left(\frac{2\pi}{\lambda}\right)$.

N_∞ is total number of modes in a straight fiber.

$$N_\infty = \frac{\alpha}{\alpha+2}(n_1\, k\, a)^2\Delta \qquad \ldots (2.4.2)$$

Microbending

- Microbending is a loss due to small bending or distortions. This small microbending is not visible. The losses due to this are temperature related, tensile related or crush related.
- The effects of microbending on multimode fiber can result in increasing attenuation (depending on wavelength) to a series of periodic peaks and troughs on the spectral attenuation curve. These effects can be minimized during installation and testing. Fig. 2.4.2 illustrates microbening.



Macrobending

- The change in spectral attenuation caused by macrobending is different to microbending. Usually there are no peaks and troughs because in a macrobending no light is coupled back into the core from the cladding as can happen in the case of microbends.

- The macrobending losses are cause by large scale bending of fiber. The losses are eliminated when the bends are straightened. The losses can be minimized by not exceeding the long term bend radii. Fig. 2.4.3 illustrates macrobending.

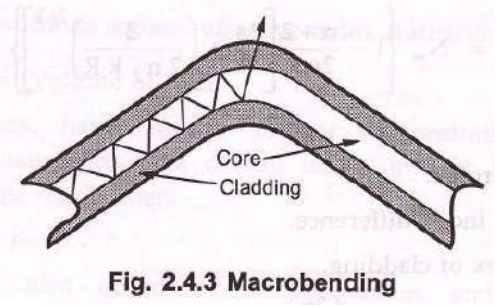


Fig. 2.4.3 Macrobending

Core and Cladding Loss

- Since the core and cladding have different indices of refraction hence they have different attenuation coefficients α_1 and α_2 respectively.
- For step index fiber, the loss for a mode order (v, m) is given by,

$$\alpha_{v\,m} = \alpha_1 \frac{P_{core}}{P} + \alpha_2 \frac{P_{cladding}}{P} \quad \ldots (2.5.1)$$

For low-order modes, the expression reduced to

$$\alpha_{v\,m} = \alpha_1 + (\alpha_2 + \alpha_1) \frac{P_{cladding}}{P} \quad \ldots (2.5.2)$$

where, $\frac{P_{core}}{P}$ and $\frac{P_{cladding}}{P}$ are fractional powers.

- For graded index fiber, loss at radial distance is expressed as,

$$\alpha(r) = \alpha_1 + (\alpha_2 - \alpha_1) \frac{n^2(0) - n^2(r)}{n^2(0) - n_2^2}$$

… (2.5.3)

The loss for a given mode is expressed by,

$$\propto_{\text{Graded Index}} = \frac{\int_0^{\infty} \propto(r)\, P(r)\, r\, dr}{\int_0^{\infty} P(r) r\, dr} \quad \ldots (2.5.4)$$

where, P(r) is power density of that model at radial distance r.

Signal Distortion in Optical Waveguide

- The pulse get distorted as it travels along the fiber lengths. Pulse spreading in fiber is referred as dispersion. Dispersion is caused by difference in the propagation times of light rays that takes different paths during the propagation. The light pulses travelling down the fiber encounter dispersion effect because of this the pulse spreads out in time domain. Dispersion limits the information bandwidth. The distortion effects can be analyzed by studying the group velocities in guided modes.

Information Capacity Determination

- Dispersion and attenuation of pulse travelling along the fiber is shown in Fig. 2.6.1.

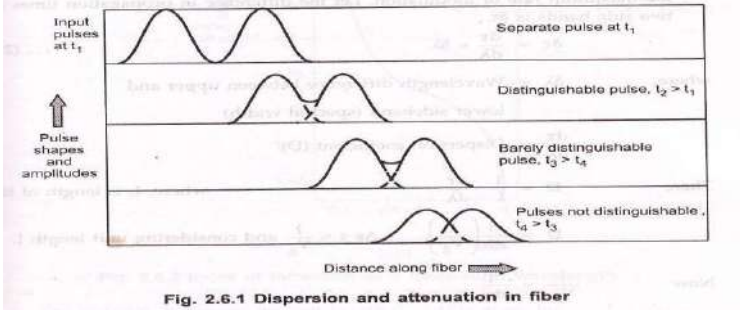


Fig. 2.6.1 Dispersion and attenuation in fiber

- Fig. 2.6.1 shows, after travelling some distance, pulse starts broadening and overlap with the neighbouring pulses. At certain distance the pulses are not even distinguishable and error will occur at receiver. Therefore the information capacity is specified by bandwidth-distance product (MHz . km). For step index bandwidth distance product is 20 MHz . km and for graded index it is 2.5 MHz . km.

Group Delay

- Consider a fiber cable carrying optical signal equally with various modes and each mode contains all the spectral components in the wavelength band. All the spectral components travel independently and they observe different **time delay** and **group delay** in the direction of propagation. The velocity at which the energy in a pulse travels along the fiber is known as **group velocity**. Group velocity is given by,

$$V_g = \frac{\partial w}{\partial \beta} \quad \ldots (2.6.1)$$

- Thus different frequency components in a signal will travel at different group velocities and so will arrive at their destination at different times, for digital modulation of carrier, this results in dispersion of pulse, which affects the maximum rate of modulation. Let the difference in propagation times for two side bands is $\delta\tau$.

$$\delta\tau = \frac{d\tau}{d\lambda} \, x \, \delta\lambda \quad \ldots (2.6.2)$$

where, $\delta\tau$ = Wavelength difference between upper and lower sideband (spectral width)

$\frac{d\tau}{d\lambda}$ = Dispersion coefficient (D)

Then, $D = \frac{1}{L} \cdot \frac{d\tau}{d\lambda}$ where, L is length of fiber.

$D = \frac{d}{d\lambda}\left(\frac{1}{V_g}\right)$ As $\tau = \frac{1}{V_g}$ and considering unit length L = 1.

Now $\frac{1}{V_g} = \frac{d\beta}{d\omega}$

$$\frac{1}{V_g} = \frac{d\lambda}{d\omega} x \frac{d\beta}{d\lambda}$$

$$\frac{1}{V_g} = \frac{-\lambda^2}{2\pi c} x \frac{d\beta}{d\lambda}$$

$\therefore$

$$D = \frac{d}{d\lambda}\left(\frac{-\lambda^2}{2\pi c} \cdot \frac{d\beta}{d\lambda}\right) \quad \ldots (2.6.3)$$

- Dispersion is measured in picoseconds per nanometer per kilometer.

Material Dispersion

- Material dispersion is also called as chromatic dispersion. Material dispersion exists due to change in index of refraction for different wavelengths. A light ray contains components of various wavelengths centered at wavelength λ_{10}. The time delay is

different for different wavelength components. This results in time dispersion of pulse at

the receiving end of fiber. Fig. 2.6.2 shows index of refraction as a function of optical wavelength.

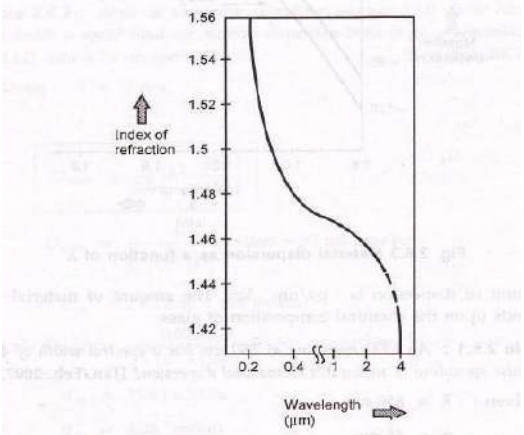


Fig. 2.6.2 Index of refraction as a function of wavelength

2. The material dispersion for unit length (L = 1) is given by

$$D_{mat} = \frac{-\lambda}{c} x \frac{d^2 n}{d\lambda^2} \quad \ldots (2.6.4)$$

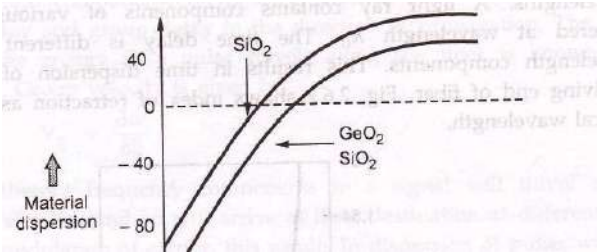
where, c = Light velocity

λ = Center wavelength

$\frac{d^2 n}{d\lambda^2}$ = Second derivative of index of refraction w.r.t wavelength

Negative sign shows that the upper sideband signal (lowest wavelength) arrives before the lower sideband (highest wavelength).

- A plot of material dispersion and wavelength is shown in



- The unit of dispersion is : ps/nm . km. The amount of material dispersion depends upon the chemical composition of glass.

Example 2.6.1 : An LED operating at 850 nm has a spectral width of 45 nm. What is the pulse spreading in ns/km due to material dispersion? **[Jan./Feb.-2007, 3 Marks]**

Solution : Given : $\lambda = 850$ nm

$\sigma = 45$ nm

R.M.S pulse broadening due to material dispersion is given by,

$$\sigma_m = \sigma \, LM$$

Considering length L = 1 metre

$$\text{Material dispersion constant } D_{mat} = \frac{-\lambda}{c} \cdot \frac{d^2n}{d\lambda^2}$$

For LED source operating at 850 nm, $\left|\lambda^2 \frac{d^2n}{d\lambda^2}\right| = 0.025$

$\therefore$ $$M = \frac{1}{c\lambda}\left|\lambda^2 \frac{d^2n}{d\lambda^2}\right| = \frac{1}{(3 \times 10^5)(850)} \times 0.025$$

$M = 9.8$ ps/nm/km

$\sigma_m =$ **441 ns/km** **… Ans.**

Example 2.6.2 : What is the pulse spreading when a laser diode having a 2 nm spectral width is used? Find the the material-dispersion-induced pulse spreading at 1550 nm for an LED with a 75 nm spectral width **[Jan./Feb.-2007, 7 Marks]**

Solutions : Given : $\lambda = 2$ nm

$\sigma = 75$

$$D_{mat} = \frac{1}{c\lambda}\left|\lambda^2 . \frac{d^2 n}{d\lambda^2}\right|$$

$$D_{mat} = \frac{1}{(3 \times 10^5) \times 2} \times 0.03 = 50 \text{ ps/nm/km}$$

$\sigma_m = 2 \times 1 \times 50 =$ **100 ns/km** **… Ans.**

For LED $D_{mat} = \frac{0.025}{(3 \times 10^5) \times 1550} = 53.76 \text{ ps nm}^{-1}\text{km}^{-1}$

$\sigma_m = 75 \times 1 \times 53.76$

$\sigma_m =$ **4.03 ns/km** **… Ans.**

Waveguide Dispersion

- Waveguide dispersion is caused by the difference in the index of refraction between the core and cladding, resulting in a 'drag' effect between the core and cladding portions of the power.
- Waveguide dispersion is significant only in fibers carrying fewer than 5-10 modes. Since multimode optical fibers carry hundreds of modes, they will not have observable waveguide dispersion.
- The group delay (τ_{wg}) arising due to waveguide dispersion.

$$(\tau_{wg}) = \frac{L}{c}\left[n_2 + n_2 \Delta \frac{d\,(kb)}{dk}\right] \quad \ldots (2.6.5)$$

Where, b = Normalized propagation constant

k = $2\pi / \lambda$ (group velocity)

Normalized frequency V,

$$V = ka(n_1^2 - n_2^2)^{\frac{1}{2}}$$

$$V = k\,a\,n_2\sqrt{2\Delta} \text{ (For small } \Delta)$$

$$\therefore \quad \tau_{wg} = \frac{L}{c}\left[n_2 + n_2 \Delta \frac{d(V_b)}{dV}\right] \qquad \text{... (2.6.6)}$$

The second term $\frac{d(V_b)}{dV}$ is waveguide dispersion and is mode dependent term..

- As frequency is a function of wavelength, the group velocity of the energy varies with frequency. The produces additional losses (waveguide dispersion). The propagation constant (b) varies with wavelength, the causes of which are independent of material dispersion.

Chromatic Dispersion

- The combination of material dispersion and waveguide dispersion is called chromatic dispersion. These losses primarily concern the spectral width of transmitter and choice of correct wavelength.
- A graph of effective refractive index against wavelength illustrates the effects of material, chromatic and waveguide dispersion.

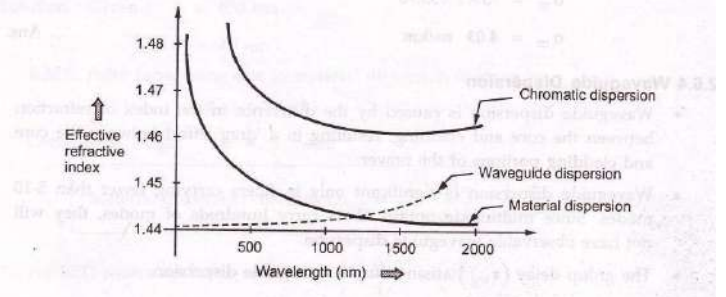


Fig. 2.6.4 Graph of effective refractive index against wavelength showing effects of chromatic, waveguide and material dispersion

- Material dispersion and waveguide dispersion effects vary in vary in opposite senses as the wavelength increased, but at an optimum wavelength around 1300 nm, two effects almost cancel each other and chromatic dispersion is at minimum. Attenuation is therefore also at minimum and makes 1300 nm a highly attractive operating wavelength.

Modal Dispersion

- As only a certain number of modes can propagate down the fiber, each of these modes carries the modulation signal and each one is incident on the boundary at a different angle, they will each have their own individual propagation times. The net effect is spreading of pulse, this form o dispersion is called modal dispersion.
- Modal dispersion takes place in multimode fibers. It is moderately present in graded index fibers and almost eliminated in single mode step index fibers.
- Modal dispersion is given by,

$$\Delta t_{modal} = \frac{n_1 Z}{c}\left(\frac{\Delta}{1-\Delta}\right)$$

where Δt_{modal} = Dispersion

n_1 = Core refractive index

Z = Total fiber length

c = Velocity of light in air

□ = Fractional refractive index $\left(\frac{n_1 - n_2}{n_1}\right)$

Putting $\Delta = \frac{(NA^2)Z}{2n_1 c}$ in above equation

$$\Delta t_{modal} = \frac{(NA^2)Z}{2n_1 c}$$

- The modal dispersion Δt_{modal} describes the optical pulse spreading due to modal effects optical pulse width can be converted to electrical rise time through the relationship.

$$t_{r\,mod} = 0.44\,(\Delta t_{modal})\pi r^2$$

Signal distortion in Single Mode Fibers

- The pulse spreading σ_{wg} over range of wavelengths can be obtained from derivative of group delay with respect t

$$\sigma_{wg} = \left|\frac{d\tau_{wg}}{d\lambda}\right| \sigma_\lambda$$

where,

$$D_{wg}(\lambda) = \frac{-n_2\Delta}{c\lambda}\left[V\frac{d^2(Vb)}{dV^2}\right] \quad \ldots (2.6.8)$$

- This is the equation for waveguide dispersion for unit length.

Example 2.6.3 : For a single mode fiber n_2 = 1.48 and = 0.2 % operating at A = 1320 nm, compute the waveguide dispersion if $V.\frac{d^2(Vb)}{dv^2} = 0.26.$

Solution : n_2 = 1.48

0.2

□ = 1320 nm

Waveguide dispersion is given by,

$$D_{wg}(\lambda) = \frac{-n_2\Delta}{c\lambda}\left[V\frac{d^2(Vb)}{dV^2}\right]$$

$$= \frac{-1.48 \times 0.2}{3 \times 10^5 \times 1320}[0.20]$$

i) -1.943 picosec/nm . km.

Higher Order Dispersion

- Higher order dispersive effective effects are governed by dispersion slope S.

$$S = \frac{dD}{d\lambda}$$

where, D is total dispersion

Also, $$S = \left(\frac{2\pi c}{\lambda^2}\right)^2 \beta_3 + \left(\frac{4\pi c}{\lambda^3}\right)\beta_2$$

where,

β_2 and β_3 are second and third order dispersion parameters.

- Dispersion slope S plays an important role in designing WDM system

Dispersion Induced Limitations

- The extent of pulse broadening depends on the width and the shape of input pulses. The pulse broadening is studied with the help of wave equation.

Basic Propagation Equation

- The basic propagation equation which governs pulse evolution in a single mode fiber is given by,

$$\frac{\partial A}{\partial z} + \beta_1 \frac{\partial A}{\partial t} + \frac{i\beta_2}{2} \cdot \frac{\partial^2 A}{\partial t^2} - \frac{\beta_3}{6} \frac{\partial^3 A}{\partial t^3} = 0$$

where,

β_1, β_2 and β_3 are different dispersion parameters.

Chirped Gaussian Pulses

- A pulse is said to b e chirped if its carrier frequency changes with time.
- For a Gaussian spectrum having spectral width σ_ω, the pulse broadening factor is given by,

$$\frac{\sigma^2}{\sigma_0^2} = \left(1 + \frac{C\beta_2 L}{2\sigma_2^2}\right)^2 + (1 + V_\omega^2)\left(\frac{\beta_2 L}{2\sigma_0^2}\right)^2 + (1 + C + V_\omega^2)^2 \left(\frac{\beta_3 L}{4\sqrt{2\sigma_0^3}}\right)\pi r^2$$

where, $V_\omega = 2\sigma_\omega\ \sigma_0$

Limitations of Bit Rate

- The limiting bit rate is given by,

$$4B\ \sigma \leq 1$$

- The condition relating bit rate-distance product (BL) and dispersion (D) is given

$$BL\,|S|\sigma_\lambda^2 \leq \frac{1}{\sqrt{8}}$$

where, S is dispersion slope.

- Limiting bit rate a single mode fibers as a function of fiber length for $\sigma\lambda = 0$, a and 5nm is shown in fig. 2.6.5.

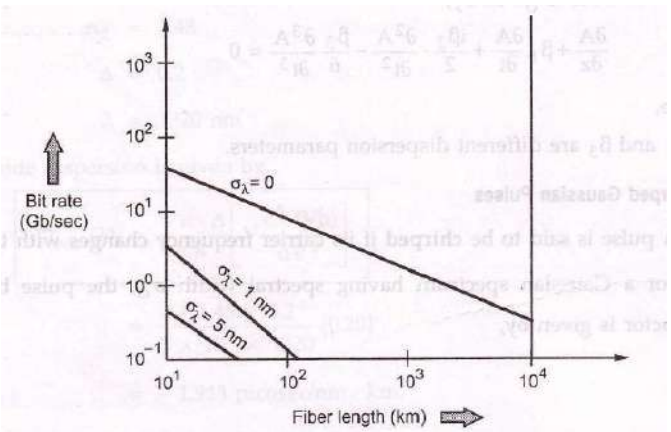


Fig. 2.6.5 Dependence of bit rate on fiber length

Polarization Mode Dispersion (PMD)

- Different frequency component of a pulse acquires different polarization state (such as linear polarization and circular polarization). This results in pulse broadening is kno**w as polarization mode dispersion (PMD).**
- PMD is the limiting factor for optical communication system at high data rates. The effects of PMD must be compensated.

Pulse Broadening in GI Fibers

- The core refractive index varies radially in case of graded index fibers, hence it supports multimode propagation with a low intermodal delay distortion and high data rate over long distance is possible. The higher order modes travelling in outer regions of the core, will travel faster than the lower order modes travelling in high refractive index region. If the index profile is carefully controlled, then the transit times of the individual modes will be identical, so eliminating modal dispersion.
- The r.m.s pulse broadening is given as:

$$\sigma = \left(\sigma^2_{\text{intermodal}} + \sigma^2_{\text{intermodal}}\right)^{1/2} \quad \ldots (2.7.1)$$

where,

$\sigma_{\text{intermodal}}$ – R.M.S pulse width due to intermodal delay distortion.

$\sigma_{\text{intermodal}}$ – R.M.S pulse width resulting from pulse broadening within each mode.

- The intermodal delay and pulse broadening are related by expression given by Personick.

$$\sigma_{\text{intermodal}} = \left(\langle\tau_g^2\rangle - \langle\tau_g\rangle^2\right)^{1/2} \quad \ldots (2.7.2)$$

Where τ_g is group delay.

From this the expression for intermodal pulse broadening is given as:

$$\propto_{\text{intermodal}} = \frac{LN_1\Delta}{2c} \cdot \frac{\propto}{\propto +1}\left(\frac{\propto +2}{3 \propto +2}\right)^{1/2} x$$

$$\left[c_1^2 + \frac{4c_1c_2(\propto+1)}{2\propto+1} + \frac{16\Delta^2c_2^2(\propto+1)^2}{(5\propto+2)(3\propto+2)}\right]^{1/2} \quad \ldots (2.7.3)$$

$$c_1 = \frac{\propto -2 - E}{\propto +2} \; and \; c_2 = \frac{3 \propto -2 - 2c}{2(\propto +2)}$$

- The intramodal pulse broadening is given as :

$$\sigma^2_{intramodal} = \left(\frac{\sigma\lambda}{\lambda}\right)^2 \left\langle\left(\lambda\frac{d\tau g}{d\lambda}\right)^2\right\rangle \quad \ldots (2.7.4)$$

Where $\sigma\lambda$ is spectral width of optical source.

Solving the expression gives :

$$\sigma^2_{intramodal} = \frac{L}{c} \cdot \frac{\sigma\lambda}{\lambda}\left[\left(-\lambda^2\frac{d^2n_1}{d\lambda^2}\right)^2 - N_1c_1\Delta\right.$$

$$\left.\left(2\lambda^2\frac{d^2n_1}{d\lambda^2} \cdot \frac{\propto}{\propto+1} - N_1c_1\Delta\frac{4\propto^2}{(\propto+2)(3\propto+2)}\right)\right]^{1/2}$$

Mode Coupling

- After certain initial length, the pulse distortion increases less rapidly because of mode coupling. The energy from one mode is coupled to other mods because of:
  - Structural imperfections.
  - Fiber diameter variations.
  - Refractive index variations.
  - Microbends in cable.
- Due to the mode coupling, average propagation delay become less and intermodal distortion reduces.
- Suppose certain initial coupling length = L_C, mode coupling length, over L_C = Z. Additional loss associated with mode coupling = h (dB/ km).

Therefore the excess attenuation resulting from mode coupling = hZ.

The improvement in pulse spreading by mode coupling is given as :

$$hZ\left(\frac{\sigma_c}{\sigma_0}\right) = C$$

where, C is constant independent of all dimensional quantities and refractive indices.

σ_C is pulse broadening under mode coupling.

σ_0 is pulse broadening in absence of mode coupling.

- For long fiber length's the effect of mode coupling on pulse distortion is significant. For a graded index fiber, the effect of distance on pulse broading for various coupling losses are shown

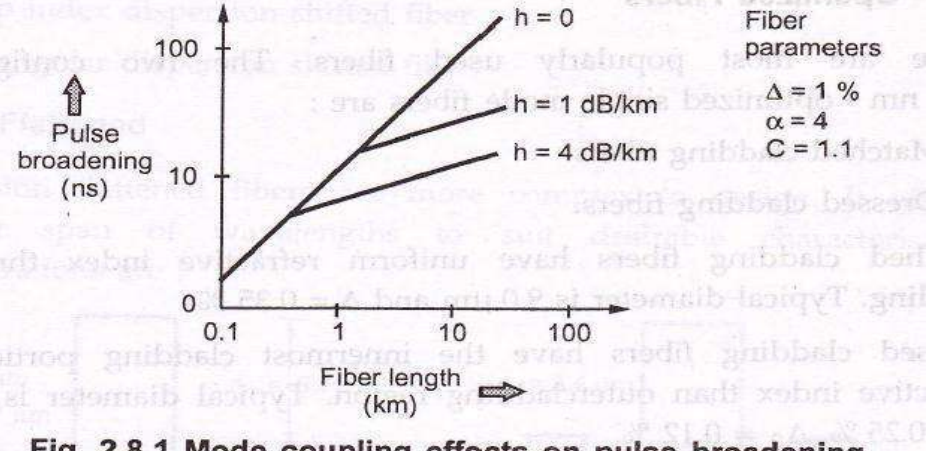


Fig. 2.8.1 Mode coupling effects on pulse broadening

Significant mode coupling occurs of connectors, splices and with other passive components of an optical link.

Design Optimization

- Features of single mode fibers are :
  - Longer life.
  - Low attenuation.
  - Signal transfer quality is good.
  - Modal noise is absent.
  - Largest BW-distance product.
- Basic design – optimization includes the following :
  - Cut-off wavelength.
  - Dispersion.
  - Mode field diameter.
  - Bending loss.
  - Refractive index profile.

Refractive Index Profile

- Dispersion of single mode silica fiber is lowest at 1300 nm while its attenuation is minimum at 1550 nm. For archiving maximum transmission distance the dispersion null should be at the wavelength of minimum attenuation. The waveguide dispersion is easier to control than the material dispersion. Therefore a variety of core-cladding refractive.

idex configuration fivers. Such as 1300 nm – optimized fibers, dispersion shifted fibers, dispersion – flattened fibers and large effective core area fibers.

- **1300 nm – Optimized Fibers**

These are most popularly used fibers. The two configurations of 1300 nm – optimized single mode fibers are :

Matched cladding fibers.

Dressed cladding fibers.

Matched cladding fibers have uniform refractive index throughout its cladding. Typical diameter is 9.0 μm and $\Delta = 0.35$ %.

Dressed cladding fibers have the innermost cladding portion has low refractive index than outrcladding region. Typical diameter is 8.4 μm and $\Delta_1 = 0.25$ %, $\Delta_2 = 0.12$ %.

Fig 2.9.1 shows both types of fibers.

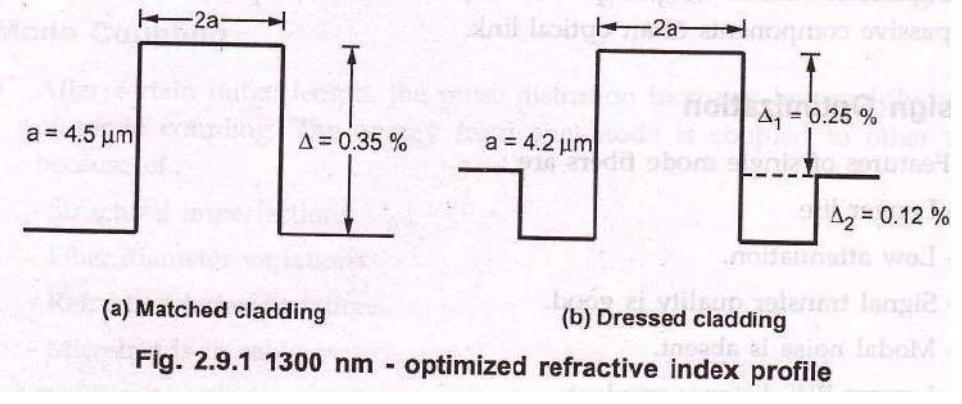


Fig. 2.9.1 1300 nm - optimized refractive index profile

2. Dispersion Shifted Fibers

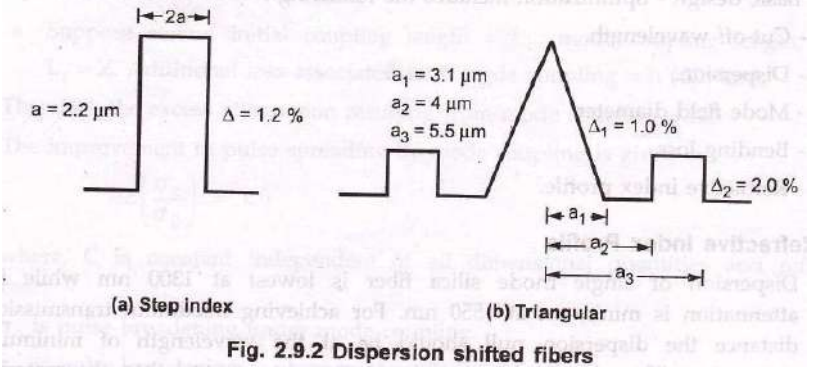


Fig. 2.9.2 Dispersion shifted fibers

3. The addition of wavelength and material dispersion can shift the zero dispersion point of longer wavelength. Two configurations of dispersion shifted fibers are

Step index dispersion shifted fiber.

Triangular dispersion shifted fiber.

□ Dispersion Flattened

Dispersion flattened fibers are more complex to design. It offers much broader span of wavelengths to suit desirable characteristics. Two configurations are :

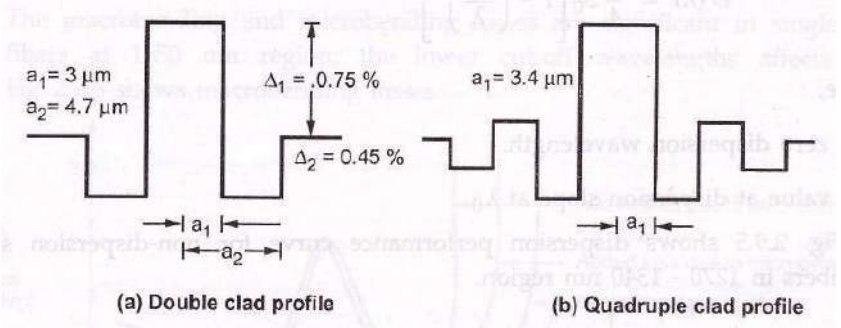


Fig. 2.9.3 Dispersion flattened

- Fig 2.9.4 shows total resultant dispersion.

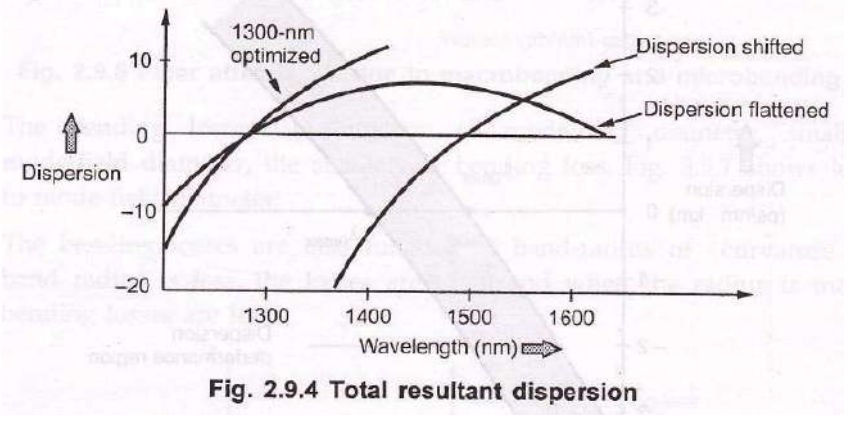


Fig. 2.9.4 Total resultant dispersion

Dispersion Calculations

- The total dispersion consists of material and waveguide dispersions. The resultant intermodal dispersion is given as,

$$D(\lambda) = \frac{d\tau}{d\lambda}$$

where, τ is group delay per unit length of fiber.

- The broadening σ of an optical pulse is given

$$\sigma = D(\lambda)\, L\sigma\lambda$$

where, $\sigma\lambda$ is half power spectral width of source.

= As the dispersion varies with wavelength and fiber type. Different formulae are used to calculate dispersions for variety of fiber at different wavelength.

= For a non – dispersion shifted fiber between 1270 nm to 1340 nm wavelength, the expression for dispersion is given as :

$$D(\lambda) = \frac{\lambda}{4} S_0 \left[1 - \left(\frac{\lambda_0}{\lambda}\right)^4\right]$$

where,

λ_0 is zero dispersion wavelength.

S_0 is value at dispersion slop at λ_0.

ιιι) Fig 2.9.5 shows dispersion performance curve for non-dispersion shifted fibers in 1270 – 1340 nm region.

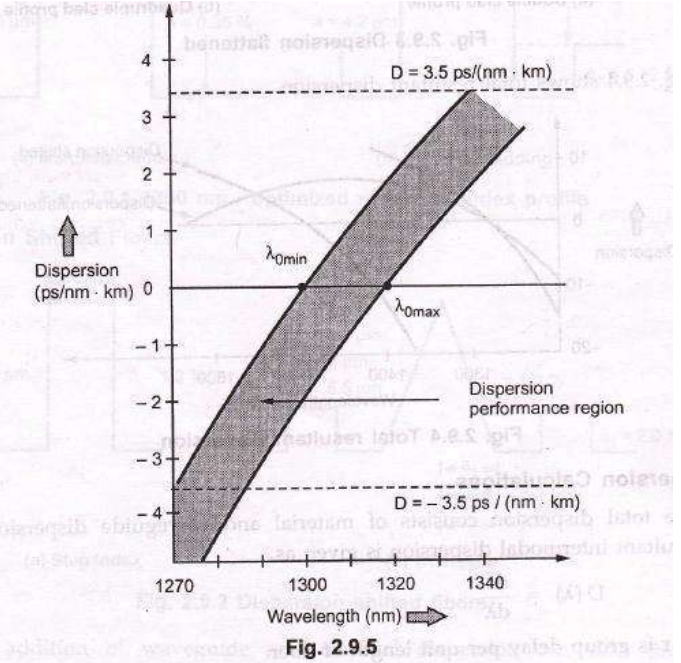


Fig. 2.9.5

- Maximum dispersion specified as 3.5 ps/(nm . km) marked as dotted line in Fig. 2.9.5.

The cut-off frequency of an optical fiber

- The cut-off frequency of an optical fiber is determined not only by the fiber itself (modal dispersion in case of multimode fibers and waveguide dispersion in case of single mode fibers) but also by the amount of material dispersion caused by the spectral width of transmitter.

Bending Loss Limitations

- The macrobending and microbending losses are significant in single mode fibers at 1550 nm region, the lower cut-off wavelengths affects more. Fig. 2.9.6 shows macrobending losses.

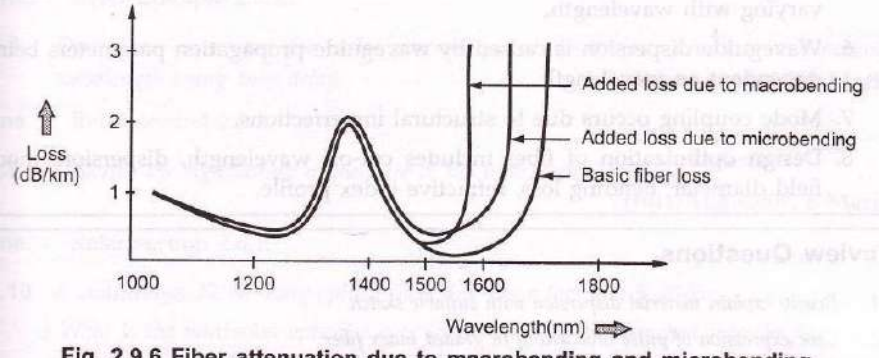


Fig. 2.9.6 Fiber attenuation due to macrobending and microbending

〈 The bending losses are function of mode-filed diameter, smaller the mode-field diameter, the smaller the bending loss. Fig. 2.9.7 shows loss due to mode-field diameter.

〈 The bending losses are also function of bend-radius of curvature. If the bend radius is less, the losses are more and when the radius is more, the bending losses are less.

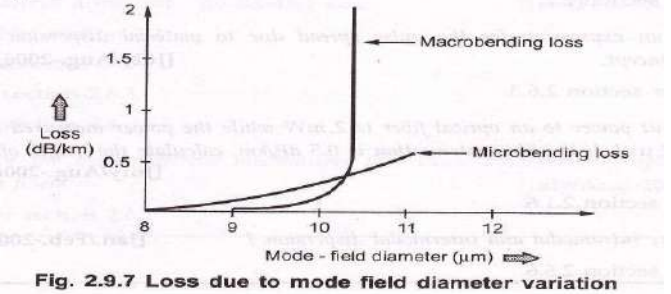


Fig. 2.9.7 Loss due to mode field diameter variation

Recommended Questions:

d) Briefly explain material dispersion with suitable sketch.

e) Give expression of pulse broadening in graded index fiber.

f) State the significance of mode coupling in optic fiber communication.

g) Explain in detail the design optimization of single mode fibers.

h) Elaborate dispersion mechanism in optical fibers.

UNIT - 3

OPTICAL SOURCES AND COUPLING

Optical Sources

3. Optical transmitter coverts electrical input signal into corresponding optical signal. The optical signal is then launched into the fiber. Optical source is the major component in an optical transmitter.
4. Popularly used optical transmitters are Light Emitting Diode (LED) and semiconductor Laser Diodes (LD).

Characteristics of Light Source of Communication

- To be useful in an optical link, a light source needs the following characteristics:
  - It must be possible to operate the device continuously at a variety of temperatures for many years.
  - It must be possible to modulate the light output over a wide range of modulating frequencies.
  - For fiber links, the wavelength of the output should coincide with one of transmission windows for the fiber type used.
  - To couple large amount of power into an optical fiber, the emitting area should be small.
  - To reduce material dispersion in an optical fiber link, the output spectrum should be narrow.
  - The power requirement for its operation must be low.

- The light source must be compatible with the modern solid state devices.
- The optical output power must be directly modulated by varying the input current to the device.
- Better linearity of prevent harmonics and intermodulation distortion.
- High coupling efficiency.
- High optical output power.
- High reliability.
- Low weight and low cost.

Two types of light sources used in fiber optics are light emitting diodes (LEDs) and laser diodes (LDs).

Light Emitting Diodes(LEDs)

p-n Junction

Conventional p-n junction is called as **homojunction** as same semiconductor material is sued on both sides junction. The electron-hole recombination occurs in relatively

layer = 10 μm. As the carriers are not confined to the immediate vicinity of junction, hence high current densities can not be realized.

- The carrier confinement problem can be resolved by sandwiching a thin layer (= 0.1 μm) between p-type and n-type layers. The middle layer may or may not be doped. The carrier confinement occurs due to bandgap discontinuity of the junction. Such a junction is call **heterojunction** and the device is called double **heterostructure.**
- In any optical communication system when the requirements is –
  1. Bit rate f 100-2—Mb/sec.
  2. Optical power in tens of micro watts.

  LEDs are best suitable optical source.

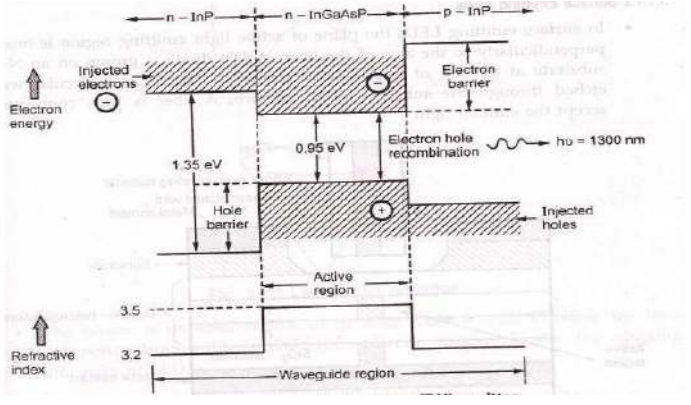
LED Structures

Heterojuncitons

- A heterojunction is an interface between two adjoining single crystal semiconductors with different bandgap.
- Heterojuctions are of two types, Isotype (n-n or p-p) or Antisotype (p-n).

Double Heterojunctions (DH)

In order to achieve efficient confinement of emitted radiation double **heterojunctions** are used in LED structure. A heterojunciton is a junction formed by dissimilar semiconductors. Double heterojunction (DH) is formed by two different semiconductors on each side of active region. Fig. 3.1.1 shows double heterojunction (DH) light emitter.



- The crosshatched regions represent the energy levels of freecharge. Recombination occurs only in active InGaAsP layer. The two materials have different bandgap energies and different refractive indices. The changes in bandgap energies create potential barrier for both holes and electrons. The free charges can recombine only in narrow, well defined active layer side.
- A double heterjuction (DH) structure will confine both hole and electrons to a narrow active layer. Under forward bias, there will be a large number of carriers injected into active region where they are efficiently confined. Carrier recombination occurs in small active region so leading to an efficient device. Antoer advantage DH structure is that the active region has a higher refractive index than the materials on either side, hence light emission occurs in an optical waveguide, which serves to narrow the output beam.

LED configurations

- At present there are two main types of LED used in optical fiber links –
  Surface emitting LED.
  Edge emitting LED.
  Both devices used a DH structure to constrain the carriers and the light to an active layer.

Surface Emitting LEDs

In surface emitting LEDs the plane of active light emitting region is oriented perpendicularly to the axis of the fiber. A DH diode is grown on an N-type substrate at the top of the diode as shown in Fig. 3.1.2. A circular well is etched through the substrate of the device. A fiber is then connected to accept the emitted

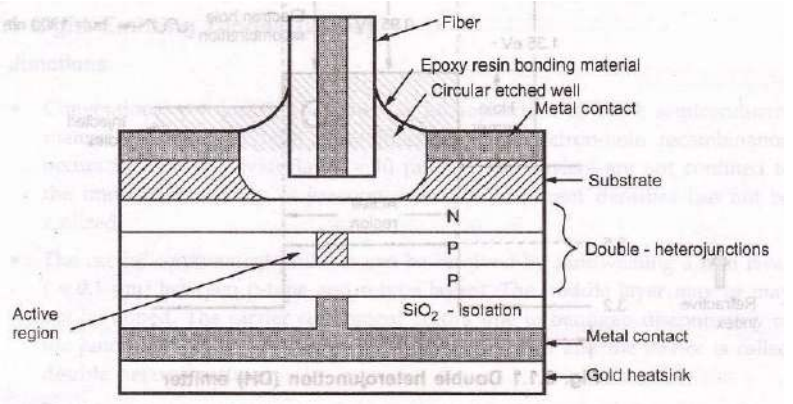


Fig. 3.1.2 Cross-section through a typical surface emitting LED

- At the back of device is a gold heat sink. The current flows through the p-type material and forms the small circular active region resulting in the intense beam of light.
  Diameter of circular active area = 50 μm
  Thickness of circular active area = 2.5 μm
  Current density = 2000 A/cm^2 half-power
  Emission pattern = Isotropic, 120^o beamwidth.
- The isotropic emission pattern from surface emitting LED is of Lambartian pattern. In Lambartian pattern, the emitting surface is uniformly bright, but its projected area diminishes as cos θ, where θ is the angle between the viewing direction and the normal to the surface as shown in Fig. 3.1.3. The beam intensity is maximum along the normal.

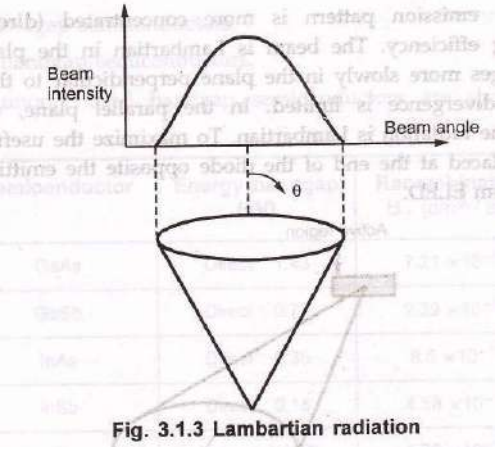


Fig. 3.1.3 Lambartian radiation

- The power is reduced to 50% of its peak when θ = 60^o, therefore the total half-power beamwidth is 120^o. The radiation pattern decides the coupling efficiency of LED.

Edge Emitting LEDS (ELEDs)

- In order to reduce the losses caused by absorption in the active layer and to make the beam more directional, the light is collected from the edge of the LED. Such a device is known as **edge emitting** LED or ELED.

It consists of an active junction region which is the source of incoherent light and two guiding layers. The refractive index of guiding layers is lower than active region but higher than outer surrounding material. Thus a waveguide channel is form and optical radiation is directed into the fiber. Fig. shows structure of LED

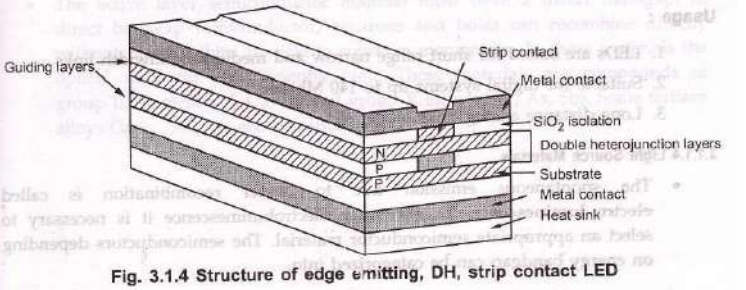


Fig. 3.1.4 Structure of edge emitting, DH, strip contact LED

Edge emitter's emission pattern is more concentrated (directional) providing improved coupling efficiency. The beam is Lambartian in the plane parallel to the junction but diverges more slowly in the plane perpendicular to the junction. In this plane, the beam divergence is limited. In the parallel plane, there is no beam confinement and the radiation is Lambartian. To maximize the useful output power, a reflector may be placed at the end of the diode opposite the emitting edge. Fig. 3.1.5 shows radiation from ELED.

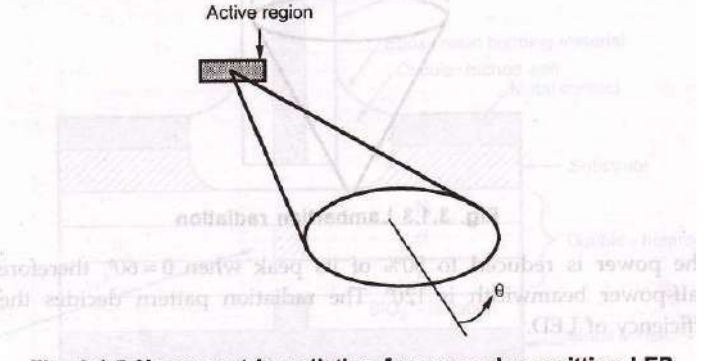


Fig. 3.1.5 Unsymmetric radiation from an edge emitting LED

Features of ELED:

- Linear relationship between optical output and current.
- Spectral width is 25 to 400 nm for $\lambda = 0.8 – 0.9$ µm.
- Modulation bandwidth is much large.

- Not affected by catastrophic gradation mechanisms hence are more reliable.
- ELEDs have better coupling efficiency than surface emitter.
- ELEDs are temperature sensitive.

Usage :

7. LEDs are suited for short range narrow and medium bandwidth links.
8. Suitable for digital systems up to 140 Mb/sec.
9. Long distance analog links

Light Source Materials

10. The spontaneous emission due to carrier recombination is called **electro luminescence**. To encourage electroluminescence it is necessary to select as appropriate semiconductor material. The semiconductors depending on energy bandgap can be categorized into,
    - Direct bandgap semiconductors.
    - Indirect bandgap semiconductors.
11. Some commonly used bandgap semiconductors are shown in following table 3.1.1

Semiconductor	Energy bandgap (eV)	Recombination B_r (cm^3 / sec)
GaAs	Direct : 1.43	7.21×10^{-10}
GaSb	Direct : 0.73	2.39×10^{-10}
InAs	Direct : 0.35	8.5×10^{-11}
InSb	Direct : 0.18	4.58×10^{-11}
Si	Indirect : 1.12	1.79×10^{-15}
Ge	Indirect : 0.67	5.25×10^{-14}
GaP	Indirect : 2.26	5.37×10^{-14}

Table 3.1.1 Semiconductor material for optical sources

15. Direct bandgap semiconductors are most useful for this purpose. In direct bandgap semiconductors the electrons and holes on either side of bandgap have same value of crystal momentum. Hence direct recombination is possible. The recombination occurs within 10^{-8} to 10^{-10} sec.
16. In indirect bandgap semiconductors, the maximum and minimum energies occur at

different values of crystal momentum. The recombination in these semiconductors is quite slow i.e. 10^{-2} and 10^{-3} sec.

The active layer semiconductor material must have a **direct bandgap**. In direct bandgap semiconductor, electrons and holes can recombine directly without need of third particle to conserve momentum. In these materials the optical radiation is sufficiently high. these materials are compounds of group III elements (Al, Ga, In) and group V element (P, As, Sb). Some tertiary allos Ga_{1-x} Al_x As are also used.

5. Emission spectrum of Ga_{1-x} Al_xAs LED is shown in Fig. 3.1.6.

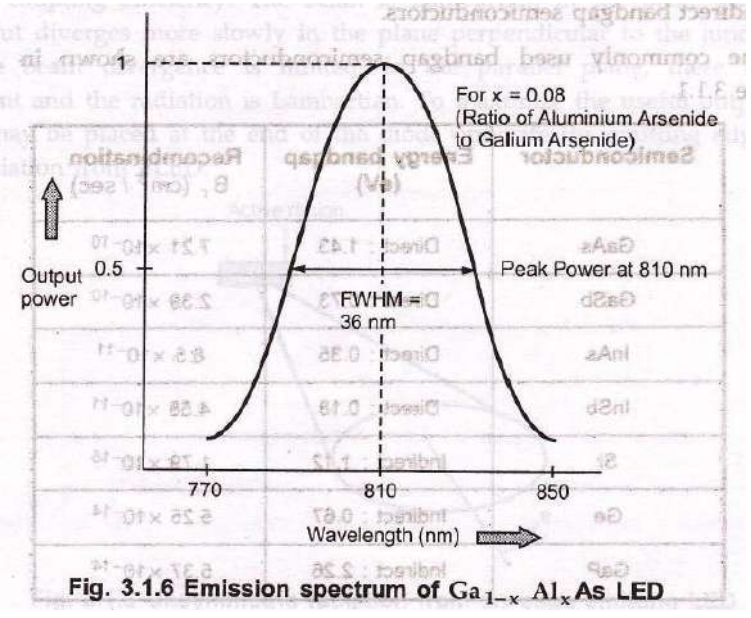


Fig. 3.1.6 Emission spectrum of Ga_{1-x} Al_x As LED

8. The peak output power is obtained at 810 nm. The width of emission spectrum at half power (0.5) is referred as full width half maximum (FWHM) spectral width. For the given LED FWHM is 36 nm.
9. The fundamental quantum mechanical relationship between gap energy E and frequency v is given as –

$$E = hv$$

$$E = h\frac{c}{\lambda}$$

$$\Rightarrow \quad \lambda = \frac{hc}{E}$$

where, energy (E) is in joules and wavelength (λ) is in meters. Expressing the gap energy (E_g) in electron volts and wavelength (λ) in micrometers for this application.

$$\lambda(\mu m) = \frac{1.24}{E_g(eV)}$$

Different materials and alloys have different band gap energies

3. The bandgap energy (E_g) can be controlled by two compositional parameters x and y, within direct bandgap region. The quartenary alloy $In_{1-x}\ Ga_x\ As_y\ P_{1-y}$ is the principal material sued in such LEDs. Two expression relating E_g and x,y are –

$$E_g = 1.424 + 1.266\,x + 0266\,x^2 \quad \ldots 3.1.3$$

$$E_g = 1.35 - 0.72\,y + 0.12\,y^2 \quad \ldots 3.1.4$$

Example 3.1.1 : Compute the emitted wavelength from an optical source having x = 0.07.

Solution : x = 0.07

$$E_g = 1.424 + 1.266\,x + 0266\,x^2$$

$$E_g = 1.424 + (1.266 \times 0.07) + 0266 \times (0.07)^2$$

$E_g = 1.513$ eV

Now $\lambda = \dfrac{1.24}{E_g}$

$$\lambda = \frac{1.24}{1.513}$$

$\lambda = 0.819\ \mu m$

$$\lambda = \mathbf{0.82\ \mu m}$$ **…Ans.**

Example 3.1.2 : For an alloy $In_{0.74}\ Ga_{0.26}\ As_{0.57}\ P_{0.43}$ to be sued in Led. Find the wavelength emitted by this source.

Solution : Comparing the alloy with the quartenary alloy composition.

$In_{1-x}\ Ga_x\ As\ P_{1-y}$ it is found that

x = 0.26 and y= 0.57

$$E_g = 1.35 - 0.72\,y + 0.12\,y^2$$

Using

$$E_g = 1.35-(0.72 \times 0.57) + 0.12 \times 0.57^2$$

$$E_g = 0.978 \text{ eV}$$

Now $$\lambda = \frac{1.24}{E_g}$$

$$\lambda = \frac{1.24}{0.978}$$

$$\lambda = 1.2671\ \mu m$$

$$\lambda = \mathbf{1.27\ \mu m}$$ **… Ans.**

Quantum Efficiency and Power

- The internal quantum efficiency (η_{int}) is defined as the ratio of radiative recombination rate to the total recombination rate.

$$\eta_{int} = \frac{R_r}{R_r + R_{nr}} \quad \text{… 3.1.5}$$

Where,

R_r is radiative recombination rate.

R_{nr} is non-radiative recombination rate.

If n are the excess carriers, then radiative life time, $\tau_r = \frac{n}{R_r}$ and

non-radiative life time, $\tau_r = \frac{n}{R_{nr}}$

The internal quantum efficiency is given

6 The recombination time of carriers in active region is τ. It is also known as bulk recombination life time.

$$\frac{1}{\tau} = \eta_{int} = \frac{1}{1 + \frac{R_{nr}}{R_r}} \quad \text{… 3.1.7}$$

$$\eta_{int} = \frac{1}{1 + \frac{\tau_r}{}}$$

Therefore internal quantum efficiency is given as –

$$\eta_{int} = \frac{\tau}{\tau_r} \qquad \ldots 3.1.8$$

- If the current injected into the LED is I and q is electron charge then total number of recombinations per second is –

$R_r = R_{nr} = \frac{I}{q}$ From equation 3.1.5

$$\eta_{int} = \frac{R_r}{I/q}$$

$\therefore$

$$R_r = \eta_{int} \times \frac{I}{q} \qquad \ldots 3.1.9$$

- Optical power generated internally in LED is given as –

$$P_{int} = R_r . h\nu$$

$$P_{int} = \left(\eta_{int} \times \frac{I}{q}\right) . h\nu$$

$$P_{int} = \left(\eta_{int} \times \frac{I}{q}\right) . h\frac{c}{\lambda}$$

$\therefore$ $$\boxed{P_{int} = \eta_{int} . \frac{hc\, I}{q\lambda}} \qquad \ldots 3.1.10$$

 Not all internally generated photons will available from output of device. The external quantum efficiency is used to calculate the emitted power. The external quantum

efficiency is defined as the ratio of photons emitted from LED to the number of photons generated internally. It is given by equation

$$\eta_{ext} = \frac{1}{n(n+1)^2} \qquad \ldots 3.1.11$$

- The optical output power emitted from LED is given as –

$$P = \eta_{ext} \cdot P_{int}$$

$$\boxed{P = \frac{1}{n\,(n+1)^2} \cdot P_{int}}$$

Example 3.1.3 : The radiative and non radiative recombination life times of minority carriers in the active region of a double heterojunction LED are 60 nsec and 90 nsec respectively. Determine the total carrier recombination life time and optical power generated internally if the peak emission wavelength si 870 nm and the drive currect is 40 mA. **[July/Aug.-2006, 6 Marks]**

Solutions : Given : $\lambda = 870$ nm 0.87×10^{-6} m

$\tau_r = 60$ nsec.

$\tau_{nr} = 90$ nsec.

$I = 40$ mA $= 0.04$ Amp.

i) Total carrier recombination life time:

$$\boxed{\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}}}$$

$$\frac{1}{\tau} = \frac{1}{60} + \frac{1}{90}$$

$$\frac{1}{\tau} = \frac{150}{5400}$$

$\therefore$ $\tau = 36$ **nsec.** **… Ans.**

ii) Internal optical power

$$P_{int} = \eta_{int} \cdot \frac{hc\,I}{q\lambda}$$

$$P_{int} = \left(\frac{\tau}{\tau_r}\right)\left(\frac{hc\,I}{q\lambda}\right)$$

$$P_{int} = \left(\frac{30}{60}\right)\left[\frac{(6.625 \times 10^{-34})(3 \times 10^{8}) \times 0.04}{(1.602 \times 10^{-19})(0.87 \times 10^{-6})}\right]$$

$$P_{int} = 34.22\ mW$$

Example 3.1.4 : A double heterjunciton InGaAsP LED operating at 1310 nm has radiative and non-radiative recombination times of 30 and 100 ns respectively. The current injected is 40 Ma. Calculate –

- Bulk recombination life time.
- Internal quantum efficiency.
- Internal power level.

Solution : $\lambda = 1310$ nm $= (1.31 \times 10^{-6}$ m)

$\tau_r = 30$ ns

$\tau_{nr} = 100$ ns

I = 40 MA – 0.04 Amp.

- **Bulk Recombination Life time (τ) :**

$$\eta_{int} = \frac{\tau}{\tau_r}$$

$$\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}}$$

∴ $\tau = 23.07$ **nsec.** **… Ans.**

- **Internal euqntum efficienty (η_{int})**

$$\eta_{int} = \frac{23.07}{30}$$

$$\eta_{int} = \mathbf{0.769}$$

… Ans.

iii) Internal pwer level (P_{int}) :

$$P_{int} = \eta_{int} \cdot \frac{hc\,I}{q\lambda}$$

Advantages and Disadvantages of LED

Advantages of LED

- Simple design.
- Ease of manufacture.
- Simple system integration.
- Low cost.
- High reliability.

Disadvantages of LED

- Refraction of light at semiconductor/air interface.
- The average life time of a radiative recombination is only a few nanoseconds, therefore nodulation BW is limited to only few hundred megahertz.
- Low coupling efficiency.
- Large chromatic dispersion.

Comparison of Surface and Edge Emitting LED

LED type	Maximum modulation frequency (MHz)	Output power (mW)	Fiber coupled power (mW)
Surface emitting	60	< 4	< 0.2
Edge emitting	200	< 7	< 1.0

Injection Laser Diode (ILD)

- The laser is a device which amplifies the light, hence the LASER is an acronym for light amplification by stimulated emission of radiation.

  The operation of the device may be described by the formation of an electromagnetic standing wave within a cavity (optical resonator) which provides an output of monochromatic highly coherent radiation.

Principle :

Material absorb light than emitting. Three different fundamental process occurs between the two energy states of an atom.

Absorption 2) Spontaneous emission 3) Stimulated emission.

- Laser action is the result of three process absorption of energy packets (photons) spontaneous emission, and stimulated emission. (These processes are represented by the simple two-energy-level diagrams).

  Where E_1 is the lower state energy level.

  E_2 is the higher state energy level.

- Quantum theory states that any atom exists only in certain discrete energy state, absorption or emission of light causes them to make a transition from one state to another. The frequency of the absorbed or emitted radiation f is related to the difference in energy E between the two states.

  If E_1 is lower state energy level.

  and E_2 is higher state energy level.

  $E = (E_2 - E_1) = h.f.$

  Where, $h = 6.626 \times 10^{-34}$ J/s (Plank's constant).

- An atom is initially in the lower energy state, when the photon with energy $(E_2 - E_1)$ is incident on the atom it will be excited into the higher energy state E_2 through the absorption of the photon

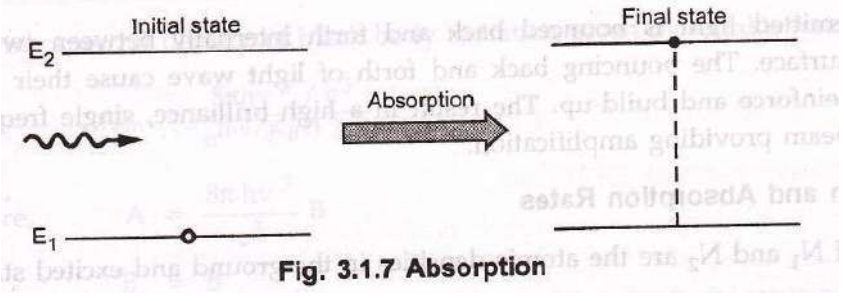


Fig. 3.1.7 Absorption

- When the atom is initially in the higher energy state E_2, it can make a transition to the lower energy state E_1 providing the emission of a photon at a frequency corresponding to $E = h.f.$ The emission process can occur in two ways.

  By spontaneous emission in which the atom returns to the lower energy state in random manner.

  By stimulated emission when a photon having equal energy to the difference between the two states $(E_2 - E_1)$ interacts with the atom causing it to the lower state with the creation of the second photon

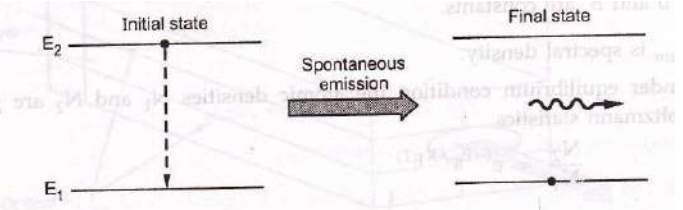


Fig. 3.1.8 Spontaneous emission

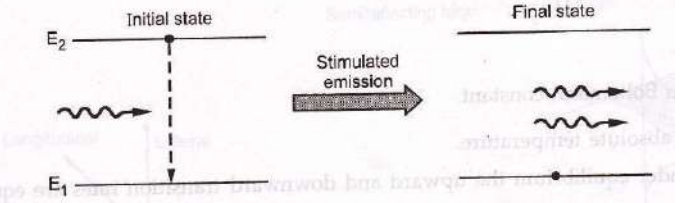


Fig. 3.1.9 Stimulated emission

- Spontaneous emission gives incoherent radiation while stimulated emission gives coherent radiation. Hence the light associated with emitted photon is of same frequency of incident photon, and in same phase with same polarization.
- It means that when an atom is stimulated to emit light energy by an incident wave, the liberated energy can add to the wave in constructive manner. The emitted light is bounced back and forth internally between two reflecting surface. The bouncing back and forth of light wave cause their intensity to reinforce and build-up. The result in a high brilliance, single frequency light beam providing amplification.

Emission and Absorption Rates

3. It N_1 and N_2 are the atomic densities in the ground and excited states.

Rate of spontaneous emission

$$R_{spon} = AN_2 \qquad \ldots 3.1.13$$

Rate of stimulated emission

$$R_{stim} = BN_2 \rho_{em} \qquad \ldots 3.1.14$$

Rate of absorption

$$R_{abs} = B' N_1 \rho_{em} \qquad \ldots 3.1.15$$

where,

A, B and B' are constants.

ρ_{em} is spectral density.

- Under equilibrium condition the atomic densities N_1 and N_2 are given by Boltzmann statistics.

$$\frac{N_2}{N_1} = eg^{(-E_B / K_B T)} \quad \ldots 3.1.16$$

$$\frac{N_2}{N_1} = eg^{(-h_v / K_B T)} \quad \ldots 3.1.17$$

where,

K_B is Boltzmann constant.

T is absolute temperature.

- Under equilibrium the upward and downward transition rates are equal.

$AN_2 + BN_2 \rho_{em} = B' N_1 \rho_{em}$ … 3.1.18

Spectral density ρ_{em}

… 3.1.19

Comparing spectral density of black body radiation given by Plank's formula,

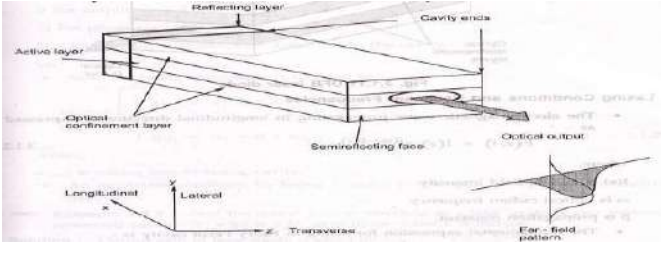
… 3.1.20

Therefore, … 3.1.21

- A and B are called Einstein's coefficient.

Fabry – Perot Resonator

- Lasers are oscillators operating at frequency. The oscillator is formed by a resonant cavity providing a selective feedback. The cavity is normally a Fabry-Perot resonator i.e. two parallel plane mirrors separated by distance L,



Light propagating along the axis of the interferometer is reflected by the mirrors back to the amplifying medium providing optical gain. The dimensions of cavity are 25-500 μm longitudinal 5-15 μm lateral and 0.1-0.2 μm transverse. Fig. 3.1.10 shows Fabry-Perot resonator cavity for a laser diode.

- The two heterojunctions provide carrier and optical confinement in a direction normal to the junction. The current at which lasing starts is the threshold current. Above this current the output power increases sharply.

Distributed Feedback (DFB) Laser

In DFB laster the lasing action is obtained by periodic variations of refractive index along the longitudinal dimension of the diode. Fig. 3.1.11 shows the structure of DFB laser diode

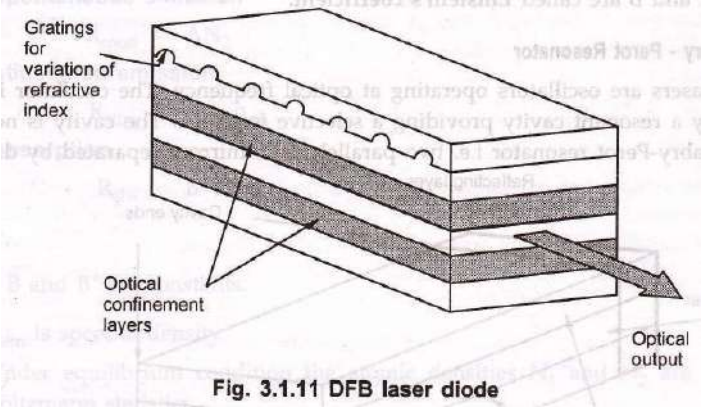


Fig. 3.1.11 DFB laser diode

Lasing conditions and resonant Frequencies

- The electromagnetic wave propagating in longitudinal direction is expressed as –

$$E(z, t) = I(z)\, e^{j(\omega t - \beta z)} \qquad \ldots 3.1.23$$

where,

I(z) is optical field intensity.

is optical radian frequency.

β is propagation constant.

The fundamental expression for lasing in Fabry-Perot cavity is –

$$I(z) = I(0)e^{[\{\Gamma\, g(h\,v)-\alpha(h\,v)\}z]} \quad \ldots 3.1.24$$

where,

□is optical field confinement factor or the fraction of optical power in the active layer.

α is effective absorption coefficient of material.

g is gain coefficient.

h v is photon energy.

z is distance traverses along the lasing cavity.

The condition of lasing threshold is given as –

- □ For amplitude : I (2L) = I (0)
- □ For phase : $e\text{-}j^{2\beta L} = 1$
- □ Optical gain at threshold = Total loss in the cavity.

i.e. $\Gamma\, g_{th} = \alpha_t$

- Now the lasing expression is reduced to –

$$\boxed{\Gamma\, g_{th} = a_t = \alpha + \frac{1}{2L}\ln\left(\frac{1}{R_1 R_2}\right)} \quad \ldots 3.1.26$$

$$\Gamma g_{th} = \alpha_t = \alpha + \alpha_{end} \quad \ldots 3.1.27$$

where,

A_{end} is mirror loss in lasing cavity.

- An important condition for lasing to occur is that gain, g ≥ g th i.e. threshold gain.

Example 3.1.5 : Find the optical gain at threshold of a laser diode having following parametric values – $R_1 = R_2 = 0.32$, $\alpha = 10cm^{-1}$ and L = 500 μm.

Solution : Optical gain in laser diode is given by –

$$\Gamma g_{th} = 10 + \frac{1}{2 \times (500 \times 10^{-4})} \ln\left(\frac{1}{0.32 \times 0.32}\right)$$

$$\Gamma g_{th} = 33.7 \text{ cm}^{-1}$$

… Ans.

Power Current Characteristics

The output optic power versus forward input current characteristics is plotted in Fig. 3.1.12 for a typical laser diode. Below the threshold current (I_{th}) only spontaneous emission is emitted hence there is small increase in optic power with drive current. at threshold when lasing conditions are satisfied. The optical power increases sharply after the lasing threshold because of stimulated emission.

- The lasing threshold optical gain (g_{th}) is related by threshold current density (J_{th}) for stimulated emission by expression –

$$g_{th} = \beta J_{th} \quad \ldots 3.1.28$$

where, β is constant for device structure.

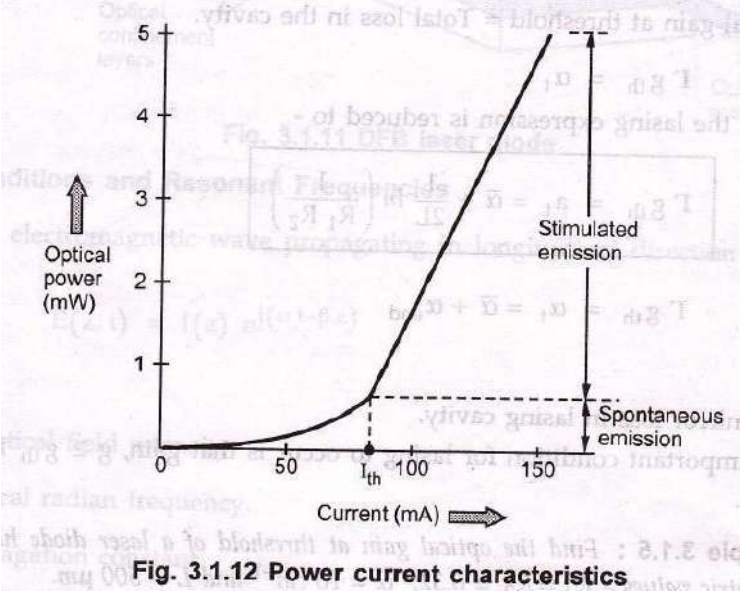


Fig. 3.1.12 Power current characteristics

External Quantum Efficiency

- The external quantum efficiency is defined as the number of photons emitted per electron hole pair recombination above threshold point. The external quantum efficiency η_{ext} is given by –

$$\eta_{ext} = \frac{\eta_i(g_{th} - \alpha)}{g_{th}}$$

… 3.1.29

where,

η_i = Internal quantum efficiency (0.6-0.7).

g_{th} = Threshold gain.

α = Absorption coefficient

- Typical value of η_{ext} for standard semiconductor laser is ranging between 15-20 %.

Resonant Frequencies

- At threshold lasing

$$2\beta L = 2\pi m$$

where, $\beta = \frac{2\pi m}{\lambda}$ (propagation constant)

m is an integer.

$\therefore$ $m = 2L.\frac{n}{\lambda}$ ...3.1.30

Since $c = v\lambda$

$\therefore$ $\lambda = \frac{c}{v}$

Substituting λ in 3.1.30

$$m = 2L\,\frac{nv}{c}$$ =z… 3.1.31

- Gain in any laser is a function of frequency. For a Gaussian output the gain and frequency are related by expression –

$$g(\lambda) = g(0)e^{\left[-\frac{(\lambda-\lambda_0)^2}{2\sigma^2}\right]} \quad \ldots 3.1.32$$

where,

g(0) is maximum gain.

λ_0 is center wavelength in spectrum.

 is spectral width of the gain.The frequency spacing between the two successive modes is –

$$\Delta v = \frac{c}{2Ln}$$

$$\Delta\lambda = \frac{\lambda^2}{2Ln} \quad \ldots 3.1.34$$

Optical Characteristics of LED and Laser

- The output of laser diode depends on the drive current passing through it. At low drive current, the laser operates as an inefficient Led, When drive current crosses threshold value, lasing action beings. Fig. 3.1.13 illustrates graph comparing optical powers of LED operation (due to spontaneous emission) and laser operation (due to stimulated emission).

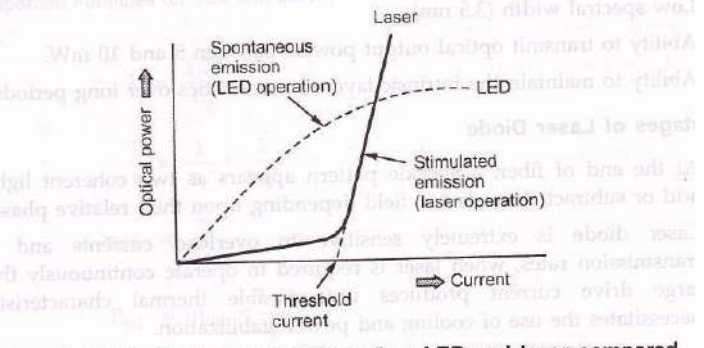


Fig. 3.1.13 Optical characteristics of an LED and laser compared

Spectral and Spatial Distribution of Led and Laser

 At low current laser diode acts like normal LED above threshold current, stimulated emission i.e. narrowing of light ray to a few spectral lines instead of broad spectral distribution, exist. This enables the laser to easily couple to single mode fiber and reduces the amount of uncoupled light (i.e. spatial radiation distribution). Fig. 3.1.14 shows spectral and spatial distribution difference between two diodes

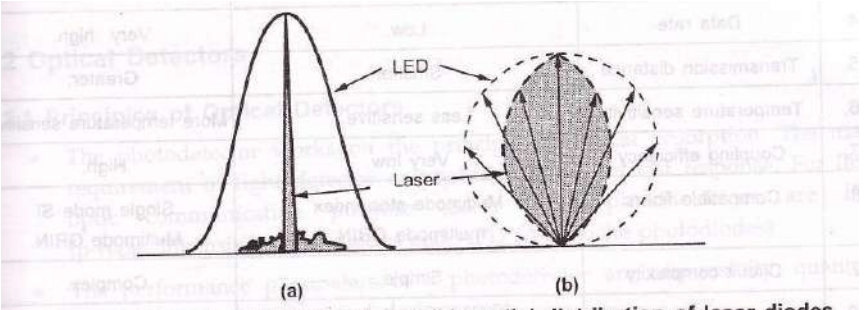


Fig. 3.1.14 Comparing (a) spectral and (b) spatial distribution of laser diodes

Advantages and Disadvantages of Laser Diode

Advantages of Laser Diode

- Simple economic design.
- High optical power.
- Production of light can be precisely controlled.
- Can be used at high temperatures.
- Better modulation capability.
- High coupling efficiency.
- Low spectral width (3.5 nm)
- Ability to transmit optical output powers between 5 and 10 mW.
- Ability to maintain the intrinsic layer characteristics over long periods.

Disadvantages of Laser Diode

- At the end of fiber, a speckle pattern appears as two coherent light beams add or subtract their electric field depending upon their relative phases.
- Laser diode is extremely sensitive to overload currents and at high transmission rates, when laser is required to operate continuously the use of large drive current produces unfavourable thermal characteristics and necessitates the use of cooling and power stabilization.

Comparison of LED and Laser Diode

Sr. No.	Parameter	LED	LD (Laser Diode)
1.	Principle of operation	Spontaneous emission.	Stimulated emission.
2.	Output beam	Non – coherent.	Coherent.
3.	Spectral width	Board spectrum (20 nm – 100 nm)	Much narrower (1-5 nm).

4.	Data rate	Low.	Very high.
5.	Transmission distance	Smaller.	Greater.
6.	Temperature sensitivity	Less sensitive.	More temperature sensitive.
7.	Coupling efficiency	Very low.	High.

8.	Compatible fibers	Multimode step index multimode GRIN.	Single mode Sl Multimode GRIN.
9.	Circuit complexity	Simple	Complex
10.	Life time	10^5 hours.	10^4 hours.
11.	Cost	Low.	High.
12.	Output power	Linearly proportional to drive current.	Proportional to current above threshold.
13.	Current required	Drive current 50 to 100 mA peak.	Threshold current 5 to 40 mA.
14.	Wavelengths available	0.66 to 1.65 µm.	0.78 to 1.65 µm.
15.	Applications	Moderate distance low data rate.	Long distance high data rates.

Important Formulae for LED and Laser

LED

1. $\lambda = \frac{1.24}{E_g}$
2. $\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_{nr}}$
3. $\eta_{int} = \frac{\tau}{\tau_r}$
4. $P_{int} = \eta_{int} \times \frac{hcI}{q\lambda}$

LASER

1. $\Gamma g_{th} = \alpha + \frac{1}{2L} \ln\left(\frac{1}{R_1 R_2}\right)$
2. $\Delta v = \frac{c}{2Ln}$
3. $\Delta\lambda = \frac{\lambda^2}{2Ln}$

Optical Detectors

Principles of Optical Detectors

- The photodetector works on the principle of optical absorption. The main requirement of light detector or photodector is its fast response. For fiber optic communication purpose most suited photodetectors are PIN (p-type- Instrinsic-n-type) diodes and APD (Avalanche photodiodes)
- The performance parameters of a photodetector are responsivity, quantum efficiency, response time and dark current.

Cut-off Wavelength (λ_c)

- Any particular semiconductor can absorb photon over a limited wavelength range. The highest wavelength is known as cut-off wavelength (λ_c). The cut-off wavelength is determined by bandgap energy E_g of material.

$$\lambda_c = \frac{hc}{E_g} = \frac{1.24}{E_g} \quad \ldots 3.2.1$$

where,

E_g inelectron volts (eV) and

λ_c cut-off wavelength is in µm.

Typical value of λ_c for silicon is 1.06 µm and for germanium it is 1.6 µm.

Quantum Efficiency (η)

- The quantum efficiency is define as the number of electron-hole carrier pair generated per incident photon of energy h v and is given as –

$$\eta = \frac{\text{Number of electron hole pairs generated}}{\text{Number of incident photons}}$$

$$\eta = \frac{I_p / q}{P_{in} / h\,v} \quad \ldots 3.2.2$$

where, I_p is average photocurrent.

P_{in} is average optical power incident on photo detectors.

- Absorption coefficient of material determines the quantum efficiency. Quantum efficiency $\eta < 1$ as all the photons incident will not generate e-h pairs. It is normally expressed in percentage.

Fiber Alignment

5. In any fiber optic communication system, in order to increase fiber length there is need to joint the length of fiber. The interconnection of fiber causes some loss of optical power. Different techniques are used to interconnect fibers. A permanent joint of cable is referred to as **splice** and a temporary joint can be done with the connector.
6. The fraction of energy coupled from one fiber to other proportional to common mode volume M_{common}. The fiber – to – fiber coupling efficiency is given as –

$$\eta_F = \frac{M_{common}}{M_E}$$

where,

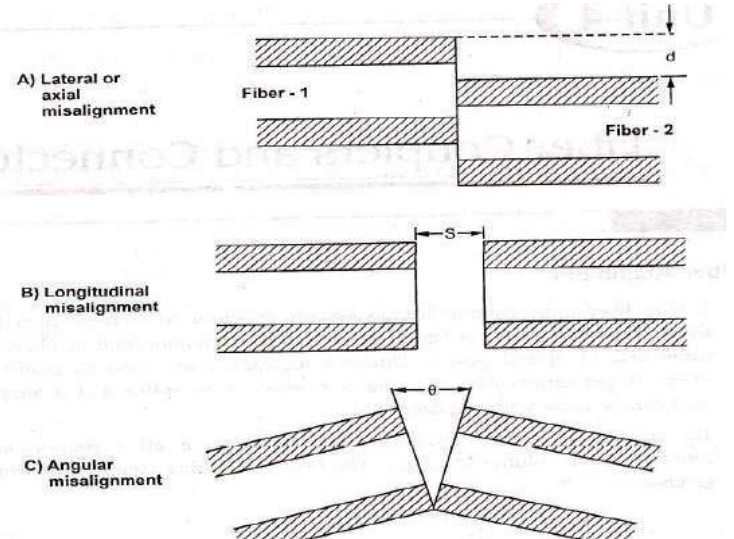
M_E is number of modes in fiber which launches power into next fiber.

- The fiber – to – fiber coupling loss L_F is given as –

$$L_F = -10\log \eta_F$$

Mechanical Misalignment

The diameter of fiber is few micrometer hence the microscopic alignment is required. If the radiation cone of emitting fiber does not match the acceptance cone of receiving fiber, radiation loss takes place. The magnitude of radiation loss depends on the degree of misalignment. Different types of mechanical misalignments are shown in Fig. 4.1.1.



Lateral misalignment

Lateral or axial misalignment occurs when the axes of two fibers are separated by distance 'd'.

Longitudinal misalignment

Longitudinal misalignment occurs when fibers have same axes but their end faces are separated by distance 'S'.

Angular misalignment

Angular misalignment occurs when fiber axes and fiber end faces are no longer parallel. There is an angle 'θ' between fiber end faces.

The axial or lateral misalignment is most common in practice causing considerable power loss. The axial offset reduces the common core area of two fiber end faces as shown in

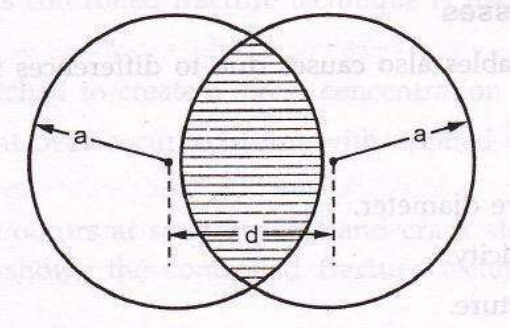


Fig. 4.1.2 Axial offset of fibers

- The optical power coupled is proportional to common area of two fiber cores. The common area is given by expression –

$$A_{common} = 2a^2 \text{ across}\frac{d}{2a} - d\left(a^2 - \frac{d^2}{n}\right)^{1/2} \quad \ldots (4.1.3)$$

where,

a is core radius of fiber.

d is separation of core axes.

- The coupling efficiency for step index fiber is the ratio of common core area to the end-face area.

$$\eta_{step} = \frac{A_{common}}{\pi a^2}$$

$$\eta_{step} = \frac{2}{\pi} \operatorname{across} \frac{d}{2a} - \frac{d}{\pi a}\left[1 - \left(\frac{d}{2a}\right)^2\right]^{1/2}$$

- For graded index fiber, the total received power for axial misalignment is given by –

$$P_T = \frac{2}{\pi} P \left\{ \operatorname{across} \frac{d}{2a} - \left[1 - \left(\frac{d}{2a}\right)^2\right]^{1/2} \frac{d}{6a}\left(5 - \frac{d^2}{2a^2}\right)\right\}$$ .1.5)

where,

P is the power in emitting fiber.

When, d << a, the above expression reduces … (4.1.6)

Fiber Related Losses

- Losses in fiber cables also causes due to differences in geometrical and fiber characteristics.
  These includes,
  1) Variation in core diameter.
  2) Core area ellipticity.
  3) Numerical aperture.
  4) Refractive – index profile.
  5) Core-cladding concentricity.

  The user have less control over these variations since they are related to manufacturing process.

- Coupling loss when emitter fiber radius a_E and receiving fiber radius a_R is not same, is given as –

$$L_F(a) = \begin{cases} -10 \log \left(\frac{a_R}{a_E}\right)^2 & for\ a_R < a_E \\ 0 & for\ a_R \geq a_E \end{cases} \quad (4.1.7)$$

where,

a_E is emitter fiber radius.

a_R is receiver fiber radius.

- Coupling loss when numerical apertures of two fibers are not equal, to expressed as –

$$L_F(NA) = \begin{cases} -10 \log \left[\frac{NA_R\ (0)}{NA_E\ (0)}\right]^2, & for\ NA_R < NA_E \\ 0, & for\ NA_R \geq NA_E \end{cases} \quad (4.1.8)$$

- Coupling loss when core refractive index of two fibers are not same, is expressed as

$$L_F(\propto) = \begin{cases} -10 \log \frac{\propto_R (\propto_E + 2)}{\propto_E (\propto_R + 2)}, & for\ \propto_R < \propto_E \\ 0, & for\ \propto_R \geq \propto_E \end{cases} \quad (4.1.9)$$

Fiber Splices

- A permanent or semipermanent connection between two individual optical fibers is known as **fiber splice**. And the process of joining two fibers is called as **splicing**.
- Typically, a splice is used outside the buildings and connectors are used to join the cables within the buildings. Splices offer lower attenuation and lower back reflection than connectors and are less expensive.

Types of Splicing

- There are two main types of splicing
  Fusion splicing.
  Mechanical splicing / V groove

Fusion Splicing

10. Fusion splicing involves butting two cleaned fiber end faces and heating them until they melt together or fuse.
11. Fusion splicing is normally done with a fusion splicer that controls the alignment of the two fibers to keep losses as low as 0.05 dB.

6. Fiber ends are first prealigned and butted together under a microscope with micromanipulators. The butted joint is heated with electric arc or laser pulse to melt the fiber ends so can be bonded together. Fig. 4.2.1 shows fusion splicing of optical fiber

Mechanical Splicing / V Groove

12. Mechanical splices join two fibers together by clamping them with a structure or by epoxying the fibers together.
13. Mechanical splices may have a slightly higher loss and back reflection. These can be reduced by inserting index matching gel.
14. V groove mechanical splicing provides a temporary joint i.e fibers can be disassembled if required. The fiber ends are butted together in a V – shaped groove as shown in Fig..

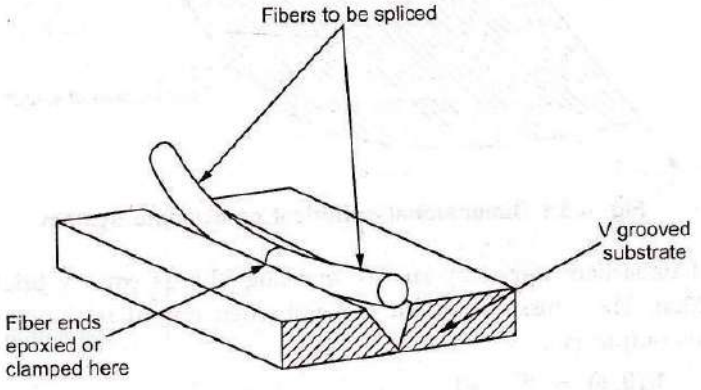


Fig. 4.2.2 V groove optical fiber splicing technique

17. The splice loss depends on fiber size and eccentricity.

Source-to-Fiber Power Launching

Optical output from a source is measured in radiance (B). Radiance is defined as the optical power radiated into a solid angle per unit emitting surface area. Radiance is specified in Watts/cm^2/Steradian. Radiance is important for defining source to fiber coupling efficiency.

Source Output Pattern

10. Spatial radiation pattern of source helps to determine the power accepting capability of fiber.
11. Fig. 4.3.1 shows three dimensional spherical co-ordinate system for characterizing the emission pattern from an optical source. Where the polar axis is normal to the emitting surface and radiance is a function of θ and ϕ.

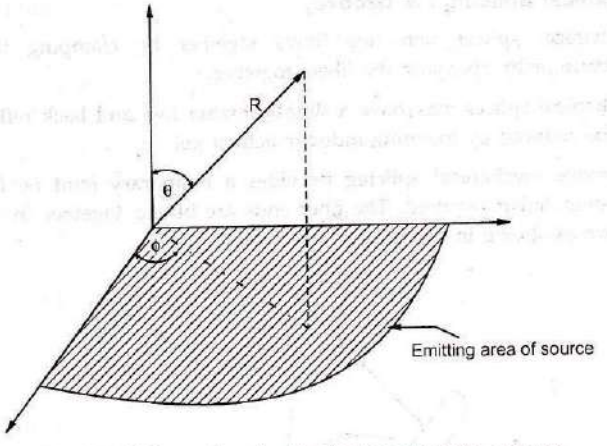


Fig. 4.3.1 Dimensional spherical co-ordinate system

4. The Lambartian output by surface emitting LED is equally bright from any direction. The emission pattern of Lambartian output is shown in Fig. 4.3.2 and its output is –

$$B(\theta,\phi) = B_0 \cos\theta$$

where, B_0 is the radiance along the normal to the radiating surface.

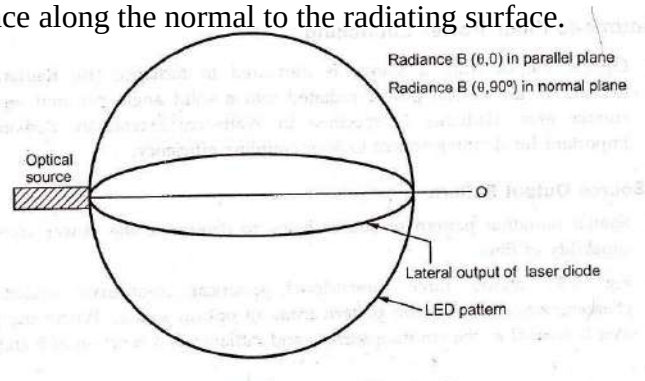


Fig. 4.3.2 Radiance pattern of Lambartian source

- Both radiations in parallel and normal to the emitting plane are approximated by expression –

$$\frac{1}{B(\theta,\phi)} = \frac{\sin^2\phi}{B_0 \cos^T\theta} + \frac{\cos^2\phi}{B_0 \cos^L\theta} \quad \ldots (4.3.2)$$

where,

T and L are transverse and lateral power distribution coefficients.

Power Coupling Calculation

- To calculate power coupling into the fiber, consider an optical source launched into the fiber as shown in Fig. 4.3.3.

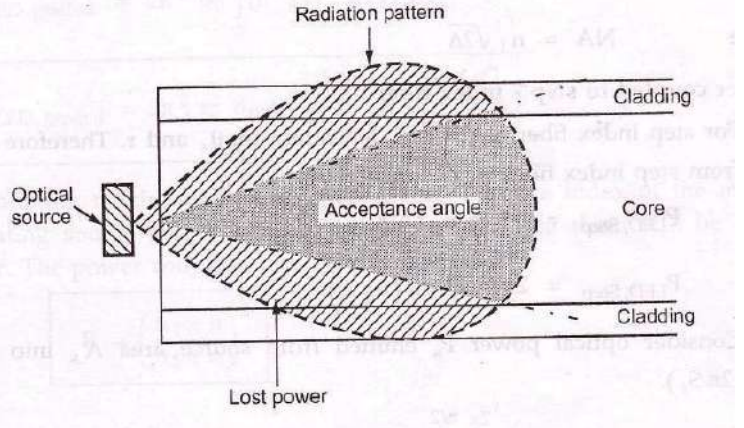


Fig. 4.3.3 Optical source coupled to fiber

Brightness of source is expressed as $B(A_S, \Omega_S)$,

Where, A_S is area of source

Ω_S is solid emission angle of source.

The coupled power P can be calculated as –

$$P = \int_{A_f} dA_s \int_{\Omega_2} d\Omega_s \; B(A_s \Omega_s)$$

$$P = \int_0^r \int_0^{2\pi} \left[\int_0^{2\pi^\phi} \int_0^{0_{max}} B(\theta,\phi) \sin\theta \, d\theta \, d\phi \right] d\theta_s \, r dr \quad \ldots (4.3.3)$$

The integral limits are area of source and solid acceptance angle (θ_{0max}).

Here $d\theta_s$ rdr is incremental emitting area.

- Let the radius of surface emitting LED is rs, and for Lambartian emitter, $B(\theta,\phi) = B_0 \cos\theta$, then

$$P = \int_0^r \int_0^{2\pi} \left(2\pi B_0 \int_0^{0_{max}} \cos\theta \sin\theta \, d\theta\right) d\theta_s \, rdr$$

$$P = B_0 . \pi \int_0^{r_s} \int_0^{2\pi} \sin^2\theta \; 0_{max} \, d\theta_s \, rdr$$

$$P = B_0 . \pi \int_0^{r_s} \int_0^{2\pi} NA^2 \, d\theta_s \, rdr \qquad \ldots (4.3.4)$$

Since $NA = n_1\sqrt{2\Delta}$

Power coupled to step – index fiver

- For step index fiber NA is not dependent on θ_S and r. Therefore LED power from step index fiber is,

$$P_{LED}, Step = \pi^2 r_s^2 B_0 (NA)^2$$

$$P_{LED}, Step = 2\pi^2 r_s^2 B_0 (n_1^2 \Delta) \qquad \ldots (4.3.5)$$

- Consider optical power P_S emitted from source are A_S into hemisphere($2\pi S_r$).

$$P = A_s \int_0^{2\pi} \int_0^{\pi/2} B(\theta,\phi) \sin\theta \, d\theta \, d\phi$$

$$P = \pi r_s^2 2\pi B_o \int_0^{\pi/2} \cos\theta \sin\theta \, d\theta$$

- When source radius r_S <a, the fiber core radius, the LED output power is given from equation (4.3.5).

$$P_s, Step = P_s (NA)^2 \qquad \ldots (4.3.7)$$

- When r_S > a equation (4.3.5) becomes,

$$\boxed{P_{LED}, Step = \left(\frac{a}{r_s}\right)^2 . P_s (NA)^2} \qquad \ldots (4.3.8)$$

Power coupled to graded index fiber

- In graded index fiber, the index of refraction varies radially from fiber axis. Numerical aperture for graded index fiber is given by,

$$P_{LED,Step} = 2\pi^2 B_0 \int_0^{r_s} [n^2(r) - n_2^2] r dr$$

Is source radius (rs) is less than fiber core radius (a) i.e. $r_s < a$, the power coupled from surface emitting LED is given as –

$$P_{LED,Step} = 2\pi^2 B_0 r_s^2 n_1^2 \Delta \left[\frac{2}{\alpha+2}\left(\frac{r_s}{a}\right)^{\alpha}\right]$$

- For coupling maximum power to fiber, the refractive index of the medium separating source and fiber must be same, otherwise there will be loss of power. The power couple is reduced by factor,

$$R = \left(\frac{n_1 - n}{n_1 + n}\right)^2$$

where,

n is the refractive index of medium.

n_1 is the refractive index of fiber core. R is the Fresnel reflection or reflectivity

Lensing Schemes for Coupling Improvement

- When When the emitting area of the source is smaller than the core area of fiber, the power coupling efficiency becomes poor. In order to improve the coupling efficiency miniature lens is placed between source and fiber. Microlens magnifies the emitting area of source equal to core area. The power coupled increases by a factor equal to magnification factor of lens.

- Important types of lensing schemes are :
  Rounded – end fiber.
  Spherical – surfaced LED and Spherical-ended fiber.
  Taper ended fiber.
  Non imaging microsphere.
  Cylindrical lens,
  Imaging sphere.

- There are some drawbacks of using lens.

Complexity increases.
Fabrication and handling difficulty.

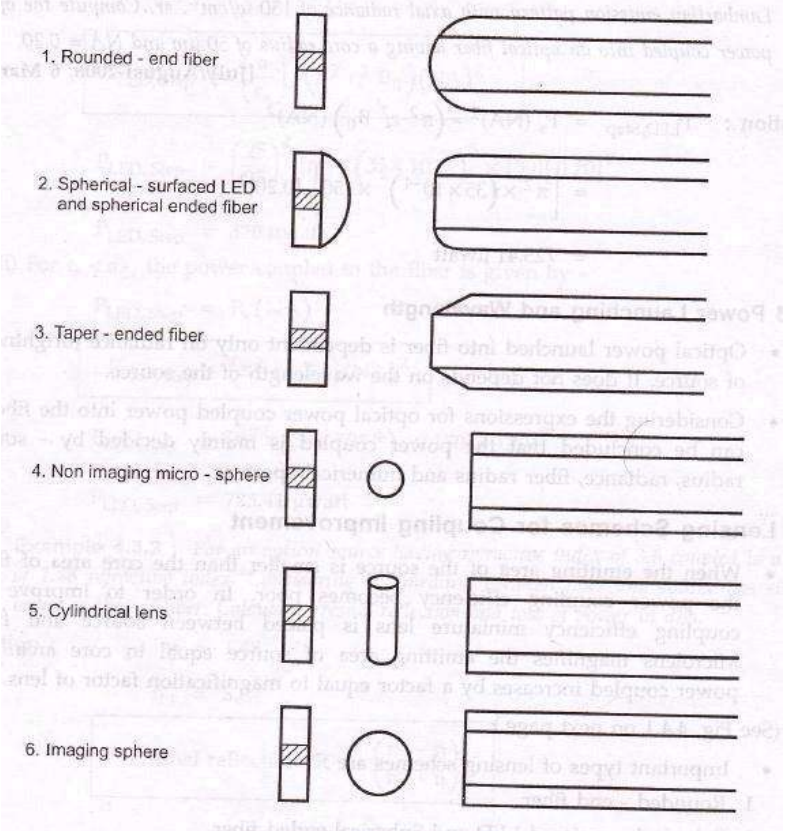


Fig. 4.4.1 Lensing schemes

UNIT-IV

Fiber Optical Receivers

Detector Responsivity ($\Re$)

- The responsivity of a photodetector is the ratio of the current output in amperes to the incident optical power in watts. Responsivity is denoted by $\Re$.

$$\Re = \frac{I_p}{P_{in}} \qquad \ldots 3.2.3$$

But $$\eta = \frac{I_p - q}{P_{in} - h\nu} = \frac{I_p}{q}\frac{h\nu}{P_{in}}$$

$\therefore$ $$\frac{I_p}{P_{in}} = \frac{\eta q}{h\nu} \qquad \ldots 3.2.4$$

Therefore $$\boxed{\Re = \frac{\eta q}{h\nu} = \frac{\eta q^{\lambda}}{h\nu}} \qquad \therefore \nu = \frac{c}{\lambda} \qquad \ldots 3.2.5$$

- Responsivity gives transfer characteristics of detector i.e. photo current per unit incident optical power.
- Typical responsivities of pin photodiodes are –
  Silicon pin photodiode at 900 nm →0.65 A/W.
  Germanium pin photodiode at 1.3 μm →0.45 A/W.
  In GaAs pin photodiode at 1.3 μm →0.9 A/W.

$$\eta = \frac{5.4 \times 10^6}{6 \times 10^6}$$

$$\eta = 0.9 = 90\ \% \qquad \textbf{… Ans.}$$

- r photodetectors are sued. As the intensity of optical signal at the receiver is very low, the detector has to meet high performance specifications.

The conversion efficiency must be high at the operating wavelength.

The speed of response must be high enough to ensure that signal distortion does not occur

The detection process introduce the minimum amount of noise.

It must be possible to operate continuously over a wide range of temperatures for many years.

The detector size must be compatible with the fiber dimensions.

4. At present, these requirements are met by reverse biased p-n photodiodes. In these devices, the semiconductor material absorbs a photon of light, which excites an electron from the valence band to the conduction band (opposite of photon emission). The photo
5. generated electron leaves behind it a hole, and so each photon generates two charge carriers. The increases the material conductivity so call **photoconductivity** resulting in an increase in the diode current. The diode equation is modified as –

$$I_{diode} = (I_d + I_s)\left(e^{V_q/\eta k T} - 1\right) \quad \ldots 3.2.6$$

where,

I_d is dark current i.e. current that flows when no signal is present.

I_s is photo generated current due to incident optical signal.

Fig. 3.2.1 shows a plot of this equation for varying amounts of incident optical power.

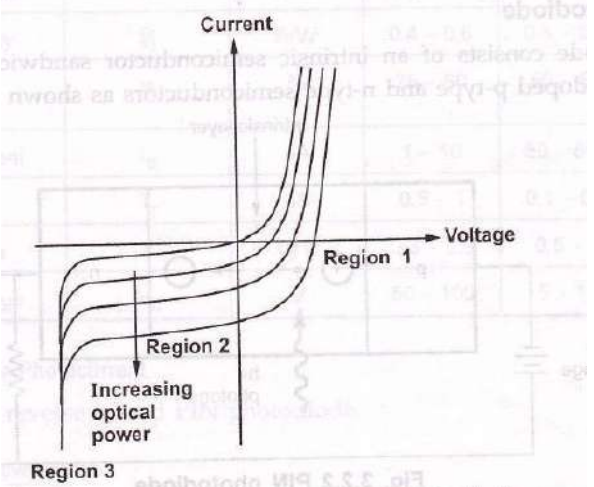


Fig. 3.2.1 V-I characteristics of photodiode

- Three regions can be seen forward bias, reverse bias and avalanche breakdown.

Forward bias, region 1 : A change in incident power causes a change in terminal voltage, it is called as **photovoltaic mode**. If the diode is operated in this mode, the frequency response of the diode is poor and so photovoltaic operation is rarely used in optical links

Reverse bias, region 2 : A change in optical power produces a proportional change in diode current, it is called as **photoconductive mode** of operation which most detectors use. Under these condition, the exponential term in equation 3.2.6 becomes insignificant and the reverse bias current is given by –

$$I_{diode} = -(I_0 + I_L)$$

- **Responsivity** of photodiode is defined as the change in reverse bias current per unit change in optical powr, and so efficient detectors need large responsivities.

Avalanche breakdown, region 3 : When biased in this region, a photo generated electron-hole pair causes avalanche breakdown, resulting in large diode for a single incident photon. Avalance photodiodes (APDs) operate in this region APDs exhibit carrier multiplication. They are usually very sensitive detectors. Unfortunately V-I characteristic is very steep in this region and so the bias voltage must be tightly controlled to prevent spontaneous breakdown.

PIN Photodiode

- PIN diode consists of an intrinsic semiconductor sandwiched between two heavily doped p-type and n-type semiconductors as shown in Fig. 3.2.2

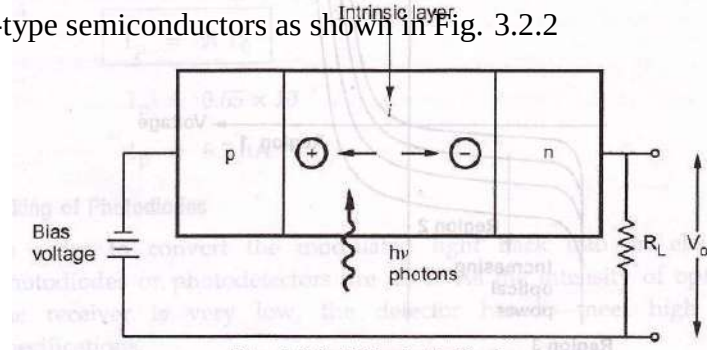


Fig. 3.2.2 PIN photodiode

- Sufficient reverse voltage is applied so as to keep intrinsic region free from carries, so its resistance is high, most of diode voltage appears across it, and the electrical forces are strong within it. The incident photons give up their energy and excite an electron from valance to conduction band. Thus a free electron hole pair is generated, these are

as **photocarriers**. These carriers are collected across the reverse biased junction resulting in rise in current in external circuit called **photocurrent**.

In the absence of light, PIN photodiodes behave electrically just like an ordinary rectifier diode. If forward biased, they conduct large amount of current

PIN detectors can be operated in two modes : **Photovoltaic** and **photoconductive**. In photovoltaic mode, no bias is applied to the detector. In this case the detector works very slow, and output is approximately logarithmic to the input light level. Real world fiber optic receivers never use the photovoltaic mode.

In photoconductive mode, the detector is reverse biased. The output in this case is a current that is very linear with the input light power.

The intrinsic region some what improves the sensitivity of the device. It does not provide internal gain. The combination of different semiconductors operating at different wavelengths allows the selection of material capable of responding to the desired operating wavelength.

Characteristics of common PIN photodiodes

Sr. No.	Parameters	Symbol	Unit	Si	Ge	InGaAs
1.	Wavelength	λ	μ m	0.4 – 1.1	0.8 – 1.8	1.0– 1.7
2.	Reponsivity	$\Re$	A/W	0.4 – 0.6	0.5 – 0.7	0.6– 0.9
3.	Quantum efficiency	H	%	75 -90	50 – 55	60 – 70
4.	Darl current	I_d	nA	1 – 10	50 – 500	1 - 20
5.	Rise time	T_r	nS	0.5 – 1	0.1 – 0.5	0.02 – 0.5
6.	Bandwidth	B	GHz	0.3 – 0.6	0.5 – 3	1 – 10
7.	Bias voltage	V_b	V	50 – 100	5 – 10	5 - 6

Depletion Layer Photocurrent

ιϖ) Consider a reverse biased PIN photodiode.

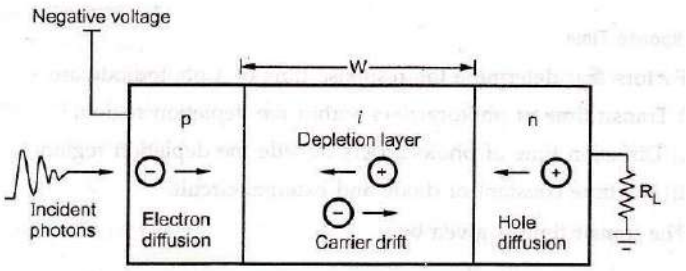


Fig. 3.2.3 Reverse biased PIN diode

- The total current density through depletion layer is –

$$J_{tot} = J_{dr} + J_{diff} \quad \ldots 3.2.7$$

Where,

J_{dr} is drift current densioty due to carriers generated in depletion region.

J_{diff} is diffusion current density due to carriers generated outside depletion region.

- The drift current density is expressed as –

$$J_{dr} = \frac{I_p}{A}$$
$$J_{dr} = q\,\phi_0\,(1 - e^{-\alpha_s w})$$

where,

A is photodiode area.

ϕ_0 is incident photon flux per unit area.

- The diffusion current density is expressed as –

$$J_{diff} = q\,\phi_0\,\frac{\alpha_s L_p}{1+\alpha_s L_p}\,e^{-\alpha_s w} + qP_{n0}\frac{D_p}{L_p} \quad \ldots 3.2.$$

where,

D_p is hole diffusion coefficient

P_n is hole concentration in n-type material.

P_{n0} is equilibrium hole density.

Substituting in equation 3.2.7, total current density through reverse biased depletion layer is –

$$J_{tot} = q\,\phi_0\left[1 - \frac{e^{-\alpha_s W}}{1+\alpha_s L_p}\right] + qP_{n0}\frac{D_p}{L_p}$$

Response Time

- Factors that determine the response time of a photodiode are –
  - Transit time of photocarriers within the depletion region.
  - Diffusion time of photocarriers outside the depletion region.
  - RC time constant of diode and external circuit.
- The transit time is given by –

$$t_d = \frac{w}{v_d}$$

ι) The diffusion process is slow and diffusion times are less than carrier drift time. By considering the photodiode response time the effect of diffusion can be calculated. Fig. 3.2.4 shows the response time of photodiode which is not fully depleted.

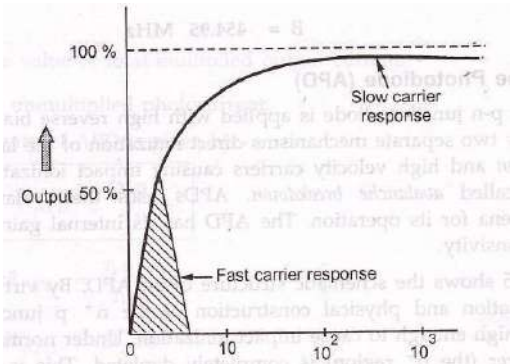


Fig. 3.2.4

- The detector behaves as a simple low pass RC filter having passband of

$$N = \frac{1}{2\pi R_R C_T}$$

where

R_T, is combination input resistance of load and amplifier.

C_T is sum of photodiode and amplifier capacitance.

Example 3.2.5 : Compute the bandwidth of a photodetector having parameters as –

Photodiode capacitance = 3 pF

Amplifier capacitance = 4 pF

Load resistance = 50 Ω

Amplifier input resistance = 1 MΩ

Solution : Sum of photodiode and amplifier capacitance

$$C_T = 3 + 4 = 7 \text{ pF}$$

Combination of load resistance and amplifier and input resistance

$$R_T = 50\Omega \parallel 1\ M\Omega \approx 50\ \Omega$$

$$\text{Bandwidth of photodetector } B = \frac{1}{2\pi\ R_R\ C_T}$$

$$B = \frac{1}{2\pi\ 50\ x\ 7\ x\ 10^{-12}}$$

B = 454.95 MHz **… Ans.**

Avalanche Photodiode (APD)

- When a p-n junction diode is applied with high reverse bias breakdown can occur by two separate mechanisms direct ionization of the lattice atoms, zener breakdown and high velocity carriers impact ionization of the lattice atoms called avalanche breakdown. APDs uses the avalanche breakdown phenomena for its operation. The APD has its internal gain which increases its responsivity.

  Fig. 3.2.5 shows the schematic structure of an APD. By virtue of the doping concentration and physical construction of the n^+ p junction, the electric filed is high enough to cause impact ionization. Under normal operating bias, the I-layer (the p^- region) is completely depleted. This is known as **reach through** condition, hence APDs are also known as **reach through APD** or **RAPDs**

- Similar to PIN photodiode, light absorption in APDs is most efficient in I-layer. In this region, the E-field separates the carriers and the electrons drift into the avalanche region where carrier multiplication occurs. If the APD is biased close to breakdown, it will result

in reverse leakage current. Thus APDs are usually biased just below breakdown, with the bias voltage being tightly controlled.

- The multiplication for all carriers generated in the photodiode is given as –

$$M = \frac{I_M}{I_P} \quad \ldots$$

where,

I_M = Average value of total multiplied output current.

I_P = Primary unmultiplied photocurrent.

- Responsivity of APD is given by –

$$\boxed{\Re_{APD} = \frac{\eta q}{h v} M}$$

$$\Re_{APD} = \frac{\eta q \lambda}{h v} M \qquad \because v = \frac{c}{\lambda}$$

$$\Re_{APD} = \Re_0 M \quad \ldots$$

where, $\Re_0$ = Unity gain responsivity

MSM Photodetector

- Metal-semiconductor-metal (MSM) photodetector uses a sandwiched semiconductor between two metals. The middle semiconductor layer acts as optical absorbing layer. A Schottky barrier is formed at each metal semiconductor interface (junction), which prevents flow of electrons.
- When optical power is incident on it, the electron-hole pairs generated through photo absorption flow towards metal contacts and causes photocurrent.
- MSM photodetectors are manufactured using different combinations of semiconductors such as – GaAs, InGaAs, InP, InAIAs. Each MSM photodetectors had distinct features e.g. responsivity, quantum efficiency, bandwidth etc.

- With InAIAs based MSM photodetector, 92 % quantum efficiency can be obtained at 1.3 μm with low dark current. An inverted MSM photodetector shows high responsivity when illuminated from top.
- A GaAs based device with travelling wave structure gives a bandwidth beyond 500 GHz.

Optical Detector

- With a proper sketch briefly explain the structure of PIN diode.
- Explain the following term relating to PIN photodiode with proper expressions.
  - Cut-off wavelength.
  - Quantum efficiency.
  - Responsivity.
- Explain the structure and principle of working of APD.
- Deduce the expression for total current density for APD.
- How the response time of APD is estimated?
- Give expression for passband of APD detector.
- Compare the performance parameters of PIN and APD.

- **UNIT-V**
DESIGN OF DIGITAL SYSTEMS

Digital Links:

System Design Considerations:

- In optical system design major consideration involves
    1. Transmission characteristics of fiber (attenuation & dispersion).
    2. Information transfer capability of fiber.
    3. Terminal equipment & technology.
    4. Distance of transmission.
- In long-haul communication applications repeaters are inserted at regular intervals as shown in Fig. 6.2.1

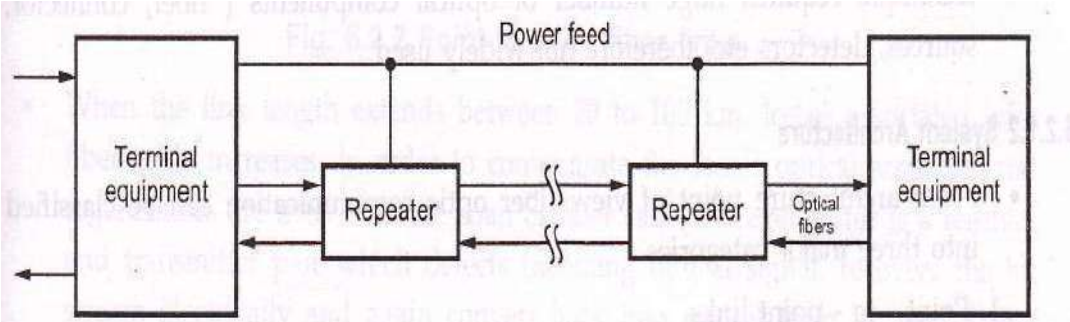


Fig. 6.2.1 Repeaters in long-haul communication system

- Repeater regenerates the original data before it is retransmitted as a digital optical signal. The cost of system and complexity increases because of installation of repeaters.
- An optical communication system should have following basic required specifications –
    7. Transmission type (Analog / digital).
    8. System fidelity (SNR / BER)
    9. Required transmission bandwidth
    10. Acceptable repeater spacing
    11. Cost of system
    12. Reliability
    13. Cost of maintenance.

Multiplexing

- Multiplexing of several signals on a single fiber increases information transfer rate of communication link. In Time Division Multiplexing (TDM) pulses from multiple channels are interleaved and transmitted sequentially, it enhance the bandwidth utilization of a single fiber link. In Frequency Division Multiplexing (FDM) the optical

channel bandwidth is divided inot various nonoverlapping frequency bands and each signal is assigned one of these bands of frequencies. By suitable filtering the combined FDM signal can be retrieved.

- When number of optical sources operating at different wavelengths are to be sent on single fiber link Wavelength Division Multiplexing (WDM) is used. At receiver end, the separation or extraction of optical signal is performed by optical filters (interference filters, differaction filters prism filters).
- Another technique called Space Division Multiplexing (SDM) used separate fiber within fiber bundle for each signal channel. SDM provides better optical isolation which eliminates cross-coupling between channels. But this technique requires huge number of optical components (fiber, connector, sources, detectors etc) therefore not widely used.

System Architecture

- From architecture point of view fiber optic communication can be classified into three major categories.
  - Point – to – point links
  - Distributed networks
  - Local area networks.

Point-to-Point Links:

- A point-to-point link comprises of one transmitter and a receiver system. This is the simplest form of optical communication link and it sets the basis for examining complex optical communication links.
- For analyzing the performance of any link following important aspects are to be considered.
  - Distance of transmission
  - Channel data rate
  - Bit-error rate
- All above parameters of transmission link are associated with the characteristics of various devices employed in the link. Important components and their characteristics are listed below.
- When the link length extends between 20 to 100 km, losses associated with fiber cable increases. In order to compensate the losses optical amplifier and regenerators are used over the span of fiber cable. A regenerator is a receiver and transmitter pair which detects incoming optical signal, recovers the bit stream electrically and again convert back into optical from by modulating an optical source. An optical amplifier amplify the optical bit stream without converting it into electrical form.
- The spacing between two repeater or optical amplifier is called as repeater spacing (L). The repeater spacing L depends on bit rate B. The bit rate-distance product (BL) is a measure of system performance for point-to-point links.

Two important analysis for deciding performance of any fiber link are –

- Link power budget / Power budget
- Rise time budget / Bandwidth budget

- The Link power budget analysis is used to determine whether the receiver has sufficient power to achieve the desired signal quality. The power at receiver is the transmitted power minus link losses.

- The components in the link must be switched fast enough and the fiber dispersion must be low enough to meet the bandwidth requirements of the application. Adequate bandwidth for a system can be assured by developing a rise time budget.

System Consideration:

- Before selecting suitable components, the operating wavelength for the system is decided. The operating wavelength selection depends on the distance and attenuation. For shorter distance, the 800-900 nm region is preferred but for longer distance 100 or 1550 nm region is preferred due to lower attenuations and dispersion.
- The next step is selection of photodetector. While selecting a photodetector following factors are considered –
  - Minimum optical power that must fall on photodetector to satisfy BER at specified data rate.
  - Complexity of circuit.
  - Cost of design.
  - Bias requirements.
- Next step in system consideration is choosing a proper optical source, important factors to consider are –
  - Signal dispersion.
  - Data rate.
  - Transmission distance.
  - Cost.
  - Optical power coupling.
  - Circuit complexity.
- The last factor in system consideration is to selection of optical fiber between single mode and multimode fiber with step or graded index fiber. Fiber selection depends on type of optical source and tolerable dispersion. Some important factors for selection of fiber are :
  - Numerical Aperture (NA), as NA increases, the fiber coupled power increases also the dispersion.
  - Attenuation characteristics.
  - Environmental induced losses e.g. due to temperature variation, moisture and dust etc.

MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

OME 752 SUPPLY CHAIN MANAGEMENT

Semester - 07

Notes

MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

Vision

To excel in providing value based education in the field of Electronics and Communication Engineering, keeping in pace with the latest technical developments through commendable research, to raise the intellectual competence to match global standards and to make significant contributions to the society upholding the ethical standards.

Mission

- ✓ To deliver Quality Technical Education, with an equal emphasis on theoretical and practical aspects.
- ✓ To provide state of the art infrastructure for the students and faculty to upgrade their skills and knowledge.
- ✓ To create an open and conducive environment for faculty and students to carry out research and excel in their field of specialization.
- ✓ To focus especially on innovation and development of technologies that is sustainable and inclusive, and thus benefits all sections of the society.
- ✓ To establish a strong Industry Academic Collaboration for teaching and research, that could foster entrepreneurship and innovation in knowledge exchange.
- ✓ To produce quality Engineers who uphold and advance the integrity, honour and dignity of the engineering.

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

1. To provide the students with a strong foundation in the required sciences in order to pursue studies in Electronics and Communication Engineering.
2. To gain adequate knowledge to become good professional in electronic and communication engineering associated industries, higher education and research.
3. To develop attitude in lifelong learning, applying and adapting new ideas and technologies as their field evolves.
4. To prepare students to critically analyze existing literature in an area of specialization and ethically develop innovative and research oriented methodologies to solve the problems identified.
5. To inculcate in the students a professional and ethical attitude and an ability to visualize the engineering issues in a broader social context.

PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO1: Design, develop and analyze electronic systems through application of relevant electronics, mathematics and engineering principles.

PSO2: Design, develop and analyze communication systems through application of fundamentals from communication principles, signal processing, and RF System Design & Electromagnetics.

PSO3: Adapt to emerging electronics and communication technologies and develop innovative solutions for existing and newer problems.

Table of Contents

Unit I INTRODUCTION

1.1. Role of Logistics and Supply chain Management

What is Supply Chain?

A supply chain consists of all parties involved, directly or indirectly, in fulfilling a customer request. The supply chain includes not only the manufacturer and suppliers, but also transporters, warehouses, retailers, and even customers themselves. Supply chain is otherwise defined as organizations, people, activities, information, and resources involved in supplying a product or service to a consumer. Within each organization, such as a manufacturer, the supply chain includes all functions involved in receiving and filling a customer request. These functions include new product development, marketing, operations, distribution, finance, and customer service as shown in Figure 1.

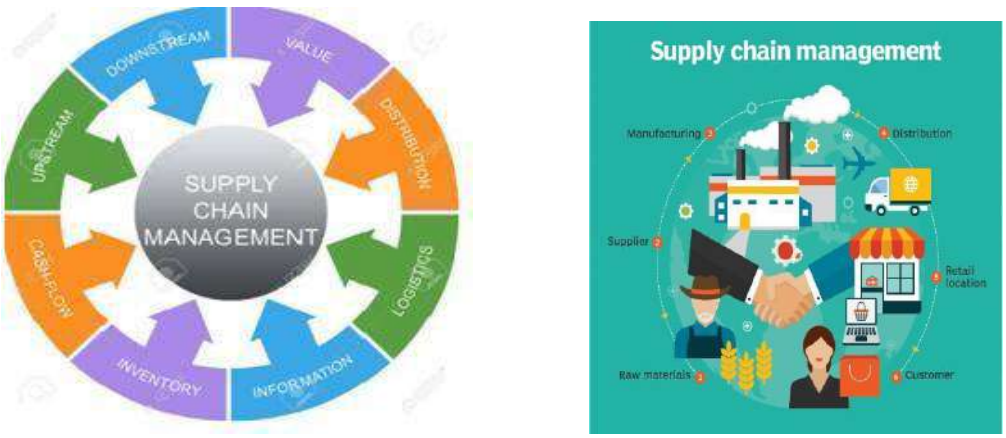


Figure 1. Stages of a General Supply Chain

Some of the examples are: health care, manufacturing industry, student recruitment process, agriculture and many more.

In a student career planning process, students decides their values, learn about the yourself, identify their own skills, finding various carrier options, mapping their skills with carrier options to decide. The following diagram Figure 2 shows the whole process.

Figure 2. Career Planning Supply Chain

In a hospital supply chain, the various stages are patient, doctor, pharmacist, pharmacy, medicines, nurse, primary and emergency care devices, patient monitoring system, patient data analytics and other sensor devices connected to service a patient as shown in Figure 3.

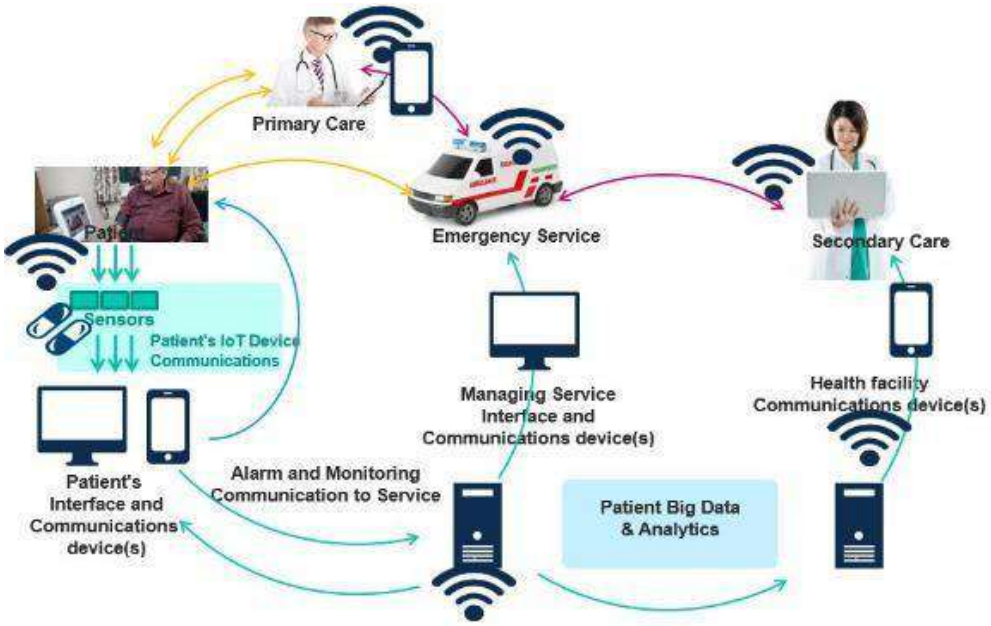


Figure 3. Healthcare Supply Chain

In a manufacturing industry, raw materials supplied by the supplier to the manufacturer and the finished goods are distributed to various distribution centers, and to the customers. Figure 4 shows the supply chain of industry.

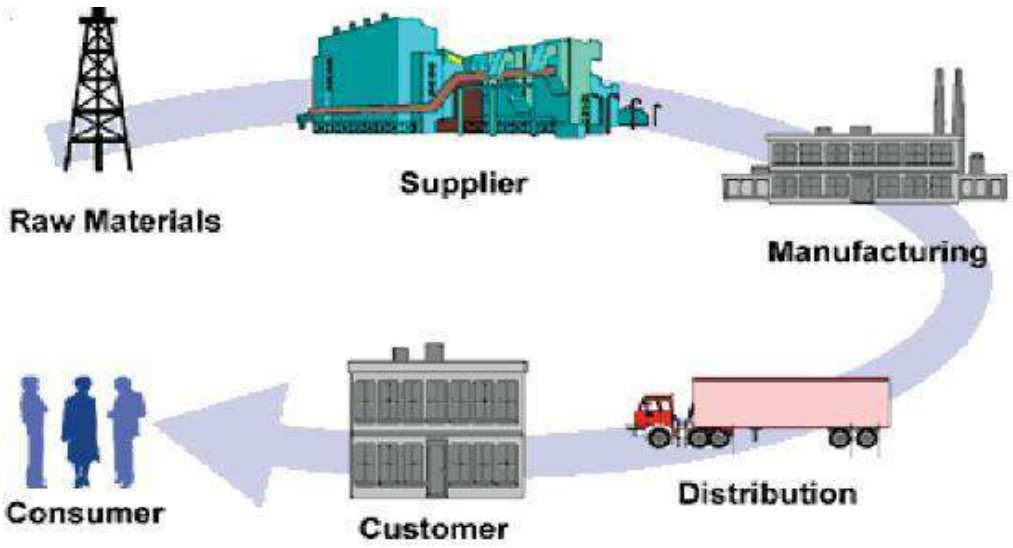


Figure 4 Supply Chain in Manufacturing Industry

Supply chain is dynamic and involves the constant flow of information, product, and funds between different stages. An example of **Wal-Mart** is considered. The customers are buying products and paying amount for that. So customers transfer funds. The ordered items are transferred from manufacturer to distributor and then to the retailers. The replacement orders are also transferred back to the manufacturer to be recycled.

Dell Computer's supply chain includes the customer, Dell's Web site, the Dell assembly plant, and all of Dell's suppliers and their suppliers. The Web site provides the customer with information regarding pricing, product variety, and product availability. The customer enters the order information and pays for the product. The customer may check the status of the order. The supply chain use customer order information to fill the request. That process involves an additional flow of information, product, and funds among various stages of the supply chain.

The primary purpose of any supply chain is to satisfy customer needs, generate profit and service or product to the consumers. A manufacturer may receive material from several suppliers and then supply several distributors. Thus, most supply chains are actually networks. It may be more accurate to use the term supply network or supply web to describe the structure of most supply chains.

What is Logistics?

Logistics is globalization of resource management, from every local unit to the entire network of production points. Supply chain management involves logistics and thus performs end-to-end optimization. The purpose of efficient logistics management is to achieve maximum competitiveness and profitability of the company, as well as the entire network structure of supply chains, including the end-user.

Logistics is concerned with both materials flow and information flow. Logistics can be seen as a link between the manufacturing and marketing operations of a company. The functional areas of the network design are: transportation and inventory management. Manufacturing plants, warehouses, stores etc. are all facilities which form key components in the network design. The cost and consistency required out of the transportation network determines the type and mode of the movement of goods and also affects the inventory as in Figure 5.

Figure 5 Logistics of a firm

The objective behind logistics is to make sure the customer receives the desired product at the right time and place with the right quality and price. This process can be divided into two subcategories: inbound logistics and outbound logistics.

Inbound logistics covers the activities concerned with obtaining materials and then handling, storing and transporting them. Outbound logistics covers the activities concerned with the collection, maintenance and distribution to the customer. Other activities, such as packing and fulfilling orders, warehousing, managing stock and maintaining the equilibrium between supply and demand also factor into logistics.

Differences between Supply chain management and logistics are:

Supply chain management	**Logistics**
A way to link major business processes within and across companies into a high-performance business model that drives competitive advantage.	Movement, storage, and flow of goods, services and information inside and outside the organization.
Main focus of supply chain is a competitive advantage.	The main focus of logistics is meeting customer requirements.
Causing many of the functional areas such as logistics, transportation and inventory to intersect with one another	Logistics is a term that has been around for a long time, emerging from its military roots

Supply chain management	Logistics
Supply chain management incorporates the field of logistics	Logistics is an activity within the supply chain.
Multiple functions and companies to ensure that a finished product not only gets to the end consumer but meets all requirements as well.	Logistics is just one small part of the larger, all-encompassing supply chain network.

Stages in Supply Chain

A typical supply chain may involve a variety of stages: shown in Figure 6.

- Component/raw material
- suppliers Manufacturers
- Wholesalers/distributors
- Retailers
- Customers

Each stage in a supply chain is connected through the flow of products, information, and funds. These flows often occur in both directions and may be managed by one of the stages or an intermediary.

Dell has two supply chain structures that it uses to serve its customers. For its corporate clients and also some individuals who want a customized personal computer (PC), Dell builds to order; that is, a customer order initiates manufacturing at Dell. For these customers, Dell does not have a separate retailer, distributor, or wholesaler in the supply chain. Since 2007, Dell has also sold its PCs through Wal-Mart in the United States

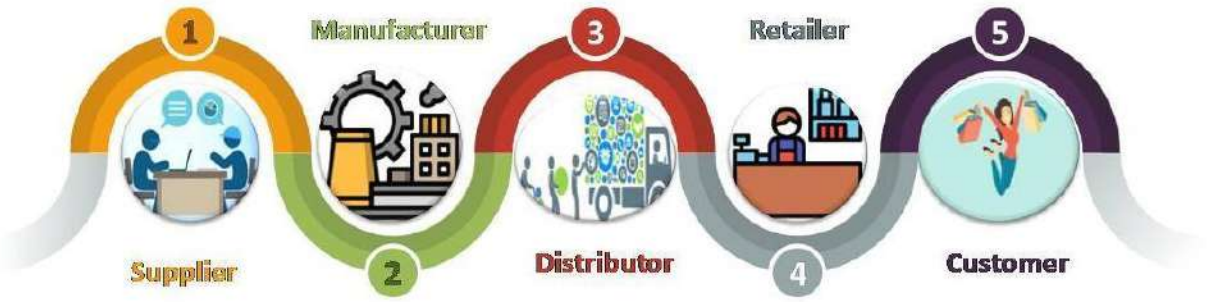


Figure 6 Stages in Supply Chain

This supply chain thus contains an extra stage (the retailer) compared to the direct sales model also used by Dell. In the case of other retail stores, the supply chain may also contain a wholesaler or distributor between the store and the manufacturer.

1.2. Scope and Importance

The scope of supply chains extends through the organization from the demand end to the supply end. However, the core supply chain functions primarily relate to the demand and supply management processes directly controlled by the enterprise. On the supply end, Supplier Relationship Management (SRM) processes extend the supply chains by managing sourcing and suppliers to ensure reliable sources for fulfilling the demand. The scope of supply chain management is to monitor and control the activities right from customer's customer to supplier's supplier.

- Minimizes Operating Cost
- Boost Customer Service
- Enhance Financial
- Position Manages
- Distribution Coordination
- Among Partners Inventory
- Management Supplier

  Management

Minimizes Operating Cost

Supply chain management focuses on reducing the overall operating cost of the organization. It aims at bringing efficiency and raising the profitability of organizations. By developing a proper chain it brings down the purchasing cost, production cost and delivery cost. It enables smooth flow of raw materials from the supplier to an organization which reduces the holding period of materials with the supplier and avoids any losses due to delay in production.

Boosts Customer Service

Supply chain management helps in providing better service to customers. All production strategies are framed in accordance with requirements of customers to manufacture right product. Supply managers monitor all operations of business and ensure that quality products are produced using best combination of resources. Right product available to right cost provide better satisfaction to customers. This will boost their confidence level in company's products also.

Enhance Financial Position

Management of supply chain has an effective role on the financial position of business. It improves the efficiency of the organization, cut down the excessive cost and avoids any shortage. Supply chain manager bring down the cost by reducing the use of fixed assets like plants, transportation vehicles, warehouses etc. Proper supply chain results in speedy flow of products which minimizes the blockage of funds in inventories. It ensures that optimum funds are always available which helps in improving financial position.

Manages Distribution

Distribution of products at the right time and the right location is a complex task for every organization. It coordinates with various transportation channels and warehouses for attaining faster movement of goods. Supply chain managers ensure that all products get delivered at the right location within the time limit.

Bring Coordination among Partners

Proper coordination among all partners of business increase productivity and profitability. It develops a proper channel through which employees, supplier and customers can easily interact with business. Managers can easily control the activities of their subordinates by communicating them all the required information. Employees in case of any problem or error can contact their supervisors. Customers can also access their brands for any information through self-portals which are developed as a part of the customer support system.

Inventory Management

Maintaining an optimum inventory is a must for uninterrupted operation of every business. It keeps record of all inventories that is raw materials, spare parts and finished goods. Supply chain managers ensure that the proper amount of inventory is always maintained within the organization.

Supplier Management

Supply chain management works on strengthening the relationships between business and suppliers. It tracks and records every interactions or transaction with the suppliers. Proper supply chain enables timely procurement of all required raw materials from suppliers. Supply chain management solutions provide a self- service portal through which suppliers can contact the company in case of any issues or problems.

Objectives

The objective of every supply chain should be to maximize the overall value generated.

Supply Chain Surplus = Customer Value – Supply Chain Cost

The difference between the value of the product and its price remains with the customer as consumer surplus. The **supply chain surplus** or **supply chain profitability**, the difference between the revenue generated from the customer and the overall cost across the supply chain. Supply chain profitability is the total profit to be shared across all supply chain stages and intermediaries. The higher the supply chain profitability, the more successful is the supply chain.

For any supply chain, there is only one source of revenue: the customer. The customer is the only one providing positive cash flow for any organization. All other cash flows are simply fund exchanges that occur within the supply chain, given that different stages have different owners. All flows of information, product, or funds generate costs within the supply chain. Thus, the appropriate management of these flows is a key to supply chain success. Effective supply chain management involves the management of supply chain assets and product, information, and fund flows to maximize total supply chain surplus.

Supply chain design, planning, and operation decisions play a significant role in the success or failure of a firm. To stay competitive, supply chains must adapt to changing technology and customer expectations.

Some of the objectives of supply chain management are: To Minimize work in progress.

- To Reduce transportation cost.
- To increase distribution channel of the product.
- To provide social services by giving them electricity, food, medicine and etc.
- To make plans and strategy in order to maximize productivity
- To increase customer satisfaction by delivering products to consumers on time and providing fast service.

1.3. Evolution of Supply Chain

Over the years, most firms have focused their attention on the effectiveness and efficiency of separate business functions such as purchasing, production, marketing, financing, and logistics. To capture the synergy of inter-functional and inter-organizational integration and coordination across the supply chain and to subsequently make better strategic decisions, a growing number of firms have begun to realize the strategic importance of planning, controlling, and designing a supply chain as a whole.

Over the last 100 plus years of the history of supply chain management has evolved from an initial focus on improving relatively simple, but very labor- intensive processes to the present day engineering and managing of extraordinarily complex global networks. Evolution is shown in Figure 7

Figure 7 Evolution of Supply Chain

Creation Era

The term supply chain management was first coined by an American industry consultant in the early 1980s. However the concept of supply chain in management, was of great importance long before in the early 20th century.

- Assembly line concept was used for mass production in automotive industries which were developed more during that period.
- Concentrates on Up-scaling where the quantity need to be produced from 50 to 500.
- Re-engineering is the concept of converting manual operations into automated to increase the production and also process flow needs to be changed
- At the same time, Down-sizing, reducing the resources for cost control is significant.

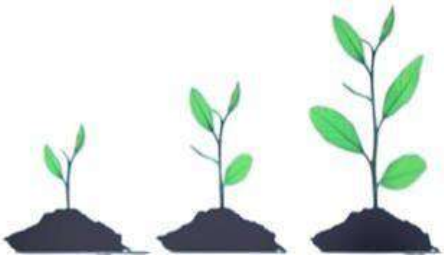


Figure 8 Creation Era

- Assembly line concept was used for mass production in automotive industries which were developed more during that period.
- Concentrates on Up-scaling where the quantity need to be produced from 50 to 500.
- Re-engineering is the concept of converting manual operations into automated to increase the production and also process flow needs to be changed
- At the same time, Down-sizing, reducing the resources for cost control is significant.

Integration Era

This era of supply chain management studies was highlighted with the development of Electronic Data Interchange (EDI) systems in the 1960s and developed through the 1990s by the introduction of Enterprise Resource Planning (ERP) systems.

Figure 9 Integration Era

- This era of supply-chain evolution is characterized by both increasing value added and reducing costs through integration.
- A supply chain can be classified as a stage 1, 2 or 3 network.

- In a stage 1–type supply chain, systems such as production, storage, distribution, and material control are not linked and are independent of each other.
- In a stage 2 supply chain, these are integrated under one plan and enterprise resource planning (ERP) is enabled.
- A stage 3 supply chain is one that achieves vertical integration with upstream suppliers and downstream customers.

Vertical Integration is that multiple components of the supply chain coming under
same upper management.

Globalization Era

This era is characterized by the globalization of supply chain management in organizations with the 32 goal of increasing competitive advantage, creating more value-added, and reducing costs through global sourcing.

- Attention given to global systems of supplier relationships and the expansion of supply chains beyond national boundaries and into other

continents.

Figure 10 Integration Era

Specialization Era Phase - I

(Outsourced Manufacturing and Distribution)

In the 1990s industries began to focus on “core competencies” and adopted a specialization model. Companies abandoned vertical integration, sold off non-core operations, and outsourced those functions to other companies.

Attention given to global systems of supplier relationships and the expansion of supply chains beyond national boundaries and into other continents.

- This era is characterized by the globalization of supply-chain management in organizations with the goal of increasing their competitive advantage, adding value, and reducing costs through global sourcing.

Figure 11 Specialization Era Phase I

Specialization Era Phase II

(Supply Chain Management as a Service)

Specialization within the supply chain began in the 1980s with the inception of transportation brokerages, ware house management, and non asset based carriers and has matured beyond transportation and logistics into aspects of supply planning, collaboration, execution and performance management.

Figure 12 Specialization Era Phase II

- Outsourced manufacturing and distribution has done; it allows them to focus on their core competencies and assemble networks of specific, best-in-class partners to contribute to the overall value chain itself, thereby increasing overall performance and efficiency.
- Overheads are distributed and can be reduced.
- Having advantages in the cost point of view.

SCM 2.0

SCM 2. 0 is defined as a trend in the use of the World Wide Web that is meant to increase creativity, information sharing, and collaboration among users.

- Multi-suppliers and customers
- Automated Business Transaction
- Combination of processes, methodologies, tools, and delivery options to guide companies to their results quickly as the complexity and speed of the supply- chain increase due to global competition; rapid price fluctuations; short product life cycle.

Figure 13 SCM 2.0

Before the 1950s, logistics was thought of in military terms. It had to do with procurement, maintenance, and transportation of military facilities, materials, and personnel. The study and practice of physical distribution and logistics emerged in the 1960s and 1970s

The importance of logistics increased considerably, when physical distribution management in manufacturing firms was recognized as a separate organizational function

The SCM concept was coined in the early 1980s by consultants in logistics. The supply chain must have been viewed as a single entity and that strategic decision- making at the top level was needed to manage the chain in their original formulation. The evolution of SCM continued into the 1990s due to the intense global competition

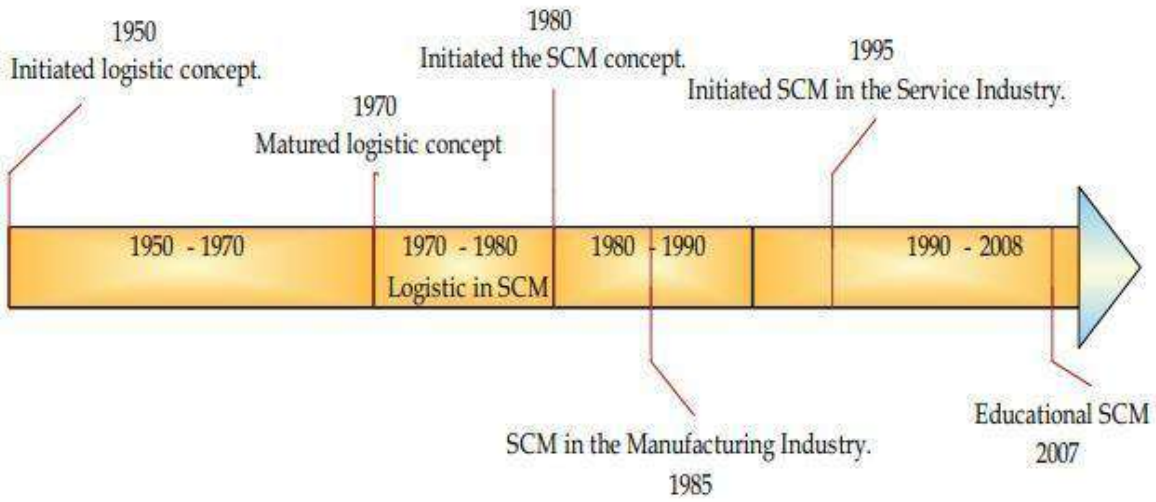


Figure 14 Evolutionary Timeline of Supply Chain Management

One of the most significant changes in paradigm of modern business management is that individual businesses no longer compete as solely autonomous entities, but rather as supply chains. Business management has entered the era of inter-network competition and the ultimate success of a single business will depend on management's ability to integrate the company's intricate network of business relationships.

1.4. Decision Phases in Supply Chain

Decision phases can be defined as the different stages involved in supply chain management for taking an action or decision related to some product or services. Successful supply chain management requires decisions on the flow of information, product, and funds. Categorize supply chain decision phases based on the frequency with which they are made and the time frame they take into account. Each decision should be made to raise the supply chain surplus. Each category of decisions must consider uncertainty over the decision horizon.

Different decision phases are:

- Supply Chain Strategy or Design (Long term
- planning) Supply Chain Planning (Mid term planning)
- Supply Chain Operation (Short term planning and control)

1. Supply Chain Strategy or Design

A company decides how to structure the supply chain over the next several years. It decides

- What the chain's configuration
- will be How resources will be allocated
- What processes each stage will perform

Strategic decisions made by companies include

- Whether to outsource or perform a supply chain function
- in-house Location and capacities of production and
- warehousing facilities Products to be manufactured or stored at various locations
- Modes of transportation to be made available and different shipping
- leg Type of information system to be utilized.

A firm must ensure that the supply chain configuration supports its strategic objectives. Supply chain design decisions are typically made for the long term and are expensive to alter on short notice. When companies make these decisions, they must take into account uncertainty in anticipated market conditions over the next few years.

2. Supply Chain Planning

For decisions made during this phase, the time frame considered is a quarter to a year. Supply chain's configuration determined in the strategic phase is fixed. This configuration establishes constraints within which planning must be done. The goal of planning is to maximize the supply chain surplus that can be generated over the planning horizon given the constraints established during the strategic or design phase. Companies start the planning phase with a forecast for the coming year of demand and other factors such as costs and prices in different markets. Planning includes making decisions regarding:

A firm must ensure that the supply chain configuration supports its strategic objectives. Supply chain design decisions are typically made for the long term and are expensive to alter on short notice. When companies make these decisions, they must take into account uncertainty in anticipated market conditions over the next few years.

2. Supply Chain Planning

For decisions made during this phase, the time frame considered is a quarter to a year. Supply chain's configuration determined in the strategic phase is fixed. This configuration establishes constraints within which planning must be done. The goal of planning is to maximize the supply chain surplus that can be generated over the planning horizon given the constraints established during the strategic or design phase. Companies start the planning phase with a forecast for the coming year of demand and other factors such as costs and prices in different markets. Planning includes making decisions regarding:

- Which markets will be supplied from which
- locations Subcontracting of manufacturing
- Inventory policies to be
- followed Timing and size of
- marketing Price promotions.

Given a shorter time frame and better forecasts than in the design phase, companies in the planning phase try to incorporate any flexibility built into the supply chain in the design phase and exploit it to optimize performance. Companies define a set of operating policies that govern short-term operations.

3. Supply Chain Operation

The time horizon here is weekly or daily. During this phase, companies make decisions regarding individual customer orders. At the operational level, supply chain configuration is considered fixed, and planning policies are already defined. The goal of supply chain operations is to handle incoming customer orders in the best possible manner. During this phase, firms allocate inventory or production to individual orders, set a date that an order is to be filled, generate pick lists at a warehouse, allocate an order to a particular shipping mode and shipment, set delivery schedules of trucks, and place replenishment orders.

Because operational decisions are being made in the short term (minutes, hours, or days), there is less uncertainty about demand information. Given the constraints established by the configuration and planning policies, the goal during the operation phase is to exploit the reduction of uncertainty and optimize performance.

The design, planning, and operation of a supply chain have a strong impact on overall profitability and success. It is fair to state that a large part of the success of firms such as Wal-Mart and Seven-Eleven Japan can be attributed to their effective supply chain design, planning, and operation.

Supply chain decision phases may be categorized as design, planning, or operational, depending on the time frame during which the decisions made apply. Design decisions constrain or enable good planning, which in turn constrains or enables effective operation.

1.5 Competitive and supply chain strategies

A company's competitive strategy defines, relative to its competitors, the set of customer needs that it seeks to satisfy through its products and services.

For example, Wal-Mart aims to provide high availability of a variety of products of reasonable quality at low prices. Most products sold at Wal-Mart are commonplace (everything from home appliances to clothing) and can be purchased elsewhere. What Wal-Mart provides is a low price and product availability.

McMaster-Carr sells maintenance, repair, and operations (MRO) products. It offers more than 500,000 products through both a catalog and a Web site. Its competitive strategy is built around providing the customer with convenience, availability, and responsiveness. With this focus on responsiveness, McMaster does not compete based on low price. Clearly, the competitive strategy at Wal- Mart is different from that at McMaster.

We can also contrast Blue Nile, with its online retailing model for diamonds, with Zales, which sells diamond jewelry through retail outlets. Blue Nile has emphasized the variety of diamonds available from its Web site and the fact that its margins are significantly lower than its bricks-and-mortar competition. Customers, however, have to wait to get their jewelry and do not have any opportunity to touch and see it before purchase (Blue Nile does provide a 30-day return period).

At Zales, in contrast, a customer can walk into the retail store, be helped by a salesperson, and leave immediately with a diamond ring. The amount of variety available at a Zales store, however, is limited. Whereas Blue Nile offers more than 70,000 stones on its site, a typical Zales store carries less than a thousand.

In each case, the competitive strategy is defined based on how the customer prioritizes product cost, delivery time, variety, and quality.

A McMaster-Carr customer places greater emphasis on product variety and response time than on cost. A Wal-Mart customer, in contrast, places greater emphasis on cost. A Blue Nile customer, purchasing online, places great emphasis on product variety and cost. A customer purchasing jewelry at Zales is most concerned with fast response time and help in product selection.

Thus, a firm's competitive strategy will be defined based on its customers' priorities. Competitive strategy targets one or more customer segments and aims to provide products and services that satisfy these customers' needs.

Value Chain for a typical Organization

In order to see the relationship between competitive and supply chain strategies, we start with the value chain for a typical organization, as shown in Figure 15.

Figure 15 The Value Chain in a Company

The value chain begins with new product development, which creates specifications for the product. Marketing and sales generate demand by publicizing the customer priorities that the products and services will satisfy. Marketing also brings customer input back to new product development. Using new product specifications, operations transforms inputs to outputs to create the

product. Distribution either takes the product to the customer or brings the customer to the product. Service responds to customer requests during or after the sale. These are core processes or functions that must be performed for a successful sale. Finance, accounting, information technology, and human resources support and facilitate the functioning of the value chain.

To execute a company's competitive strategy, all these functions play a role, and each must develop its own strategy. Here, strategy refers to what each process or function will try to do particularly well

A product development strategy specifies the portfolio of new products that a company will try to develop. It also dictates whether the development effort will be made internally or outsourced. A marketing and sales strategy specifies how the market will be segmented and how the product will be positioned, priced, and promoted. A supply chain strategy determines the nature of procurement of raw materials, transportation of materials to and from the company, manufacture of the product or operation to provide the service, and distribution of the product to the customer, along with any follow-up service and a specification of whether these processes will be performed in-house or outsourced. Supply chain strategy specifies what the operations, distribution, and service functions, whether performed in-house or outsourced, should do particularly well.

Strategies followed by organizations

For example, Dell's initial decision to sell direct, its 2007 decision to start selling PCs through resellers, Amazon's decisions to build warehouses to stock some products and to continue using distributors as a source of other products are part of its supply chain strategy. Similarly, Toyota's decision to have production facilities in each of its major markets is part of its supply chain strategy.

For a firm to succeed, all functional strategies must support one another and the competitive strategy.

Seven-Eleven Japan's success can be related to the excellent fit among its functional strategies. Marketing at Seven-Eleven has emphasized convenience in the form of easy access to stores and availability of a wide range of products and services. New product development at Seven-Eleven is constantly adding products and services, such as bill payment services that draw customers in and exploit the excellent information infrastructure and the fact that customers frequently visit Seven-Eleven. Operations and distribution at Seven-Eleven have focused on having a high density of stores, being very responsive, and providing an excellent information infrastructure. The result is a virtuous cycle in which supply chain infrastructure is exploited to offer new products and service that increase demand, and the increased demand in turn makes it easier for operations to improve the density of stores, responsiveness in replenishment, and the information infrastructure.

Achieving Strategic fit

Strategic fit requires that both the competitive and supply chain strategies of a company have aligned goals. It refers to consistency between the customer priorities that the competitive strategy hopes to satisfy and the supply chain capabilities that the supply chain strategy aims to build.

For a company to achieve strategic fit, it must accomplish the following:

1. The competitive strategy and all functional strategies must fit together to form a coordinated overall strategy. Each functional strategy must support other functional strategies and help a firm reach its competitive strategy goal.

2. The different functions in a company must appropriately structure their processes and resources to be able to execute these strategies successfully.

3. The design of the overall supply chain and the role of each stage must be aligned to support the supply chain strategy.

A company may fail either because of a lack of strategic fit or because its overall supply chain design, processes, and resources do not provide the capabilities to support the desired strategic fit.

1.6. Drivers of Supply Chain Performance

To understand how a company can, improve supply chain performance in terms of responsiveness and efficiency, we examine the logistical and cross–functional drivers of supply chain performance:

Facilities

Inventory

Transportation

Information

Sourcing and

Pricing.

These drivers interact to determine the supply chain's performance in terms of responsiveness and efficiency.The goal is to structure the drivers to achieve the desired level of responsiveness at the lowest possible cost, thus improving the supply chain surplus and the firm's financial performance.

Driver and its impact on the performance of the supply chain.

1. Facilities

Facilities are the actual physical locations in the supply chain network where product is stored, assembled, or fabricated.

The two major types of facilities are:

production sites and

storage sites

Decisions regarding the role, location, capacity, and flexibility of facilities have a significant impact on the supply chain's performance.

For example, in 2009, Amazon increased the number of warehousing facilities located close to customers to improve its responsiveness. In contrast, Blockbuster tried to improve its efficiency in 2010 by shutting down many facilities even though it reduced responsiveness. Facility costs show up under property, plant and equipment, if facilities are owned by the firm or under selling, general, and administrative if they are leased.

2. Inventory

Inventory encompasses all raw materials, work in process, and finished goods within a supply chain. The inventory belonging to a firm is reported under assets. Changing inventory policies can dramatically alter the supply chain's efficiency and responsiveness.

For example, W.W. Grainger makes itself responsive by stocking large amounts of inventory and satisfying customer demand from stock even though the high inventory levels reduce efficiency. Such a practice makes sense for Grainger because its products hold their value for a long time.

A strategy using high inventory levels can be dangerous in the fashion apparel business where inventory loses value relatively quickly with changing seasons and trends. Rather than hold high levels of inventory, Spanish apparel retailer Zara has worked hard to shorten new product and replenishment lead times. As a result, the company is very responsive but carries low levels of inventory. Zara thus provides responsiveness at low cost.

3. Transportation

Transportation entails moving inventory from point to point in the supply chain. Transportation can take the form of many combinations of modes and routes, each with its own performance characteristics. Transportation choices have a large impact on supply chain responsiveness and efficiency.

For example, a mail-order catalog company can use a faster mode of transportation such as FedEx to ship products, thus making its supply chain more responsive, but also less efficient given the high costs associated with using FedEx. McMaster-Carr and W.W. Grainger, however, have structured their supply chain to provide next-day service to most of their customers using ground transportation. They are providing a high level of responsiveness at lower cost.

Outbound transportation costs of shipping to the customer are typically included in selling, general, and administrative expense, while inbound transportation costs are typically included in the cost of goods sold.

4. Information

Information consists of data and analysis concerning facilities, inventory, transportation, costs, prices, and customers throughout the supply chain. Information is potentially the biggest driver of performance in the supply chain because it directly affects each of the other drivers. Information presents management with the opportunity to make supply chains more responsive and more efficient.

For example, Seven-Eleven Japan has used information to better match supply and demand while achieving production and distribution economies. The result is a high level of responsiveness to customer demand while production and replenishment costs are lowered.

Information technology–related expenses are typically included under either operating expense (typically under selling, general, and administrative expense) or assets. For example, in 2009, Amazon included $1.24 billion in technology expense under operating expense and another $551 million under fixed assets to be depreciated.

5. Sourcing

Sourcing is the choice of who will perform a particular supply chain activity such as production, storage, transportation, or the management of information. At the strategic level, these decisions determine what functions a firm performs and what functions the firm outsources. Sourcing decisions affect both the responsiveness and efficiency of a supply chain.

After Motorola outsourced much of its production to contract manufacturers in China, it saw its efficiency improve but its responsiveness suffers because of the long distances. To make up for the drop in responsiveness, Motorola started flying in some of its cell phones from China even though this choice increased transportation cost.

Flextronics, an electronics contract manufacturer, is hoping to offer both responsive and efficient sourcing options to its customers. It is trying to make its production facilities in high-cost locations very responsive while keeping its facilities in low-cost countries efficient. Flextronics hopes to become an effective source for all customers using this combination of facilities. Sourcing costs show up in the cost of goods sold, and monies owed to suppliers are recorded under accounts payable.

6. Pricing

Pricing determines how much a firm will charge for the goods and services that it makes available in the supply chain. Pricing affects the behavior of the buyer of the good or service, thus affecting supply chain performance.

For example, if a transportation company varies its charges based on the lead time provided by the customers, it is likely that customers who value efficiency will order early and customers who value responsiveness will be willing to wait and order just before they need a product transported.

Differential pricing provides responsiveness to customers that value it and low cost to customers that do not value responsiveness as much. Any change in pricing impacts revenues directly but could also affect costs based on the impact of this change on the other drivers.

Supply chain management includes the use of logistical and cross-functional drivers to increase the supply chain surplus. It is important to realize that these drivers do not act independently but interact to determine the overall supply chain performance. Good supply chain design and operation recognize this interaction and make the appropriate trade-offs to deliver the desired level of responsiveness.

Obstacles

In order to achieve a strategic fit, a company needs to strike a balance between efficiency and responsiveness. In its endeavor to achieve this strategic fit, the company needs to understand what the customer wants, on the basis of which the company should place itself on the responsiveness spectrum.

The obstacles are becoming dynamic creating more difficulties for companies to create a proper balance. On the other hand, they have also helped the companies with increased opportunities to improve on the supply chain management. Thus, managers have to play a very important role in tackling these obstacles in order to turn it to an advantage and in turn increase the profitability of their supply chain.

Obstacles can be of various types some of which are as under:

Increase in product variety: The demand of customers has been continuously increasing. There has been a continuous increase in the demand for customized products.

Increase in demanding customers: 'Customer is King' in today's world. There has been an increase in the number of customers who constantly demand improved services like timely delivery, cost, product performance, discounts and shorter lead times.

Smaller product lifecycles: With the increase in the number and types of products demanded, there has been a decrease in the life cycles of a large number of products. Today, there are products whose lifecycles can be measured in terms of months, like mobile phones and computers unlike products that would remain in the market for years and years.

Effect of globalization: The removal of trade restrictions by various governments has enabled increased Global Trade. The effects of globalization on supply chains have been tremendous and have also provided supply chains a broader environment to work in.

Difficulty in execution of strategies: The creation of a successful supply chain strategy is not very easy. The formulation of the strategy may be easy;

however, the execution of this strategy is most difficult. The successful execution of a supply chain strategy involves the skillful ability of employees and managers at each and every level of the organization.

It is observed that these obstacles for make it more difficult any company to achieve a strategic fit by creating a balance between responsiveness and efficiency in the supply chain.

Part-A Questions and Answers

1. What is supply chain? (K1, CO1)

All functions involved in receiving and filling a customer request - new product development, marketing, operations, distribution, finance, and customer service.

2. List out the primary purpose of the supply chain (K1, CO1)

- Satisfy customer needs
- Generate profit
- Product or service to a consumer

3. Mention any two scope of the SCM (K1, CO1)

Minimizes Operating Cost, Boost Customer Service, Enhance Financial Position, Manages Distribution, Coordination Among Partners, Inventory Management, Supplier Management

4. What is overall operational cost? (K1, CO1)

- ✔ Purchasing cost
- ✔ Production cost
- ✔ Delivery cost

5. What is a customer support system? (K1, CO1)

Customer's can interact with distributers or business management for any enquiries or complaints about the product through a forum which is called as customer support system.

6. How will you calculate supply chain profitability? (K1, CO1)

Supply Chain Surplus/Value/Supply chain profitability = Customer Value – Supply Chain Cost

7. What are the objectives of the supply chain? (K1, CO1)

Customer satisfaction and maximum profitability

8. What are the various stages in SCM? (K1, CO1)

- Customers
- Retailers
- Wholesalers/distributors
- Manufacturers
- Component/raw material suppliers

9. How the stages in supply chain are connected with one another? (K1, CO1)

Each stage in a supply chain is connected through the flow of products, information, and funds. Product flow is in the direction from supplier to

customer. Information flow in both direction, and funds flow from customer to supplier.

10. **What are the functions of the SCM? (K1, CO1)**
    - Functions on Strategic Level
    - Functions on Tactical Level
    - Functions on Operational Level

11. **What up-scaling and down sizing in SCM? (K1, CO1)**
    - Up-scaling – increases the quantity of the product to be produced
    - Down-sizing – Reduces the resources in terms of employees and other materials

12. **List out the different decision phases in the SCM (K1, CO1)**
    - Supply Chain Strategy or Design-Long term planning
    - Supply Chain Planning - Mid term planning
    - Supply Chain Operation - Short term planning and control

13. **Write about the long term decision phase (K1, CO1)**

    Supply chain design decisions are typically made for the long term and are expensive to alter on short notice. When companies make these decisions, they must take into account uncertainty in anticipated market conditions over the next few years.

14. **What is strategy? (K1, CO1)**

    Strategy refers to what each process or function will try to do particularly well.

15. **What is value chain of an organization? (K1, CO1)**

    The relationship between competitive and supply chain strategies.

Part-B Questions

Q. No.	Questions		
1	Explain the scope and importance of the Supply chain management		
2	Differentiate Logistics and supply chain management		
3	Discuss various stages of supply chain		
4	With neat examples explain the evolution of supply chain management		
5	Explain the various decision phases in supply chain management		
6	Differentiate the supply chain strategies and competitive strategies		
7	Discuss the various drivers of supply chain with examples		
8	Write a short note on obstacles in the supply chain.		

Table of Contents

Unit 2 SUPPLY CHAIN NETWORK DESIGN

UNIT 2- SUPPLY CHAIN NETWORK DESIGN

2. WHAT IS SUPPLY CHAIN NETWORK DESIGN?

The success of any supply chain network will depend on the plants, suppliers, warehouses and how the product flows from each of the origins to the final customer. For any successful supply chain, the number of facilities and their locations are a critical factor. In fact, 80 percent of the operational costs of the supply chain network depend on where the facilities are located and the product flows between them. To cut the costs, you need a more systematic engineering approach so that you can plan and design the network efficiently. That is why Supply Chain Network Design is so important.

Supply Chain Network Design is a systematic approach to determining the best location and optimal size of the facilities to be included in the supply chain, and to ensure an optimal flow of products using advanced mathematical modelling.

2.1 Role of Network Design in Supply Chain

The supply chain network structure of any company will determine how efficient its processes are and whether it is able to provide its customers with a great experience. Designing the most efficient supply chain network structure requires the network to satisfy the strategic objectives of the company over an extended period of time.

The role of network design in supply chain involves the following:

- Defining the business objectives
- Defining the project scope
- Determining the analyses to be performed
- Determining the tools to be utilized
- Completion of the project in accordance with the design.

The Supply Chain Network Design determines the path ahead for the business. If the supply chain network is completed in accordance with the design, the business can expect to gain in a substantial way.

What is Distribution Management?

Distribution management includes forecasting, transportation, warehousing, and delivery within the larger universe of logistics and supply chain management. These require precise tracking, real-time information, and highly-skilled staffing to execute effectively.

Distribution management is an integral part of logistics.

While continually facing a spectrum of variables in their daily business, distribution managers are charged with resolving the three fundamental questions:

- When?
- Where?
- How Much?

Successful distribution management utilizes:

- State-of-the-art information systems
- Logistics software
- Highly efficient equipment
- Forecasting tools
- Warehouse inventory management systems (WMS)
- Excellent safety and training programs

2.2 Factors Influencing Distribution Network Design

Why do you require a distribution network between manufacturing facility and customer location?

The performance of no distribution network system or a distribution network in place or proposed has to be evaluated on two major dimensions.

1. The customer needs that are being met.
2. Cost of the network or costs incurred in meeting those needs.

The distribution network can change the satisfaction of the following customer needs that differ from product to product as well as from distribution outlet to distribution outlet.

Response time

Product variety

Product availability

Customer experience

Order visibility

Returnability

When customers demand less response time, the firm needs more outlets close to the customer. When customers are happy with larger response times, the firm can more centralized facilities.

Changing the distribution network design affects the following supply chain costs:

- Inventory cost
- Transportation cost
- Facilities and handling related cost
- Information system cost

As the number of facilities in a supply chain increases, the inventory and resulting inventory costs also increase. For example, Amazon has fewer facilities and therefore is able to turn its inventory about twelve times a year. Borders has about 400 facilities and it achieves only about two turns per year.

As long as inbound transportation costs to warehouses are kept the same, increasing the number of facilities decreases total transportation cost. But, if the number of facilities is increased to a point where there is a significant loss of economies of scale in inbound transportation (as full truck loads are not employed), increasing the number of facilities increases total transportation cost. A distribution network with more than one warehouse allows initially to reduce transportation cost relative to a network with a single warehouse. Total logistics costs are the sum of inventory, transportation, and facility costs for a supply chain network. As the number of facilities is increased, total logistics costs first decrease and then increase. Each firm should have at least the number of facilities that minimize total logistics costs.

As a firm wants to further reduce the response time to its customers, it may have to increase the number of facilities beyond the point that minimizes logistics costs.

A firm should add facilities beyond the cost- minimizing point only if managers are confident that the increase in revenues because of better responsiveness is greater than the increase in costs because of the additional facilities.

There are two key decisions when designing a distribution network:

1. Will product be delivered to the customer location or picked up from a preordained site (door delivery or a retail facility delivery)?
2. Will product flow through an intermediary or a distribution channel separate from retailer (or intermediate location)?

Based on the choices for the two decisions, there are six distinct distribution network designs that are classified as follows:

1. Manufacturer storage with direct shipping
2. Manufacturer storage with direct shipping and in-transit merge (cross docking)

3. Distributor storage with package carrier delivery
4. Distributor storage with last mile delivery
5. Manufacturer / distributor storage with costumer pickup
6. Retail storage with customer pickup

While the book gives above categories We can Manufacturer, Distributor, Retailer as three entities. Customer pickup or door delivery as two options. If the door delivery options is used the mode of door delivery. Also there is transport between manufacturer and distributor, distributor and retailer and between manufacturer and retailers. The customer preference for each alternative, resulting demand for the product or products and cost of the distribution arrangement come into the picture to take the distribution system decision.

Only niche companies will end up using a single type of distribution network. Most companies are employ a combination of different types for different products, different customers and different usage situations. In a company, fast moving and emergency items are stocked locally and customers can either pick them up directly or have them shipped depending upon the urgency. Slower moving items are stocked at a national distribution centre from where they are shipped to the customer within a day or two. Very slow moving items are typically drop shipped from the manufacturer and involve a longer lead time.

Firms that target customers who can tolerate a long response time require only a few locations that may be far from the customer. These companies can focus on increasing the capacity of each location. In contrast, firms that target customers who value short response times need to locate facilities close to them.

These firms must have many facilities, each with a low capacity. Thus, a decrease in the response time customers desire increases the number of facilities required in the network, as shown in Figure below.

Changing the distribution network design affects the following supply chain costs Inventories

Transportation

Facilities and

handling

Information

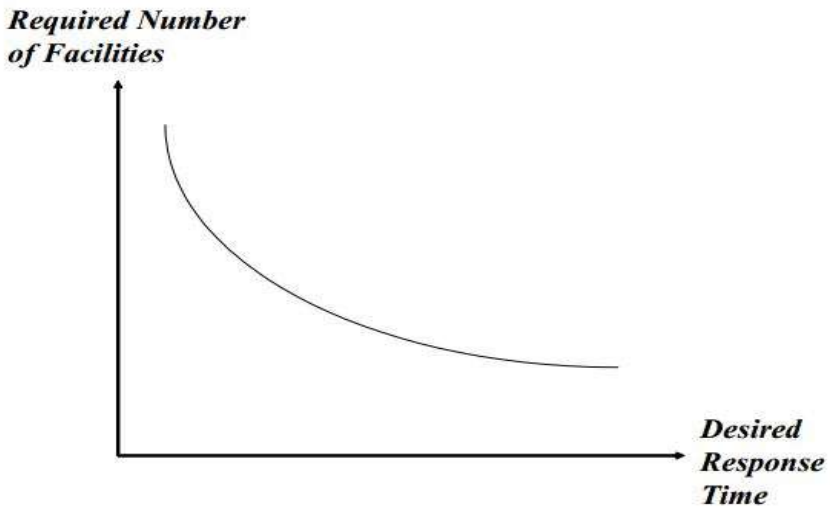


Figure 1 Response Time Vs Number of Facilities

The other two drivers, sourcing and pricing, also affect the choice of the distribution system; the link will be discussed when relevant. As the number of facilities in a supply chain increases, the inventory and resulting inventory costs also increase, as shown in Figure 2.

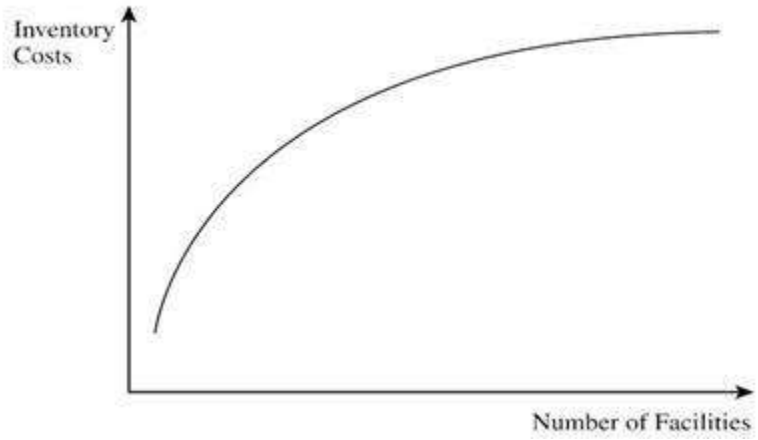


Figure 2 Number of Facilities Vs Inventory Costs

To decrease inventory costs, firms try to consolidate and limit the number of facilities in their supply chain network. For example, with fewer facilities, Amazon is able to turn its inventory about 12 times a year, whereas Borders, with about 400 facilities, achieves only about two turns per year.

Inbound transportation costs are the costs incurred in bringing material into a facility. Outbound transportation costs are the costs of sending material out of a facility. Outbound transportation costs per unit tend to be higher than inbound costs because inbound lot sizes are typically larger. For example, the Amazon warehouse receives full truckload shipments of books on the inbound side, but ships out small packages with only a few books per customer on the outbound side. Increasing the number of warehouse locations decreases the average outbound distance to the customer and makes outbound transportation distance a smaller fraction of the total distance travelled by the product. Thus, as long as inbound transportation economies of scale are maintained, increasing the number of facilities decreases total transportation cost, as shown in Figure 3.

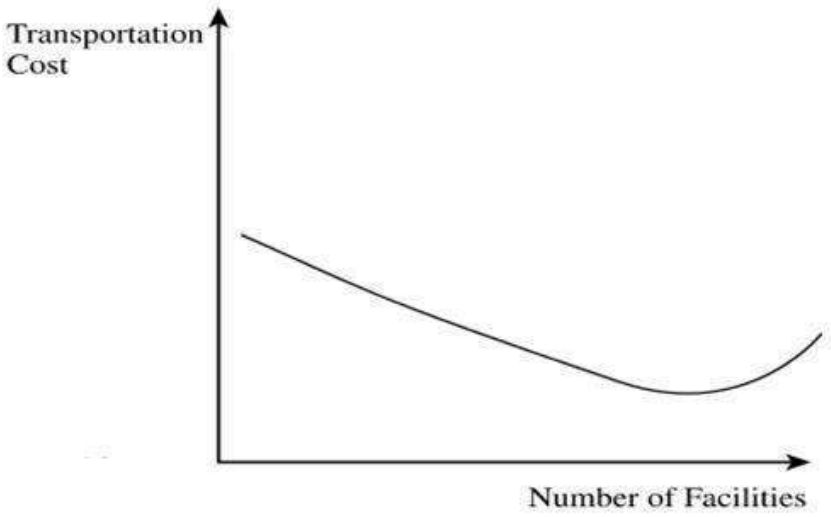


Figure 3 Number of Facilities Vs Transportation Costs

Facility costs decrease as the number of facilities is reduced, as shown in Figure 4, because a consolidation of facilities allows a firm to exploit economies of scale

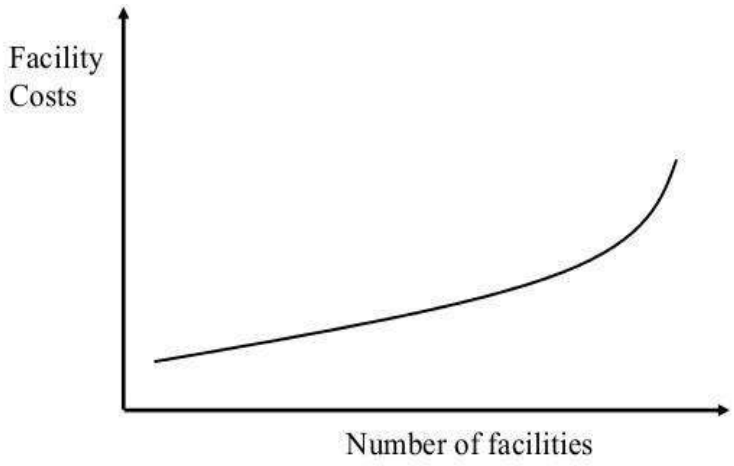


Figure 4 Number of Facilities Vs Facility Costs

Total logistics costs are the sum of inventory, transportation, and facility costs for a supply chain network. As the number of facilities increases, total logistics costs first decrease and then increase as shown in Figure 4-5. Each firm should have at least the number of facilities that minimize total logistics costs.

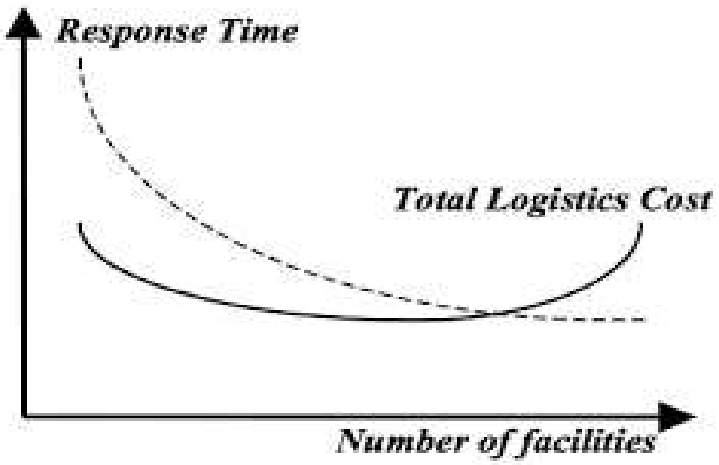


Figure 5 Number of facilities Vs Total Logistics Cost

2.3 Design Options for a Distribution Network

Distribution network choices from the manufacturer to the end consumer Two key decisions to be considered when designing a distribution

- network: Will product be delivered to the customer location or picked up from a prearranged site?
- Will product flow through an intermediary (or intermediate location)?

Based on the firm's industry and the answers to these two questions, one of six distinct distribution network designs may be used to move products from factory to customer.

These designs are classified as follows:

1. Manufacturer storage with direct shipping
2. Manufacturer storage with direct shipping and in-transit merge
3. Distributor storage with carrier delivery
4. Distributor storage with last-mile delivery
5. Manufacturer/distributor storage with customer pickup
6. Retail storage with customer pickup

1. Manufacturer Storage with Direct Shipping

The product is shipped directly from the manufacturer to the end customer, bypassing the retailer (who takes the order and initiates the delivery request). This option is also referred to as drop-shipping. The retailer carries no inventory. Information flows from the customer, via the retailer, to the manufacturer, and product is shipped directly from the manufacturer to customers.

It is best suited for a large variety of low-demand, high-value items for which customers are willing to wait for delivery and accept several partial shipments.

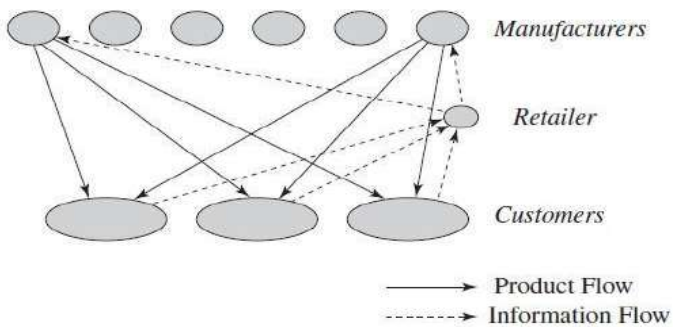


Figure 6 Manufacturer Storage with Direct Shipping

Inventory: Centralized inventories at the manufacturer who can aggregate demand across all retailers that it supplies. High level of product availability with lower levels of inventory.

Transportation: It is high because the average outbound distance to the end consumer is large, and package carriers are used to ship the product. Package carriers have high shipping costs per unit compared to truckload or less-than- truckload carriers.

Facilities and Handling: Fixed cost of facilities when using drop-shipping because all inventories are centralized at the manufacturer. This eliminates the need for other warehousing space in the supply chain.

Information: Good information infrastructure is needed between the retailers and the manufacturer so that the retailer can provide product availability information to the customer, even though the inventory is located at the manufacturer.

Response time: Response time tend to be long when drop-shipping is used because the order has to be transmitted from the retailer to the manufacturer and shipping distances are generally longer from the manufacturer's centralized site.

Product Variety and Availability: Manufacturer storage allows a high level of product variety to be available to the customer.

Customer Experience: Drop-shipping provides a good customer experience in the form of delivery to the customer location. However, suffers when a single order containing products from several manufacturers is delivered in partial shipments.

Time to Market: Drop-shipping allows a new product to be available to the market the day the first unit is produced.

Order visibility: Order visibility is difficult as two stages are involved in every customer order. Order tracking is harder to implement because it requires complete integration of information systems at both the retailer and the manufacturer.

Returnability: Handling of returns is more expensive and difficult because each order may involve shipments from more than one manufacturer. Returns can be handled in two ways.

- Customer return the product directly to the manufacturer. It incurs high transportation and coordination costs
- Retailer set up a separate facility (across all manufacturers) to handle returns. It requires investment in a facility to handle returns.

Advantages:

Drop-shipping offers the manufacturer the opportunity to postpone customization until after a customer has placed an order. **Postponement** lowers inventories by aggregating to the component level.

Performance Characteristics of Manufacturer Storage with Direct Shipping Network

Cost Factor	Performance
Inventory	Lower costs because of aggregation. Benefits of aggregation are highest for low-demand, high-value items. Benefits are large if product customization can be postponed at the manufacturer.
Transportation	Higher transportation costs because of increased distance and disaggregate shipping.
Facilities and handling	Lower facility costs because of aggregation. Some saving on handling costs if manufacturer can manage small shipments or ship from production line.
Information	Significant investment in information infrastructure to integrate manufacturer and retailer.

Service Factor	Performance
Response time	Long response time of one to two weeks because of increased distance and two stages for order processing. Response time may vary by product, thus complicating receiving.
Product variety	Easy to provide a high level of variety.
Product availability	Easy to provide a high level of product availability because of aggregation at manufacturer.
Customer experience	Good in terms of home delivery but can suffer if order from several manufacturers is sent as partial shipments.
Time to market	Fast, with the product available as soon as the first unit is produced.
Order visibility	More difficult but also more important from a customer service perspective.
Returnability	Expensive and difficult to implement.

2. Manufacturer Storage with Direct Shipping and In-Transit Merge

Unlike pure drop shipping, under which each product in the order is sent directly from its manufacturer to the end customer, in-transit merge combines pieces of the order coming from different locations so that the customer gets a single delivery.

For e.g., when a customer orders a PC from Dell along with a Sony monitor, the package carrier picks up the PC from the Dell factory and the monitor from the Sony factory; it then merges the two together at a hub before making a single delivery to the customer.

It is best suited for low-to medium demand, high value items the retailer is sourcing from a limited number of manufacturers.

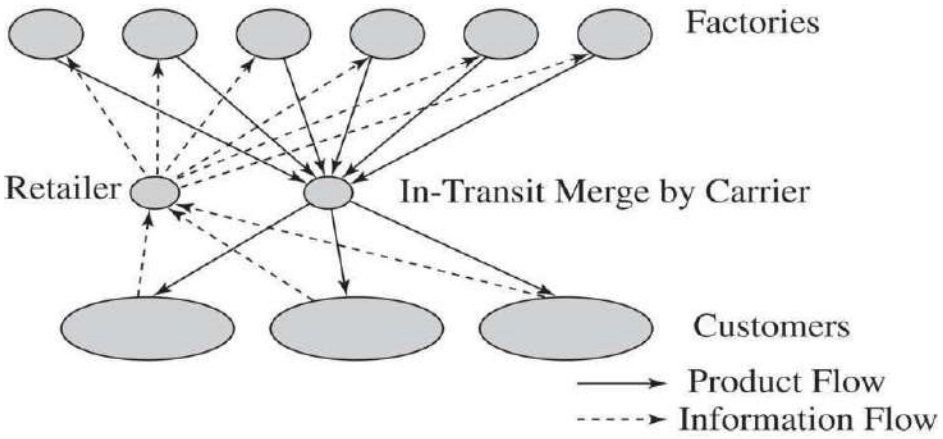


Figure 7 Manufacturer Storage with Direct Shipping and In-Transit Merge

Inventory: The ability to aggregate inventories and postpone product customization is a significant advantage of in-transit merge.

Transportation: Merge in-transit decreases transportation costs relative to drop- shipping by aggregating the final delivery.

Facilities and Handling: Facility and handling costs are somewhat higher than with drop-shipping. The party performing the in-transit merge has higher facility costs because of the merge capability required. Receiving costs at the customer are lower because a single delivery is received.

Information: Information infrastructure is higher than that for drop-shipping to allow in-transit merge. Operations at the retailer, manufacturers, and the carrier must be coordinated.

Response time: Response time is higher than the drop shipping network because of the need to perform the merge.

Product Variety and Availability: It is similar to drop shipping network.

Customer Experience: Response time is higher than the drop shipping network because of the need to perform the merge.

Time to Market: It is similar to drop shipping network.

Order visibility: Order visibility is better as tracking the shipment is comparatively easy from the carrier hub than drop shipping.

Returnability: Returnability is expensive and difficult to implement, compared with drop-shipping.

Advantages:

- Lower transportation cost
- Improved customer experience

Disadvantages:

- Additional effort during the merge
- Difficult to coordinate and implement when customer orders from too many sources.

Performance Characteristics of In-Transit Merge

Cost Factor	Performance
Inventory	Similar to drop-shipping. Ability to aggregate inventories and postpone product customization is a significant advantage of in-transit merge.
Transportation	Somewhat lower transportation costs than drop- shipping.
Facilities and handling	Handling costs higher than drop-shipping at carrier; receiving costs lower at customer.
Information	Investment is somewhat higher than for drop-shipping

Service Factor	Performance
Response time	Similar to drop-shipping; may be marginally higher because of the need to perform the merge.
Product variety	Similar to drop-shipping.
Product availability	Similar to drop-shipping.
Customer experience	Better than drop-shipping because only a single delivery has to be received.
Time to market	Similar to drop-shipping.
Order visibility	Similar to drop-shipping.
Returnability	Similar to drop-shipping. Problems in handling returns are likely, and the reverse supply chain will continue to be expensive and difficult to implement

3. Distributor Storage with Carrier Delivery

Inventory is not held by manufacturers at the factories but is held by distributors/retailers in intermediate warehouses, and package carriers are used to transport products from the intermediate location to the final customer.

It is well suited for slow- to fast-moving items. It also makes sense when customers want delivery faster than is offered by manufacturer storage but do not need it immediately.

Inventory: Relative to manufacturer storage, distributor storage requires a higher level of inventory because of a loss of aggregation.

Transportation: Transportation costs are lower for distributor storage compared to manufacturer storage.

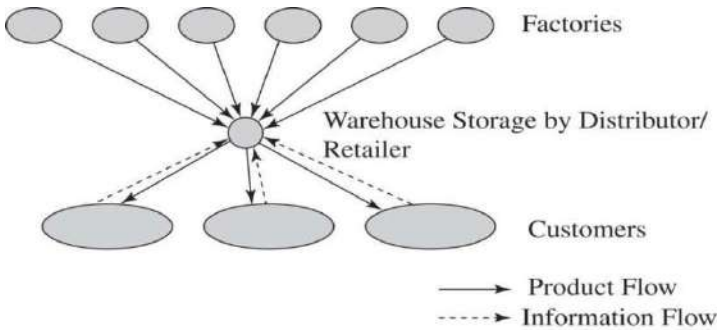


Figure 8 Distributor Storage with Carrier Delivery

Trucks are used for inbound shipments to the warehouse and outbound orders to the customer are bundled into a single shipment thereby reducing transportation cost. Whereas manufacturer storage, multiple shipments go out for a single customer order with multiple items thereby increase in transportation cost.

Facilities and Handling: Compared to manufacturer storage, facility costs (of warehousing) are higher with distributor storage because of a loss of aggregation. Distributor storage has higher processing and handling costs.

Information: Information infrastructure less complex than that needed with manufacturer storage. The distributor warehouse serves as a buffer between the customer and the manufacturer.

Response time: Response time is better than that of manufacturer storage because distributor warehouses are closer to customers and the entire order is aggregated at the warehouse before being shipped.

Product Variety: Distributor's Warehouse storage limits to some extent the variety of products that can be offered.

Product Availability: Higher cost is needed to provide the same level of availability as manufacturer storage.

Customer Experience: Customer convenience is high with distributor storage because a single shipment reaches the customer in response to an order.

Time to market: Time to market under distributor storage is higher than under manufacturer storage.

Order visibility: Order visibility is easier than with manufacturer storage because there is a single shipment from the warehouse to the customer.

Returnability: Returnability is better than with manufacturer storage because all returns are processed at the warehouse itself. The customer can return only one package, even if the items are from several manufacturers.

Performance Characteristics of Distributor Storage with Carrier Delivery

Cost Factor	Performance
Inventory	Higher than manufacturer storage. Difference is not large for faster moving items but can be large for very slow-moving items.
Transportation	Lower than manufacturer storage. Reduction is highest for faster moving items. Economic mode of transportation
Facilities and handling	Somewhat higher than manufacturer storage. The difference can be large for very slow-moving items. loss of aggregation
Information	Simpler infrastructure compared to manufacturer storage.

Service Factor	Performance
Response time	Faster than manufacturer storage. distributor warehouses are closer to customers, and the entire order is aggregated at the warehouse before being shipped
Product variety	Lower than manufacturer storage.
Product availability	Higher cost to provide the same level of availability as manufacturer storage.
Customer experience	Better than manufacturer storage with drop-shipping. Single shipment
Time to market	Higher than manufacturer storage. Only one stage (stock) need to be added in supply chain
Order visibility	Easier than manufacturer storage. Only one stage in supply chain
Returnability	Easier than manufacturer storage. all returns can be processed at the warehouse itself. The customer also has to return only one package, even if the items are from several manufacturers.

4. Distributor Storage with Last Mile Delivery

Last-mile delivery refers to the distributor/retailer delivering the product to the customer's home instead of using a package carrier.

The automotive spare parts industry is one in which distributor storage with last-mile delivery is the dominant model. It is too expensive for dealers to carry all spare parts in inventory. Thus, original equipment manufacturers (OEMs) tend to carry most spare parts at a local distribution center typically located no more than a couple of hours' drive from their dealers and often managed by a third party. The local distribution center is responsible for delivering needed parts to a set of dealers and makes multiple deliveries per day.

Unlike package carrier delivery, last-mile delivery requires the distributor warehouse to be much closer to the customer. Given the limited radius that can be served with last-mile delivery, more warehouses are required compared to when package delivery is used.

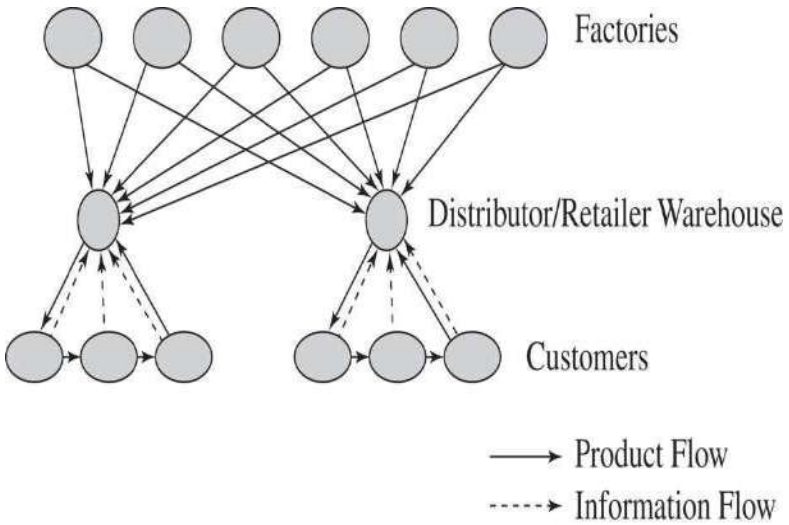


Figure 9 Distributor Storage with Last-Mile Delivery

Inventory: Higher levels of inventory than the other options (except for retail stores) because it has a lower level of aggregation. Warehouse stores fast-moving items that are needed quickly. Auto parts required by car dealers fit this description. **Transportation:** Transportation costs are highest for last-mile delivery, when delivering to individuals.

Facilities and Handling: Facility costs are lower than for a network with retail stores but higher than for either manufacturer storage or distributor storage with package carrier delivery.

Information: Information infrastructure with last-mile delivery is similar to that for distributor storage with package carrier delivery.

Response time: Response time is faster than using package carriers.

Product Variety: Product variety is lower than for distributor storage with carrier delivery.

Product Availability: Product Availability is higher than for every option other than retail stores

Customer Experience: Customer Experience is good using this option, particularly for bulky, hard-to-carry items.

Time to Market: Time to market is higher than for distributor storage with package carrier delivery.

Order Visibility: Order visibility is less of an issue as the deliveries are made within 24 hours.

Returnability: Returnability is best with last-mile delivery, because trucks making deliveries can also pick up returns from customers.

Disadvantages:

In areas with high labour costs, it is very hard to justify distributor storage with last mile delivery on the basis of efficiency or improved margin. It can only be justified if there is a large enough customer segment willing to pay for this convenience. An effort should be made to couple last mile delivery with an existing distribution network to exploit economies of scale and improve utilization.

Performance Characteristics of Distributor Storage with Last-Mile Delivery

Cost Factor	Performance
Inventory	Higher than distributor storage with package carrier delivery
Transportation	Very high cost given minimal scale economies. Higher than any other distribution option.
Facilities and handling	Facility costs higher than manufacturer storage or distributor storage with package carrier delivery, but lower than a chain of retail stores.
Information	Similar to distributor storage with package carrier delivery.

Service Factor	Performance
Response time	Very quick. Same day to next-day delivery.
Product variety	Somewhat less than distributor storage with package carrier delivery but larger than retail stores.
Product availability	More expensive to provide availability than any other option except retail stores.
Customer experience	Very good, particularly for bulky items.
Time to market	Slightly higher than distributor storage with package carrier delivery.
Order visibility	Less of an issue and easier to implement than manufacturer storage or distributor storage with package carrier delivery.
Returnability	Easier to implement than other previous options. Harder and more expensive than a retail network.

5. Manufacturer or Distributor Storage with Customer Pickup

Inventory is stored at the manufacturer or distributor warehouse but customers place their orders online or on the phone and then travel to designated pickup points to collect their merchandise. Orders are shipped from the storage site to the pickup points as needed.

Inventory: Inventory costs can be kept low, with either manufacturer or distributor storage.

Transportation: Transportation cost is lower than using package carriers as significant aggregation is possible when delivering orders to a pickup site. This allows the use of truckload or less-than-truckload carriers to transport orders to the pickup site.

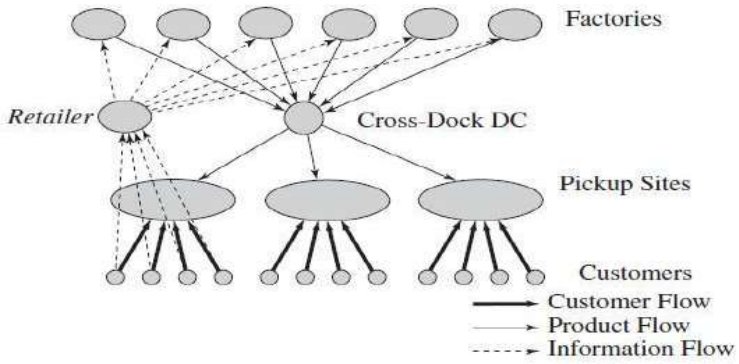


Figure 10 Manufacturer or Distributor Warehouse Storage with Consumer Pickup

Facilities and Handling: Facility costs are high if new pickup sites have to be built. using existing sites can lower the additional facility costs.

Information: Significant information infrastructure is needed to provide visibility of the order until the customer picks it up. Good coordination is needed among the retailer, the storage location, and the pickup location.

Response time: Response time comparable to that using package carriers can be achieved.

Product Variety and Availability: Product Variety and Product availability is comparable to any manufacturer or distributor storage option.

Customer Experience: Customer Experience is not appreciable as customers must pick up their own orders.

Time to Market: Time to market for new products can be as short as with manufacturer storage.

Order Visibility: Order visibility is comparatively difficult for customer pickups. The customer must be informed when the order has arrived, and the order should be easily identified once the customer arrives to pick it up.

Returnability: Returns can be handled at the pickup site, making it easier for customers.

Advantages:

- It reduces the delivery cost and expand the set of products sold and customers served online.

Disadvantages:

- Increased handling cost and complexity at the pickup site. So, if existing retail locations are used as pickup sites, it improves the economies.

Performance Characteristics of Network with Consumer Pickup Sites

Cost Factor	Performance
Inventory	Can match any other option, depending on the location of inventory
Transportation	Lower than the use of package carriers, especially if using an existing delivery network.
Facilities and handling	Facility costs can be high if new facilities have to be built. Costs are lower if existing facilities are used. The increase in handling cost at the pickup site can be significant.
Information	Significant investment in infrastructure required

Service Factor	Performance
Response time	Similar to package carrier delivery with manufacturer or distributor storage. Same-day delivery possible for items stored locally at pickup site.
Product variety	Similar to other manufacturer or distributor storage options.
Product availability	Similar to other manufacturer or distributor storage options.
Customer experience	Lower than other options because of the lack of home delivery. Experience is sensitive to capability of pickup location.
Time to market	Similar to manufacturer storage options.
Order visibility	Difficult but essential.
Returnability	Somewhat easier given that pickup location can handle returns.

6. Retail Storage with Customer Pickup

Inventory is stored locally at retail stores. Customers walk into the retail store or place an order online or by phone and pick it up at the retail store.

It is best suited for fast-moving items or items for which customers value rapid response.

Inventory: Local storage increases inventory costs because of the lack of aggregation.

Transportation: Transportation cost is much lower than with other solutions because inexpensive modes of transport can be used to replenish product at the retail store.

Facilities and Handling: Facility costs are high because many local facilities are required.

Information: Minimal information infrastructure is needed if customers walk into the store and place orders. A significant information infrastructure is needed to provide visibility of the online order. **Response time:** Good response times is achieved because of local storage.

Product Variety: Product variety stored locally is lower than under other options. **Product Availability:** It is more expensive to provide a high level of product availability.

Customer Experience: Customer experience depends on whether or not the customer likes to shop.

Time to Market: Time to market is the time taken because the new product has to penetrate through the entire supply chain before it is available to customers.

Order Visibility: Order visibility is extremely important for customer pickups when orders are placed online or by phone.

Returnability: Returns can be handled at the pickup site.

Advantages:

- Lower delivery costs
- Provide a faster response than other networks

Disadvantages:

- Increased inventory and facility costs

Performance Characteristics of Retail Storage at Consumer Pickup Sites

Cost Factor	Performance
Inventory	Higher than all other options.
Transportation	Lower than all other options.
Facilities and handling	Higher than other options. The increase in handling cost at the pickup site can be significant for online and phone orders.
Information	Some investment in infrastructure required for online and phone orders.

Service Factor	Performance
Response time	Same-day (immediate) pickup possible for items stored locally at pickup site.
Product variety	Lower than all other options.
Product availability	More expensive to provide than all other options.
Customer experience	Related to whether shopping is viewed as a positive or negative experience by customer
Time to market	Highest among distribution options.
Order visibility	Trivial for in-store orders. Difficult, but essential, for online and phone orders.
Returnability	Easier than other options because retail store can provide a substitute.

Selecting a Distribution Network Design

A network designer needs to consider product characteristics as well as network requirements when deciding on the appropriate delivery network.

The strengths and weaknesses of different networks are shown in Table below

Comparative Performance of Delivery Network Designs

	Retail Storage with Customer Pickup	Manufacturer Storage with Direct Shipping	Manufacturer Storage with In-Transit Merge	Distributor Storage with Package Carrier Delivery	Distributor Storage with Last-Mile Delivery	Manufacturer Storage with Pickup
Response time	1	4	4	3	2	4
Product variety	4	1	1	2	3	1
Product availability	4	1	1	2	3	1
Customer experience	Varies from 1 to 5	4	3	2	1	5
Time to market	4	1	1	2	3	1
Order visibility	1	5	4	3	2	6
Returnability	1	5	5	4	3	2
Inventory	4	1	1	2	3	1
Transportation	1	4	3	2	5	1
Facility and handling	6	1	2	3	4	5
Information	1	4	4	3	2	5

Key: 1 corresponds to the strongest performance and 6 the weakest performance.

Very few companies end up using a single distribution network. Most companies are best served by a combination of delivery networks. The combination used depends on product characteristics and the strategic position that the firm is targeting.

The suitability of different delivery designs (from a supply chain perspective) in various situations is shown in Table.

Delivery Networks for Different Product/Customer Characteristics

	Retail Storage with Customer Pickup	Manufacturer Storage with Direct Shipping	Manufacturer Storage with In-Transit Merge	Distributor Storage with Package Carrier Delivery	Distributor Storage with Last-Mile Delivery	Manufacturer Storage with Pickup
High-demand product	+2	–2	–1	0	+1	–1
Medium-demand product	+1	–1	0	+1	0	0
Low-demand product	–1	+1	0	+1	–1	+1
Very low-demand product	–2	+2	+1	0	–2	+1
Many product sources	+1	–1	–1	+2	+1	0
High product value	–1	+2	+1	+1	0	+2
Quick desired response	+2	–2	–2	–1	+1	–2
High product variety	–1	+2	0	+1	0	+2
Low customer effort	–2	+1	+2	+2	+2	–1

Key: +2 = very suitable; +1 = somewhat suitable; 0 = neutral; –1 = somewhat unsuitable; –2 = very unsuitable.

2.4 Distribution Networks in Practice

The ownership structure of the distribution network can have as big an impact as the type of distribution network.

- A manufacturer that owns its distribution network can control the network's action, if the manufacturer does not own the distribution network, a wide variety of issues need to be optimized over the network. To optimize a distribution network with multiple enterprises requires great skill in coordinating everyone in supply chain.

It is important to have adaptable distribution networks.

- Distribution networks must be able to adapt to changing technology and environments. An inability to adapt can be very damaging in these times of rapid change.

Product price, commoditization, and criticality affect the type of distribution system preferred by customers.

- Interactions between a buyer and a seller take time and resources.
- Mostly customers prefer a single enterprise that can deliver number of products. For low-value products like office supplies, most customers prefer a one-stop shop.
- For high-value, specialized, or critical products, customers are willing to have a relationship solely around that particular product. For high value products like laptops, customers prefer directly from the manufacturer.

Integrate the Internet with the existing physical network.

- To extract maximum benefit from the online channel for physical goods, firms should integrate it with their existing supply chain networks. Separating the two networks often results in inefficiencies within the supply chain. This coupling of the online channel with the existing physical network is referred to as "clicks-and- mortar."

2.5 The Role of Network Design in The Supply Chain

Supply chain network design decisions include the assignment of facility role; location of manufacturing-, storage-, or transportation-related facilities; and the allocation of capacity and markets to each facility. Supply chain network design decisions are classified as follows:

Facility role: What role should each facility play? What processes are performed

at each facility?

Facility location: Where should facilities be located?

Capacity allocation: How much capacity should be allocated to each facility? **Market and supply allocation:** What markets should each facility serve? Which supply sources should feed each facility?

Network design decisions have a significant impact on performance because they determine the supply chain configuration and set constraints within which the other supply chain drivers can be used either to decrease supply chain cost or to increase responsiveness. All network design decisions affect one another and must be made taking this fact into consideration. Decisions concerning the role of each facility are significant because they determine the amount of flexibility the supply chain has in changing the way it meets demand. For example, Toyota has plants located worldwide in each market that it serves. Before 1997, each plant was capable of serving only its local market. This hurt Toyota when the Asian economy went into a recession in the late 1990s. The local plants in Asia had idle capacity that could not be used to serve other markets that were experiencing excess demand. Toyota has added flexibility to each plant to be able to serve markets other than the local one.

This additional flexibility helps Toyota deal more effectively with changing global market conditions. Similarly, the flexibility of Honda's U.S. plants to produce both SUVs and cars in the same plant was helpful in 2008 when SUV demand dropped but small car demand did not.

Facility location decisions have a long-term impact on a supply chain's performance because it is expensive to shut down a facility or move it to a different location. A good location decision can help a supply chain be responsive while keeping its costs low. Toyota, for example, built its first U.S. assembly plant in Lexington, Kentucky, in 1988 and has continued to build new plants in the United States since then. The U.S. plants proved profitable for Toyota when the yen strengthened and cars produced in Japan were too expensive to be cost competitive with cars produced in the United States. Local plants allowed Toyota to be responsive to the U.S. market while keeping costs low.

Whereas capacity allocation can be altered more easily than location, capacity decisions do tend to stay in place for several years. Allocating too much capacity to a location results in poor utilization and, as a result, higher costs. Allocating too little capacity results in poor responsiveness if demand is not satisfied or high cost if demand is filled from a distant facility.

The allocation of supply sources and markets to facilities has a significant impact on performance because it affects total production, inventory, and transportation costs incurred by the supply chain to satisfy customer demand. This decision should be reconsidered on a regular basis so that the allocation can be changed as production and transportation costs, market conditions, or plant capacities change. Of course, the allocation of markets and supply sources can be changed only if the facilities are flexible enough to serve different markets and receive supply from different sources.

Network design decisions must be revisited as market conditions change or when two companies merge. For example, as its subscriber base grew, Netflix added about 60 DCs by 2010 across the United States to lower transportation cost and improve responsiveness. With the growth in video streaming and the corresponding drop in DVD rentals, Netflix anticipated closing some of its DCs as

DVD rental demand started to drop. Changing the location and demand allocation of DCs with changing demand has been critical to maintaining low cost and responsiveness at Netflix. Following a merger, consolidating some facilities and changing the location and role of others can often help reduce cost and improve responsiveness because of the redundancies and differences in markets served by either of the two separate firms. Network design decisions may also need to be revisited if factor costs such as transportation have changed significantly. In 2008, P&G announced that it would rethink its distribution network, which was implemented when the "cost of oil was $10 per barrel."

We focus on developing a framework as well as methodologies that can be used for network design in a supply chain.

2.6 Framework For Network Design Decisions

The goal when designing a supply chain network is to maximize the firm's profits while satisfying customer needs in terms of demand and responsiveness. To design an effective network, a manager must consider all the factors described and Global network design decisions are made in four phases.

Phase I: Define a Supply Chain Strategy/Design

The objective of the first phase of network design is to define a firm's broad supply chain design. This includes determining the stages in the supply chain and whether each supply chain function will be performed

in-house or outsourced .

Phase I starts with a clear definition of the firm's competitive strategy as the set of customer needs that the supply chain aims to satisfy. The supply chain strategy then specifies what capabilities the supply chain network must have to support the competitive strategy. Next, managers must forecast the likely evolution of global competition and whether competitors in each market will be local or global

players. Managers must also identify constraints on available capital and whether growth will be accomplished by acquiring existing facilities, building new facilities, or partnering.

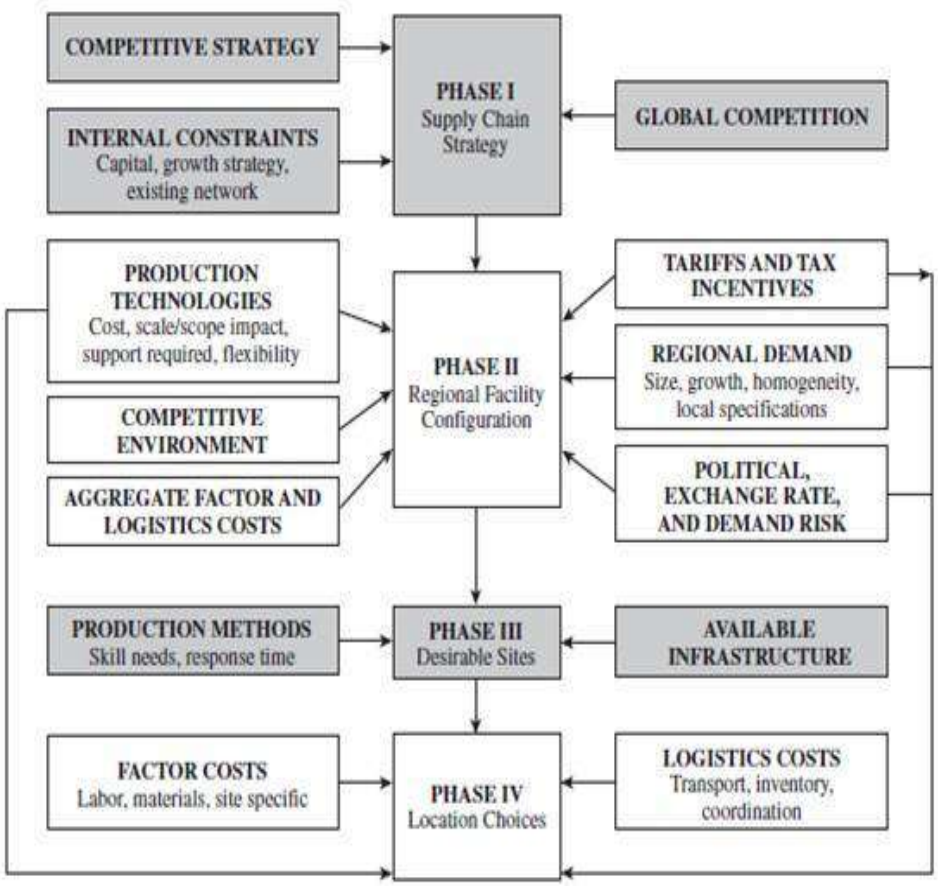


Figure 11 Framework for Network Design Decisions

Based on the competitive strategy of the firm, its resulting supply chain strategy, an analysis of the competition, any economies of scale or scope, and any constraints, managers must deter- mine the broad supply chain design for the firm.

Phase II: Define the Regional Facility Configuration

The objective of the second phase of network design is to identify regions where facilities will be located, their potential roles, and their approximate capacity. An analysis of Phase II starts with a forecast of the demand by country or region. Such a forecast must include a measure of the size of the demand and a determination of the homogeneity or variability of customer requirements across different regions. Homogeneous requirements favor large consolidated facilities, whereas requirements that vary across countries favor smaller, localized facilities. The next step is for managers to identify whether economies of scale or scope can play a significant role in reducing costs, given available production technologies. If economies of scale or scope are significant, it may be better to have a few facilities serving many markets. For example, semiconductor manufacturers such as Advanced Micro Devices have few plants for their global markets, given the economies of scale in production. If economies of scale or scope is not significant, it may be better for each market to have its own facility.

Next, managers must identify demand risk, exchange-rate risk, and political risk associated with regional markets. They must also identify regional tariffs, any requirements for local production, tax incentives, and any export or import restrictions for each market.

The tax and tariff information is used to identify the best location to extract a major share of the profits. In general, it is best to obtain the major share of profits at the location with the lowest tax rate.

Managers must identify competitors in each region and make a case for whether a facility needs to be located close to or far from a competitor's facility. The desired response time for each market and logistics costs at an aggregate level in each region must also be identified.

Based on all this information, managers identify the regional facility configuration for the supply chain network using network design models discussed in the next section. The regional configuration defines the approximate number of facilities in the network, regions where facilities will be set up, and whether a facility will produce all products for a given market or a few products for all markets in the network.

Phase III: Select a Set of Desirable Potential Sites

The objective of Phase III is to select a set of desirable potential sites within each region where facilities are to be located. Sites should be selected based on an analysis of infrastructure availability to support the desired production methodologies. Hard infrastructure requirements include the availability of suppliers, transportation services, communication, utilities, and ware- housing facilities. Soft infrastructure requirements include the availability of a skilled workforce, workforce turnover, and the community receptivity to business and industry.

Phase IV: Location Choices

The objective of Phase IV is to select a precise location and capacity allocation for each facility. Attention is restricted to the desirable potential sites selected in Phase III. The network is designed to maximize total profits, taking into account the expected margin and demand in each market, various logistics and facility costs, and the taxes and tariffs at each location.

Part-A Questions and Answers

1. **What is supply chain network design? (K1,CO3)**

   Supply Chain Network Design is a systematic approach to determining the best location and optimal size of the facilities to be included in the supply chain, and to ensure an optimal flow of products using advanced mathematical modelling.

2. **What are the role of network design in supply chain? (K1,CO3)** The role of network design in supply chain involves the following: Defining the business objectives
   Defining the project scope
   Determining the analyses to be performed Determining the tools to be utilized
   Completion of the project in accordance with the design.

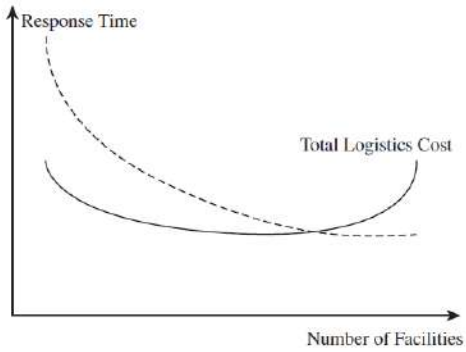
3. **What is distribution management? (K1,CO2)**

   Distribution management includes forecasting, transportation, warehousing, and delivery within the larger universe of logistics and supply chain management. These require precise tracking, real-time information, and highly- skilled staffing to execute effectively.

4. **What are the two key decisions when designing a distribution network? (K1,CO2)**

   1. The customer needs that are being met.
   2. Cost of the network or costs incurred in the meeting those needs.

5. **What is the relationship between total logistic cost and the number of facilities in the distribution network? (K1,CO2)**



6. **Outbound transportation cost per unit tend to be higher than inbound costs. Why? (K1,CO2)**

Outbound transportation may include customized service. Customization will have high impact on cost. Mass package is not possible in outbound transportation.

7. **What is Drop shipping? (K1,CO2)**

The product is shipped directly from the manufacturer to the end customer, bypassing the retailer (who takes the order and initiates the delivery request). This is referred as drop-shipping.

8. **What is Last-mile delivery? (K1,CO2)**

Last-mile delivery refers to the distributor/retailer delivering the product to the
customer's home instead of using a package carrier.

9. List the various distribution network designs. (K1,CO2)

The designs are classified as follows:

1. Manufacturer storage with direct shipping
2. Manufacturer storage with direct shipping and in-transit merge
3. Distributor storage with carrier delivery
4. Distributor storage with last-mile delivery
5. Manufacturer/distributor storage with customer pickup
6. Retail storage with customer pickup

10. What are the advantages of drop shipping method? (K1,CO2)

Drop-shipping offers the manufacturer the opportunity to postpone customization until after a customer has placed an order. **Postponement** lowers inventories by aggregating to the component level.

Centralized inventories at the manufacturer who can aggregate demand across all retailers that it supplies. **High level of product availability** with **lower levels of inventory**. This reduces the need of inventories and thereby inventory cost.

11. What is distribution? (K1,CO2)

- Distribution refers to the steps taken to move and store a product from the supplier stage to a customer stage in the supply chain.
- Distribution occurs between every pair of stages in the supply chain. Raw materials and components are moved from suppliers to manufacturers, whereas finished products are moved from the manufacturer to the end consumer.
- Distribution is a key driver of the overall profitability of a firm because it affects both the supply chain cost and the customer value directly.

12. What are the main factors to be considered in designing distribution network? (K1,CO2)

Cost Factor

Inventory

Transportation

Facilities and

Handling

Information

Service Factor

Response time

Product variety

Product

availability

Customer

Experience Time

to market Order

visibility

Returnability

13. What are the advantages and disadvantages of in-transmit merge method? (K1,CO2)

Advantages:

Lower transportation cost

Improved customer

experience

Disadvantages:

Additional effort during the merge

Difficult to coordinate and implement when customer orders

from too many sources.

14. Give a note on Distributor Storage with Carrier Delivery method. ? (K1,CO2)

Inventory is not held by manufacturers at the factories but is held by distributors/retailers in intermediate warehouses, and package carriers are used to transport products from the intermediate location to the final customer.

15. What are the supply chain decisions? (K1,CO2)

Facility role: What role should each facility play? What processes are performed at each facility?

Facility location: Where should facilities be located?

Capacity allocation: How much capacity should be allocated to each facility? **Market and supply allocation:** What markets should each facility serve? Which supply sources should feed each facility?

16. Define Facility location decisions. (K1,CO2)

Facility location decisions have a long-term impact on a supply chain's performance because it is expensive to shut down a facility or move it to a different location.

17. What is the first phase of network design? (K1,CO3)

The objective of the first phase of network design is to define a firm's broad supply chain design. This includes determining the stages in the supply chain and whether each supply chain function will be performed in-house or outsourced.

18. Define in bound and out bound transportation. (K1,CO2)

Inbound transportation costs are the costs incurred in bringing material into a facility.

Outbound transportation costs are the costs of sending material out of a facility.

19. What is response time? (K1,CO2)

Response time is the amount of time it takes for a customer to receive an order.

20. What do you mean by product variety ? (K1,CO2)

Product variety is the number of different products/configurations that are offered

by the distribution network.

21. Define product availability ? (K1,CO2)

Product availability is the probability of having a product in stock when a customer order arrives.

22. What is time to market? (K1,CO2)

Time to market is the time it takes to bring a new product to the market.

23. What is Order visibility? (K1,CO2)

Order visibility is the ability of customers to track their orders from placement to delivery.

24. Define returnability. (K1,CO2)

Returnability is the ease with which a customer can return unsatisfactory merchandise and the ability of the network to handle such returns.

25. What do you mean by customer experience ? (K1,CO2)

Customer experience includes the ease with which customers can place and receive orders and the extent to which this experience is customized. It also includes purely experiential aspects, such as the possibility of getting a cup of coffee and the value that the sales staff provides.

Part-B Questions

Q. No.	Questions		
1	Explain in detail about Factors influencing Distribution network design.		
2	Explain the various distribution network designs with diagram and its performance factors.		
3	List out the advantages and disadvantages of various distribution design options and which one is best suited for Amazon online shopping. Justify.		
4	Enumerate the distribution networks in practice.		
5	Discuss in detail about the role of network design in supply chain.		
6	Enumerate the framework for network design decisions.		

Table of Contents

UNIT 3 – LOGISTICS IN SUPPLY CHAIN

3.1 THE ROLE OF TRANSPORTATION IN SUPPLY CHAIN

Transportation refers to the movement of product from one location to another as it makes its way from the beginning of a supply chain to the customer. Transportation is an important supply chain driver because products are rarely produced and consumed in the same location. Transportation is a significant component of the costs incurred by most supply chains. According to the Bureau of Transportation Statistics (BTS), "over 19 billion tons of freight, valued at $13 trillion, was carried over 4.4 trillion ton-miles in the United States in 2002."[1] Only three sectors— housing, health care, and food—contributed a larger share to the gross domestic product (GDP) than transportation. Transportation-related jobs employed nearly 20 million people in 2002, accounting for 16 percent of U.S. total occupational em

Transportation is "Moving of Goods from the point of origin to the point of destination across the supply chain through its members". The role of transportation in logistics has changed with transport alternatives and value added services to support the supply chain. Private transport companies have become part of the supply chain. Third party logistics companies offer a wide variety of services like to improve the efficiency of the supply chain:

- Product sorting
- Sequencing
- Customized freight delivery

Advancement in technology has improved the services of the logistics with real time visibility of tracking the vehicles using GPRS systems and the integration of delivery systems.

Development of an integrated transport management

- system has Improved the product delivery.
- Reduced the inventory and material handling resulting in reduced transportation cost.
- Improved value of transportation for the supply chain

Transportation is the back bone of an economy

Improved transportation infrastructure helps the economy to grow.
Transportation is an important element of logistics.

- It helps the supply chain in Movement of
- products Storage of products

Movement of product from the suppliers to the customers across the supply chain has become the primary value preposition of transportation in a supply chain.

Its performance is based

- on Time
- Cost
- Impact on the environment

- Information technology has improved the access to data on the movement of the transport and the time of delivery.
- It helps the supply chain to plan the exact time of delivery to the customers.
- On an average transportation cost is about 30% of the total logistics cost.
- Transportation is a non-value adding activity.
- The focus of the supply chain is to reduce the cost of transportation to decrease the total supply chain cost.
- Transportation industry is the largest consumer of fuel and oil across the globe.
- Efforts have been on to bring in fuel efficient vehicles which are environment friendly as they affect the environment through air pollution and noise pollution.

The role of transportation is even more significant in global supply chains. According to the BTS, the U.S. freight transportation network carried export and import merchandise worth more than $2.2 trillion in 2004, an increase of 168 percent from $822 billion in 1990. During the same period, the ratio of exports from and imports into the United States to the GDP increased from 12 percent to 21 percent. Transportation-related jobs employed nearly 20 million people. The role of transportation is even more significant in global supply chains. According to the BTS, the U.S. freight transportation network carried export and import merchandise.

Any supply chain's success is closely linked to the appropriate use of transportation. IKEA, the Scandinavian home furnishings retailer, has built a global network with about 270 stores in 26 countries primarily on the basis of effective transportation. IKEA's sales for the year ending August 2009 reached 21.5 billion euros. Its strategy is built around providing good-quality products at low prices. Its goal is to cut prices by 2 to 3 percent each year. As a result, IKEA works hard to find the most inexpensive global source for each of its products. Modular design of its furniture allows IKEA to transport its goods worldwide much more cost effectively than a traditional furniture manufacturer.

The large size of IKEA stores and shipments allows inexpensive transportation of home furnishings all the way to the retail store. Effective sourcing and inexpensive transportation allow IKEA to provide high-quality home furnishings at low prices globally.

Seven-Eleven Japan is another firm that has used transportation to achieve its strategic goals. The company has a goal of carrying products in its stores to match the needs of customers as they vary by geographic location or time of day. To help achieve this goal, Seven-Eleven Japan uses a responsive transportation system that replenishes its stores several times a day so that the products available match customers' needs. Products from different suppliers are aggregated on trucks according to the required temperature to help achieve frequent deliveries at a reasonable cost. Seven-Eleven Japan uses a responsive transportation system along with aggregation to decrease its transportation and receiving costs while ensuring that product availability closely matches customer demand.

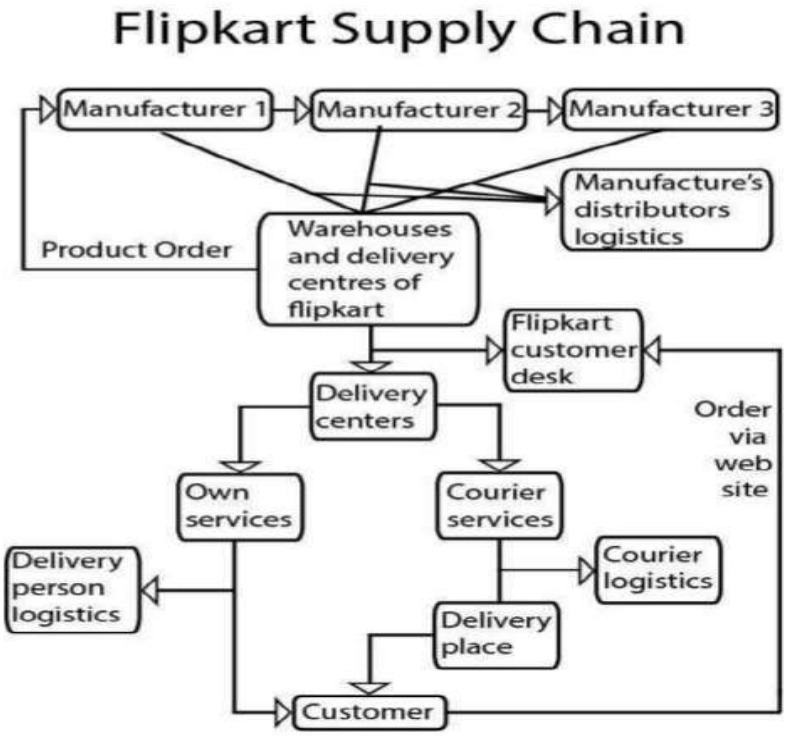
Supply chains also use responsive transportation to centralize inventories and operate with fewer facilities. For example, Amazon relies on package carriers and the postal system to deliver customer orders from centralized warehouses.

Transportation has allowed Netflix to operate a movie rental business without any stores. The company uses responsive transportation provided by the postal system along with suitably located warehouses to allow its customers to receive and return movies they want to watch.

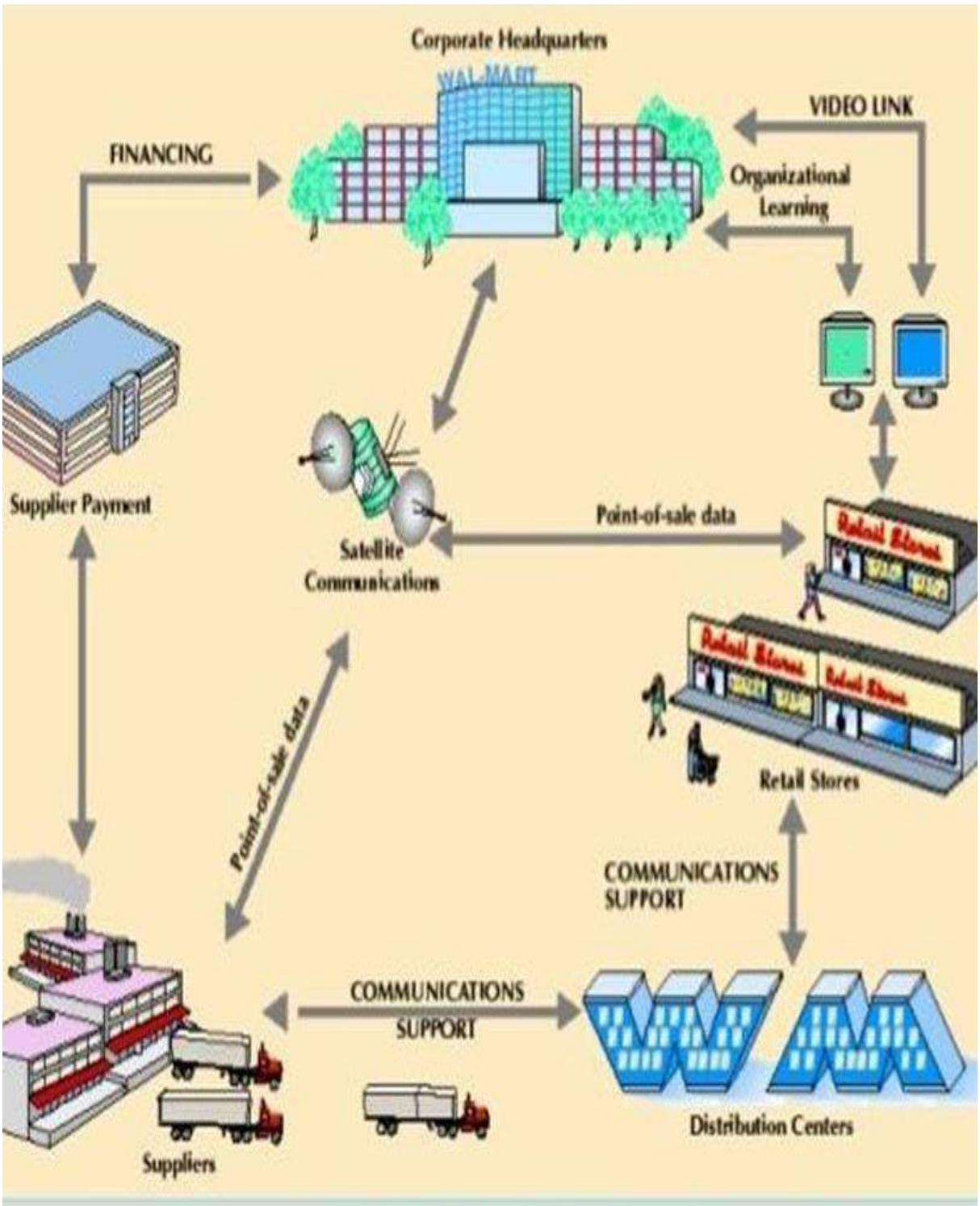
The shipper is the party that requires the movement of the product between two points in the supply chain. The carrier is the party that moves or transports the product. For example, when Netflix uses USPS to ship its DVDs from the warehouse to the customer, Netflix is the shipper and USPS is the carrier. Besides the shipper and the carrier, two other parties have a significant impact on

transportation: (1) the owners and operators of transportation infrastructure such as roads, ports, canals, and airports and (2) the bodies that set transportation policy worldwide. Actions by all four parties influence the effectiveness of transportation.

FLIPKART'S SUPPLY CHAIN

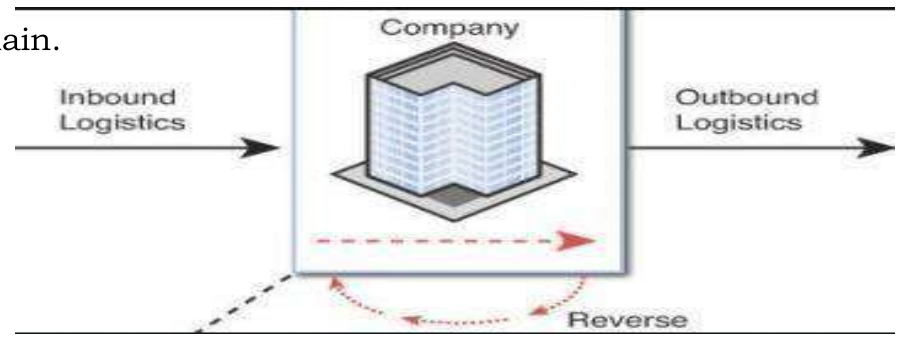


WALMARTS SUPPLY CHAIN



To understand transportation in a supply chain, it is important to consider the perspectives of all four parties. A carrier makes investment decisions regarding the transportation equipment (locomotives, trucks, airplanes, etc.) and in some cases infrastructure (rail) and then makes operating decisions to try to maximize the return from these assets. A shipper, in contrast, uses transportation to minimize the total cost (transportation, inventory, information, sourcing, and facility) while providing an appropriate level of responsiveness to the customer. The effectiveness of carriers is influenced by infrastructure such as ports, roads, waterways, and airports. Most transportation infrastructure throughout the world is owned and managed as a public good. It is important that infrastructure be managed in such a way that monies are available for maintenance and investment in further capacity as needed. Transportation policy sets direction for the amount of national resources that go into improving transportation infrastructure. Transportation policy also aims to prevent abuse of monopoly power; promote fair competition; and balance environmental, energy, and social concerns in transportation.

Transportation is a key supply chain process that must be included in supply chain strategy development, network design and total cost management. Transportation provides the critical links between supply chain partners, permitting goods to flow between their facilities. Transportation service availability is critical to demand fulfillment in the supply chain. Transportation efficiency promotes the competitiveness of a supply chain.



Transportation involves the movement of product, service/transit time, and cost traditional key issues of effective supply chains. It also impacts with the other issues of movement of information and integration within and among suppliers, customers and carriers. A transportation strategy, to be effective in supply chain management, is fitting the movement of goods to the corporate supply chain. It is not playing one carrier off against another. Rather it is a way to respond to the dynamics of the business, its customers, suppliers' and operation.

Actions by all the below parties influence the effectiveness of transportation.

- The shipper is the party that requires the movement of the product between two points in the supply chain. Shipper is the person or company who is usually the supplier or owner of commodities shipped. Also called Consignor.
- The carrier is the party that moves or transports the product. Carrier is a person or company that transports goods or people for any person or company and that is responsible for any possible loss of the goods during transport.
- The owners and operators of transportation infrastructure such as roads, ports, canals, and airports
- The bodies that set transportation policy worldwide.
- Consignee (party that receives the shipment)

May have certain responsiveness needs.

The transportation strategy should recognize:

- **Segment:** Each shipment does not have the same priority. Products, suppliers, customers, time of the year, and other factors can affect the importance and urgency of transport movements. The strategy cannot be one-dimensional. It should be segmented to reflect urgencies.

- **Customer requirements:** The supply chain involves continuous and efficient movement of product from vendor to manufacturer to customer. Therefore, the transportation program must reflect and meet customer needs. The time and service aspects of transportation are vital.

- **Shipments must move timely:** Customers demand their shipments be delivered as they require – on the date needed, by the carrier preferred, in the proper shipping packaging method and complete, both shipped complete and delivered complete and in good order. Being able to have a transportation program which can do this provides customer satisfaction and can give your company a competitive advantage.

- **Mode selection:** How will products move, by air versus surface? What modes will be used? What roles do transit time play in your supply chain? How will the inventory and service impacts be measured as compared to the freight charges?

- **Carrier relationships:** Volume creates carrier/forwarder attention. Even if there is no strategy, the number of carriers trying to get business will make firms develop one. Infrequent shipping dictates another approach.

- **Carrier mergers and alliances and closings:** This is an important and difficult issue. Firms should understand what is happening within each mode and align the strategy with carriers who will still be viable in the future—often five years since strategic plans may extend that far. A great strategy with a carrier who is taken over or goes out of business is suddenly not a good strategy.

- **Flexibility/Adaptability:** Change is happening. It is not a question of whether or not it happens. The only question is how quickly it occurs. The strategy should be able to change. New customers. New products. New businesses. New suppliers. New corporate emphasis. Each of these can dramatically impact the strategy. The times they are a changing--and so will the strategy.

- **Measuring/Metrics.** It is important to know how well the strategy and carriers are performing. This takes two approaches. One is measuring. Measuring means comparing performance versus agreed standards. What is the actual delivery to customer performance, on a macro basis, carrier and customer by customer basis? A macro measure can hide a problem even if the overall measure is good. And, with supply chain management, this means realizing primary customers and delivery locations. A test of measuring costs is how well the transport spend is being managed. Transport performance metrics can provide a way to view the value of the spend.

Transportation is a key logistics function and is critical to supply chain performance. To meet the vigorous requirements of the supply chain, the strategy should be dynamic. It must be responsive, both as to service and cost demands.

3.2 FACTORS AFFECTING TRANSPORTATIONS DECISION

From earlier, we know that there may be three separate parties involved. All of them have factors to consider:

1. Carrier (party that moves or transports the product) – Vehicle-related costs, Fixed operating costs, Trip-related costs – Often incurs huge investments (new fleets, etc…)

2. Shipper (party that requires the movement of the product between two points in the supply chain) – May need to balance Transportation costs with Inventory and Facility costs

3. Consignee (party that receives the shipment) – May have certain responsiveness needs We should also consider:

4. The owners of the infrastructure (Ports, highways, railroads)

5. Government and/or bodies that set worldwide transportation policy

The selection of a mode of transportation or service offering within a mode of transportation depends on a variety of service characteristics

- Freight rates (cost of service)
- Reliability
- Transit time
- Loss, damage, claims processing tracing
- Shipper market considerations
- Carrier considerations.

Other factors

- Capability
- Availability & adequacy of
- equipment Availability of service
- Frequency of
- service Security
- Claims
- handling

  Shipment

  tracing
- Problem-solving assistance

Speed & dependability affect both the seller's & buyer's inventory level, as well as the inventory that is in transit. Slower, less reliable modes require more inventory in the distribution channel. When alternative modes are available, the one chosen should be the one that offers the lowest total cost consistent with customer service goals.

Transportation influences or is influenced by many logistics activities to include:

- **Transportation costs** which are directly affected by the location of the firm's manufacturing facilities, warehouses, suppliers, retailers and customers.
- The **transportation mode** selected has an impact on the packing required, and carrier classification rules determine package choice.
- **Inventory requirements** which are influenced by the mode of transport selected for use. When high speed, high priced transportation systems are used, the inventories required to be maintained in the logistics system would be smaller as compared to that when slow, less expensive transportation systems are used.
- **Customer service goals** which influence the type and quality of carrier source selected by the firm.

- The firm's **materials handling equipments** are determined by the type of carrier used for transportation, for example, the handling equipments for loading and unloading the carrier and the design of the receiving and shipping docks depend on the type of carriers used.
- An **order management methodology** which encourage maximum consolidation of shipments between common points facilities larger shipments and advantages of volume discounts.

Location decisions are a basic determinant of profitability in logistics. Decisions on where to manufacture, to assemble, to store, to transship and to consolidate can make the difference between profit and loss. Because of differences in basic factor costs and because of exchange rate movements, location decisions are very important. Location decisions can have a continuing impact over time on the company's financial and competitive position.

Facility Location Models

- Break Even Analysis
- Transportation Models

Factors Affecting Transportation Decisions:

- Modes of transportation
- Size and weight
- Destination
- Routes
- Carrier
- Shipper

MODES OF TRANSPORTATION

Supply chains use a combination of the following modes of

- transportation: Air
- Package carriers
- Truck
- Rail
- Water
- Pipelin e
- Intermodal

AIRLINES

They have three cost components:

- a fixed cost of infrastructure and equipment
- cost of labor and fuel that is independent of the passengers or cargo on a flight but is fixed for a flight
- a variable cost that depends on the passengers or cargo carried.

Air carriers offer a fast and fairly expensive mode of transportation for cargo. Small, high-value items or time- sensitive emergency shipments that have to travel a long distance are best suited for air transport.

Key issues that air carriers face include identifying the location and number of hubs, assigning planes to routes, setting up maintenance schedules for planes, scheduling crews, and managing prices and availability at different prices.

Advantages

- Fastest mode of transport
- Transporting goods to area which is not easily accessible by other
- means Reduces lead time
- Improved service levels

Disadvantages

- Expensive
- Not suitable for transporting heavy and bulky goods Not suitable for short distance travel

PACKAGE CARRIERS

Package carriers are transportation companies such as FedEx, UPS, and the U.S. Postal Service. They use air, truck, and rail to transport time-critical smaller packages. They also provide other value-added services such as package tracking and in some cases processing and assembly of products. They are the preferred mode of transport for online businesses

Advantages

- Shippers is rapid and reliable
- delivery. E-Business
- Consolidation of shipments

Disadvantages

- Expensive
- Small and time-sensitive shipments

TRUCK

Truck have complete freedom to use roads. It support flexibility of changes in location, direction, speed and timing of travel.

Significant fraction of the goods are moved by

- **Truckload (TL)**
  - Low fixed cost
  - Imbalance between flows

- **Less than truckload (LTL)**
  - Small lots
  - Hub and spoke
  - system May take longer than TL

Advantages

- Cheaper
- Flexible
- Any place can be reachable

Disadvantages

- Not economical for long distance
- High cost for bulk and heavy loads

RAIL

- Uses freight rails
- Bulk shipment of products from production plant to warehouses
- Move commodities over large distances
- High fixed costs in equipment and facilities
- Scheduled to maximize utilization
- Transportation time can be long – Trains 'built' not scheduled

Advantages

- Faster
- Suitable for carrying heavy goods
- Cost effective

Disadvantages

- Expensive for carrying goods in short distance
- Not available for remote areas
- Fixed time schedule and not flexible for loading and unloading of goods at any place.

WATER

- Limited to certain geographic areas
- Ocean, inland waterway system, coastal waters certain geographic areas.
- Very large loads at very low cost – Lowest energy/emission intensity per tonne- km, though some concern about port pollution
- Slowest – Also subject to bottlenecks at Ports
- Dominant in overseas trade (autos, grain, apparel, etc.)
- Containers
- Example of successful usage: IKEA - IKEA makes very strong use of water and other low cost transport

PIPELINE

- High fixed cost
- Primarily for crude petroleum, refined petroleum products, natural gas
- Best for large and predictable demand
- Would be used for getting crude oil to a port or refinery, but not for getting refined gasoline to a gasoline station
- Pricing structure encourages use for predicable component of demand

INTERMODAL

- Use of more than one mode of transportation to move a shipment to its destination – rail/truck, water/rail/truck or water/truck
- Grown considerably with increased use of containers
- Increased global trade has also increased use of intermodal transportation
- More convenient for shippers (one entity can provide the complete service)
- Key issue involves the exchange of information to facilitate transfer between different transport modes

DIFFERENCE BETWEEN CARRIER AND SHIPPER

Carrier

A carrier makes investment decisions regarding the transportation. Equipment and in some cases infrastructure and then makes operating decisions to try to maximize the return from these assets.

- Vehicle related cost
- Fixed operating cost
- Trip related cost

Shipper

A shipper uses transportation to minimize the total cost (transportation, inventory, information, sourcing, and facility) while providing an appropriate level of responsiveness to the customer.

Most transportation infrastructure throughout the world is owned and managed as a public good.

3.3 DESIGN OPTION FOR A TRANSPORTATION NETWORK

The design of a transportation network affects the performance of a supply chain by establishing the infrastructure within which operational transportation decisions regarding scheduling and routing are made. A well- designed transportation network allows a supply chain to achieve the desired degree of responsiveness at a low cost. Three basic questions need to be considered when designing a transportation network between two stages of a supply chain:

1. Should transportation be direct or through an intermediate site?
2. Should the intermediate site stock product or only serve as a cross-docking location?
3. Should each delivery route supply a single destination or multiple destinations (milk run)?

Based on the answers to these questions, the supply chain ends up with a variety of transportation networks. We discuss these options and their strengths and weaknesses in the context of a buyer with multiple locations sourcing from several suppliers.

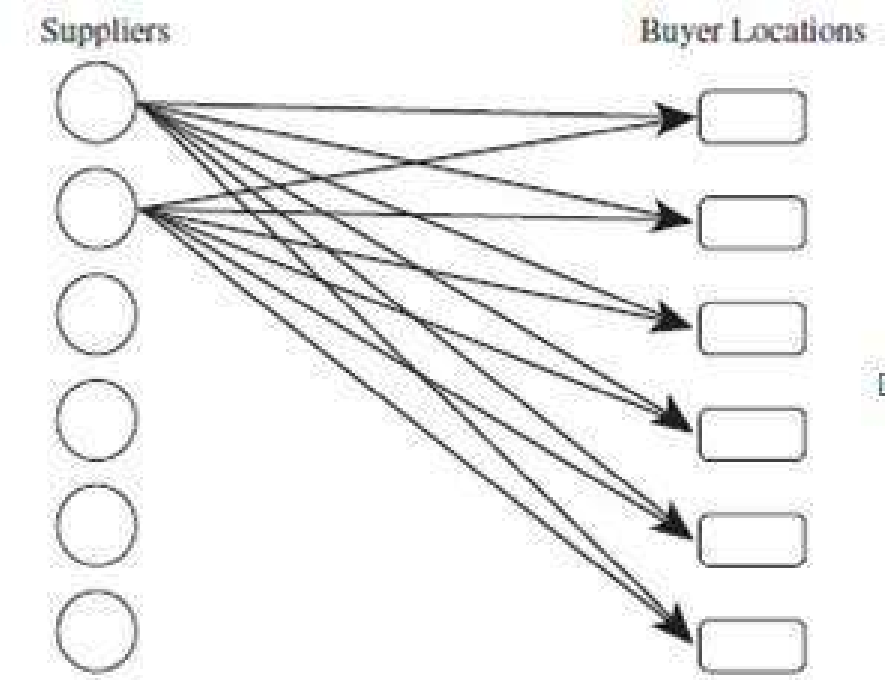
Direct Shipment Network to Single Destination

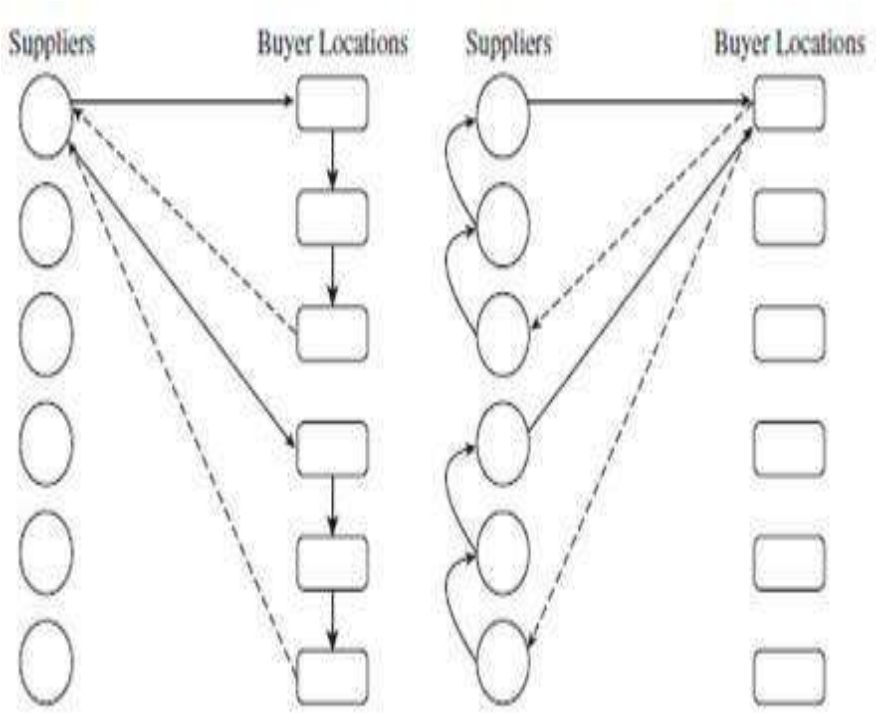
With the direct shipment network to a single destination option, the buyer structures the transportation network so that all shipments come directly from each supplier to each buyer location, as shown in Figure. With a direct shipment network, the routing of each shipment is specified, and the supply chain manager needs to decide only the quantity to ship and the mode of transportation to use. This decision involves a trade-off between transportation and inventory costs, as discussed later in the chapter.

The major advantage of a direct shipment transportation network is the elimination of intermediate warehouses and its simplicity of operation and coordination. The shipment decision is completely local, and the decision made for one shipment does not influence others. The transportation time from supplier to buyer location is short because each shipment goes direct.

A direct shipment network to single destination is justified only if demand at buyer locations is large enough that optimal replenishment lot sizes are close to a truckload from each supplier to each location. Home Depot started with a direct shipment network, given that most of the stores it opened until about 2002 were large stores. The stores ordered in quantities that were large enough that

ordering was managed locally within the store and delivery to the store arrived directly from the supplier. The direct shipment network to single destination, however, proved to be problematic as Home Depot started to open smaller stores that did not have large enough orders to justify a direct shipment.





Direct Shipping with Milk Runs

A milk run is a route on which a truck either delivers product from a single supplier to multiple retailers or goes from multiple suppliers to a single buyer location, as shown in Figure. In direct shipping with milk runs, a supplier delivers directly to multiple buyer locations on a truck or a truck picks up deliveries destined for the same buyer location from many suppliers. When using this option, a supply chain manager has to decide on the routing of each milk run.

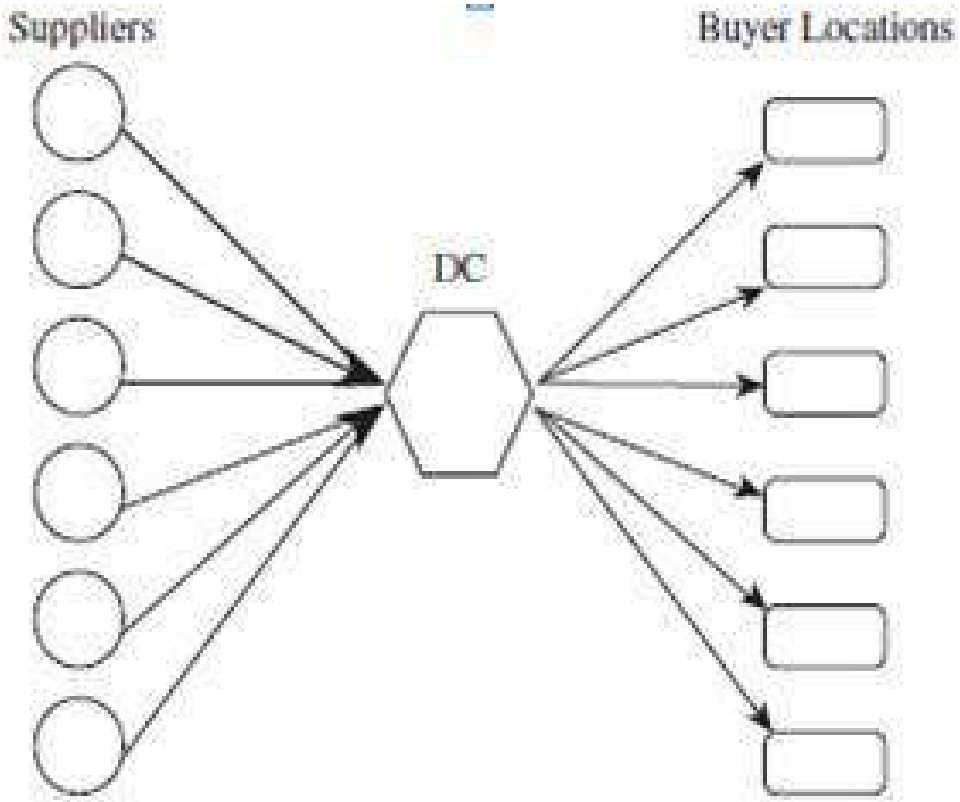
Direct shipping provides the benefit of eliminating intermediate warehouses, whereas milk runs lower transportation cost by consolidating shipments to multiple locations on a single truck. Milk runs make sense when the quantity destined for each location is too small to fill a truck but multiple locations are close enough to each other such that their combined quantity fills the truck.

Companies such as Frito-Lay that make direct store deliveries use milk runs to lower their transportation cost. If frequent small deliveries are needed on a regular basis and either a set of suppliers or a set of retailers is in geographic proximity, the use of milk runs can significantly reduce transportation costs. For example, Toyota uses milk runs from suppliers to support its just-in-time (JIT) manufacturing system in both Japan and the United States. In Japan, Toyota has many assembly plants located close together and thus uses milk runs from a single supplier to many plants. In the United States, however, Toyota uses milk runs from many suppliers to each assembly plant given the large distance between assembly plants.

All Shipments via Intermediate Distribution Center with Storage

Under this option, product is shipped from suppliers to a central distribution center where it is stored until needed by buyers when it is shipped to each buyer location, as shown in Figure Storing product at an intermediate location is justified if transportation economies require large shipments on the inbound side or shipments on the outbound side cannot be coordinated. In such a situation, product comes into a DC in large quantities where it is held in inventory and sent to buyer locations in smaller replenishment lots when needed.

The presence of a DC allows a supply chain to achieve economies of scale for inbound transportation to a point close to the final destination, because each supplier sends a large shipment to the DC that contains product for all locations the DC serves. Because DCs serve locations nearby, the outbound transportation cost is not very large. For example, W.W. Grainger has its suppliers ship products to one of nine DCs (typically in large quantities), with each DC in turn replenishing stores in its vicinity with the smaller quantities they need.



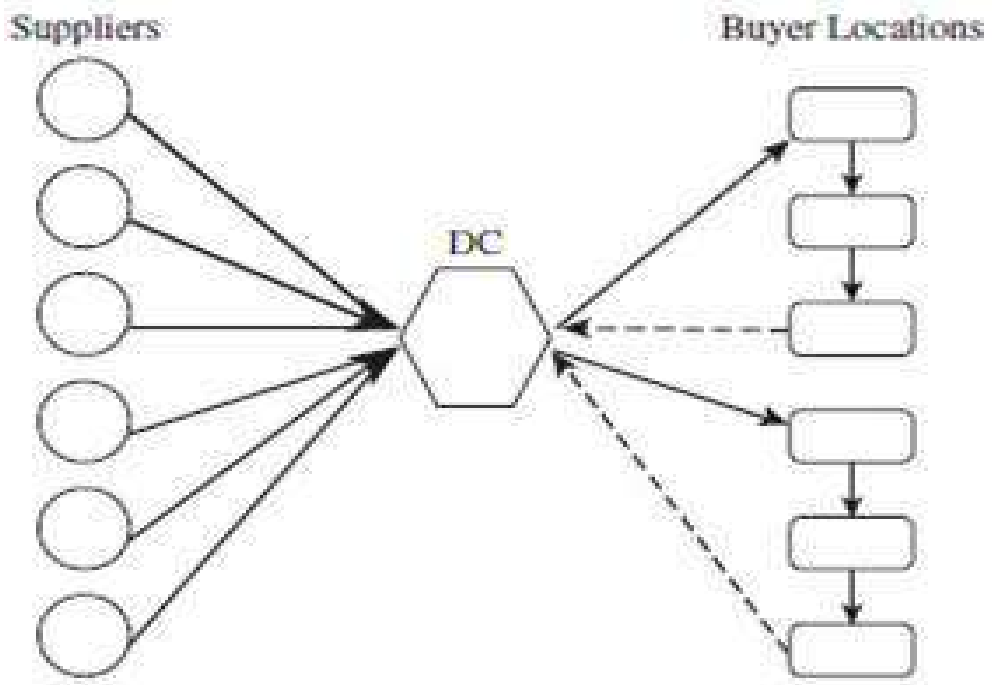
It would be expensive for suppliers to try and serve each store directly. Similarly, when Home Depot sources from an overseas supplier, the product is held in inventory at the DC because the lot size on the inbound side is much larger than the sum of the lot sizes for the stores served by the DC.

All Shipments via Intermediate Transit Point with Cross-Docking

Under this option, suppliers send their shipments to an intermediate transit point (could be a DC) where they are cross-docked and sent to buyer locations without storing them. The product flow is similar to that shown in Figure except that there is no storage at the intermediate facility. When a DC cross-docks product, each inbound truck contains product from suppliers for several buyer locations, whereas each outbound truck contains product for a buyer location from several suppliers. Major benefits of cross-docking are that little inventory needs to be held and product flows faster in the supply chain. Cross-docking also saves on handling cost because product does not have to be moved into and out of storage. Cross-docking is appropriate when economies of scale in transportation can be achieved on both the inbound and outbound sides and both inbound and outbound shipments can be coordinated.

Wal-Mart has used cross-docking successfully to decrease inventories in the supply chain without incurring excessive transportation costs. Wal-Mart builds many large stores in a geographic area supported by a DC. As a result, the total lot size to all stores from each supplier fills trucks on the inbound side to achieve economies of scale. On the outbound side, the sum of the lot sizes from all suppliers to each retail store fills up the truck to achieve economies of scale.

Another good example of the use of a transit point with cross-docking comes from Peapod in the Chicago area. Peapod has a DC in Lake Zurich from which it delivers to its customers using milk runs. This approach proved effective for customers in the northern and western sub-urbs of Chicago. Peapod, however, wanted to increase its reach to the city of Chicago and the city of Milwaukee. Both
are far enough from the Lake Zurich DC that a milk run wasted about two hours in transit making no productive deliveries. These markets were also small enough that they did not justify a local DC. Peapod's response has been to set up a cross- docking facility (which tends to be cheaper than a DC because no storage is involved) at each location. Peapod then sends out all deliveries to the local cross- dock facility in a larger truck and uses smaller trucks for local deliveries. The use of cross-docking at a transit point has allowed Peapod to increase the reach of the Lake Zurich DC without significantly increasing transportation expense.



Shipping via DC Using Milk Runs

As shown in Figure milk runs can be used from a DC if lot sizes to be delivered to each buyer location are small. Milk runs reduce outbound transportation costs by consolidating small shipments. For example, Seven-Eleven Japan cross-docks deliveries from its fresh-food suppliers at its DCs and sends out milk runs to the retail outlets because the total shipment to a store from all suppliers does not fill a truck. The use of cross-docking and milk runs allows Seven-Eleven Japan to lower its transportation cost while sending small replenishment lots to each store. The use of cross-docking with milk runs requires a significant degree of coordination and suitable routing and scheduling.

The online grocer Peapod uses milk runs from DCs when making customer deliveries to help reduce transportation costs for small shipments to be delivered to homes. OshKosh B'Gosh, a manufacturer of children's wear, has used this idea to virtually eliminate LTL shipments from its DC in Tennessee to retail stores.

Tailored Network

The tailored network option is a suitable combination of previous options that reduces the cost and improves responsiveness of the supply chain. Here transportation uses a combination of cross-docking, milk runs, and TL and LTL carriers, along with package carriers in some cases. The goal is to use the appropriate option in each situation. High-demand products to high- demand retail outlets may be shipped directly, whereas low-demand products or shipments to low-demand retail outlets are consolidated to and from the DC. The complexity of managing this transportation network is high because different shipping procedures are used for each product and retail outlet. Operating a tailored network requires significant investment in information infrastructure to facilitate the coordination.

Such a network, however, allows for the selective use of a shipment method to minimize the transportation as well as inventory costs.

Selecting a Transportation Network

A retail chain has eight stores in a region supplied from four supply sources. Trucks have a capacity of 40,000 units and cost $1,000 per load plus $100 per delivery. Thus, a truck making two deliveries charges $1,200. The cost of holding one unit in inventory at retail for a year is $0.20.

The vice president of supply chain is considering whether to use direct shipping from suppliers to retail stores or setting up milk runs from suppliers to retail stores. What network do you recommend if annual sales for each product at each retail store are 960,000 units? What network do you recommend if sales for each product at each retail store are 120,000 units?

Analysis:

We provide a detailed analysis when annual sales of each product at each retail store are 960,000 units. Our analysis assumes that all trucks travel full. A more sophisticated analysis can be performed for which the optimal load on each truck is calculated and used in the analysis.

We first analyze the direct shipping network and assume that full truckloads will be shipped from suppliers to retail stores. In this case, we have the following:

Batch size shipped from each supplier to each store = 40,000 units

Number of shipments / year from each supplier to each store = 960,000 / 40,000
= 24

Annual trucking cost for direct network = 24 × 1,100 x 4 × 8 = $844,800

Average inventory at each store for each product = 40,000 / 2 = 20,000 units
Annual inventory cost for direct network = 20,000 × 0.2 × 4 x 8 = $128,000

Total annual cost of direct network = $844,800 + $128,000 = $972,800

Now we analyze the network in which suppliers run milk runs to retail stores. Milk runs increase the transportation cost but decrease the level of inventory each store has to hold. We provide a detailed analysis for the instance of suppliers running milk runs to two stores on each truck. In this case, we have the following:

Batch size shipped from each supplier to each store = 40,000 / 2 = 20,000 units

Number of shipments / year from each supplier to each store = 960,000 / 20,000
= 48

Transportation cost per shipment per store (two stores / truck) = 1,000 / 2 + 100
= $600

Annual trucking cost for milk run network = 48 × 600 x 4 × 8 =

$921,600 Average inventory at each store for each product = 20,000 /

2 = 10,000 units Annual inventory cost for direct network = 10,000 ×

0.2 x 4 × 8 = $64,000 Total annual cost of direct network = $921,600 +

$64,000 = $985,600

This analysis shows that when demand per product per store is 960,000 units, the direct network is cheaper than running milk runs with two stores per route. Increasing the number of stores on a milk run ends up costing even more because it raises transportation costs more than it saves in holding costs.

When demand per product per store is 120,000, we first provide the detailed costs for the direct shipping network as follows (assuming all trucks travel full):

Batch size shipped from each supplier to each store = 40,000 units

Number of shipments / year from each supplier to each store = 120,000 / 40,000
= 3

Annual trucking cost for direct network = 3 × 1,100 x 4 x 8 = $105,600

Average inventory at each store for each product = 40,000 / 2 =

20,000 units Annual inventory cost for direct network = 20,000 × 0.2 x

4 × 8 = $128,000 Total annual cost of direct network = $105,600 +

$128,000 = $233,600

For the direct network, it turns out that it is better not to fill each truck but to send only 36,332 units per truck to minimize total annual costs. The optimal loading increases transportation costs a bit but decreases total costs to
$232,524 per year.

Now we analyze the network in which suppliers use milk runs to retail stores. We provide a detailed analysis for the instance of suppliers running milk runs to four stores on each truck and each truck travels full. In this case, we have the following:

Batch size shipped from each supplier to each store = 40,000 / 4 = 10,000 units

Number of shipments / year from each supplier to each store = 120,000 / 10,000

= 12

Transportation cost per shipment per store (four stores / truck) = 1,000 / 4 + 100

= $350

Annual trucking cost for milk run network = 12 × 350 x 4 x 8 =

$134,400 Average inventory at each store for each product = 10,000

/ 2 = 5,000 units Annual inventory cost for direct network = 5,000 ×

0.2 x 4 × 8 = $32,000 Total annual cost of direct network = $134,400

+ $32,000 = $166,400

This analysis shows that when demand per product per store is 120,000 units, the milk run network with four stores per route is cheaper than the direct network (even when truck loads are optimized). The direct network ends up costing more because of increased inventory holding costs even though transportation is cheaper. Observe that milk runs become more attractive as the amount flowing through the system decreases. In the next section, we discuss a variety of trade-offs that supply chain managers need to consider when designing and operating a transportation network.

3.4 TAILORED TRANSPORTATION

Tailored transportation is the use of different transportation networks and modes based on customer and product characteristics. Most firms sell a variety of products and serve many different customer segments. Products vary in size and value, and customers vary in the quantity purchased, responsiveness required, uncertainty of the orders, and distance. Given these differences, a firm should not design a common transportation network to meet all needs.

A firm can meet customer needs at a lower cost by using tailored transportation to provide the appropriate transportation choice based on customer and product characteristics.

The various forms of tailored transportation in supply chains are

- Tailored Transportation by Customer Density and Distance
- Tailored Transportation by Size of Customer
- Tailored Transportation by Product Demand and Value

Tailored Transportation by Customer Density and Distance

Firms must consider customer density and distance from warehouse when designing transportation networks. The ideal transportation options based on density and distance are shown in Table 3.4.1

Table 3.4.1 Transportation Options Based on Customer Density and Distance

	Short Distance	**Medium Distance**	**Long Distance**
High density	Private fleet with milk runs	Cross-dock with milk runs	Cross-dock with milk runs
Medium density	Third-party milk runs	LTL carrier	LTL or package carrier
Low density	Third-party milk runs	LTL or package carrier	Package carrier

	or LTL carrier		

When a firm serves a high density of customers close to the DC, it is often best for the firm to own a fleet of trucks that are used with milk runs originating at the DC to supply customers, it allows good use of the vehicles and provides customer contact.

If customer density is high but distance from the warehouse is large, it is better to use a public carrier with large trucks to haul the shipments to a cross-dock center close to the customer area, where the shipments are loaded onto smaller trucks that deliver product to customers using milk runs. In this situation, it may not be ideal for a firm to own its own fleet to avoid empty trucks on return trip.

As customer density decreases, use of an LTL carrier or a third party doing milk runs is more economical because the third-party carrier can aggregate shipments across many firms.

If a firm wants to serve an area with a low density of customers far from the warehouse, even LTL carriers may not be feasible and the use of package carriers may be the best option as long as loads are small.

Firms should serve areas with high customer density more frequently because these areas are likely to provide sufficient economies of scale in transportation, making temporal aggregation less valuable. To lower transportation costs, firms should use a higher degree of temporal aggregation and aim for somewhat lower responsiveness when serving areas with a low customer density.

Tailored Transportation by Size of Customer

Firms must consider customer size and location when designing transportation networks. Large customers can be supplied using a TL carrier, whereas smaller customers will require an LTL carrier or milk runs.

When using milk runs, a shipper incurs two types of costs:

- Transportation cost based on total route distance
- Delivery cost based on number of deliveries

The transportation cost is the same whether going to a large or small customer. If a delivery is to be made to a large customer, including other small customers on the same truck can save on transportation cost. For each small customer, however, the delivery cost per unit is higher than for large customers. Thus, it is not optimal to deliver to small and large customers with the same
frequency at the same price. So firms either have to charge a higher delivery cost for smaller customers or to tailor milk runs so that they visit larger customers with a higher frequency than smaller customers.

Firms can partition customers into large (L), medium (M), and small (S) based on the demand at each. The optimal frequency of visits can be evaluated based on the transportation and delivery costs.

If large customers are to be visited every milk run, medium customers every other milk run, and low-demand customers every third milk run, suitable milk runs can be designed by combining large, medium, and small customers on each run.

Medium customers would be partitioned into two subsets (M1, M2), and small customers would be partitioned into three subsets (S1, S2, S3).

The firm can sequence the following six milk runs to ensure that each customer is visited with the appropriate frequency:

(L, M1, S1), (L, M2, S2), (L, M1, S3), (L, M2, S1), (L, M1, S2), (L, M2, S3).

Advantages:

- each truck carries about the same load
- larger customers are provided more frequent delivery than smaller customers
- consistent with their relative costs of delivery.

Tailored Transportation by Product Demand and Value

The degree of inventory aggregation and the modes of transportation used in a supply chain network should vary with the demand and value of a product, as shown in the following Table 3.4.2.

Table 3.4.2 Transportation by Product Demand and Value

Product Type	High Value	Low Value
High demand	Disaggregate cycle inventory. Aggregate safety inventory. Inexpensive mode of transportation for replenishing cycle inventory and fast mode when using safety inventory.	Disaggregate all inventories and use inexpensive mode of transportation for replenishment.
Low Demand	Aggregate all inventories. If needed, use fast mode of transportation for filling customer orders	Aggregate only safety inventory. Use inexpensive mode of transportation for replenishing cycle inventory

For high-value products with high demand the inventory is disaggregated to save on transportation costs because this allows replenishment orders to be transported less expensively. For high-demand products with low value, all inventories should be disaggregated and held close to the customer to reduce transportation costs. For low-demand, high-value products, all inventories should be aggregated to save on inventory costs.

For low-demand, low-value products, cycle inventories can be held close to the customer and safety inventories aggregated to reduce transportation costs.

3.5 ROUTING AND SCHEDULING IN TRANSPORTATION

The most important operational decision related to transportation in a supply chain is the routing and scheduling of deliveries. Selecting the best paths for the transport mode to follow to minimize travel time or distance reduces transportation costs and improves customer service.

Managers must decide on the customers to be visited by a particular vehicle and the sequence in which they will be visited. For example, an online grocer such as Peapod is built on delivering customer orders to their homes. The success of its operations hinges on its ability to decrease transportation and delivery costs while providing the promised level of responsiveness to the customer.

Given a set of customer orders, the goal is to route and schedule delivery vehicles such that the costs incurred to meet delivery promises are as low as possible.

Typical objectives when routing and scheduling vehicles include a combination of minimizing cost by

- decreasing the number of vehicles needed
- the total distance travelled by vehicles
- the total travel time of vehicles
- eliminating service failures such as a delays in shipments

In this note, routing and scheduling problems are discussed from the point of view of the manager of a Peapod distribution center (DC). After customers place orders for groceries online, staff at the DC has to pick the items needed and load them on trucks for delivery. The manager must decide which trucks will deliver to which customers and the route that each truck will take when making deliveries. The manager must also ensure that no truck is overloaded and that promised delivery times are met.

For example, the DC manager at Peapod has delivery orders from thirteen different customers. The DC's location, each customer on the grid, and the order size from each customer are shown in Table 3.5.1. The manager has four trucks, each capable of carrying up to 200 units. The manager believes that the delivery costs are strongly linked to the total distance the trucks travel, and that the distance between two points on the grid is correlated with the actual distance a vehicle will travel between those two points. The manager thus decides to assign customers to trucks and identify a route for each truck, with a goal of minimizing the total distance traveled.

Table 3.5.1: Customer Location and Demand for Peapod

	X-Coordinate	Y-Coordinate	Order Size ai
Warehouse	0	0	
Customer 1	0	12	48
Customer 2	6	5	36
Customer 3	7	15	43
Customer 4	9	12	92
Customer 5	15	3	57
Customer 6	20	0	16
Customer 7	17	-2	56
Customer 8	7	-4	30
Customer 9	1	-6	57
Customer 10	15	-6	47
Customer 11	20	-7	91
Customer 12	7	-9	55
Customer 13	2	-15	38

The DC manager must first assign customers to be served by each vehicle and then decide on each vehicle's route. After the initial assignment, route sequencing and route improvement procedures are used to decide on the route for each vehicle.

The DC manager decides to use the following computational procedures to support his decision:

- The Savings Matrix method
- The Generalized Assignment method

SAVINGS MATRIX METHOD

This method is simple to implement and can be used to assign customers to vehicles even when delivery time windows or other constraints exist. The major steps in the savings matrix method are:

1. Identify the distance matrix.
2. Identify the savings matrix.
3. Assign customers to vehicles or routes.
4. Sequence customers within routes.

The first three steps result in customers being assigned to vehicles, and the fourth step is used to route each vehicle to minimize the distance traveled.

Identify the Distance Matrix

The distance matrix identifies the distance between every pair of locations to be visited. The distance is used as a surrogate for the cost of traveling between the pair of locations. If the transportation costs between every pair of locations are known, the costs can be used in place of distances. The distance Dist(A, B) on a grid between a point A with coordinates (x_A, y_A) and a point B with coordinates (x_B, y_B) is evaluated as:

$$\text{Dist(A,B)} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2} \quad (1)$$

The distance between every pair of locations for Peapod is shown in Table 3.5.2. The distances between every pair of locations are next used to evaluate the savings matrix.

Table 3.5.2: Distance Matrix for Peapod Deliveries

	DC	Cust 1	Cust 2	Cust 3	Cust 4	Cust 5	Cust 6	Cust 7	Cust 8	Cust 9	Cust 10	Cust 11	Cust 12	Cust 13
DC	0													
Cust 1	12	0												
Cust 2	8	9	0											
Cust 3	17	8	10	0										
Cust 4	15	9	8	4	0									
Cust 5	15	17	9	14	11	0								
Cust 6	20	23	15	20	16	6	0							
Cust 7	17	22	13	20	16	5	4	0						
Cust 8	8	17	9	19	16	11	14	10	0					
Cust 9	6	18	12	22	20	17	20	16	6	0				
Cust 10	16	23	14	22	19	9	8	4	8	14	0			
Cust 11	21	28	18	26	22	11	7	6	13	19	5	0		
Cust 12	11	22	14	24	21	14	16	12	5	7	9	13	0	
Cust 13	15	27	20	30	28	22	23	20	12	9	16	20	8	0

Identify the Savings Matrix

The savings matrix represents the savings that accrue on consolidating two customers on a single truck. Savings may be evaluated in terms of distance, time, or money. The manager at the Peapod DC constructs the savings matrix in terms of distance. A trip is identified as the sequence of locations a vehicle visits. The trip DC
-> Cust x -> DC starts at the DC, visits customer x, and returns to the DC. The savings S(x,y) is the distance saved if the trips DC -> Cust x -> DC and DC -> Cust y -> DC are combined to a single trip, DC -> Cust x -> Cust y -> DC. These savings can be calculated by the following formula:

$$S(x,y) = Dist(DC, x) + Dist(DC, y) – Dist(x, y) \qquad (2)$$

For example, using Table 3.5.2 the manager evaluates S(1,2) = 12 + 8 – 9 = 11.

The savings matrix for the Peapod deliveries is shown in Table 3.5.3. The savings matrix is then used to assign customers to vehicles or routes.

Table 3.5.3: Savings Matrix for Peapod Deliveries

	Cust 1	Cust 2	Cust 3	Cust 4	Cust 5	Cust 6	Cust 7	Cust 8	Cust 9	Cust 10	Cust 11	Cust 12	Cust 13
Cust 1	0												
Cust 2	11	0											
Cust 3	21	15	0										
Cust 4	18	15	28	0									
Cust 5	10	14	18	19	0								
Cust 6	9	13	17	19	29	0							
Cust 7	7	12	14	16	27	33	0						
Cust 8	3	7	6	7	12	14	15	0					
Cust 9	0	2	1	1	4	6	7	8	0				
Cust 10	5	10	11	12	22	28	29	16	8	0			
Cust 11	5	11	12	14	25	34	32	16	8	32	0		
Cust 12	1	5	4	5	12	15	16	14	10	18	19	0	
Cust 13	0	3	2	2	8	12	12	11	12	15	16	18	0

Assign Customers to Vehicles or Routes

When assigning customers to vehicles, the manager attempts to maximize savings. An iterative procedure is used to make this assignment. Initially each customer is assigned to a separate route. Two routes can be combined into a feasible route if the total deliveries across both routes do not exceed the vehicle's capacity. At each iterative step, the Peapod manager attempts to combine routes with the highest savings into a new feasible route. The procedure is continued until no more combinations are feasible. At the first step, the highest savings of 34 results on combining truck Routes 6 and 11. The combined route is feasible because the total load is 16 + 91 = 107, which is less than 200. The two customers are thus combined on a single route, as shown in Figure 3.5.1, and the saving of 34 is eliminated from further consideration.

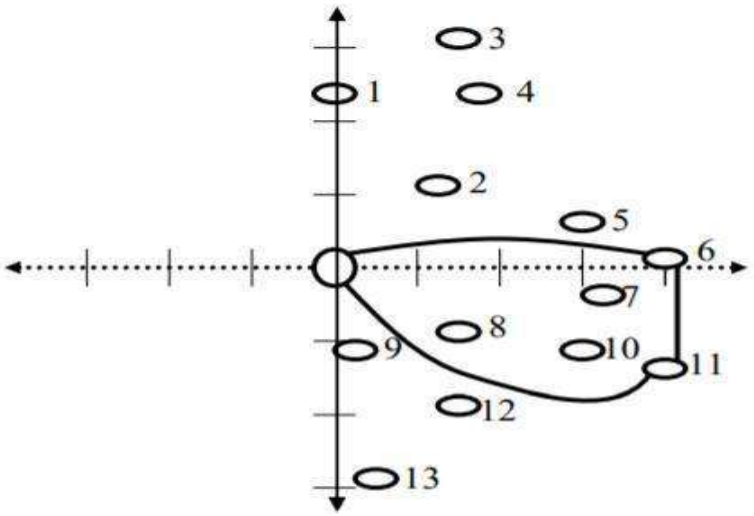
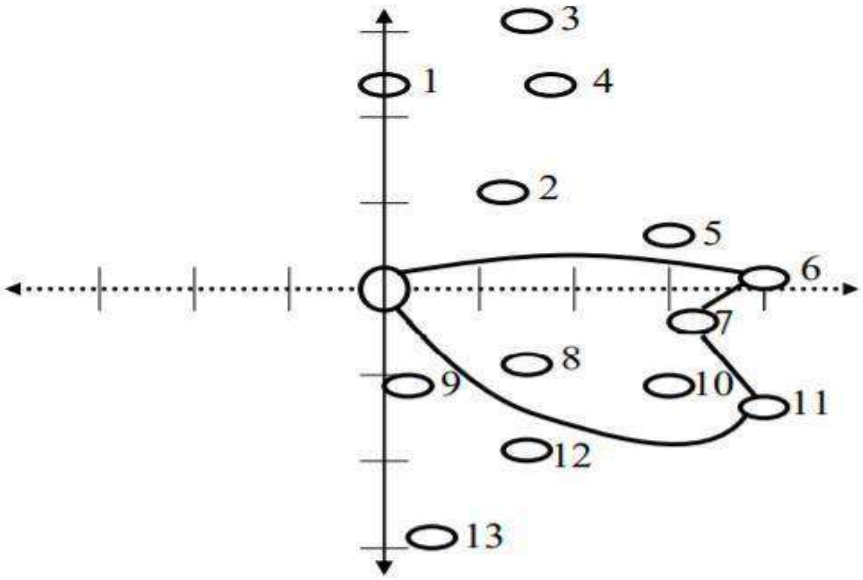


Figure 3.5.1: Delivery Route by Assigning 6 and 11 to a Common Route

The next highest saving is 33, which results from adding Customer 7 to the route for Customer 6. This is feasible because the resulting load is 107 + 56 = 163, which is less than 200. Thus, Customer 7 is also added to Route 6,



as shown in Figure 3.5.2.

Figure 3.5.2: Delivery Route by Assigning 6, 7, and 11 to a Common Route

The next highest saving now is 32, which results from adding Customer 10 to Route 6 (we need not consider the saving of 32 on combining Customer 7 with Customer 11 because both are already in Route 6). This, however, is not feasible, as Customer 10 has a delivery totaling 47 units and adding this amount to the deliveries already on Route 6 would exceed the vehicle capacity of 200. The next highest saving is 29, which results from adding either Customer 5 or 10 to Route 6. Each of these is also infeasible because of the capacity constraint. The next highest saving is 28, which results from combining Routes 3 and 4, which is feasible. The two routes are combined into a single route, as shown in Figure 3.5.3.

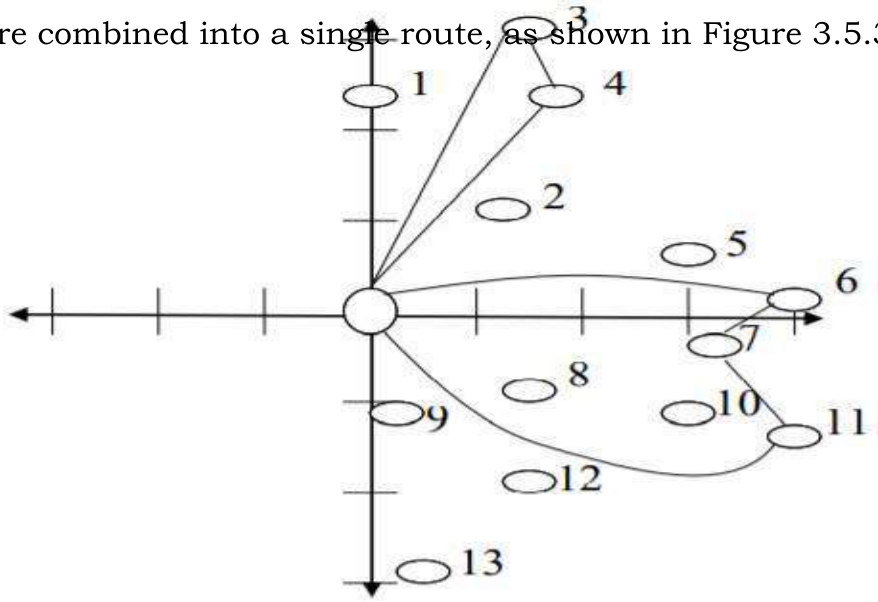


Figure 3.5.3: Delivery Route by Assigning 3 and 4 to a Common Route

Continuing the iterative procedure, the manager partitions customers into four groups {1, 3, 4}, {2, 9}, {6, 7, 8, 11}, {5, 10, 12, 13}, with each group assigned to a single vehicle. The next step is to identify the sequence in which each vehicle will visit customers.

Sequence Customers within Routes

At this stage the manager's goal is to sequence customer visits so as to minimize the distance each vehicle must travel. Changing the sequence in which deliveries are made can have a significant impact on the distance traveled by the vehicles. Consider the truck that has been assigned deliveries to Customers 5, 10, 12, and 13. If the deliveries are in the sequence 5, 10, 12, 13, the total distance traveled by the truck is 15 + 9 + 9 + 8 + 15 = 56 (distances are obtained from Table 3.5.2). In contrast, if deliveries are in the sequence 12, 5, 13, 10, the truck
covers a larger distance of 11 + 14 + 22 + 16 + 16 = 79. Delivery sequences are determined by obtaining an initial route sequence and then using route improvement procedures to obtain delivery sequences with a lower transportation distance or cost.

ROUTE SEQUENCING PROCEDURES

The Peapod manager can use route sequencing procedures to obtain an initial trip for each vehicle. The initial trip is then improved using the route improvement procedure discussed later in this note. All route sequencing procedures are illustrated for the vehicle assigned to Customers 5, 10, 12, and 13.

Farthest Insert

Given a vehicle trip (including a trip consisting of only the DC) for each remaining customer, find the minimum increase in length for this customer to be inserted from all the potential points in the trip that they could be inserted. Then choose to actually insert the customer with the largest minimum increase to obtain a new trip. This step is referred to as a farthest insert because the customer farthest from the current trip is inserted. The process is continued until all remaining customers to be visited by the vehicle are included in a trip.

For the Peapod example, the manager is seeking a trip starting at the DC and visiting Customers 5, 10, 12, 13. The initial trip consists of just the DC with a length of 0. Including Customer 5 in the trip adds 30 to its length, including Customer 10 adds 32, including Customer 12 adds 22, and including Customer 13 adds 30 (see Table 3.5.2). Using the farthest insert, the manager adds Customer 10 to obtain a new trip (DC, 10, DC) of length 32.

At the next step, inserting Customer 5 in the trip raises the length of the trip to a minimum of 40, inserting Customer 12 raises it to 36, and inserting Customer 13 raises it to 47. The manager thus inserts the farthest Customer 13 to obtain the new trip (DC, 10, 13, DC) of length 47. This still leaves Customers 5 and 12 to be inserted. The minimum cost insertion for Customer 5 is (DC, 5, 10, 13, DC) for a length of 55, and the minimum cost insertion for Customer 12 is (DC, 10, 12, 13, DC) for a length of 48. The manager thus inserts Customer 5 to obtain a trip (DC, 5, 10, 13, DC) of length 55. Customer 12 is then inserted between Customers
10 and 13 to obtain a trip (DC, 5, 10, 12 13, DC) of length 56.

Nearest Insert

Given a vehicle trip (including a trip consisting of only the DC) for each remaining customer, find the minimum increase in length for this customer to be inserted from all the potential points in the trip that they could be inserted. Insert the customer with the smallest minimum increase to obtain a new trip. This step is referred to as a nearest insert because the customer closest to the current trip is inserted. The process is continued until all remaining customers the vehicle will visit are included in a trip.

For the Peapod example, the manager applies the nearest insert to the vehicle serving Customers 5, 10, 12, and 13. Starting at the DC, the nearest customer is 12.

Inserting Customer 12 results in the trip (DC, 12, DC) of length 22. At the next step, inserting Customer 5 results in a trip of length 40, inserting Customer 10 results in a trip of length 36, and inserting Customer 13 results in a trip of length 34. Customer 13 results in the smallest increase and is inserted to obtain a trip (DC, 12, 13, DC) of length 34. The next nearest insertion is Customer 10 resulting in a trip (DC, 10, 12, 13, DC) of length 48, and the final insertion of Customer 5 results in a trip (DC, 5, 10, 12, 13, DC) of length 56.

Nearest Neighbor

Starting at the DC, this procedure adds the closest customer to extend the trip. At each step, the trip is built by adding the customer closest to the point last visited by the vehicle until all customers have been visited.

For the Peapod example, the customer closest to the DC is 12 (see Table 3.5.2). This results in the path (DC, 12). The customer closest to Customer 12 is 13, extending the path to (DC, 12, 13). The nearest neighbor of Customer 13 is 10 and the nearest neighbor of Customer 10 is 5. The Peapod manager thus obtains a trip (DC, 12, 13, 10, 5, DC) of length 59.

Sweep

In the sweep procedure, any point on the grid is selected (generally the DC itself) and a line is swept either clockwise or counterclockwise from that point. The trip is constructed by sequencing customers in the order they are encountered during the sweep.

The Peapod manager uses the sweep procedure with the line centered at the DC. Customers are encountered in the sequence 5, 10, 12, 13 to obtain the trip (DC, 5,
10, 12 13, DC) for a length of 56.

The initial trips resulting from each route sequencing procedure and their lengths are summarized in Table 3.5.4.

Table 3.5.4: Initial Trips Using Different Route Sequencing Procedures at Peapod

Route Sequencing Procedure	Resulting Trip	Trip Length
Farthest insert	DC, 5, 10, 12, 13, DC	56
Nearest insert	DC, 5, 10, 12, 13, DC	56
Nearest neighbor	DC, 12, 10, 5, 13, DC	59
Sweep	DC, 5, 10, 12, 13, DC	56

ROUTE IMPROVEMENT PROCEDURES

Route improvement procedures start with a trip obtained using a route sequencing procedure and improve the trip to shorten its length. The Peapod manager next applies route improvement procedures to alter the sequence of customers visited by a vehicle and to shorten the distance a vehicle must travel. The two route improvement procedures discussed are illustrated on the trip obtained as a result of the nearest neighbor procedure.

2-OPT

The 2-OPT procedure starts with a trip and breaks it at two places. This results in the trip breaking into two paths, which can be reconnected in two possible ways. The length for each reconnection is evaluated, and the smaller of the two is used to define a new trip. The procedure is continued on the new trip until no further improvement results.

For example, the trip (DC, 12, 10, 5, 13, DC) resulting from the nearest neighbor procedure can be broken into two paths (13, DC) and (12, 10, 5) and reconnected into the trip (DC, 5, 10, 12, 13, DC), as shown in Figure 3.5.4. The new trip has length 56, which is an improvement over the existing trip.

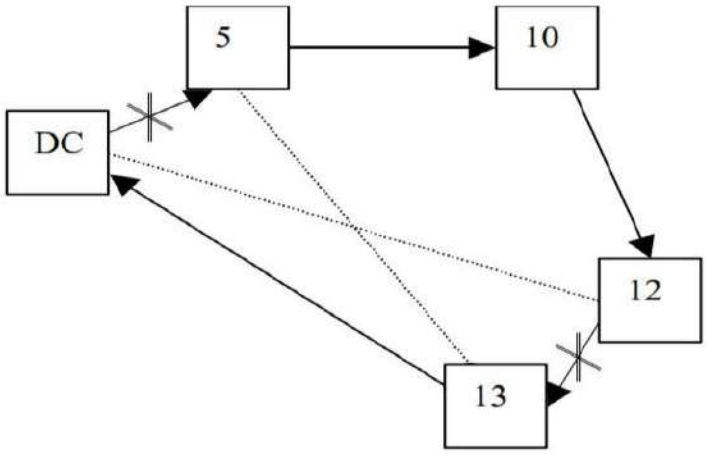


Figure 3.5.4: Improving Route Sequencing Using 2-OPT

3-OPT

The 3-OPT procedure breaks a trip at three points to obtain three paths that can be reconnected to form up to eight different trips. The length of each of the eight possible trips is evaluated and the shortest trip is retained. The procedure is continued on the new trip until no further improvement results.

The trip (DC, 5, 10, 12, 13, DC) resulting from the 2-OPT procedure is broken up into three paths (DC), (5, 10), and (12, 13). The various resulting trips on reconnecting the three paths are (DC, 12, 13, 5, 10, DC) of length 65, (DC, 12, 13,
10, 5, DC) of length 81, and (DC, 13, 12, 5, 10, DC) of length 61. All other trips correspond to one of these four trips reversed. This application of the 3- OPT procedure does not improve the trip because the current trip is the shortest. At this stage the Peapod manager can form three new paths from the trip and repeat the procedure.

The Peapod manager uses route sequencing and improvement procedures to obtain delivery trips for each of the four trucks, as shown in Table 3.5.5 and Figure 3.5.5. The total travel distance for the delivery schedule is 185.

Table 3.5.5: Peapod Delivery Schedule Using Saving Matrix Method

Truck	Trip	Length of Trip	Load on Truck
1	**DC, 2, 9, DC**	**32**	**93**
2	DC, 1, 3, 4, DC	39	183
3	DC, 8, 11, 6, 7, DC	58	193
4	DC, 5, 10, 12, 13, DC	56	197

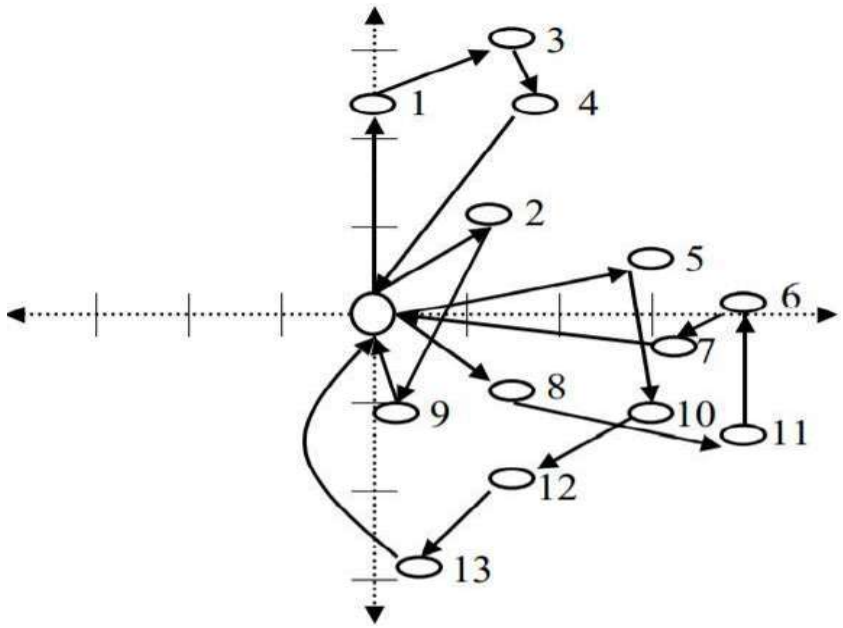


Figure 3.5.5: Delivery Routes at Peapod Using Savings Matrix Method

GENERALIZED ASSIGNMENT METHOD

The generalized assignment method is more sophisticated than the savings matrix method and usually results in better solutions when there are few delivery constraints to be satisfied. The procedure for routing and sequencing of vehicles consists of the following steps:

1. Assign seed points for each route.
2. Evaluate insertion cost for each customer.
3. Assign customers to routes.
4. Sequence customers within routes.

The first three steps result in customers being assigned to a vehicle, and the fourth step identifies a route for each vehicle to minimize the distance traveled. We discuss each step in greater detail in the context of the delivery decision at Peapod.

Assign Seed Points for Each Route

The goal of this step is to determine a seed point corresponding to the center of the trip taken by each vehicle using the following procedure:

1. Divide the total load to be shipped to all customers by the number of trucks to obtain L_{seed}, the average load allocated to each seed point.

2. Starting at any customer, use a ray starting at the DC to sweep clockwise to obtain cones assigned to each seed point. Each cone is assigned a load of L_{seed}.

3. Within each cone, the seed point is located in the middle (in terms of angle) at a distance equal to that of the customer (with a partial or complete load allocated to the cone) farthest from the DC.

The manager at Peapod uses the procedure described earlier to obtain seed points for the deliveries described in Table 3.5.1. Given four vehicles and a total delivery load across all customers of 666 units, the manager obtains an average load per vehicle of L_{seed} = 666 / 4 = 166.5 units.

The next step is to sweep clockwise with a ray emanating from the DC to obtain four cones, one for each vehicle, including all customers. The first step in defining the cones is to obtain the angular position of each customer. The angular position θ_i of customer i with coordinates (x_i, y_i) is the angle made relative to the x axis by the line joining the customer i to the origin (DC), as shown in Figure 3.5.6.

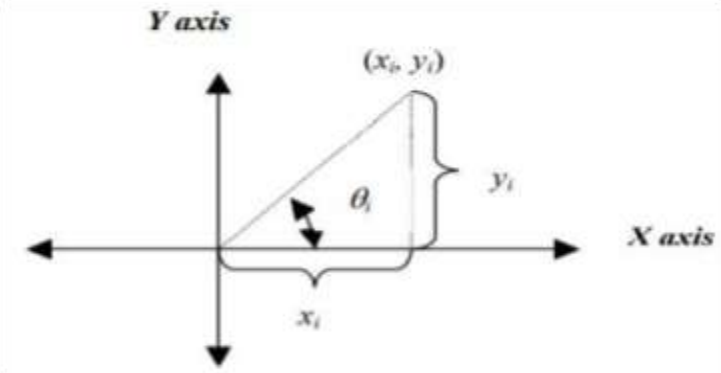


Figure 3.5.6: Angular Position of Customer *i*

The angular position of each customer is obtained as the inverse tangent of the ratio of its y coordinate to the x coordinate.

$$\theta_i = \tan^{-1}(y_i/x_i) \qquad (3)$$

The inverse tangent can be evaluated using the Excel function ATAN() as

$$\theta_i = \mathrm{ATAN}(y_i/x_i) \qquad (4)$$

The angular position of each customer is obtained using Equation 4, as shown in Table 3.5.6.

Table 3.5.6: Angular Positions of Peapod Customers

	X Coordinate	Y Coordinate	Angular Position (Radians)	Demand
DC	0	0		
Customer 1	0	12	1.57	48
Customer 2	6	5	0.69	36
Customer 3	7	15	1.13	43
Customer 4	9	12	0.93	92
Customer 5	15	3	0.20	57
Customer 6	20	0	0.00	16
Customer 7	17	-2	-0.12	56
Customer 8	7	-4	-0.52	30
Customer 9	1	-6	-1.41	57
Customer 10	15	-6	-0.38	47
Customer 11	20	-7	-0.34	91
Customer 12	7	-9	-0.91	55
Customer 13	2	-15	-1.44	38

The next step is to sweep clockwise and order the customers as encountered. For Peapod, a clockwise sweep encounters customers in the order 1, 3, 4, 2, 5, 6, 7, 11, 10, 8, 12, and 9. Starting with Customer 1, four cones, each representing a load of L_{seed} = 166.5 units, are formed. Customers 1 and 3 combine to load 91 units on the truck. Customer 4 is encountered next in the sweep. Adding the entire load for Customer 4 would result in a load of 183, which is larger than L_{seed} = 166.5. To get a load of 166.5, only 166.5 – 91 = 75.5 units of the load should be included. Thus, the first cone extends to a point that is 75.5 / 92 of the angle between Customers 3 and 4. Customer 3 has an angular position of 1.13 and Customer 4 has an angular position of 0.93, resulting in an angle between them of

1.13 – 0.93 = 0.20. The first cone thus extends to an angle (75.5 / 92) x 0.20 beyond Customer 3 with a resulting angle of 1.13 – (75.5 / 92) x 0.20 = 0.97. The first cone thus has one end at Customer 1 (angle of 1.57) and the other at an angle of 0.97, as shown in Figure 3.5.7.

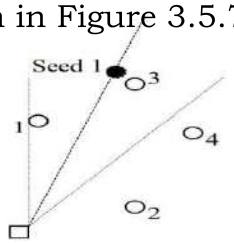


Figure 3.5.7: Sweep Method to Locate Seed 1

The seed point is then located at an angle $\alpha_1 = (0.97 + 1.57) / 2 = 1.27$ in the middle of the cone at a distance equal to that of the farthest customer included. Customer 3, at a distance $d_1 = \sqrt{(7-0)^2 + (15-0)^2} = 17$, is the farthest customer in the first cone. Given the distance d_1, the coordinates (X_1, Y_1) of the Seed Point 1 are thus given by:

$$X_1 = d_1 \cos(\alpha_1) = 17 \cos(0.95) = 10, \text{ and } Y_1 = d_1 \sin(\alpha_1) = 17 \sin(0.95) = 14$$

The second cone starts at the angle 0.97 and includes 92 – 75.5 = 16.5 units of the Customer 4 load. On sweeping clockwise, Customers 2, 5, 6, and 7 are encountered before a load of 166.5 is exceeded. To get a load of exactly 166.5, only 41/56 of Customer 7 load is needed. The angular position of the end of the cone is thus 41/56 between Customers 6 and 7. Customer 6 is at angle of 0.00 and Customer 7 is at the angle of –0.12. The second cone thus ends at an angle of 0.00 – 0.12 x (41 / 56) = –0.09. The second cone has one end at an angle of 0.33 and the other at an angle of –0.09. The seed point is thus located at an angle α_2 in the middle of the cone; that is, $\alpha_2 = (0.33 - 0.09) / 2 = 0.12$. The distance d_2 of the seed point for the second cone is the same as Customer 6, the farthest customer in the cone. This corresponds to a distance of $d_2 = 20$ (see Table 3.5.2). The coordinates (X_2, Y_2) of the Seed Point 2 are thus given by:

$$X_2 = d_2 \cos(\alpha_2) = 20 \cos(0.12) = 20, \text{ and } Y_2 = d_2 \sin(\alpha_2) = 20 \sin(0.12) = 2.$$

Proceeding in the same manner, the Peapod manager forms four cones to determine the four seed points, as shown in Table 3.5.8.

Table 3.5.8: Seed Point Coordinates for Peapod Deliveries

Seed Point	X Coordinate	Y Coordinate
S_1	5	16
S_2	20	2
S_3	19	-5
S_4	5	-5

Evaluate Insertion Cost for Each Customer

For each Seed Point S_k and Customer i, the insertion cost c_{ik} is the extra distance that would be traveled if the customer is inserted into a trip from the DC to the seed point and back and is given by:

$$c_{ik} = Dist(\text{DC}, i) + Dist(i, S_k) - Dist(\text{DC}, S_k),$$

where the Dist() function is evaluated as in Equation 1. For Customer 1 and Seed Point 1, the insertion cost is given by:

$$c_{11} = Dist(\text{DC}, 1) + Dist(1, S_1) - Dist(\text{DC}, S_1) = 12 + 10 - 17 = 5.$$

The Peapod manager evaluates all insertion costs c_{ik}, as shown in Table 3.5.9.

Table 3.5.9: Insertion Costs for Peapod Deliveries for Each Customer and Seed Point

Customer	Seed Point 1	Seed Point 2	Seed Point 3	Seed Point 4
1	2	14	18	23
2	2	2	5	11
3	2	15	21	30
4	4	10	15	25
5	15	0	4	21
6	25	2	5	29
7	22	2	1	22
8	11	2	0	3
9	12	7	4	3
10	24	5	0	19
11	32	10	4	29
12	20	8	4	8
13	30	20	15	18

Assign Customers to Routes

The manager next assigns customers to each of the four vehicles to minimize total insertion cost while respecting vehicle capacity constraints.

The assignment problem is formulated as an integer program and requires the following input:

c_{ik} = insertion cost of Customer i and Seed

Point k, a_i = order size from Customer i,

b_k = capacity of Vehicle k.

Define the following decision variables:

y_{ik} = 1 if Customer i is assigned to Vehicle k, 0 otherwise.

The integer program for assigning customers to vehicles is given by:

$$Min \sum_{k=1}^{K} \sum_{i=1}^{n} c_{ik} y_{ik}$$

subject to:

$$\sum_{k=1}^{K} y_{ik} = 1, i = 1,...,n,$$

$$\sum_{i=1}^{n} a_i y_{ik} \le b_k, k = 1,...,K,$$

$$y_{ik} = 0 \text{ or } 1, \text{ for all } i \text{ and } k.$$

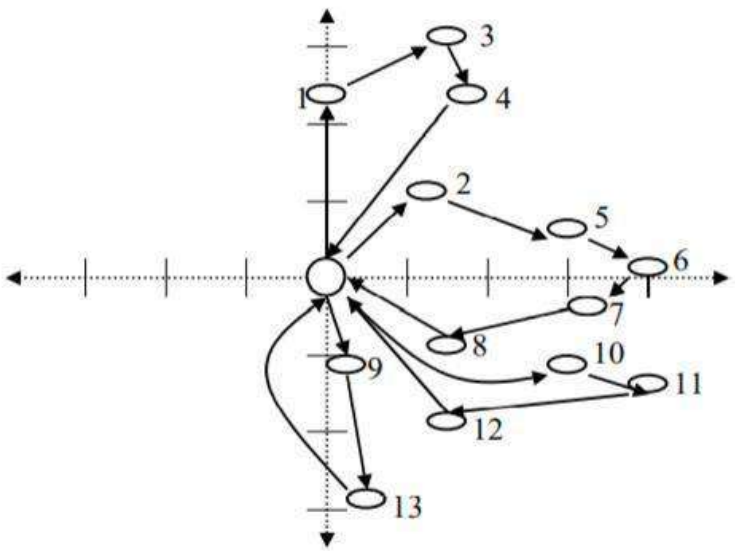
For Peapod, the order size for each customer is given in Table 3.5.1, the insertion cost c_{ik} is obtained from Table 3.5.9, and the capacity of each vehicle is 200 units. The manager at Peapod solves the integer program using the Solver tool in Excel to obtain the assignment of customers to vehicles as shown in Table 3.5.10 and Figure
3.5.8. The sequencing of customers within each trip is obtained using the route sequencing and route improvement procedures discussed earlier. The total distance traveled for the delivery schedule is 159.

Table 3.5.10: Peapod Delivery Schedule Using Generalized Assignment

Truck	Trip	Length of Trip	Load on Truck
1	DC, 1, 3, 4, DC	39	183
2	DC, 2, 5, 6, 7, 8, DC	45	195
3	DC, 10, 11, 12, DC	45	193
4	DC, 9, 13, DC	30	95

Method

Figure 3.5.8: Delivery Routes at Peapod Using Generalized Assignment Method



Applicability of Routing and Scheduling Methods

The delivery schedule for Peapod resulting from the generalized assignment method in Table 3.5.10 is superior to the solution obtained from the savings matrix method in Table 3.5.5. The generalized assignment method is more sophisticated and generally gives a better solution than the savings matrix method when the delivery schedule has no constraints other than vehicle capacity.

The main disadvantage of the generalized assignment method is that it has difficulty generating good delivery schedules as more constraints are included. For example, if Peapod has fixed time windows within which deliveries must be made to customers, it is difficult to use the generalized assignment method to generate a delivery schedule. The generalized assignment method is recommended if the constraints are limited to vehicle capacity or total travel time.

The main strength of the savings matrix method is its simplicity and robustness. The method is simple enough to be easily modified to include delivery time windows and other constraints and robust enough to give a reasonably good solution that can be implemented in practice.

Its main weakness is the quality of the solution. It is often possible to find better delivery schedules using more sophisticated methods. The savings matrix method is recommended in case there are many constraints that need to be satisfied by the delivery schedule. Software packages for transportation planning and routing and scheduling of deliveries are available from many supply chain software companies.

Part-A Questions and Answers

1. What is Logistics and Supply Chain Management? (K1, CO4)

Logistics typically refers to activities that occur within the boundaries of a single organization and Supply Chain refers to networks of companies that work together and coordinate their actions to deliver a product to market. Also, traditional logistics focuses its attention on activities such as procurement, distribution, maintenance, and inventory management. Supply Chain Management (SCM) acknowledges all of traditional logistics and also includes activities such as marketing, new product development, finance, and customer service.

2. What is Logistics? (K1, CO4)

Logistics is about getting the right product, to the right customer, in the right quantity, in the right condition, at the right place, at the right time, and at the right cost.

3. Differentiate Inbound & Outbound Logistics? (K1, CO4)

Inbound Logistics refers to movement of goods and raw materials from suppliers to your company. In contrast, Outbound Logistics refers to movement of finished goods from your company to customers.

4. What is Transport and Logistics? (K1, CO4)

Transport and Logistics refers to 2 types of activities, namely, traditional services such as air/sea/land transportation, warehousing, customs clearance and value- added services which including information technology and consulting.

5. What is International Logistics? (K1, CO4)

"International Logistics focuses on how to manage and control overseas activities effectively as a single business unit. Therefore, companies should try to harness the value of overseas product, services, marketing, R&D and turn them into competitive advantage“.

6. What is Third Party Logistics or 3PL? (K1, CO4)

The concept of 3PL appeared on the scene in the 1980s as the way to reduce costs and improve services which can be defined as below,

"Third Party Logistics or 3PL refers to the outsourcing of activities, ranging from a specific task, such as trucking or marine cargo transport to broader activities serving the whole supply chain such as inventory management, order processing and consulting."

7. What is Supply Chain Network? (K1, CO4)

Many companies have the department that controls supply chain activity so they believe that SCM is a "function". Some companies think SCM is a kind of management system under IT (information system or enterprise resource planning.) In fact, SCM is actually a "network" consists of many players. A generic supply chain structure is as simple as Supplier, Manufacturer, Wholesaler and Retailer (it's more complex in the real world but a simple illustration serves the purpose.)

8. What is Supply Chain Coordination? (K1, CO4)

Information sharing requires a certain degree of "coordination" (it's also referred to as collaboration or integration in scholarly articles). Do you wonder when people started working together as a network? In 1984, companies in the apparel business worked together to reduce overall lead-time. In 1995, companies in the automotive industry used Electronic Data Interchange to share information. So, working as a "chain" is the real-world practice.

9. What is the Cost/Service Trade-off? (K1, CO4)

The concept of Cost/Service Trade-off appeared as early as in 1985 but it seems that people really don't get it. When you want to improve service, the cost goes up. When you want to cut cost, service suffers. It's like a "seesaw", the best way you can do is to try to balance both sides.

10. What are Conflicting Objectives? (K1, CO4)

Working as a network requires the same objective, but this is often not the case (even with someone in the same company). "Conflicting Objectives" is the term used to describe the situation when each function wants something that won't go well together. For example, purchasing people always place the orders to the cheapest vendors (with a very long lead-time) but production people or project manager need material more quickly. To avoid conflicting objectives, you need to decide if you want to adopt a time-based strategy, low-cost strategy or differentiation strategy. A clear direction is needed so people can make the decisions accordingly.

11. What is Supply Chain Relationship? (K1, CO4)

To work as the same team, long-term relationship is key. Otherwise, you're just a separate company with a different strategy/agenda. So academia keeps preaching about the importance of relationship-building but is not for everyone.

12. Define Transportation. (K1, CO4)

Transportation involves the physical movement of goods between origin and destination points.

- The transportation system links geographically separated partners and facilities in a company's supply.
- Transportation facilitates the creation of time and place utility in the supply chain.
- Transportation also has a major economic impact on the financial performance of businesses.

13. What are the role of transportation in supply chain? (K1, CO4)

- Transportation provides the critical links between these organizations, permitting goods to flow between their facilities.
- Transportation service availability is critical to demand fulfillment in the supply chain.
- Transportation efficiency promotes the competitiveness of a supply chain.

14. What are the factors affecting transportation decisions? (K1, CO4)

- Carrier (party that moves or transports the product)
- Vehicle-related cost
- Fixed operating cost
- Trip-related cost
- Shipper (party that requires the movement of the product between two points in the supply chain)
- Transportation cost
- Inventory cost
- Facility cost

15. List out the Transportation Modes?(K1,CO4)

- Trucks
- TL (Truck Load)
- LTL(Less than Truck Load)
- Rail
- Air
- Package Carriers
- Water
- Pipeline
- Intermodal

16. List out the Package Carriers? (K1,CO4)

- Companies like FedEx, UPS, USPS, that carry small packages ranging from letters to shipments of about 150 pounds
- Expensive
- Rapid and reliable delivery
- Small and time-sensitive shipments
- Preferred mode for e-businesses (e.g., Amazon, Dell, McMaster-Carr)
- Consolidation of shipments (especially important for package carriers that use air as a primary method of transport)

17. What are the various forms of tailored transportation in supply chain? (K1, CO4)

The various forms of tailored transportation in supply chains are

- Tailored Transportation by Customer Density and Distance
- Tailored Transportation by Size of Customer
- Tailored Transportation by Product Demand and Value

18. What are the objectives of routing and scheduling in transportation? (K1, CO4)

Typical objectives when routing and scheduling vehicles include a combination
of minimizing cost by

- decreasing the number of vehicles needed
- the total distance travelled by vehicles
- the total travel time of vehicles
- eliminating service failures such as a delays in shipments

19. What are the two methods used for routing and scheduling in transportation? (K1, CO4)

The two methods used for routing and scheduling in transportation are

- The Savings matrix method
- The Generalized Assignment method

20. What is Sweep Procedure? (K1, CO4)

In the sweep procedure, any point on the grid is selected (generally the DC itself) and a line is swept either clockwise or counterclockwise from that point. The trip is constructed by sequencing customers in the order they are encountered during the sweep.

21. Define Farthest Insert. (K1, CO4)

Farthest Insert

Given a vehicle trip (including a trip consisting of only the DC) for each remaining customer, find the minimum increase in length for this customer to be inserted from all the potential points in the trip that they could be inserted. Then choose to actually insert the customer with the largest minimum increase to obtain a new trip. This step is referred to as a farthest insert because the customer farthest from the current trip is inserted. The process is continued until all remaining customers to be visited by the vehicle are included in a trip.

22. Define Nearest Insert. (K1, CO4)

Nearest Insert

Given a vehicle trip (including a trip consisting of only the DC) for each remaining customer, find the minimum increase in length for this customer to be inserted from all the potential points in the trip that they could be inserted. Insert the customer with the smallest minimum increase to obtain a new trip. This step is referred to as a nearest insert because the customer closest to the current trip is inserted. The process is continued until all remaining customers the vehicle will visit are included in a trip.

23. Define 2-OPT. (K1, CO4)

2-OPT

The 2-OPT procedure starts with a trip and breaks it at two places. This results in the trip breaking into two paths, which can be reconnected in two possible ways. The length for each reconnection is evaluated, and the smaller of the two is used to define a new trip. The procedure is continued on the new trip until no further improvement results.

24. Define 3-OPT. (K1, CO4)

3-OPT

The 3-OPT procedure breaks a trip at three points to obtain three paths that can be reconnected to form up to eight different trips. The length of each of the eight possible trips is evaluated and the shortest trip is retained. The procedure is continued on the new trip until no further improvement results.

25. What are the steps used in Generalized Assignment Method? (K1, CO4)

The procedure for routing and sequencing of vehicles consists of the following steps:

1. Assign seed points for each route.
2. Evaluate insertion cost for each customer.
3. Assign customers to routes.
4. Sequence customers within routes.

26. What are the steps used in Savings Matrix Method? (K1, CO4)

The major steps in the savings matrix method are:

1. Identify the distance matrix.
2. Identify the savings matrix.
3. Assign customers to vehicles or routes.
4. Sequence customers within routes.

Part-B Questions

Q. No.	Questions		
1	Discuss in details about Role of transportation in supply chain.		
2	Explain the factors affecting transportations decision.		
3	Discuss in detail Design option for transportation network.		
4	Explain the various forms of tailored transportation in detail.		
5	Explain the importance of routing and Scheduling in Transportation with an example.		
6	Explain the routing and scheduling in transportation of Amazon		
7	Explain The Saving's Matrix method in detail with an example.		
8	Explain the Generalized Assignment method in detail with an example.		

Table of Contents

4.1 The Role of Sourcing in a Supply Chain

Purchasing, also called **procurement**, is the process by which companies acquire raw materials, components, products, services, or other resources from suppliers to execute their operations. **Sourcing** is the entire set of business processes required to purchase goods and services. For any supply chain function, the most significant decision is whether to outsource the function or perform it in- house. **Outsourcing** results in the supply chain function being performed by a third party. Outsourcing is one of the most important issues facing a firm, and actions across industries tend to be varied.

It is important to clarify the distinction between outsourcing and off- shoring before we proceed. A firm off-shores a supply chain function if it maintains ownership but moves the production facility offshore. In contrast, a firm outsources if the firm hires an outside firm to perform an operation rather than executing the operation within the firm. We address the outsourcing of supply chain activities based on the following two questions:

1. Will the third party increase the supply chain surplus relative to performing the activity in house?

2. To what extent do risks grow upon outsourcing?

Recall that the supply chain surplus is the difference between the value of a product for the customer and the total cost of all supply chain activities involved in bringing the product to the customer. The supply chain surplus is the total size of the pie that all supply chain participants (including the customer) get to share. Our basic premise is that outsourcing makes sense if it increases the supply chain surplus without significantly affecting risks.

Once a decision to outsource has been made, sourcing processes include the selection of suppliers, design of supplier contracts, product design collaboration, procurement of material or services, and evaluation of supplier performance, as shown in Figure 4-1.

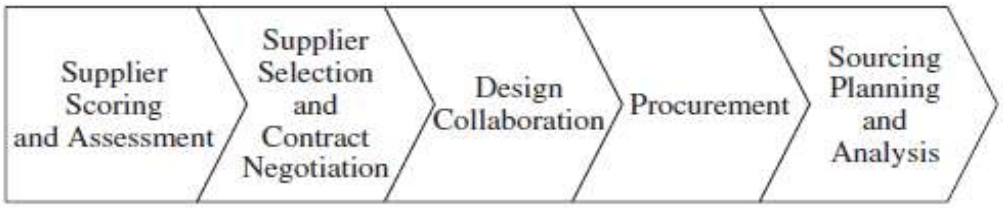


Fig 4.1 Key Sourcing-Related Processes

Supplier scoring and assessment is the process used to rate supplier performance. Suppliers should be compared based on their impact on the supply chain surplus and total cost. Unfortunately, sourcing decisions are often driven based solely on the price charged by a supplier. Many other supplier characteristics, such as lead time, reliability, quality, and design capability, also affect the total cost of doing business with a supplier.

A good supplier scoring and assessment process must identify and track performance along all dimensions that affect the total cost of using a supplier. Supplier selection uses the output from supplier scoring and assessment to identify the appropriate supplier(s). A supply contract is then negotiated with the supplier. A good contract should account for all factors that affect supply chain performance and should be designed to increase supply chain profits in a way that benefits both the supplier and the buyer.

Given that about 80 percent of the cost of a product is determined during design, it is crucial that suppliers be actively involved at this stage. Design collaboration allows the supplier and the manufacturer to work together when designing components for the final product. Design collaboration also ensures that any design changes are communicated effectively to all parties involved with designing and manufacturing the product. Once the product has been designed, procurement is the process whereby the supplier sends product in response to orders placed by the buyer. The goal of procurement is to enable orders to be placed and delivered on schedule at the lowest possible overall cost.

Effective sourcing processes within a firm can improve profits for the firm and total supply chain surplus in a variety of ways. It is important that the drivers of improved profits be clearly identified when making sourcing decisions. Some of the benefits from effective sourcing decisions are the following:

- Better economies of scale can be achieved if orders within a firm are aggregated.

- More efficient procurement transactions can significantly reduce the overall cost of purchasing. This is most important for items for which a large number of low value transactions occur.

- Design collaboration can result in products that are easier to manufacture and distribute, resulting in lower overall costs. This factor is most important for supplier products that contribute a significant amount to product cost and value.

- Good procurement processes can facilitate coordination with the supplier and improve forecasting and planning. Better coordination lowers inventories and improves the matching of supply and demand.

- Appropriate supplier contracts can allow for the sharing of risk, resulting in higher profits for both the supplier and the buyer.

- Firms can achieve a lower purchase price by increasing competition through the use of auctions.

When designing a sourcing strategy, it is important for a firm to be clear on the factors that have the greatest influence on performance and target improvement on those areas. For example, if most of the spending for a firm is on materials with only a few high value transactions, improving the efficiency of procurement transactions will provide little value, whereas improving design collaboration and coordination with the supplier will provide significant value. In contrast, when sourcing items with many low-value transactions, increasing the efficiency of procurement transactions will be very valuable.

IN-HOUSE OR OUTSOURCE

The decision to outsource is based on the growth in supply chain surplus provided by the third party and the increase in risk incurred by using a third party. A firm should consider outsourcing if the growth in surplus is large with a small increase in risk. Performing the function in-house is preferable if the growth in surplus is small or the increase in risk is large.

How Do Third Parties Increase The Supply Chain Surplus

Third parties increase the supply chain surplus if they either increase value for the customer or decrease the supply chain cost relative to a firm performing the task in- house.

1. Capacity aggregation.

A third party can increase the supply chain surplus by aggregating demand across multiple firms and gaining production economies of scale that no single firm can on its own.

2. Inventory aggregation.

A third party can increase the supply chain surplus by aggregating inventories across a large number of customers. Aggregation allows them to significantly lower overall uncertainty and improve economies of scale in purchasing and transportation.

3. Transportation aggregation by transportation intermediaries.

A third party may increase the surplus by aggregating the transportation function to a higher level than any shipper can on its own. UPS, FedEx, and a host of LTL carriers are examples of transportation intermediaries that increase the supply chain surplus by aggregating transportation across a variety of shippers.

4. Transportation aggregation by storage intermediaries.

A third party that stores inventory can also increase the supply chain surplus by aggregating inbound and outbound transportation. On the inbound side they are able to aggregate shipments from several manufacturers onto a single truck. This results in a lower transportation cost than could be achieved by each manufacturer independently. On the outbound side they aggregate packages for customers at a common destination,

resulting in a significantly lower transportation cost than can be achieved by each customer separately.

5. Warehousing aggregation.

A third party may increase the supply chain surplus by aggregating warehousing needs over several customers. The growth in surplus is achieved in terms of lower real estate costs as well as lower processing costs within the warehouse. Savings through warehousing aggregation arise if a supplier's warehousing needs are small or if its needs fluctuate over time.

6. Procurement aggregation.

A third party increases the supply chain surplus if it aggregates procurement for many small players and facilitates economies of scale in production and inbound transportation. Procurement aggregation is most effective across many small buyers. A good example is FleetXchange, a firm that offers small truck fleets lower prices for truck equipment and services through aggregate buying. Procurement aggregation is not likely to be a big factor in a situation with a few large customers.

7. Information aggregation.

A third party may increase the surplus by aggregating information to a higher level than can be achieved by a firm performing the function in-house. All retailers aggregate information on products from many manufacturers in a single location. This information aggregation reduces search costs for customers.

8. Receivables aggregation.

A third party may increase the supply chain surplus if it can aggregate the receivables risk to a higher level than the firm or it has a lower collection cost than the firm. Bright Star is a distributor for Motorola in most Latin American countries other than Brazil. Cell phones in the area are sold through many small, independently owned retail outlets. Collecting receivables from each retail outlet is a very expensive proposition for a manufacturer.

9. Relationship aggregation.

An intermediary can increase the supply chain surplus by decreasing

the number of relationships required between multiple buyers and sellers. Without an intermediary, connecting a thousand sellers to a million buyers requires a billion relationships. The presence of an intermediary lowers the number of relationships required to just over a million.

10. Lower costs and higher quality.

A third party can increase the supply chain surplus if it provides lower cost or higher quality relative to the firm. If these benefits come from specialization and learning, they are likely to be sustainable over the longer term. A specialized third party that is further along the learning curve for some supply chain activity is likely to maintain its advantage over the long term.

Three important factors that affect the increase in surplus that a third party provides: **scale, uncertainty, and the specificity of assets.**

RISKS OF USING A THIRD PARTY

Firms must evaluate the following risks when they move any function to a third party.

1. The process is broken
2. Underestimation of the cost of coordination.
3. Reduced customer/supplier contact.
4. Loss of internal capability and growth in third-party power.
5. Leakage of sensitive data and information.
6. Ineffective contracts.

SUPPLIER SCORING AND ASSESSMENT

When comparing suppliers, many firms make the fundamental mistake of focusing only on the quoted price, ignoring the fact that suppliers may differ on other important dimensions that affect the total cost of using a supplier. For instance, suppliers have different replenishment lead times. Does it pay to select a more expensive supplier with a shorter lead time? Or consider suppliers that have

different on-time performance. Is the more reliable supplier worth the few extra pennies it charges per piece?

In each of the aforementioned instances, the price charged by the supplier is only one of many factors that affect the supply chain surplus. When scoring and assessing suppliers, the following factors other than quoted price must be considered:

- Replenishment lead time
- On-time performance
- Supply flexibility
- Supply quality
- Inbound transportation cost
- Pricing terms
- Information coordination capability
- Design collaboration capability
- Exchange rates, taxes, and duties
- Supplier viability

Supplier performance must be rated on each of these factors because they all affect the total supply chain cost. Next we discuss how each factor affects total supply chain cost and how a supplier's rating on the factor can be used to infer a total cost of using the supplier.

1. **Replenishment lead time.** As the replenishment lead time from a supplier grows, the amount of safety inventory that needs to be held by the buyer also grows proportional to the square root of the replenishment lead time. Scoring the performance of suppliers in terms of replenishment lead time thus allows the firm to evaluate the impact each supplier has on the cost of holding safety inventory.

2. **On-time performance.** On-time performance affects the variability of the lead time. A reliable supplier has low variability of lead time, whereas an unreliable supplier has high variability. As the variability of lead time grows, the required safety inventory at the firm grows very rapidly.

3. **Supply flexibility.** Supply flexibility is the amount of variation in order quantity that a supplier can tolerate without letting other

performance factors deteriorate. The less flexible a supplier is, the more lead-time variability it will display as order quantities change. Supply flexibility thus affects the level of safety inventory that the firm will have to carry.

4. **Delivery frequency/minimum lot size.** The delivery frequency and the minimum lot size offered by a supplier affect the size of each replenishment lot ordered by a firm. As the replenishment lot size grows, the cycle inventory at the firm grows, thus increasing the cost of holding.

5. **Supply quality.** A worsening of supply quality increases the variability of the supply of components available to a firm. Quality affects the lead time taken by the supplier to complete the replenishment order and also the variability of this lead time because follow-up orders often need to be fulfilled to replace defective products. As a result, the firm has to carry more safety inventory from a low- quality supplier compared to a high-quality supplier.

6. **Inbound transportation cost.** The total cost of using a supplier includes the inbound transportation cost of bringing material in from the supplier. Sourcing a product overseas may have lower product cost but generally incurs a higher inbound transportation cost, which must be accounted for when comparing suppliers. The distance, mode of transportation, and delivery frequency affect the inbound transportation cost associated with each supplier.

7. **Pricing terms.** Pricing terms include the allowable time delay before payment has to be made and any quantity discounts offered by the supplier. Allowable time delays in payment to suppliers save the buyer working capital. The cost of working capital savings for each supplier can be quantified. Price terms also include discounts for purchases above certain quantities. Quantity discounts lower the unit cost but tend to increase the required batch size and as a result the cycle inventory.

8. **Information coordination capability.** The information coordination capability of a supplier is harder to quantify, but it affects the ability of a firm to match supply and demand. Good coordination results in better replenishment planning, thus decreasing both the inventory carried as well as the sales lost because of lack of availability. Good information coordination also decreases the bullwhip effect and results in lower production,

inventory, and transportation costs while improving responsiveness to the customer.

9. **Design collaboration capability.** Given that a large part of product cost is fixed at design, collaboration capability of a supplier is significant. Good design collaboration for manufacturability and supply chain can also decrease required inventories and transportation cost. As manufacturers are increasingly outsourcing both the design and manufacture of components, their ability to coordinate design across many suppliers is critical to the ultimate success of the product and the speed of introduction. As a result, design collaboration capability of suppliers is becoming increasingly important.

10. **Exchange rates, taxes, and duties.** Although exchange rates, taxes, and duties are not supplier dependent, they can be significant for a firm with a global manufacturing and supply base. In many instances, currency fluctuations affect component price more than all other factors put together. Financial hedges can be put into place to counter exchange-rate fluctuations.

11. **Supplier viability.** Given the impact that suppliers have on a company's performance, an important factor in picking a supplier is the likelihood that it will be around to fulfil the promises it makes. This consideration can be especially important if the supplier is providing mission -critical products and it would be difficult to find a replacement for. Note that this is not necessarily a bias for larger companies-many small companies, and even some start-ups, can provide an acceptable level of viability.

The factors in Table 4-3 allow a firm to rate and compare various suppliers with different performance on each dimension. We have discussed how performance along most of the factors can be quantified . in terms of impact on cost. The overall performance of each supplier can thus be characterized in terms of total cost and a rating on the nonquantifiable factors.

The impact of each factor on total cost is summarised in the following

	Purchase Price of Component	Inventory		Transportation Cost	Product Introduction Time
		Cycle	Safety		
Replenishment Lead Time			x		
On time Performance			x		
Supply Flexibility			x		
Delivery Frequency		x	x	x	
Supply Quality	x		x		
Inbound Transport Cost				x	
Pricing Terms	x	x			
Information Coordination			x	x	
Design Collaboration	x	x	x	x	x
Exchange Rates and Taxes	x				
Supplier Viability			x		x

15

4.2 SUPPLIER SELECTION-AUCTIONS AND NEGOTIATIONS

Before selecting suppliers, a firm must decide whether to use single sourcing or multiple suppliers. Single sourcing guarantees the supplier sufficient business when the supplier has to make a significant buyer-specific investment. The buyer- specific investment may take the form of plant and equipment designed to produce a part that is specific to the buyer or may take the form of expertise that needs to be developed.

Single sourcing is also used in the automotive industry for parts such as seats that must arrive in the sequence of production. Coordinating such sequencing is impossible with multiple sources.

As a result, auto companies have a single seat source for each plant but multiple seat sources across their manufacturing network. Having multiple sources ensures a degree of competition and also the possibility of a backup should a source fail to deliver.

A good test of whether a firm has the right number of suppliers is to analyze what impact deleting or adding a supplier will have. Unless each supplier has a somewhat different role, it is very likely that the supply base is too large. In contrast, unless adding a supplier with a unique and valuable capability clearly adds to total cost, the supply base may be too small.

The selection of suppliers is done using a variety of mechanisms,

including offline competitive bids, reverse auctions, or direct negotiations. No matter what mechanism is used, supplier selection should be based on the total cost of using a supplier and not just the purchase price. Next we discuss some auction mechanisms that are often used in practice and highlight some of their properties.

AUCTIONS IN THE SUPPLY CHAIN

When outsourcing to a third party, firms have historically obtained competitive bids and in recent years have used reverse auctions on the Internet. Competitive bids are a form of auction in which the bids are not revealed to the other bidders. In many supply chain settings, a buyer looks to outsource a supply chain function such as production or transportation. Potential suppliers are first qualified and then allowed to bid on how much they would charge to perform the function. The qualification process is important because there are multiple attributes of that the buyer cares about.

When conducting an auction based primarily on unit price, it is thus important for the buyer to specify performance expectations along all dimensions other than price.

In reality a buyer may be better off with a multi attribute auction, but in most cases buyers end up with specifications on various attributes and a price-only auction. The qualification process is used to identify suppliers that meet performance expectations along the nonprice attributes.

From the buyer's perspective, the purpose of the auction is to get bidders to reveal their underlying cost structure so that the buyer can select the supplier with the lowest costs. Commonly used mechanisms for these auctions are as follows.

- **Sealed-bid first-price auctions** require each potential supplier to submit a sealed bid for the contract by a specified time. These bids are then opened and the contract is assigned to the lowest bidder.

- In **English auctions**, the auctioneer starts with a price and suppliers can make bids as long as each successive bid is lower than the previous bid. The supplier with the last (lowest) bid receives the contract. The difference in this case is that all suppliers get to see the current lowest bid as the auction unfolds.

- In **Dutch auctions**, the auctioneer starts with a low price and then raises it slowly until one of the suppliers agrees to the contract at that price.

- In **second-price (Vickrey) auctions**, each potential supplier submits a bid. The contract is assigned to the lowest bidder but at the price quoted by the second-lowest bidder.

When identifying the auction to use, the firm wants to minimize the price it pays. The firm may also care about ending up with the supplier with the lowest underlying costs because it makes it more likely that the supplier will actually be able to supply at the price it has committed to. A related issue is whether suppliers have any incentive to make false bids that are_not consistent with their cost structure. Such bids may increase what the firm pays and also lead to the contract being given to a firm that does not have the lowest costs.

An important issue with the sealed-bid first-price auction is what is known as the winner's curse. Once selected based on sealed bids, the winner quickly realizes that it could have raised its bid slightly and still won, because other suppliers bid at a higher level. In this sense, winning the bid leads the winner to realize that it left money on the table. Thus, bidders adjust their initial sealed bids upward, taking this phenomenon into account. This issue does not arise in any open auction, where bidders see the current best bid when planning their next bid. This issue also does not arise in the second price auction because the winner gets the price quoted by the second-lowest bidder and thus has no incentive to hide its true cost.

Let us start with the cost structures for suppliers. In most instances it is reasonable to assume that part of the supplier's cost arises from how it has structured its processes and part of its cost arises from market factors such as raw material and labor cost that are common across suppliers.

The following factors influence the performance of an auction:

- Is the supplier's cost structure private (not affected by factors that are common to other bidders)?
- Are suppliers symmetric or not, that is, ex ante, are they expected to have similar

cost structures?

- Do suppliers have all the information they need to estimate their cost structure?
- Does the buyer specify a maximum price it is willing to pay for the supply chain?

Thus, it is in the buyer's interest not only to reveal all public information before bidding but also to convince potential suppliers that all information has been revealed.

A very significant factor that must be accounted for when designing an auction is the possibility of collusion among bidders. Second-price auctions are particularly vulnerable to collusion among bidders. Consider an agreement among bidders under which the bidder with the lowest cost agrees to bid its true cost, with all other bidders bidding a high number (say, the cost of the most expensive bidder or the reserve price of the buyer). In a second-price auction, the lowest-cost bidder gets to perform the supply chain function but the buyer has to a pay a higher price than the cost of the second-lowest-cost supplier. This collusion strategy is an equilibrium because none of the other bidders has anything to gain by deviating from the collusion agreement. Observe that this collusion strategy can be· avoided with any first-price auction, either sealed bid or English. In either case, a collusion agreement with a very high price will not hold, because many bidders will have the temptation to join the bidding if they have a lower cost. Ultimately, any first-price auction will bring more than the lowest-cost bidder in to the auction.

Collusion results in suppliers suppressing their desire to provide the supply chain function and raising their bids from what would be appropriate given their cost.

The price is lowered slowly until suppliers have committed to all units of goods or services desired by the buyer. In this auction, each unit is supplied at a different price. In a multiunit English auction,

the buyer starts at a high price and bidders announce the quantity they are willing to supply. If the total quantity that suppliers are willing to supply exceeds the desired quantity, the buyer lowers the price until the quantity for which suppliers bid equals the desired quantity.

All suppliers then get to supply at this price. This auction is also referred to as the uniform-price auction. Suppliers in either auction can raise the final price by colluding and forming a bidding ring that assigns only one bidder to enter the auction process for the entire ring. After the initial auction the ring then has a separate auction to divide up the quantity they have been assigned among themselves.

BASIC PRINCIPLES OF NEGOTIATION

In some instances, the third party that will perform a given supply chain function has been identified and the firm enters into a negotiation to set the terms of the contract. Negotiation is likely to result in a positive outcome only if the value the buyer places on outsourcing the supply chain function to this supplier is at least as large as the value the supplier places on performing the function for the buyer. The value that a supplier places on performing a function is influenced by its cost as well as other alternatives that are available for its existing capacity. Similarly, the value that the buyer places is influenced by the cost of performing the function in-house and the price available from alternative suppliers. The difference between the values of the buyer and seller is referred to as the bargaining surplus. The goal of each negotiating party is to capture as much of the bargaining surplus as possible.

An excellent discussion on negotiations is available in Thompson (2005). We mention some of the highlights from her discussion. The first recommendation is to have a clear idea of your own value and as good an estimate of the third party's value as possible. A good estimate of the bargaining surplus improves the chance of a successful outcome. Suppliers of Toyota have often mentioned that "Toyota knows our costs better than we do," which leads to better negotiations. The second recommendation is to look for a fair outcome based on equally or equitably dividing the bargaining surplus or dividing it based on needs. Equity here refers to a division of the surplus in proportion to the contribution by each party.

The key to a successful negotiation, however, is to make it a win-win

outcome. It is impossible to obtain a win-win outcome if the two parties are negotiating on a single dimension such as price. In this setting, one party can only "win" at the expense of the other. To create a win-win negotiation, the two parties have to identify more than one issue to negotiate. Identifying multiple issues allows the opportunity to expand the pie if the two parties have different preferences. This is often easier than it seems in a supply chain setting. A buyer typically cares not just about the price of performing the supply chain function but also about the responsiveness and quality (two of the dimensions identified in Table 14-3). If the supplier finds it harder to lower the price but easier to reduce the response time, there is an opportunity for a win-win resolution in which the supplier offers better responsiveness without changing the price. Thompson discusses many hurdles in the negotiation process and also suggests effective strategies.

CONTRACTS AND SUPPLY CHAIN PERFORMANCE

A supply contract specifies parameters governing the buyer-supplier relationship. In addition to making the terms of the buyer-supplier relationship explicit, contracts have significant impact on the behaviour and performance of all stages in a supply chain. Contracts should be designed to facilitate desirable supply chain outcomes and minimize actions that hurt performance.

A manager should ask the following three questions when designing a supply chain contract:

1. How will the contract affect the firm's profits and total supply chain profits?

2. Will the incentives in the contract introduce any information distortion?

3. How will the contract influence supplier performance along key performance measures?

Ideally, a contract should be structured to increase the firm's profits and supply chain profits, discourage information distortion, and offer incentives to the supplier to improve performance along key dimensions. Many shortcomings in supply chain performance occur because the buyer and supplier are two different entities, each trying to optimize its own profits.

CONTRACTS FOR PRODUCT AVAILABILITY AND SUPPLY CHAIN PROFITS

Actions taken by the two parties in the supply chain often result in profits that are lower than what could be achieved if the supply chain were to coordinate its actions with a common objective of maximizing supply chain profits. Consider a product whose demand is significantly affected by the retail price. The retailer decides its price (and thus sales quantity) based on its margin. The retailer's margin is only a fraction of the supply chain margin, leading to a retail price that is higher than optimal and a sales quantity that is lower than optimal for the supply chain.

This phenomenon is referred to as double marginalization. The supplier can increase supply chain profits by offering a volume discount, where the retailer pays a lower price if the total quantity purchased exceeds a threshold.

Another example of double marginalization arises in the presence of demand uncertainty. A manufacturer wants the retailer to carry a large inventory of its product to ensure that any surge in demand can be satisfied. The retailer, on the other hand, loses money on any unsold inventory. As a result, the retailer prefers to carry a lower level of inventory. This tension leads to a supply chain outcome that is suboptimal.

In a contract in which the supplier specifies a fixed price and the buyer decides on the quantity to be purchased, the most common cause for suboptimal supply chain performance is double marginalization. The retailer makes its buying decision before demand is realized and thus bears all the demand uncertainty. If demand is less than the retailer's inventory, the retailer has to liquidate unsold product at a discount. Given uncertain demand, the retailer decides on the purchase quantity based on its margin and the cost of overstocking. The retailer's margin, however, is lower than the contribution margin for the entire supply chain, whereas its cost of overstocking is higher than that for the entire supply chain. As a result, the retailer is conservative and aims for a lower level of product availability than is optimal for the supply chain.

Consider a music store that sells compact discs. The supplier buys (or manufactures) compact discs at $1 per unit and sells them to the music store at
$5 per unit. The retailer sells each disc to the end consumer at $10.

At this retail price, market demand is normally distributed, with a mean of 1,000 and a standard deviation of 300.

The retailer has a margin of $5 per disc and can potentially lose $5 for each unsold disc. Using Equation 12.1, it is optimal for the retailer to aim for a service level of 0.5 and order 1,000 discs. From Equation 12.3, the retailer's expected profits are $3,803 and the manufacturer makes $4,000 from selling 1,000 discs. For the supply chain, however, the supplier and the retailer together have a margin of $9 and can lose a maximum of only $1 per unsold disc. For the entire supply chain it is thus optimal to aim for a service level of 0.9 and stock 1,384 discs. The expected supply chain profit in this case is $8,474. The music store is thus conservative and carries fewer discs than are optimal for the supply chain. As a result, the supply chain makes $670 less than it would expect to if the retailer and the supplier worked together.

To improve overall profits, the supplier must design a contract that encourages the buyer to purchase more and increase the level of product availability. This requires the supplier to share in some of the buyer's demand uncertainty. Three contracts that increase overall profits by making the supplier share some of the buyer's demand uncertainty are as follows:

1. Buyback or returns contracts

2. Revenue-sharing contracts

3. Quantity flexibility contracts

We illustrate each of the three contracts using the example of the music store and discuss their performance in terms of the three questions raised earlier.

Buyback Contracts

A buy-back or returns clause in a contract allows a retailer to return unsold inventory up to a specified amount, at an agreed-upon price. In a buy-back contract, the manufacturer specifies a wholesale price c along with a buy-back price b at which the retailer can return any unsold units at the end of the season. The manufacturer can salvage $sM for any units that the retailer returns.

The optimal order quantity o* for a retailer in response to a

buy-back contract is evaluated, where the salvage value for the retailer is s = b. The expected profit at the manufacturer depends on the overstock at the retailer that is returned. We obtain

Expected manufacturer profit = O*(c - v) - (b - sM) X expected overstock at retailer

Revenue-Sharing Contracts

In revenue-sharing contracts, the manufacturer charges the retailer a low wholesale price c, and shares a fraction f of the retailer's revenue. Even if no returns are allowed, the lower wholesale price decreases the cost to the retailer in case of an overstock. The retailer thus increases the level of product availability resulting in higher profits for both the manufacturer and the retailer.

Assume that the manufacturer has a production cost v; the retailer charges a retail price p and can salvage any leftover units for sR. The optimal order quantity o* ordered by the retailer is evaluated, where the cost of understacking is Cu = (1 - f)p - c and the cost of overstocking is Cu = c - sR.

Expected manufacturer's profits = (c - v) o* + fp(o* - expected overstock at retailer)

Quantity Flexibility Contracts

Under quantity flexibility contracts, the manufacturer allows the retailer to change the quantity ordered after observing demand. If a retailer orders 0 units, the manufacturer commits to providing Q = (1 + α)O units, whereas the retailer is committed to buying at least q = (1 - β)0 units. Both α and β are between 0 and
1. The retailer can purchase up to Q units, depending on the demand it observes. These contracts are similar to buy-back contracts in that the manufacturer now bears some of the risk of having excess inventory. Because no returns are required, these contracts can be more effective than buy-back contracts when the cost of returns is high. Quantity flexibility contracts increase the average amount the retailer purchases and may increase total supply chain profits.

CONTRACTS TO COORDINATE SUPPLY CHAIN COSTS.

Differences in costs at the buyer and supplier also lead to decisions

that increase total supply chain costs. An example is the replenishment lot size decision typically made by the buyer. The buyer decides on its optimal lot size based on its fixed cost per lot and the cost of holding inventory. The buyer does not account for the supplier's costs. If the supplier has a high fixed cost per lot, the optimal lot size for the buyer increases total cost for the supplier and the supply chain. In such a situation, the supplier can use a quantity discount contract to encourage the buyer to order in lot sizes that minimize total costs. The objective of such a contract is to encourage the retailer to buy in larger lot sizes that lower cost for the supplier and the entire supply chain.

A quantity discount contract decreases overall costs but leads to higher lot sizes and thus higher levels of inventory in the supply chain. It is typically justified only for commodity products for which the supplier has high fixed costs per lot. It is important to modify the terms of the contract as operational improvements are made at the supplier, resulting in lower fixed costs per batch.

CONTRACTS TO INCREASE AGENT EFFORT

In many supply chains, agents act on behalf of a principal and the agents' effort affects the reward for the principal. As an example, consider a car dealer (the agent) selling cars for DaimlerChrysler (the principal). The dealer also sells other brands and used cars. Every month the dealer allocates its sales effort (advertising, promotions, etc.) across all brands it sells and the used cars. Earnings for DaimlerChrysler are based on sales of its brands, which in turn are affected by the effort exerted by the dealer. Sales can be observed directly, whereas effort is hard to observe and measure. Given double marginalization, the dealer always exerts less effort than is optimal from the perspective of DaimlerChrysler and the supply chain. Thus, DaimlerChrysler must offer an incentive contract that encourages the dealer to increase effort.

In theory, a two-part tariff offers the right incentives for the dealer to exert the appropriate amount of effort. In a two-part tariff, DaimlerChrysler extracts its profits up front as a franchise fee and then sells cars to the dealer at cost. The dealer's margin is then the same as the supply chain margin, and the dealer exerts the right amount of effort.

CONTRACTS TO INDUCE PERFORMANCE IMPROVEMENT

In many instances a buyer wants performance improvement from a supplier that has little incentive to do so. A buyer with sufficient power in the supply chain may be able to force the supplier to comply. A buyer without sufficient power requires an appropriate contract to induce the supplier to improve perfomance. Even for a powerful buyer, however, an appropriate contract designed to encourage supplier cooperation results in a better outcome.

As an example, consider a buyer that wants the supplier to improve performance by reducing lead time for a seasonal item. This is an important component of all quick response (QR) initiatives in a supply chain. With a shorter lead time, the buyer hopes to have better forecasts and be better able to match supply and demand. Most of the work to reduce lead time has to be done by the supplier, whereas most of the benefit accrues to the buyer. In fact, the supplier will lose sales because the buyer will now carry less safety inventory because of shorter lead times and better forecasts. To induce the supplier to reduce lead time, the buyer can use a shared-savings contract, with the supplier getting a fraction of the savings that result from reducing lead time. As long as the supplier's share of the savings compensates for any effort it has to put in, its incentive will be aligned with that of the buyer, resulting in an outcome that benefits both parties.

4.3 DESIGN COLLABORATION

Two important statistics highlight the importance of design collaboration between a manufacturer and suppliers. Today, typically between 50 and 70 percent of the spending at a manufacturer is through procurement, compared to only about 20 percent several decades ago. It is generally accepted that about 80 percent of the cost of a purchased part is fixed during the design stage. Thus, it is crucial for a manufacturer to collaborate with suppliers during the design stage if product costs are to be kept low. Design collaboration can lower the cost of purchased material and also lower logistics and manufacturing costs. Design collaboration is also important for a company trying to provide a lot of variety and customization, because failure to do so can significantly raise the cost of variety.

Working with suppliers can speed up product development time significantly. This is crucial in an era when product life cycles are shrinking and bringing a product to market before the competition offers a significant competitive advantage. Finally, integrating the supplier into the design phase allows the manufacturer to focus on system integration, resulting in a higher quality product at lower cost. For example, auto manufacturers are increasingly playing the role of system integrators rather than component designers. This is an approach that has been used even more extensively in the high-tech industry.

As suppliers take on a bigger design role, it is important for manufacturers also to become design coordinators for the supply chain. Common part descriptions should be available to all parties involved in the design, and any design changes by one party should be communicated to all suppliers affected. A good database of existing parts and designs can save significant amounts of money and time. For example, when Johnson Controls finds a seat frame from its database that fulfills all customer requirements, it saves the customer about $20 million on the design, development, tooling, and prototyping expense.

A survey by the Procurement and Supply Chain Benchmarking Consortium at Michigan State University dramatically demonstrates the impact of successfully integrating suppliers in product design. The most successful integration efforts have seen costs decrease by 20 percent, quality improve by 30 percent, and time- to-market decrease by 50 percent.

Key themes that must be communicated to suppliers as they take greater responsibility for design are design for logistics and design for manufacturability. Design for logistics attempts to reduce transportation, handling, and inventory costs during distribution by taking appropriate actions during design. To reduce transportation and handling costs, the manufacturer must convey expected order sizes from retailers and the end consumer to the designer. Packages can then be designed so that transportation costs are lowered and handling is minimized. To reduce transportation cost, packaging is kept as compact as possible and is also designed to ensure easy stacking.

To reduce handling costs, package sizes are designed to minimize the need to break open a pack to fulfill an order. To reduce inventory costs, the primary approach is to design the product for postponement and mass customization. Postponement strategies aim to design a product and production process so that features that differentiate end products are introduced late in the manufacturing phase. As discussed, Dell designs its PCs so that all components about which customers have a choice are assembled after the customer order arrives.

This allows Dell to lower inventories by aggregating them as components. Mass customization strategies use a similar approach by designing the product so that inventory can be carried in a form that aggregates across multiple end products. The goal is to design a product so that customization occurs along a combination of the following three customization categories: modular, adjustable, and dimensional. To provide modular customization, the product is designed as an assembly of modules that fit together. All inventory is then maintained as modules that are assembled to order. A good example of modular customization is PC assembly at Dell. An example of adjustable customization is a washing machine designed by Matsushita that can automatically select from among 600 different cycles. All inventory is thus maintained as a single product, and each customer uses the machine to match its specific needs. An

example of dimensional customization given by Josepn Pine (1999) is a machine that makes custom house gutters on site, which can then be cut to fit the dimensions of the house. Another example is National Bicycle, which cuts the frame tubing to fit the body size of the customer. Design for manufacturability attempts to design products for ease of manufacture. Some of the key principles used include part commonality, eliminating right-hand and left-hand parts, designing symmetrical parts, combining parts, using catalog parts rather than designing a new part, and designing parts to provide access for other parts and tools.

A good area in which to view design collaboration efforts is in the automotive industry. Car manufacturers all over the world are asking suppliers to participate in every aspect of product development, from conceptual design to manufacturing. Ford, for example, asked suppliers for the Thunderbird not only to manufacture the components and subsystems, but also to be responsible for their design. Solid integration throughout the supply chain allowed Ford to bring the new model to market within 36 months of program approval.

To ensure effective communication, Ford required all its vendors to be on the same software platform for design. Ford also opened all its internal databases to its suppliers and collocated many of the suppliers at its offices. Ford engineers were in constant communication with the suppliers and helped coordinate the overall design. The result was a significant improvement in cost, time, and quality.

4.4 SOURCING PLANNING AND ANALYSIS

Periodically, each firm must analyze its procurement spending and supplier performance and use this as input for future sourcing decisions.

One important analysis is the aggregation of spending across and within categories and suppliers. Aggregation provides visibility into what a company is purchasing and from whom the product is being purchased. Managers can use this information to determine economic order quantities, volume discounts, and projected quantity discounts on future volumes.

A simple step is to consolidate spending and ensure that the firm's economic order quantity matches the supplier's economic production quantity. Managers can thus realize better economies of

scale and utilize resources more effectively.

The second piece of analysis relates to supplier performance. Supplier performance should be measured against plan on all dimensions that affect total cost, such as responsiveness, lead times, on-time delivery, quality, and delivery accuracy.

Spending and supplier performance analysis should be used to decide on the portfolio of suppliers to be used and the allocation of demand among the chosen suppliers. the portfolio generally should not consist of similar suppliers. The portfolio should be constructed so that one supply source performs very well on one dimension, whereas another source performs very well on a complementary dimension.

For example, a company can source more effectively using a low-cost supplier with longer lead times along with a high-cost supplier with short lead times compared to using only one type of supplier. Similarly, one should not ignore a somewhat lower-quality source if it is much cheaper than other sources. It is also not effective to use only the cheaper but lower-quality source. It may be very effective to use the cheaper but lower-quality source along with a higher-quality but more expensive source.

Once a supplier portfolio has been determined, the next question is the allocation of demand among the suppliers. The allocation should be related to the economic manufacturing quantity for each source and its cost of supply. The low-cost supplier is given large, steady orders independent of demand, whereas the flexible source is given small orders that fluctuate with demand. The flexible source has smaller economic order quantities and is better able to adjust to the fluctuations. The combination of suppliers results in a better matching of supply and demand at lower cost than using one type of supplier.

Once a supplier portfolio has been determined, the next question is the allocation of demand among the suppliers. The allocation should be related to the economic manufacturing quantity for each source and its cost of supply. The low-cost supplier is given large, steady orders independent of demand, whereas the flexible source is given small orders that fluctuate with demand. The flexible source has smaller economic order quantities and is better able to adjust to the fluctuations. The combination of suppliers results in a better matching of supply and demand at lower cost than using one type of supplier.

4.5 Supply Chain Co-ordination

Supply chain coordination occurs when all stages of a supply chain work toward the objective of maximizing total supply chain profitability based on shared information. Lack of coordination can result in a significant loss of supply chain surplus.

Coordination among different stages in a supply chain requires each stage to share appropriate information with other stages.

Supply chain coordination improves if all stages of the chain take actions that are aligned and increase total supply chain surplus. Supply chain coordination requires each stage of the supply chain to share information and take into account the impact its actions have on other stages.

A lack of coordination occurs either because different stages of the supply chain have objectives that conflict or because information moving between stages is delayed and distorted.

Different stages of a supply chain may have conflicting objectives if each stage has a different owner. As a result, each stage tries to maximize its own profits, resulting in actions that often diminish total supply chain profits.

Today, supply chains consist of stages with different owners. For example, Ford Motor Company has thousands of suppliers from Goodyear to Motorola, and each of these suppliers has many suppliers in turn. Information is distorted as it moves across the supply chain because complete information is not shared between stages.

This distortion is exaggerated by the fact that supply chains today produce a large variety of products. Ford produces different models with several options for each model. The increased variety makes it difficult for Ford to coordinate information exchange with thousands of suppliers and dealers.

The fundamental challenge today is for supply chains to achieve coordination in spite of multiple ownership and increased product variety.

4.6 Bull whip Effect

One outcome of the lack of supply chain coordination is the bullwhip effect, in which fluctuations in orders increase as they move up the supply chain from retailers to wholesalers to manufacturers to suppliers, as shown in Figure. The bullwhip effect distorts demand information within the supply chain, with each stage having a different estimate of what demand looks like.

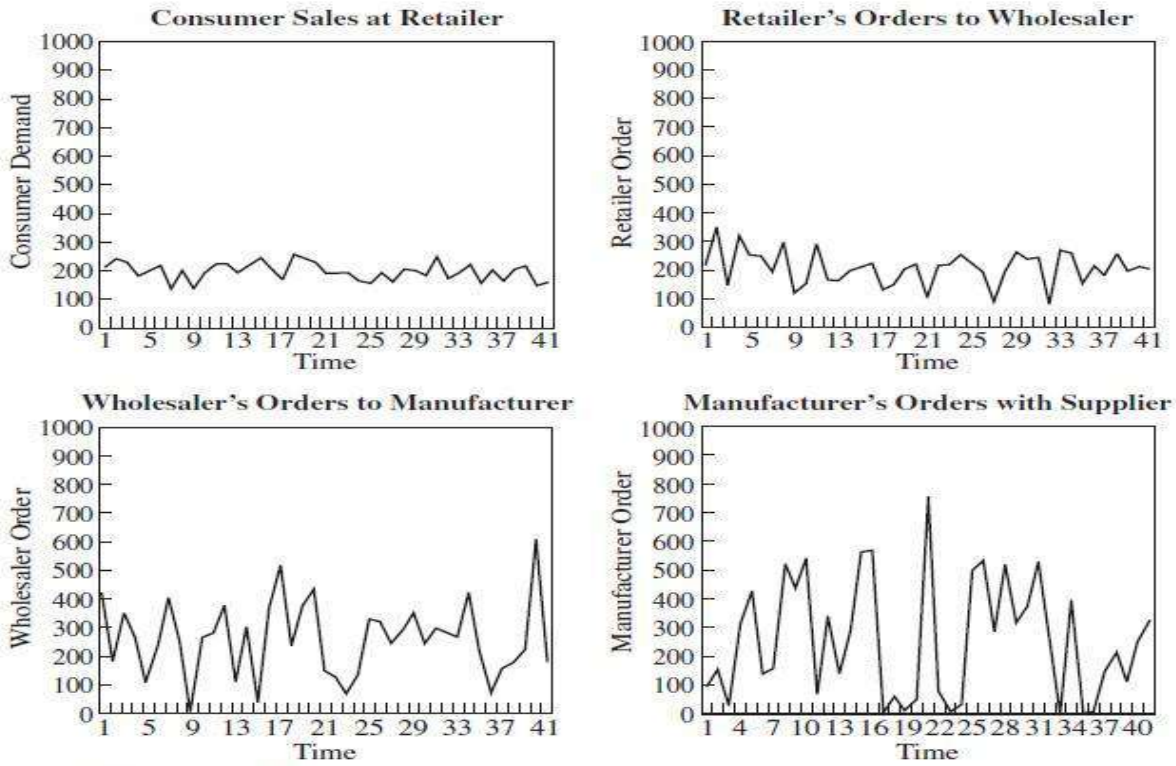


FIGURE 10-1 Demand Fluctuations at Different Stages of a Supply Chain

Procter & Gamble (P&G) has observed the bullwhip effect in the supply chain for Pampers diapers. The company found that raw material orders from P&G to its suppliers fluctuated significantly over time. Farther down the chain, when sales at retail stores were studied, the fluctuations, while present, were small. It is reasonable to assume that the consumers of diapers (babies) at the last stage of the supply chain used them at a steady rate. Although consumption of the end product was stable, orders for raw material

were highly variable, increasing costs and making it difficult to match supply and demand.

HP also found that the fluctuation in orders increased significantly as they moved from the resellers up the supply chain to the printer division to the integrated circuit division. Once again, while product demand showed some variability, orders placed with the integrated circuit division were much more variable. This made it difficult for HP to fill orders on time and increased the cost of doing so.

Studies of the apparel and grocery industry have shown a similar phenomenon: The fluctuation in orders increases as we move upstream in the supply chain from retail to manufacturing.

Barilla, an Italian manufacturer of pasta, observed that weekly orders placed by a local distribution center fluctuated by up to a factor of 70 in the course of the year, whereas weekly sales at the distribution center (representing orders placed by supermarkets) fluctuated by a factor of less than three.3 Barilla was thus facing demand that was much more variable than customer demand. This led to increased inventories, poorer product availability, and a drop in profits.

A similar phenomenon, over a longer time frame, has been observed in several industries that are quite prone to “boom and bust” cycles. A good example is the production of memory chips for personal computers. Between 1985 and 1998, at least two cycles occurred during which prices of memory chips fluctuated by a factor of more than three. These large fluctuations in price were driven by either large shortages or surpluses in capacity. The shortages were exacerbated by panic buying and over-ordering that was followed by a sudden drop in demand.

4.7 The Effect on Performance of Lack of Coordination

A supply chain lacks coordination if each stage optimizes only its local objective, without considering the impact on the complete chain. Total supply chain profits are thus less than what could be achieved through coordination. Lack of coordination also results if information distortion occurs within the supply chain.

Manufacturing Cost

The lack of coordination increases manufacturing cost in the supply chain. As a result of the bullwhip effect, P&G and its suppliers must satisfy a stream of orders that is much more variable than customer demand. P&G can respond to the increased variability by either

building excess capacity or holding excess inventory, both of which increase the manufacturing cost per unit produced.

Inventory Cost

The lack of coordination increases inventory cost in the supply chain. To handle the increased variability in demand, P&G has to carry a higher level of inventory than would be required if the supply chain were coordinated. As a result, inventory costs in the supply chain increase. The high levels of inventory also increase the warehousing space required and thus the warehousing cost incurred.

Replenishment Lead Time

Lack of coordination increases replenishment lead times in the supply chain. The increased variability as a result of the bullwhip effect makes scheduling at P&G and supplier plants much more difficult than when demand is level. There are times when the available capacity and inventory cannot supply the orders coming in. This results in higher replenishment lead times.

Transportation Cost

The lack of coordination increases transportation cost in the supply chain. The transportation requirements over time at P&G and its suppliers are correlated with the orders being filled. As a result of the bullwhip effect, transportation requirements fluctuate significantly over time. This raises transportation cost because surplus transportation capacity needs to be maintained to cover high- demand periods.

Labor Cost for Shipping and Receiving

The lack of coordination increases labor costs associated with shipping and receiving in the supply chain. Labor requirements for shipping at P&G and its suppliers fluctuate with orders. A similar fluctuation occurs for the labor requirements for receiving at distributors and retailers. The various stages have the option of carrying excess labor capacity or varying labor capacity in response to the fluctuation in orders. Either option increases total labor cost.

Level of Product Availability

Lack of coordination hurts the level of product availability and results in more stockouts in the supply chain. The large fluctuations in orders make it harder for P&G to supply all distributor and retailer orders on time. This increases the likelihood that retailers will run out of stock, resulting in lost sales for the supply chain.

Relationships across the Supply Chain

Lack of coordination has a negative effect on performance at every stage and thus hurts the relationships among different stages of the supply chain. The lack of coordination thus leads to a loss of trust among different stages of the supply chain and makes any potential coordination efforts more difficult.

Lack of coordination has a significant negative impact on the supply chain's performance by increasing cost and decreasing responsiveness. The lack of coordination hurts both responsiveness and cost in a supply chain by making it more expensive to provide a given level of product availability.

Obstacles to Coordination in a Supply Chain

Any factor that leads to either local optimization by different stages of the supply chain or an increase in information delay, distortion, and variability within the supply chain is an obstacle to coordination. If managers in a supply chain are able to identify the key obstacles, they can then take suitable actions to help achieve coordination.

We divide the major obstacles into five categories:

- Incentive obstacles
- Information-processing obstacles
- Operational obstacles
- Pricing obstacles
- Behavioral obstacles

1. Incentive Obstacles

Incentive obstacles occur in situations when incentives offered to different stages or participants in a supply chain lead to actions that increase variability and reduce total supply chain profits.

Local Optimization within Functions or Stages of a Supply Chain

Incentives that focus only on the local impact of an action result in decisions that do not maximize total supply chain surplus.

If the compensation of a transportation manager at a firm is linked to the average transportation cost per unit, the manager is likely to take actions that lower transportation costs even if they increase inventory costs or hurt customer service.

It is natural for any participant in the supply chain to take actions that optimize performance measures along which they are evaluated.

Sales Force Incentives

Improperly structured sales force incentives are a significant obstacle to coordination in a supply chain.

In many firms, sales force incentives are based on the amount the sales force sells during an evaluation period of a month or quarter.

The sales typically measured by a manufacturer are the quantity sold to distributors or retailers (sell-in), not the quantity sold to final customers.

2. Information-Processing Obstacles

Information-processing obstacles occur when demand information is distorted as it moves between different stages of the supply chain, leading to increased variability in orders within the supply chain.

Forecasting Based on Orders and not Customer Demand

When stages within a supply chain make forecasts that are based on orders they receive, any variability in customer demand is magnified as orders move up the supply chain to manufacturers and suppliers. In supply chains where the fundamental means of communication among different stages are the orders that are placed, information is distorted as it moves up the supply chain. A small change in customer demand becomes magnified as it moves up the supply chain in the form of customer orders. The retailer may interpret part of this random increase as a growth trend.

This interpretation will lead the retailer to order more than the observed increase in demand because the retailer expects growth to continue into the future and thus orders to cover for future

anticipated growth. The increase in the order placed with the wholesaler is thus larger than the observed increase in demand at the retailer. The growth trend inferred by the wholesaler will be larger than that inferred by the retailer. The wholesaler will thus place an even larger order with the manufacturer.

Lack of Information Sharing

The lack of information sharing between stages of the supply chain magnifies the information distortion. A retailer such as Wal-Mart may increase the size of a particular order because of a planned promotion. If the manufacturer is not aware of the planned promotion, it may interpret the larger order as a permanent increase in demand and place orders with suppliers accordingly. The manufacturer and suppliers thus have much inventory right after Wal-Mart finishes its promotion.

3. Operational Obstacles

Operational obstacles occur when actions taken in the course of placing and filling orders lead to an increase in variability.

- Ordering in Large Lots

When a firm places orders in lot sizes that are much larger than those in which demand arises, variability of orders is magnified up the supply chain. Firms may order in large lots because a significant fixed cost is associated with placing, receiving, or transporting an order. Large lots may also occur if the supplier offers quantity discounts based on lot size. Figure shows both the demand and the order stream for a firm that places an order every five weeks. Manufacturer supplying several retailers that batch their orders faces an order stream that is much more variable than the demand the retailers experience. If the manufacturer batches its orders to suppliers, the effect is further magnified.

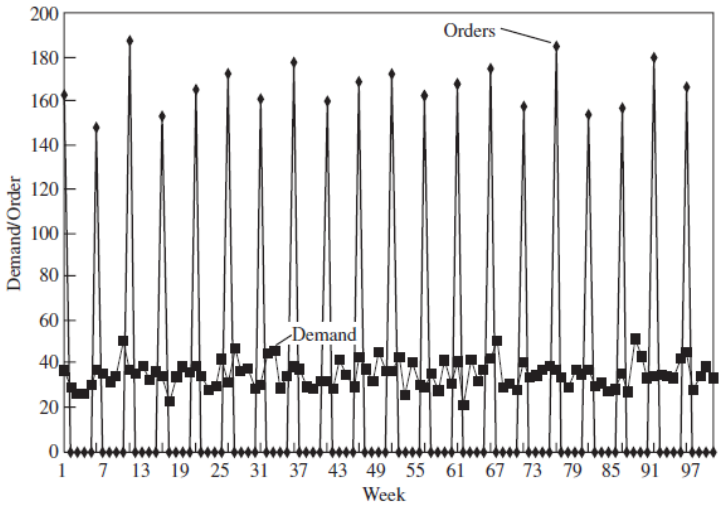


Figure Demand and Order Stream with Orders Every Five Weeks

- Large Replenishment Lead Times

Information distortion is magnified if replenishment lead times between stages are long. Consider a situation in which a retailer has misinterpreted a random increase in demand as a growth trend. If the retailer faces a lead time of two weeks, it will incorporate the anticipated growth over two weeks when placing the order. In contrast, if the retailer faces a lead time of two months, it will incorporate into its order the anticipated growth over two months.

- Rationing and Shortage Gaming

Rationing schemes that allocate limited production in proportion to the orders placed by retailers lead to a magnification of information distortion. This can occur when a high-demand product is in short supply. In such a situation, manufacturers come up with a variety of mechanisms to ration the scarce supply of product among various distributors or retailers. One commonly used rationing scheme is to allocate the available supply of product based on orders placed. Under this rationing scheme, if the supply available is 75 percent of the total orders received, each retailer receives 75 percent of its order.

If the manufacturer is using orders to forecast future demand, it will interpret the increase in orders as an increase in demand even though customer demand is unchanged. The manufacturer may

respond by building enough capacity to be able to fill all orders received. Once sufficient capacity becomes available, orders return to their normal level because they were inflated in response to the rationing scheme.

4. Pricing Obstacles

Pricing obstacles arise when the pricing policies for a product lead to an increase in variability of orders placed.

- Lot Size–Based Quantity Discounts

Lot size–based quantity discounts increase the lot size of orders placed within the supply chain because lower prices are offered for larger lots.

- Price Fluctuations

Trade promotions and other short-term discounts offered by a manufacturer result in forward buying, by which a wholesaler or retailer purchases large lots during the discounting period to cover demand during future periods. Forward buying results in large orders during the promotion period followed by very small orders. Observe that the shipments during the peak period are higher than the sales during the peak period because of a promotion offered.

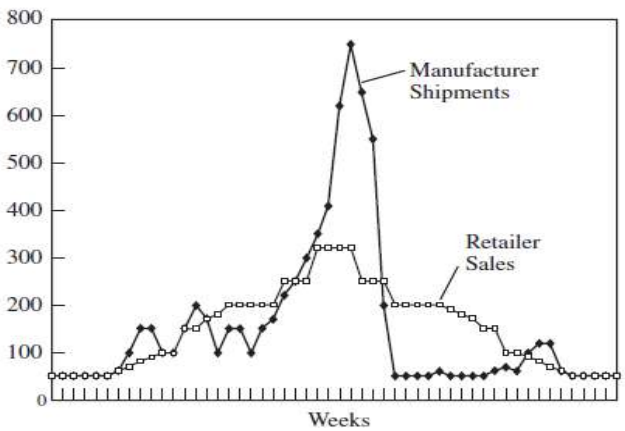


Figure Retailer Sales and Manufacturer Shipments

5. Behavioral Obstacles

Behavioral obstacles are problems in learning within organizations that contribute to information distortion. These problems are often related to the way the supply chain is structured and the communications among different stages. Some of the behavioral obstacles are as follows:

1. Each stage of the supply chain views its actions locally and is unable to see the impact of its actions on other stages.
2. Different stages of the supply chain react to the current local situation rather than trying to identify the root causes.
3. Based on local analysis, different stages of the supply chain blame one another for the fluctuations, with successive stages in the supply chain becoming enemies rather than partners.
4. No stage of the supply chain learns from its actions over time because the most significant consequences of the actions any one stage takes occur elsewhere.
5. A lack of trust among supply chain partners causes them to be opportunistic at the expense of overall supply chain performance. The lack of trust also results in significant duplication of effort.

4.8 Building Strategic Partnerships and Trust

Sharing of accurate information that is trusted by every stage results in a better matching of supply and demand throughout the supply chain and a lower cost. A better relationship also tends to lower the transaction cost between supply chain stages. For example, a supplier can eliminate its forecasting effort if it trusts orders and forecast information received from the retailer. Similarly, the retailer can lessen the receiving effort by decreasing counting and inspections if it trusts the supplier's quality and delivery. In general, stages in a supply chain can eliminate duplicated effort on the basis of improved trust and a better relationship. This lowering of transaction cost along with accurate shared information helps improve coordination. Wal-Mart and P&G have been trying to build a strategic partnership that will better coordinate their actions and be mutually beneficial.

In general, a high level of trust allows a supply chain to become more responsive at lower cost. Actions such as information sharing, changing of incentives, operational improvements, and stabilization of pricing typically help improve the level of trust. Growing the level of cooperation and trust within a supply chain requires a clear identification of roles and decision rights for all parties, effective contracts, and good conflict resolution mechanisms.

Cooperation and trust within the supply chain help improve performance for the following reasons:

- When stages trust each other, they are more likely to take the other party's objectives into consideration when making decisions, thereby facilitating win- win situations.
- Action-oriented managerial levers to achieve coordination become easier to
  implement and the supply chain becomes more agile.
- An increase in supply chain productivity results, either by elimination of duplicated effort or by allocating effort to the appropriate stage.
- Detailed sales and production information is shared; this allows the supply chain to coordinate production and distribution decisions.

Part-A Questions and Answers

1. Define procurement. (K1,CO5)

Purchasing, also called **procurement**, is the process by which companies acquire raw materials, components, products, services, or other resources from suppliers to execute their operations.

2. Define sourcing (K1,CO5)

Sourcing is the entire set of business processes required to purchase goods and services. For any supply chain function, the most significant decision is whether to outsource the function or perform it in-house.

3. Define supply chain surplus. (K1,CO5)

supply chain surplus is the difference between the value of a product for the customer and the total cost of all supply chain activities involved in bringing the product to the customer.

4. Say Some of the benefits from effective sourcing decisions. (K1,CO5)

Better economies of scale can be achieved if orders within a firm are aggregated.

More efficient procurement transactions can significantly reduce the overall cost of purchasing. This is most important for items for which a large number of low value transactions occur.

Design collaboration can result in products that are easier to manufacture and distribute, resulting in lower overall costs. This factor is most important for supplier products that contribute a significant amount to product cost and value.

Good procurement processes can facilitate coordination with the supplier and improve forecasting and planning. Better coordination lowers inventories and improves the matching of supply and demand.

Appropriate supplier contracts can allow for the sharing of risk, resulting in higher profits for both the supplier and the buyer.

Firms can achieve a lower purchase price by increasing competition through the use of auctions.

5. What are the three important factors that affect the increase in surplus. (K1,CO5)

Three important factors that affect the increase in surplus that a third party provides: **scale, uncertainty, and the specificity of assets.**

6. What are the factors influence the performance of an auction. (K1,CO5)

Is the supplier's cost structure private (not affected by factors that are common to other bidders)?

Are suppliers symmetric or not, that is, ex ante, are they expected to have similar cost structures?

Do suppliers have all the information they need to estimate their cost structure?

Does the buyer specify a maximum price it is willing to pay for the supply chain?

7. Define Buyback Contracts (K1,CO5)

A buy-back or returns clause in a contract allows a retailer to return unsold inventory up to a specified amount, at an agreed-upon price. In a buy-back contract, the manufacturer specifies a wholesale price c along with a buy-back price b at which the retailer can return any unsold units at the end of the season. The manufacturer can salvage $sM for any units that the retailer returns.

The optimal order quantity o* for a retailer in response to a buy-back contract is evaluated, where the salvage value for the retailer is s = b. The expected profit at the manufacturer depends on the overstock at the retailer that is returned. We obtain

Expected manufacturer profit = O*(c - v) - (b - sM) X expected overstock at retailer

8. Define Revenue-Sharing Contracts (K1,CO5)

In revenue-sharing contracts, the manufacturer charges the retailer a low wholesale price c, and shares a fraction f of the retailer's revenue. Even if no returns are allowed, the lower wholesale price decreases the cost to the retailer in case of an overstock. The retailer thus increases the level of product availability resulting in higher profits for both the manufacturer and the retailer.

Assume that the manufacturer has a production cost v; the retailer charges a retail price p and can salvage any leftover units for sR. The optimal order quantity o* ordered by the retailer is evaluated, where the cost of understacking is Cu = (1 - f)p - c and the cost of overstocking is Cu = c - sR.

Expected manufacturer's profits = (c - v) o* + fp(o* - expected overstock at retailer)

9. How do firms analyze its procurement spending and supplier performance? (K1,CO5)

Procurement spending should be analyzed by part and supplier to ensure appropriate economies of scale. Supplier performance analysis should be used to build a portfolio of suppliers with complementary strengths.

10. Define Supply Chain coordination? (K1,CO5)

Supply chain coordination occurs when all stages of a supply chain work toward the objective of maximizing total supply chain profitability based on shared information. Lack of coordination can result in a significant loss of supply chain surplus. Coordination among different stages in a supply chain requires each stage to share appropriate information with other stages.

11. What is the effect of lack of coordination? (K1,CO5)

A lack of coordination occurs either because different stages of the supply chain have objectives that conflict or because information moving between stages is delayed and distorted. Different stages of a supply chain may have conflicting objectives if each stage has a different owner. As a result, each stage tries to maximize its own profits, resulting in actions that often diminish total supply chain profits.

12. What is meant by bull whip effect? (K1,CO5)

The bullwhip effect occurs when changes in consumer demand causes the companies in a supply chain to order more goods to meet the new demand. The bullwhip effect usually flows up the supply chain, starting with the retailer, wholesaler, distributor, manufacturer and then the raw materials supplier. This effect can be observed through most supply chains across several industries; it occurs because the demand for goods is based on demand forecasts from companies, rather than actual consumer demand.

13. What are the various causes of the bull whip effect? (K1,CO5)

The most common causes of bull whip effect are Demand forecast, inflated orders due to lack of information sharing, long lead times, price fluctuations, order batching/large lot size.

14. What happens when the bullwhip effect hits the supply chain? (K1,CO5)

Too much Stock on hand, leading to increased inventory holding costs.
Unfulfilled orders
poor customer
service Lost
Revenue
Misguided demand
forecasts Missed
production schedules

15. How the manufacturing cost is affected by lack of coordination? (K1,CO5)

The lack of coordination increases manufacturing cost in the supply chain. Building excess capacity or holding excess inventory, both of which increase the manufacturing cost per unit produced.

16. How the inventory cost is affected by lack of coordination? (K1,CO5) The lack of coordination increases inventory cost in the supply chain. The high levels of inventory also increase the warehousing space required and thus the warehousing cost incurred.

17. How the replenishment time is affected by lack of coordination? (K1,CO5)

Lack of coordination increases replenishment lead times in the supply chain. The increased variability as a result of the bullwhip effect makes scheduling at P&G and supplier plants much more difficult than when demand is level. There are times when the available capacity and inventory cannot supply the orders coming in. This results in higher replenishment lead times.

18. How the transportation cost is affected by lack of coordination?

(K1,CO5)

The lack of coordination increases transportation cost in the supply chain. The transportation requirements over time at P&G and its

suppliers are correlated with the orders being filled. As a result of the bullwhip effect, transportation requirements fluctuate significantly over time. This raises transportation cost because surplus transportation capacity needs to be maintained to cover high-demand periods.

19. How the labour cost is affected by lack of coordination? (K1,CO5)

The lack of coordination increases labor costs associated with shipping and receiving in the supply chain. Labor requirements for shipping at P&G and its suppliers fluctuate with orders. A similar fluctuation occurs for the labor requirements for receiving at distributors and retailers. The various stages have the option of carrying excess labor capacity or varying labor capacity in response to the fluctuation in orders. Either option increases total labor cost.

20. How the product availability is affected by lack of coordination? (K1,CO5)

Lack of coordination hurts the level of product availability and results in more stockouts in the supply chain. The large fluctuations in orders make it harder for P&G to supply all distributor and retailer orders on time. This increases the likelihood that retailers will run out of stock, resulting in lost sales for the supply chain.

21. How the performance of the supply chain is affected by lack of coordination? (K1,CO5)

Lack of coordination has a negative effect on performance at every stage and thus hurts the relationships among different stages of the supply chain. The lack of coordination thus leads to a loss of trust among different stages of the supply chain and makes any potential coordination efforts more difficult.

Lack of coordination has a significant negative impact on the supply chain's performance by increasing cost and decreasing responsiveness. The lack of coordination hurts both responsiveness and cost in a supply chain by making it more expensive to provide a given level of product availability.

22. List out the obstacles to coordination in a supply chain (K1,CO5)

Incentive obstacles

Information-processing
obstacles Operational
obstacles
Pricing obstacles
Behavioral
obstacles

Part-B Questions

Q. No.	Questions		
1	Explain the role of sourcing in a supply chain in details.		
2	How do third parties increase the supply chain surplus.		
3	Explain supplier scoring and assessment in details		
4	Explain supplier selection-auctions and negotiations		
5	Summarize contracts and supply chain performance in details.		
6	Describe supply chain coordination and the bullwhip effect, and their impact on supply chain performance.		
7	How is the building of strategic partnerships and trust valuable within a supply chain?		
8	What is the impact of lack of coordination on the performance of a supply chain?		
9	Explain the different obstacles to coordination in a supply chain.		

Table of Contents

Unit 5 SUPPLY CHAIN AND INFORMATION TECHNOLOGY

5.1. THE ROLE OF IT IN A SUPPLY CHAIN

Information is a key supply chain driver because it serves as the glue that allows the other supply chain drivers to work together with the goal of creating an integrated, coordinated supply chain. Information is crucial to supply chain performance because it provides the foundation on which supply chain processes execute transactions and managers make decisions. Without information, a manager cannot know what customers want, how much inventory is in stock, and when more product should be produced or shipped. information provides supply chain visibility, allowing managers to make decisions to improve the supply chain's performance.

IT consists of the hardware, software, and people throughout a supply chain that gather, analyze, and execute upon information. IT serves as the eyes and ears (and sometimes a portion of the brain) of management in a supply chain, capturing and analyzing the information necessary to make a good decision. For instance, an IT system at a PC manufacturer may show the finished goods inventory at different stages of the supply chain and also provide the optimal production plan and level of inventory based on demand and supply information.

Using IT systems to capture and analyze information can have a significant impact on a firm's performance. For example, a major manufacturer of computer workstations and servers found that most of its information on customer demand was not being used to set production schedules and inventory levels. The manufacturing group lacked this demand information, which essentially forced it to make inventory and production decisions blindly. By installing a supply chain software system, the company was able to gather and analyze demand data to produce recommended stocking levels.

Using the IT system enabled the company to cut its inventory in half, because managers could now make decisions based on customer demand information rather than manufacturing's educated guesses.

Availability and analysis of information to drive decision making is a key to the success of a supply chain. Companies that have built their success on the availability and analysis of information include Seven-Eleven Japan, Walmart, Amazon, UPS, and Netflix. To support effective supply chain decisions, information must have the following characteristics:

1. **Information must be accurate**. Without information that gives a true picture of the state of the supply chain, it is difficult to make good decisions. That is not to say that all information must be 100 percent correct, but rather that the data available paint a picture that is at least directionally correct.

2. **Information must be accessible in a timely manner**. Accurate information often exists, but by the time it is available, it is either out of date or it is not in an accessible form. To make good decisions, a manager needs to have up-to-date information that is easily accessible.

3. **Information must be of the right kind.** Decision makers need information that they can use. Often companies have large amounts of data that are not helpful in making a decision. Companies must think about what information should be recorded so that valuable resources are not wasted collecting meaningless data while important data go unrecorded.

4. **Information must be shared.** A supply chain can be effective only if all its stakeholders share a common view of the information that they use to make business decisions. Different information with different stakeholders results in misaligned action plans that hurt supply chain performance.

Information is used when making a wide variety of decisions about each supply chain driver, as discussed next.

1. ***Facility. Determining the location, capacity, and schedules of a facility requires information*** on the trade-offs among efficiency and flexibility, demand, exchange rates, taxes, and so on (see Chapters 4, 5, and 6). Walmart's suppliers use the demand information from Walmart's stores to set their production schedules. Walmart uses demand information to determine where to place its new stores and cross-docking facilities.

2. **Inventory.** Setting optimal inventory policies requires information that includes demand patterns, cost of carrying inventory, costs of stocking out, and costs of ordering (see Chapters 11, 12, and 13). For example, Walmart collects detailed demand, cost, margin, and supplier information to make these inventory policy decisions.

3. **Transportation.** Deciding on transportation networks, routings, modes, shipments, and vendors requires information about costs, customer locations, and shipment sizes to make good decisions (see Chapter 14). Walmart uses information to tightly integrate its operations with those of its suppliers. This integration allows Walmart to implement cross-docking in its transportation network, saving on both inventory and transportation costs.

4. **Sourcing.** Information on product margins, prices, quality, delivery lead times, and so on, are all important in making sourcing decisions. Given sourcing deals with inter-enterprise transactions, a wide range of transactional information must be recorded in order to execute operations, even once sourcing decisions have been made.

5. **Pricing and revenue management.** To set pricing policies, one needs information on demand, both its volume and various customer segments' willingness to pay, and on many supply issues, such as the product margin, lead time, and availability. Using this information, firms can make intelligent pricing decisions to improve their supply chain profitability.

Information is crucial to making good supply chain decisions at all three levels of decision making (strategy, planning, and operations) and in each of the other supply chain drivers (facilities, inventory, transportation, sourcing, and pricing). IT enables not only the gathering of these data to create supply chain visibility, but also the analysis of these data so that the supply chain decisions made will maximize profitability.

5.2. THE SUPPLY CHAIN IT FRAMEWORK

IT provides access and reporting of supply chain transaction data. More advanced IT systems then layer on a level of analytics that uses transaction data to proactively improve supply chain performance. Good IT systems will record and report demand, inventory, and fulfillment information for Amazon. IT systems that provide analytics then allow Amazon to decide whether to open new distribution centers and how to stock them.

The Supply Chain Macro Processes

The emergence of supply chain management has broadened the scope across which companies make decisions. This scope has expanded from trying to optimize performance across the division, to the enterprise, and now to the entire supply chain. This broadening of scope emphasizes the importance of including processes all along the supply chain when making decisions. From an enterprise's perspective, all processes within its supply chain can be categorized into three main areas:

- Processes focused downstream,
- Processes focused internally
- Processes focused upstream.

We use this classification to define the three macro supply chain as follows:

- **Customer relationship management (CRM).** Processes that focus on downstream interactions between the enterprise and its customers.
- **Internal supply chain management (ISCM).** Processes that focus on internal operations within the enterprise. Note that the software industry commonly calls this ―supply chain management‖ (without the word ―internal‖), even though the focus is entirely within the enterprise. In our definition, supply chain management includes all three macro processes, CRM, ISCM, and SRM.
- **Supplier relationship management (SRM).** Processes that focus on upstream interactions between the enterprise and its suppliers.

All operation and analytics related to the macro processes rest on the transaction management foundation (TMF), which includes basic enterprise resource planning (ERP) systems (and its components, such as financials and human resources), infrastructure software, and integration software. TMF software is necessary for the three macro processes to function and to communicate with one another. The relationship between the three macro processes and the transaction management foundation can be seen in Figure 5.1.

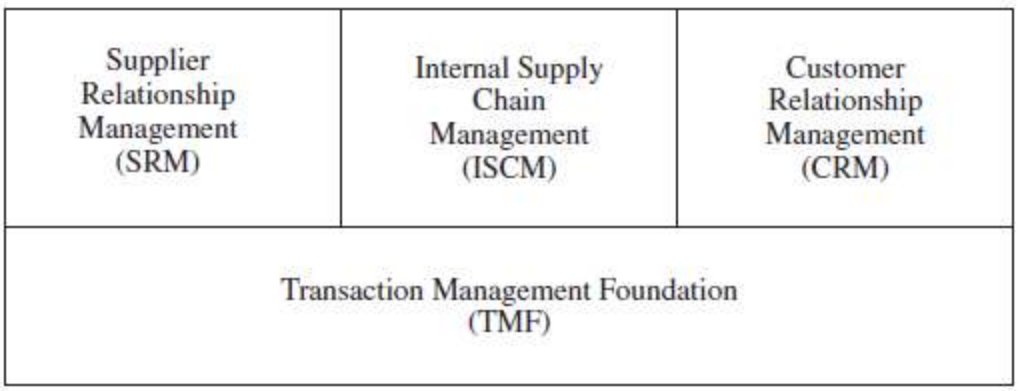


Figure 5.1. The Macro Processes in a Supply Chain

Why Focus on the Macro Processes?

As the performance of an enterprise becomes more closely linked to the performance of its supply chain, it is crucial that firms focus on these macro processes. ood supply chain management is not a zero-sum game in which one stage of the supply chain increases profits at the expense of another. Good supply chain management instead attempts to grow the supply chain surplus, which requires each firm to expand the scope beyond internal processes and look at the entire supply chain in terms of the three macro processes to achieve breakthrough performance. A good supply chain coordinates all the macro processes across all stages.

Apple is an example of a company that has coordinated all macro processes to introduce and sell blockbuster products such as the iPad2. Apple has been very successful in its interactions with customers both in designing products that meet their needs but also in operating Apple retail as a successful and profitable endeavor. All its products are designed in-house but manufactured by a third party. Despite this, Apple managed the release of the iPad2 to effectively meet huge demand. Strong coordination across all the macro processes has been fundamental for the level of success achieved by Apple.

5.3 Customer Relationship Management

The CRM macro process consists of processes that take place between an enterprise and its customers downstream in the supply chain. The goal of the CRM macro process is to generate customer demand and facilitate transmission and tracking of orders. Weakness in this process results in demand being lost and a poor customer experience because orders are not processed and executed effectively. The key processes under CRM are as follows:

- **Marketing.** Marketing processes involve decisions regarding which customers to target, how to target customers, what products to offer, how to price products, and how to manage the actual campaigns that target customers.

Good IT systems in the marketing area within CRM provide analytics that improve the marketing decisions on pricing, product profitability, and customer profitability, among other functions.

- **Sell.** The sell process focuses on making an actual sale to a customer. The sell process includes providing the sales force with the information it needs to make a sale and then to execute the actual sale. Executing the sale may require the salesperson to build and configure orders by choosing among a variety of options and features. The sell process also requires such functionality as the ability to quote due dates and access information related to a customer order.

Good IT systems support sales force automation, configuration, and personalization
to improve the sell process.

- **Order management.** The process of managing customer orders as they flow through an enterprise is important for the customer to track his order and for the enterprise to plan and execute order fulfillment. This process ties together demand from the customer with supply from the enterprise.

Good IT systems enable visibility of orders across the various stages that an order flows through before reaching the customer.

- **Call/service center.** A call/service center is often the primary point of contact between a company and its customers. A call/service center helps customers place orders, suggests products, solves problems, and provides information on order status.

Good IT systems have helped improve call/service center operations by facilitating and reducing work done by customer service representatives and by routing customers to representatives who are best suited to service their request.

Amazon has done an excellent job of using IT to enhance its CRM process. The company customizes the products presented to suit the individual customer (based on an analysis of customer preferences from past history and current clicks). Quick ordering is facilitated by systems that allow one-click orders. The order is then visible to the customer until it is delivered. In the rare instances when a customer uses the call center, systems are in place to support a positive experience including offering a callback in case the call center is heavily loaded.

The five largest CRM software providers in 2008 (as reported by Gartner) were SAP (22.5 percent), Oracle (16.1 percent), Salesforce.com (10.6 percent), Microsoft (6.4 percent), and Amdocs (4.9 percent).

5.4. Internal Supply Chain Management

ISCM is focused on operations internal to the enterprise. ISCM includes all processes involved in planning for and fulfilling a customer order. The various processes included in ISCM are as follows.

- **Strategic planning**. This process focuses on the network design of the supply chain.

- **Demand planning.** Demand planning consists of forecasting demand and analyzing the impact on demand of demand management tools such as pricing and promotions.

- **Supply planning**. The supply planning process takes as an input the demand forecasts produced by demand planning and the resources made available by strategic planning, and then produces an optimal plan to meet this demand. Factory planning and inventory planning capabilities are typically provided by supply planning software.

- **Fulfillment**. Once a plan is in place to supply the demand, it must be executed. The fulfillment process links each order to a specific supply source and means of transportation. The software applications that typically fall into the fulfillment segment are transportation and warehousing applications.

- **Field service.** Finally, after the product has been delivered to the customer, it eventually must be serviced. Service processes focus on setting inventory levels for spare parts as well as scheduling service calls. Some of the scheduling issues here are handled in a similar manner to aggregate planning, and the inventory issues are the typical inventory management problems.

Given that the ISCM macro process aims to fulfill demand that is generated by CRM processes, there needs to be strong integration between the ISCM and CRM macro processes. When forecasting demand, interaction with CRM is essential, as the CRM applications are touching the customer and have the most data and insight on customer behavior. Similarly, the ISCM processes should have strong integration with the SRM macro process. Supply planning, fulfillment, and field service are all dependent on suppliers and therefore the SRM processes.

It is of little use for your factory to have the production capacity to meet demand if your supplier cannot supply the parts to make your product. Order management, which we discussed under CRM, must integrate closely with fulfillment and be an input for effective demand planning. Again, extended supply
chain management requires that we integrate across the macro processes.

Successful ISCM software providers have helped improve decision making within ISCM processes. Good integration with CRM and SRM, however, is still largely inadequate at both the organizational and software levels. Future opportunities are likely to arise partly in improving each ISCM process, but even more so in improving integration with CRM and SRM.

Like CRM, today's ISCM landscape consists of three categories-the former best-of-breed winners, the best-of-breed start-ups, and the ERP players. The ERP players dominate this segment, although this was not always the case. There were two best-of-breed winners, i2 Technologies and Manugistics, which were ISCM pioneers and basically built the category. They showed the power of IT in supply chain management. However, they grew too quickly and spread their product lines across too many products, causing them to lose focus. This allowed the ERP players to improve their functionality relative to the best-of-breed players and eventually to take the supply chain leadership role away from them. Today, these ERP players are the only large players in supply chain IT. Manugistics, i2, and some start-ups still exist, but the landscape looks to be dominated by SAP and Oracle for the foreseeable future.

5.5 Supplier Relationship Management

Supplier Relationship Management SRM includes the processes that are focused on the interaction follows:

- **Design collaboration:** Between the enterprise and suppliers that are upstream in the supply chain. There is a natural fit between SRM processes and the ISCM processes, as integrating supplier constraints is crucial when creating internal plans. The major SRM processes are as

This software aims to improve the design of products through collaboration between manufacturers and suppliers. The software facilitates the joint selection (with suppliers) of components that have positive supply chain characteristics such as ease of manufacturability or commonality across several end products. Other design collaboration activities include the sharing of engineering change orders between a manufacturer and its suppliers. This eliminates the costly delays that occur when several suppliers are designing components for the manufacturer's product concurrently.

- **Source:**

Sourcing software assists in the qualification of suppliers and helps in supplier selection, contract management, and supplier evaluation. An important objective is to analyze the amount that an enterprise spends with each supplier, often revealing valuable trends or areas for improvement. Suppliers are evaluated along several key criteria, including lead time, reliability, quality, and price. This evaluation helps improve supplier performance and aids in supplier selection. Contract management is also an important part of sourcing, as many supplier contracts have complex details that must be tracked (such as volume-related price reductions). Successful software in this area helps analyze supplier performance and manage contracts.

- **Negotiate:**

Negotiations with suppliers involve many steps, starting with a request for quote (RFQ). The negotiation process may also include the design and execution of auctions. The goal of this process is to negotiate an effective contract that specifies price and delivery parameters for a supplier in a way that best matches the enterprise's needs. Successful software automates the RFQ process and the execution of auctions.

- **Buy:**

—Buy‖ software executes the actual procurement of material from suppliers. This includes the creation, management, and approval of purchase orders. Successful software in this area automates the procurement process and helps decrease processing cost and time.

- **Supply collaboration:**

Once an agreement for supply is established between the enterprise and a supplier, supply chain performance can be improved by collaborating on forecasts, production plans, and inventory levels. The goal of collaboration is to ensure a common plan across the supply chain. Good software in this area should be able to facilitate collaborative forecasting and planning in a supply chain.

Significant improvement in supply chain performance can be achieved if SRM processes are well integrated with appropriate CRM and ISCM processes. For instance, when designing a product, incorporating input from customers is a natural way to improve the design. This requires inputs from processes within CRM. Sourcing, negotiating, buying, and collaborating tie primarily into ISCM, as the supplier inputs are needed to produce and execute an optimal plan. However, even these segments need to interface with CRM processes such as order management. Again, the theme of integrating the three macro processes is crucial for improved supply chain

performance.

The SRM space is highly fragmented in terms of software providers and not as well defined as CRM and ISCM. Among the larger players, SAP and Oracle have SRM functionality in their software. There are many niche players, however, who focus on different aspects of SRM. All three macro processes and their processes can be seen in Figure 5.1.

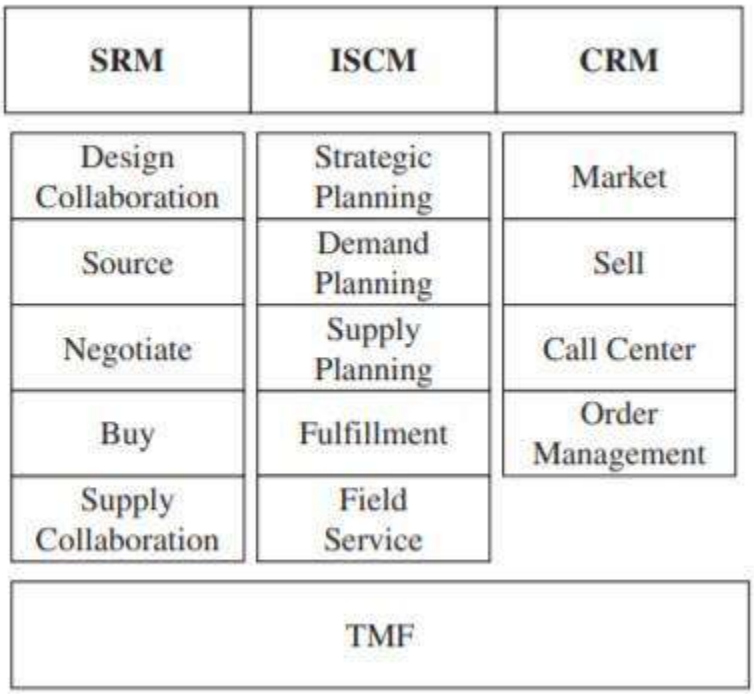


Fig. 5.1: The Macro Processes and Their Processes

5.6 The Future of IT in the Supply Chain

At the highest level, we believe that the three SCM macro processes will continue to drive the evolution of supply chain IT. While there is still plenty of room to improve the visibility and reporting of supply chain information, the relative focus on improved analysis to support decision making will continue to grow.

The following three important trends will impact IT in the supply chain:

1. The growth in software as a service (SaaS)
2. Increased availability of real-time data
3. Increased use of mobile technology

SaaS is defined as software that is owned, delivered, and managed remotely. Salesforce.com is one of the best-known pure SaaS supply chain software providers (in CRM). Gartner has predicted that SaaS (which comprised about 10 percent of the enterprise software market in 2009) will grow to about 16 percent of global software sales by 2014. This shift is likely to occur because SaaS provides lower startup and maintenance costs compared to applications that are deployed onsite. These factors are particularly important for small and midsized companies. Traditional enterprise software vendors such as SAP, Oracle, and Microsoft are increasing the availability of their software using the SaaS model.

The availability of real-time information has exploded in most supply chains. Whereas current supply chain software is primarily focused on improving strategy and planning decisions (often at the corporate level) that are revisited infrequently, significant opportunity exists to devise software that will use real-time information to help frontline supply chain staff (such as in transportation and warehousing) make smarter and faster decisions that are revisited frequently. The opportunity is to design systems that enable rapid insight based on real-time data.

The increased use of mobile technology coupled with real-time information offers some supply chains an opportunity to better match demand to supply using differential pricing. An example is an initiative by Groupon titled Groupon Now, which offers mobile users deals that are time and location specific. Businesses can improve profitability by offering deals when business is slow at specific locations. Consumers benefit from getting a deal when and where they want it. Such an approach is likely to be applicable in

many supply chain settings.

5.7 E- Business in Supply Chain

E-Business has emerged as a key enabler to drive **supply chain** integration. **Businesses** can use the Internet to gain global visibility across their extended network of trading partners and help them respond quickly to changing customer demand captured over the Internet.

E-business can be loosely defined as a business process that uses the Internet or other electronic medium as a channel to complete business transactions. As classified by Geoffrion and Krishnan (2001), e-business consists of three areas:
(1) consumeroriented activity and (2) business-oriented activity supported by (3) the e-business technology infrastructure. The consumer-oriented activities consist of business-to-consumer, consumer-to-consumer, and government-to-consumer activities. The business oriented activities comprise business-to-business, business- to-government, and government-to-business activities. The technology infrastructure relates to network infrastructure, network applications, decision technologies, and software tools and applications. Within this broad definition of e-business activities, we will restrict our attention mainly to consumer oriented and business-oriented activities, as well as decision technologies that are employed for supply chain management.

The Internet has influenced the usage of supply chain models in three ways. First, the Internet has facilitated increased use of enterprise resource planning (ERP) and advanced planning and optimization solutions (APS). Second, the ability to obtain real time information and the access to large computer systems is enabling firms to develop detailed (high granularity) supply chain models that can be utilized to make real-time decisions. Last, the Internet has created opportunities to integrate information and decision making across different functional units, thereby creating a need for supply chain models that go beyond a business unit to study the extended

enterprise.

This has elevated the role of supply chain models from being decision- making enablers for a single business unit to being enablers for driving corporate strategy. Thus, the Internet has greatly elevated the role of supply chain modeling and analysis within a firm. The advent of e-business has also created several challenges and opportunities in the supply chain environment. First and foremost, the Internet has increased the opportunity for consumers to buy products and services without going to a store. Though the practice of direct selling through catalogs and phone was in use earlier by a few firms, the Internet has made this form of sales more significant.

In a direct sales environment, the fulfillment process determines how long customers will wait between sale and delivery. This has made the back-end fulfillment process—which mostly depends on supply chain management—extremely important. Further, in the electronic environment, customer expectations in terms of quick and timely delivery have also increased. At the same time, the Internet has opened up opportunities for firms to share information and efficiently coordinate their activities with other entities in the supply chain.

This has created several new avenues in traditional supply chain areas. For example, in supplier selection and procurement, firms have to decide if they should join private or public exchanges or develop highly-integrated supply partnerships. They need to determine if they should use auction and bidding for contracts and, if so, which type would be most beneficial. In distribution, they need to decide if the firm will offer products through the Internet channel and, if so, how this method would differ from the traditional channel. This raises the question of how the synergies would be realized in terms of inventory, transportation, and distribution.

Similarly, the availability of real-time information has raised important questions such as the degree to which information sharing protocol should be standard or proprietary; the amount and type of information that should be shared with the rest of the supply chain partners; and the types of collaborative processes that may be beneficial. The degree of change in issues related to the supply chain
spans a huge spectrum from concepts and issues that have been marginally affected to a whole set of new issues that have emerged as a result of e-business. First, several issues related to supply chain management have not necessarily changed in principle, although e-business may have had an impact on some of their parameters. For example, to maintain given levels of service, a firm still needs buffer inventory or buffer capacity. This has not changed as a result of the Internet, although the uncertainty involved in the decision making may have decreased with the availability of more information. Similarly, a firm still needs to take into account the interplay between fixed and variable costs, while making decisions related to procurement or setting up additional capacity. With the prevalence of the Internet, the firm might more easily be able to obtain a lower procurement price or salvage excess capacity through market mechanisms. Next are existing supply chain issues that have become important as a result of e-business.

For example, leveraging risk-pooling concepts can greatly benefit Internet channels because products may be stored at fewer locations as compared to a traditional distribution channel. Amazon.com can store all inventory for the entire U.S. market in five warehouses as opposed to several hundred retail outlets (hence, stocking points) that would be needed for similar coverage in the traditional channel. Similarly, mass customization has gained a lot of momentum with the Internet because firms can allow customers to interactively specify customizations of their offerings.

It has become more important for firms to understand how to cope with customization in an effective manner. Finally, in the last few years, a third category of issues new to supply chain management has emerged. One example is linking the dynamic pricing of products to the inventory and capacity decisions. Another is coordinating Internet and traditional distribution channels in terms of prices as well as information and product flows. Additionally, the advent of electronic marketplaces and auctions has opened a whole new set of issues related to procurement and supplier relationships.

Part-A Questions and Answers

1. How Information is helpful in supply chain? (K1, CO6)

Information is crucial to supply chain performance because it provides the foundation on which supply chain processes execute transactions and managers make decisions.

2. List out characteristics of information. (K1, CO6)

- Information must be accurate.
- Information must be accessible in a timely manner.
- Information must be of the right kind
- Information must be shared.

3. How information is used in making decision on supply chain driver facility ? (K1, CO6)

Determining the location, capacity, and schedules of a facility requires information on the trade-offs among efficiency and flexibility, demand, exchange rates, taxes, and so on. Walmart's suppliers use the demand information from Walmart's stores to set their production schedules. Walmart uses demand information to determine where to place its new stores and cross-docking facilities.

4. How information is used in making decision on supply chain driver inventory ? (K1, CO6)

Setting optimal inventory policies requires information that includes demand patterns, cost of carrying inventory, costs of stocking out, and costs of ordering.

5. How information is used in making decision on supply chain driver transport ? (K1, CO6)

Deciding on transportation networks, routings, modes, shipments, and vendors requires information about costs, customer locations, and shipment sizes to make good decisions.

6. How information is used in making decision on supply chain driver sourcing? (K1, CO6)

Information on product margins, prices, quality, delivery lead times, and so on, are all important in making sourcing decisions. Given sourcing deals with inter-enterprise transactions, a wide range of transactional information must be recorded in order to execute operations, even once sourcing decisions have been made.

7. What are the three macro supply chain process? (K1, CO6)

- Customer relationship management (CRM).
- Internal supply chain management (ISCM).
- Supplier relationship management (SRM).

8. What is customer relationship management ? (K1, CO6)

Processes that focus on downstream interactions between the enterprise and its customers.

9. What is internal supply chain management ? (K1, CO6)

Processes that focus on internal operations within the enterprise. Note that the software industry commonly calls this ―supply chain management‖, even though the focus is entirely within the enterprise. In our definition, supply chain management includes all three macro processes, CRM, ISCM, and SRM.

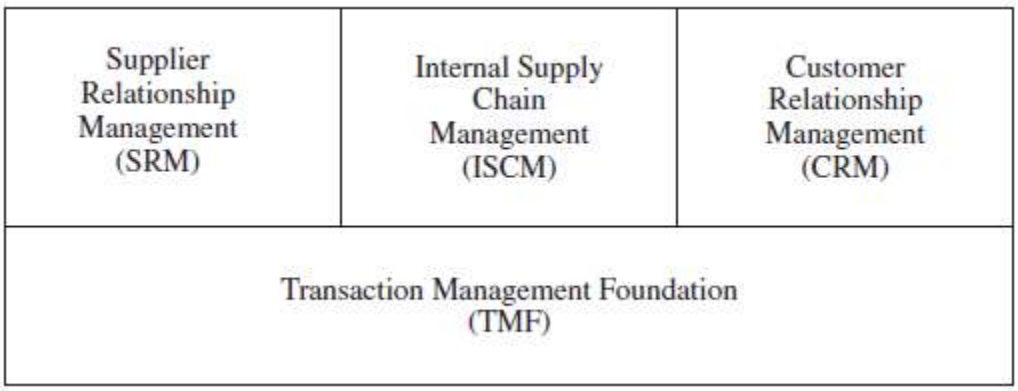
10. What is supplier relationship management ? (K1, CO6)

Supplier relationship management (SRM). Processes that focus on upstream in teractions between the enterprise and its suppliers.

11. What is Transaction Management Foundation? (K1, CO6)

All operation and analytics related to the macro processes rest on the transaction management foundation (TMF), which includes basic enterprise resource planning (ERP) systems (and its components, such as financials and human resources), infrastructure software, and integration software. TMF software is necessary for the three macro processes to function and to communicate with one another.

12. Draw the relationship between the three macro processes and the transaction management foundation (K1, CO6)



13. How the function of three macro process affects the supply chain performance ? (K1, CO6)

Good supply chain management instead attempts to grow the supply chain surplus, which requires each firm to expand the scope beyond internal processes and look at the entire supply chain in terms of the three macro processes to achieve breakthrough performance. A good supply chain coordinates all the macro processes across all stages.

12. List out various processes included in ISCM (K1, CO6)

- Strategic planning.
- Demand planning.
- Supply planning.
- Fulfillment.

- Field service.

13. What is Strategic planning ? (K1, CO6)

This process focuses on the network design of the supply chain. Key decisions include location and capacity planning of facilities.

14. What is Demand planning? (K1, CO6)

Demand planning consists of forecasting demand and analyzing the impact on demand of demand management tools such as pricing and promotions.

15. What is Supply planning ? (K1, CO6)

The supply planning process takes as an input the demand forecasts produced by demand planning and the resources made available by strategic planning; then it produces an optimal plan to meet this demand. Factory planning and inventory planning capabilities are typically provided by supply planning software.

16. What is Fulfillment ? (K1, CO6)

Once a plan is in place to supply the demand, it must be executed. The fulfillment process links each order to a specific supply source and means of transportation. The software applications that typically fall into the fulfillment segment are transportation and warehousing management applications.

17. What is Field service ? (K1, CO6)

After the product has been delivered to the customer, it eventually must be serviced. Service processes focus on setting inventory levels for spare parts as well as scheduling service calls. Some of the scheduling issues here are handled in a similar manner to aggregate planning, and the inventory issues are the typical inventory management problems.

18. What is Supplier Relationship Management? (K1, CO6)

Ans. Supplier Relationship Management SRM includes the processes that are focused on the interaction between the enterprise and suppliers that are upstream in the supply chain.

19. List the major SRM processes. (K1, CO6)

Ans. The major SRM processes are as follows:

- Design collaboration
- Source
- Negotiate
- Buy
- Supply collaboration

20. What are the three important trends that impact IT in the supply chain? (K1, CO6)

Ans. The following three important trends will impact IT in the supply chain:

- The growth in software as a service (SaaS)
- Increased availability of real-time data
- Increased use of mobile technology

21. How the Suppliers are evaluated? List them. (K1, CO6)

Ans. Suppliers are evaluated based on following key criteria. They are Lead Time, Reliability, Quality and Price. This evaluation helps improve supplier performance and aids in supplier selection.

22. Define SaaS. (K1, CO6)

SaaS is defined as software that is owned, delivered, and managed remotely. Salesforce.com is one of the best-known pure SaaS supply chain software providers (in CRM). Gartner has predicted that SaaS (which comprised about 10 percent of the enterprise software market in 2009) will grow to about 16 percent of global software sales by 2014. This shift is likely to occur because SaaS provides lower startup and maintenance costs compared to applications that are deployed onsite. These factors are particularly important for small and midsized companies. Traditional enterprise software vendors such as SAP, Oracle, and Microsoft are increasing the availability of their software using the SaaS model.

23. What is e business in supply chain? (K1,CO6)

Businesses can use the Internet to gain global visibility across their extended network of trading partners and help them respond quickly to changing customer demand captured over the Internet.

24. Define Customer Relationship Management (CRM).(K1,CO6)

The CRM macro process consists of processes that take place between an enterprise and its customers downstream in the supply chain. The goal of the CRM macro process is to generate customer demand and facilitate transmission and tracking of orders. Weakness in this process results in demand being lost and a poor customer experience because orders are not processed and executed effectively.

25. Mention the key processes under CRM. (K1,CO6)

The Key Processes under CRM are (i) Marketing (ii) Sell (iii) Order Management (iv) Call/Service Center.

26. Mention the largest CRM Software providers. (K1,CO6)

The five largest CRM software providers are SAP , Oracle , Salesforce.com,

Microsoft, and Amdocs.

Part-B Questions

Q. No.	Questions		
1	What is Supplier Relationship Management? Explain the major SRM processes in detail with diagram.		
2	Explain in detail about the important trends that will impact IT in the supply chain.		
3	Explain in detail about E business in supply chain.		
4	Discuss in detail about customer Relationship Management.		
5	Explain the role of IT in Supply Chain Management		

MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

EC8702 AD HOC AND WIRELESS SENSOR NETWORKS

Semester - 07

Notes

MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

Vision

To excel in providing value based education in the field of Electronics and Communication Engineering, keeping in pace with the latest technical developments through commendable research, to raise the intellectual competence to match global standards and to make significant contributions to the society upholding the ethical standards.

Mission

- ✓ To deliver Quality Technical Education, with an equal emphasis on theoretical and practical aspects.
- ✓ To provide state of the art infrastructure for the students and faculty to upgrade their skills and knowledge.
- ✓ To create an open and conducive environment for faculty and students to carry out research and excel in their field of specialization.
- ✓ To focus especially on innovation and development of technologies that is sustainable and inclusive, and thus benefits all sections of the society.
- ✓ To establish a strong Industry Academic Collaboration for teaching and research, that could foster entrepreneurship and innovation in knowledge exchange.
- ✓ To produce quality Engineers who uphold and advance the integrity, honour and dignity of the engineering.

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

1. To provide the students with a strong foundation in the required sciences in order to pursue studies in Electronics and Communication Engineering.
2. To gain adequate knowledge to become good professional in electronic and communication engineering associated industries, higher education and research.
3. To develop attitude in lifelong learning, applying and adapting new ideas and technologies as their field evolves.
4. To prepare students to critically analyze existing literature in an area of specialization and ethically develop innovative and research oriented methodologies to solve the problems identified.
5. To inculcate in the students a professional and ethical attitude and an ability to visualize the engineering issues in a broader social context.

PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO1: Design, develop and analyze electronic systems through application of relevant electronics, mathematics and engineering principles.

PSO2: Design, develop and analyze communication systems through application of fundamentals from communication principles, signal processing, and RF System Design & Electromagnetics.

PSO3: Adapt to emerging electronics and communication technologies and develop innovative solutions for existing and newer problems.

Unit - I

AD HOC NETWORKS INTRODUCTION AND ROUTING PROTOCOLS

Elements of Ad hoc Wireless Networks, Issues in Ad hoc wireless networks, Example commercial applications of Ad hoc networking, Ad hoc wireless Internet, Issues in Designing a Routing Protocol for Ad Hoc Wireless Networks, Classifications of Routing Protocols, Table Driven Routing Protocols – Destination Sequenced Distance Vector (DSDV), On–Demand Routing protocols –Ad hoc On–Demand Distance Vector Routing (AODV).

TABLE OF CONTENTS

1.1 Introduction

1.1.1 Computer Networks

- A computer network is the interconnection of multiple nodes through links. A node can be computer, printer, or any other device capable of sending or receiving the data. The

links connecting the nodes are known as communication channels.

- The computer network uses distributed processing in which task is divided among several computers. Instead, a single computer handles an entire task, each separate computer handles a subset

Advantages of Distributed processing

- **Security:** It provides limited interaction that a user can have with the entire system. For example, a bank allows the users to access their own accounts through an ATM without allowing them to access the bank's entire database.
- **Faster problem solving:** Multiple computers can solve the problem faster than a single machine working alone.
- **Security through redundancy:** Multiple computers running the same program at the same time can provide the security through redundancy. For example, if four computers run the same program and any computer has a hardware error, then other computers can override it.

Applications of Distributed Systems

- E-mail
- Online Ticket Reservation
- Banking, etc.,

1.1.2 Types of Communication

- Communication medium refers to the physical channel through which data is sent and received. Data is sent in the form of voltage levels which make up the digital signal. A digital signal consists of 0s and 1s. There are basically two types of networks:
  - **Wired network**
  - **Wireless network**

Wired Network

- In a wired network, data is transmitted over a physical medium.
- There are three types of physical cables used in a wired network.
  - Twisted Pair
  - Coaxial Cable
  - Fiber Optic

  Examples: Cable TV, Broadband Telephone Communication.

Wireless Network

- A wireless network uses radio waves as the sole medium for transmitting and receiving data. There are no wires involved.
- Radio waves are electromagnetic waves which are transverse in nature and they have the longest wavelength on the electromagnetic spectrum.

  Examples: Infrared, Bluetooth, WiFi.

1.2 Elements of Ad hoc Wireless Networks

- The word "ad hoc" comes from Latin Language, which means 'for this purpose only', Ad hoc Networks are the small area networks, especially designed with Wireless/Temporary connections to the different computer assisted nodes.
- A wireless ad-hoc network (WANET) is a type of local area network (LAN) that is built spontaneously to enable two or more wireless devices to be connected to each other without requiring a central device, such as a router or access point. When Wi- Fi networks are in ad-hoc mode, each device in the network forwards data to the others.
- Since the devices in the ad-hoc network can access each other's resources directly through a basic point-to-point wireless connection, central servers are unnecessary for functions such as file shares or printers.
- In a wireless ad-hoc network, a collection of devices (or nodes) is responsible for network operations, such as routing, security, addressing and key management. Figure 1.1 shows, multi-hop wireless ad hoc networks, it defined as a collection of nodes that communicate with each other wirelessly by using radio signals with a shared common channel.

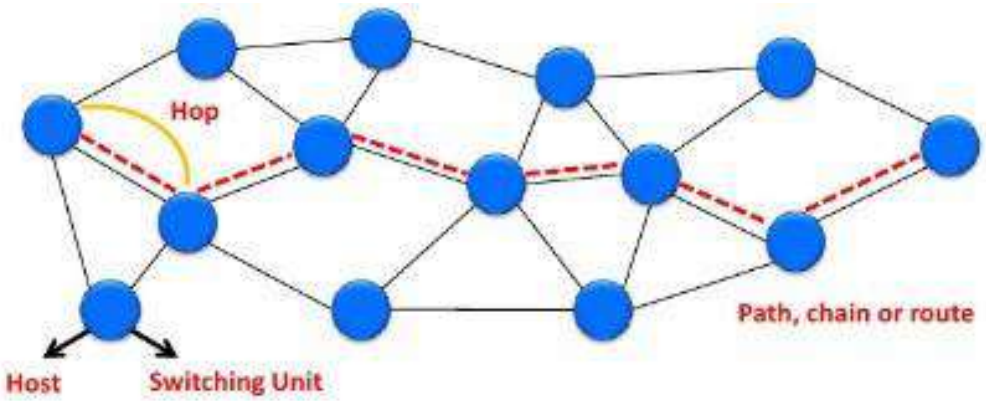


Figure 1.1 Multi- Hop Wireless Ad-Hoc Networks

Source : Ad Hoc Wireless Networks Architectures and Protocol by C. Siva Ram Murthy and B. S. Manoj

Types of Wireless Ad Hoc Networks

Wireless ad hoc networks are categorized into different classes. They are:

- **Mobile ad hoc network (MANET)**: An ad hoc network of mobile devices.
- **Vehicular ad hoc network (VANET)**: Used for communication between vehicles. Intelligent VANETs use artificial intelligence and ad hoc technologies to communicate what should happen during accidents.
- **Smartphone ad hoc network (SPAN)**: Wireless ad hoc network created on smartphones via existing technologies like Wi-Fi and Bluetooth.
- **Wireless mesh network**: A mesh network is an ad hoc network where the various nodes are in communication directly with each other to relay information throughout the total network.
- **Army tactical MENT**: Used in the army for "on-the-move" communication, a wireless tactical ad hoc network relies on range and instant operation to establish networks when needed.
- **Wireless sensor network**: Wireless sensors that collect everything from temperature and pressure readings to noise and humidity levels, can form an ad hoc network to deliver information to a home base without needing to connect directly to it.
- **Disaster rescue ad hoc network**: Ad hoc networks are important when disaster strikes and established communication hardware isn't functioning properly.

Advantages of Ad Hoc Networks

- Ad-hoc networks can have more flexibility.
- It is better in mobility.
- It can be turn up and turn down in a very short time.
- More economical
- It considered as a robust network because of its non-hierarchical distributed control and management mechanisms.

Disadvantages of Ad Hoc Networks

- Unpredictable Topology
- Limited Bandwidth
- Lose of data
- Interference
- Limited Security
- Energy Constraints

1.3 Issues in Ad hoc wireless networks

➢ The major issues that affect the design, deployment, and performance of an ad hoc wireless system are as follows:

- Medium Access Control (MAC)
- Routing
- Multicasting
- Transport layer protocol
- Quality of Service (QOS)
- Self-organization
- Security
- Energy management
- Addressing and service discovery
- Scalability
- Deployment considerations

1.3.1 Medium Access Control

➢ The purpose of this protocol is to achieve a distributed FIFO schedule among multiple nodes in an ad hoc network. When a node transmits a packet, it adds the information about the arrival time of queued packets. It provide fair access to shared broadcast radio channel. The major issues in MAC protocol are as follows:

- **Distributed Operation:** The MAC protocol design should be fully distributed involving minimum control overhead, because it need to operate in environment without centralized device.
- **Synchronization:** The synchronization is mandatory for TDMA-based systems for management of transmission and reception slots.
- **Hidden Terminals Problem:** Hidden terminals are nodes that are hidden (or not reachable) from the sender of a data transmission session, but are reachable to the receiver of the session. (Figure 1.2)

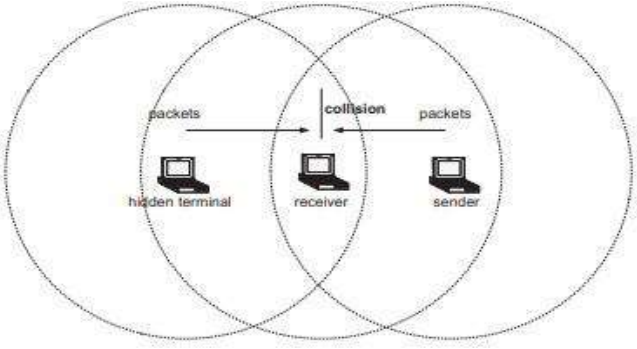


Figure 1.2 Hidden Terminal Problem

Source **:** Ad Hoc Wireless Networks Architectures and Protocol by C. Siva Ram Murthy and B. S. Manoj

- Collisions at receiver node -> inefficient bandwidth utilization, reduce throughput.

- **Exposed Terminals Problem:** The nodes that are in the transmission range of the sender of an on-going session, are prevented from making a transmission. The exposed nodes should be allowed to transmit in a controlled fashion without causing collision to the on-going data transfer. (Figure 1.3)

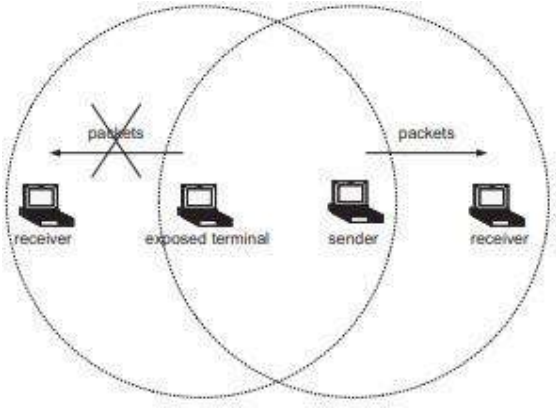


Figure 1.3 Exposed Terminal Problem

Source **:** Ad Hoc Wireless Networks Architectures and Protocol by C. Siva Ram Murthy and B. S. Manoj

- **Throughput:** The MAC protocol employed in ad hoc wireless networks should attempt to maximize the throughput of the system. The important considerations for throughput enhancement are

  - Minimizing the occurrence of collisions.
  - Maximizing channel utilization
  - Minimizing control overhead.

- **Access delay:** The average delay that any packet experiences to get transmitted. The MAC protocol should attempt to minimize the delay.

- **Fairness:** Fairness refers to the ability of the MAC protocol to provide an equal share or weighted share of the bandwidth to all competing nodes. Fairness can be either node-based or flow-based.

- **Real-time Traffic support:** In a contention-based channel access environment, without any central coordination, with limited bandwidth, and with location-dependent contention, supporting time- sensitive traffic such as voice, video, and

real-time data requires explicit support from the MAC protocol.

- **Resource reservation:** The provisioning of QoS defined by parameters such as bandwidth, delay, and jitter requires reservation of resources such as bandwidth, buffer space, and processing power.

- **Ability to measure resource availability:** In order to handle the resources such as bandwidth efficiently and perform call admission control based on their availability, the MAC protocol should be able to provide an estimation of resource availability at every node. This can also be used for making congestion control decisions.

- **Capability for power control:** The transmission power control reduces the energy consumption at the nodes, causes a decrease in interference at neighboring nodes, and increases frequency reuse.

- **Adaptive rate control:** This refers to the variation in the data bit rate achieved over a channel. A MAC protocol that has adaptive rate control can make use of a high data rate when the sender and receiver are nearby & adaptively reduce the data rate as they move away from each other.

1.3.2 Routing

➢ The responsibilities of a routing protocol include exchanging the route information; finding a feasible path to a destination. The major challenges that a routing protocol faces are as follows:

- **Mobility:** The Mobility of nodes results in frequent path breaks, packet collisions, transient loops, stale routing information, and difficulty in resource reservation.

- **Bandwidth constraint:** Since the channel is shared by all nodes in the broadcast region, the bandwidth available per wireless link depends on the number of nodes & traffic they handle.

- **Error-prone and shared channel:** The Bit Error Rate (BER) in a wireless channel is very high [10^{-5} to 10^{-3}] compared to that in its wired counterparts [10^{-12} to 10^{-9}].

- **Location-dependent contention:** The load on the wireless channel varies with the number of nodes present in a given geographical region. This makes the contention for the channel high when the number of nodes increases. The high contention for the channel results in a high number of collisions & a subsequent wastage of bandwidth.

- **Other resource constraints:** The constraints on resources such as computing power, battery power, and buffer storage also limit the capability of a routing protocol.

The major requirements of a routing protocol in ad hoc wireless networks are the following.

- **Minimum route acquisition delay**
- **Quick route reconfiguration**
- **Loop-free routing**
- **Distributed routing approach**
- **Minimum control overhead**
- **Scalability**
- **Provisioning of QoS**
- **Support for time-sensitive traffic**
- **Security and privacy**

1.3.3 Multicasting

➢ It plays important role in emergency search & rescue operations & in military communication. Use of single link connectivity among the nodes in a multicast group results in a tree-shaped multicast routing topology. Such a tree-shaped topology provides high multicast efficiency, with low packet delivery ratio due to the frequency tree breaks. The major issues in designing multicast routing protocols are as follows:

- **Robustness:** The multicast routing protocol must be able to recover & reconfigure quickly from potential mobility-induced link breaks thus making it suitable for use in high dynamic environments.
- **Efficiency:** A multicast protocol should make a minimum number of transmissions to deliver a data packet to all the group members.
- **Control overhead:** The scarce bandwidth availability in ad hoc wireless networks demands minimal control overhead for the multicast session.
- **Quality of Service:** QoS support is essential in multicast routing because, in most cases, the data transferred in a multicast session is time-sensitive.

- **Efficient group management:** Group management refers to the process of accepting multicast session members and maintaining the connectivity among them until the session expires.
- **Scalability:** The multicast routing protocol should be able to scale for a network with a large number of node
- **Security:** Authentication of session members and prevention of non-members from gaining unauthorized information play a major role in military communications.

1.3.4 Transport Layer Protocol

➢ The main objectives of the transport layer protocols include :

- Setting up & maintaining end-to-end connections,
- Reliable end-to-end delivery of packets,
- Flow control &
- Congestion control.

Examples of some transport layers protocols are,

a) UDP (User Datagram Protocol) :

- It is an unreliable connectionless transport layer protocol.
- It neither performs flow control & congestion control.
- It do not take into account the current network status such as congestion at the intermediate links, the rate of collision, or other similar factors affecting the network throughput.

b) TCP (Transmission Control Protocol):

- It is a reliable connection-oriented transport layer protocol.
- It performs flow control & congestion control.
- Here performance degradation arises due to frequent path breaks, presence of stale routing information, high channel error rate, and frequent network partitions.

1.3.5 Quality of Service (QoS)

➢ QoS is the performance level of services offered by a service provider or a network to the user.

- QoS provisioning often requires,
  - Negotiation between host & the network.
  - Resource reservation schemes.
  - Priority scheduling &
  - Call admission control.

- **QoS parameters**

Applications	Corresponding QoS parameter
1.Multimedia application	1. Bandwidth & Delay.
2.Military application	2.Security & Reliability.
3.Defense application	3.Finding trustworthy intermediate hosts & routing
4.Emergency search and rescue operations	4.Availability.
5.Hybrid wireless network	5.Maximum available link life, delay, bandwidth & channel utilization.
6.communication among the nodes in a sensor network	6.Minimum energy consumption, battery life & energy conservation

- **QoS-aware routing**
  - Finding the path is the first step toward a QoS-aware routing protocol.
  - The parameters that can be considered for routing decisions are,
    - Network throughput.
    - Packet delivery ratio.
    - Reliability.
    - Delay.
    - Delay jitter.
    - Packet loss rate.
    - Bit error rate.

1.3.6 Self-Organization

- One very important property that an ad hoc wireless network should exhibit is organizing & maintaining the network by itself.
- The major activities that an ad hoc wireless network is required to perform for self-organization are,

- Neighbour discovery.
- Topology organization &
- Topology reorganization (updating topology information)

1.3.7 Security

- Security is an important issue in ad hoc wireless network as the information can be hacked.
- Attacks against network are two types
  - Passive attack → Made by malicious node to obtain information transacted in the network without disrupting the operation.
  - Active attack → They disrupt the operation of network.
- Further active attacks are two types
  - External attack: The active attacks that are executed by nodes outside the network.
  - Internal attack: The active attacks that are performed by nodes belonging to the same network.
- The major security threats that exist in ad hoc wireless networks are as follows :
  - **Denial of service** – The attack affected by making the network resource unavailable for service to other nodes, either by consuming the bandwidth or by overloading the system.
  - **Resource consumption** – The scarce availability of resources in ad hoc wireless network makes it an easy target for internal attacks, particularly aiming at consuming resources available in the network. The major types of resource consumption attacks are,
    - Energy depletion
      - ✓ Highly constrained by the energy source
      - ✓ Aimed at depleting the battery power of critical nodes.
    - Buffer overflow
      - ✓ Carried out either by filling the routing table with unwanted routing entries or by consuming the data packet buffer space with unwanted data.

- ✓ Lead to a large number of data packets being dropped, leading to the loss of critical information.

- **Host impersonation** – A compromised internal node can act as another node and respond with appropriate control packets to create wrong route entries, and can terminate the traffic meant for the intended destination node.
- **Information disclosure** – A compromised node can act as an informer by deliberate disclosure of confidential information to unauthorized nodes.
- **Interference** – A common attack in defense applications to jam the wireless communication by creating a wide spectrum noise.

1.3.8 Addressing and Service Discovery

- Addressing & service discovery assume significance in ad hoc wireless network due to the absence of any centralised coordinator.
- An address that is globally unique in the connected part of the ad hoc wireless network is required for a node in order to participate in communication.
- Auto-configuration of addresses is required to allocate non-duplicate addresses to the nodes.

1.3.9 Energy Management

- Energy management is defined as the process of managing the sources & consumers of energy in a node or in the network for enhancing the lifetime of a network.
- Features of energy management are:
  - Shaping the energy discharge pattern of a node's battery to enhance battery life.
  - Finding routes that consumes minimum energy.
  - Using distributed scheduling schemes to improve battery life.
  - Handling the processor & interface devices to minimize power consumption.
- Energy management can be classified into the following categories:
  - **Transmission power management**
    - The power consumed by the Radio Frequency (RF) module of a mobile node is determined by several factors such as
      - ✓ The state of operation.
      - ✓ The transmission power and
      - ✓ The technology used for the RF circuitry.

- **Battery energy management**
  - The battery management is aimed at extending the battery life of a node by taking advantage of its chemical properties, discharge patterns, and by the selection of a battery from a set of batteries that is available for redundancy.
- **Processor power management**
  - The clock speed and the number of instructions executed per unit time are some of the processor parameters that affect power consumption.
  - **T**he CPU can be put into different power saving modes during low processing load conditions.
  - The CPU power can be completely turned off if the machines is idle for a long time.
- **Devices power management**
  - Intelligent device management can reduce power consumption of a mobile node significantly.
  - This can be done by the operating system (OS) by selectively powering down interface devices that are not used or by putting devices into different power saving modes, depending on their usage.

1.3.10 Scalability

- Scalability is the ability of the routing protocol to scale well in a network with a large number of nodes.
- It requires minimization of control overhead & adaptation of the routing protocol to the network size.

1.3.11 Deployment Considerations

- The deployment of a commercial ad hoc wireless network has the following benefits when compared to wired networks
  - **Low cost of deployment**
    - The use of multi-hop wireless relaying eliminates the requirement of cables & maintenance in deployment of communication infrastructure.
    - The cost involved is much lower than that of wired networks.

- **Incremental deployment**
  - Deployment can be performed incrementally over geographical regions of the city.
  - The deployed part of the network starts functioning immediately after the minimum configuration is done.
- **Short deployment time**
  - Compared to wired networks, the deployment time is considerably less due to the absence of any wired links.
- **Reconfigurability**
  - The cost involved in reconfiguring a wired network covering a Metropolitan Area Network (MAN) is very high compared to that of an ad hoc wireless network covering the same service area.

1.4 Commercial Applications of Ad Hoc Networking

➢ Ad Hoc wireless networks, due to their quick and economically less demanding deployment, find applications in several areas. Some important applications are:

- Military Applications
- Collaborative and Distributed computing
- Energy Operations
- Wireless Mesh Networks
- Wireless Sensor Networks
- Hybrid Wireless Networks

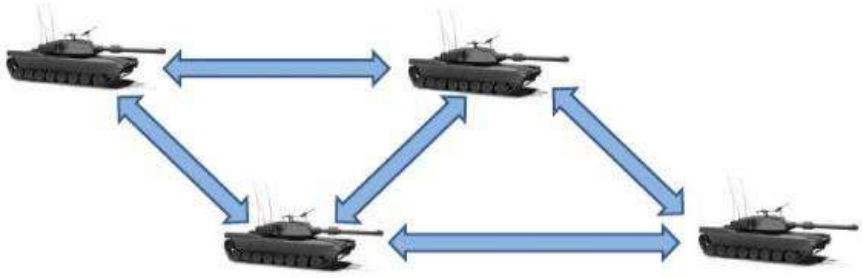
1.4.1 Military Applications

➢ Ad hoc wireless networks can be very useful in establishing communication among a group of soldiers for tactical operations.

➢ Setting up of a fixed infrastructure for communication among group of soldiers in enemy territories or in inhospitable terrains may not be possible.

➢ In such a case, adhoc wireless networks provide required communication mechanism quickly.

- The primary nature of the communication required in a military environment enforces certain important requirements on adhoc wireless networks namely, Reliability, Efficiency, Secure communication & Support for multicast routing.



Source **:** Ad Hoc Wireless Networks Architectures and Protocol by C. Siva Ram Murthy and B. S. Manoj

1.4.2 Collaborative & Distributed computing

- Adhoc wireless network helps in collaborative computing, by establishing temporary communication infrastructure for quick communication with minimal configuration among a group of people in a conference.
- In distributed file sharing application reliability is of high importance which would be provided by adhoc network.
- Other applications such as streaming of multimedia objects among participating nodes in ad hoc wireless networks require support for soft real-time communication
- Devices used for such applications could typically be laptops with add -on wireless interface cards, enhanced personal digital assistants (PDAs) or mobile devices with high processing power

1.4.3 Emergency Operations

- Ad hoc wireless networks are very useful in emergency operations such as search and rescue, crowd control and commando operations.
- The major factors that favour ad hoc wireless networks for such tasks are self-configuration of the system with minimal overhead, independent of fixed or centralised infrastructure, the freedom and flexibility of mobility, and unavailability of conventional communication infrastructure.

- In environments, where the conventional infrastructure based communication facilities are destroyed due to a war or due to natural calamities, immediate deployment of adhoc wireless networks would be a good solution for co-ordinating rescue activities.
- They require minimum initial network configuration with very little or no delay

1.4.4 Wireless Mesh Network

- Wireless mesh networks are adhoc wireless network that are formed to provide an alternate communication infrastructure for mobile or fixed nodes/users, without the spectrum reuse constraint & requirement of network planning of cellular network.(Figure 1.4)

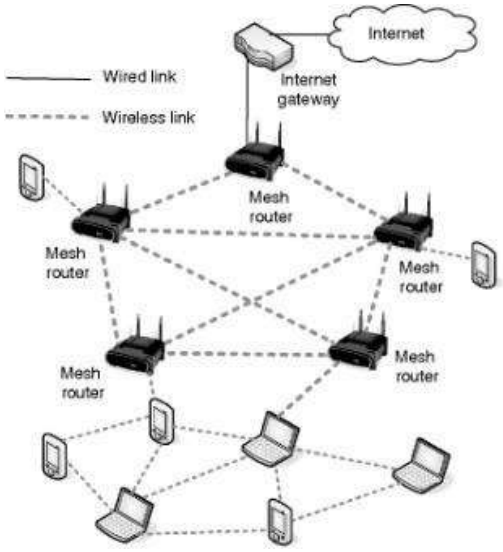


Figure 1.4 Wireless Mesh Netowrks

Source **:** Ad Hoc Wireless Networks Architectures and Protocol by C. Siva Ram Murthy and B. S. Manoj

- It provides many alternate paths for a data transfer session between a source & destination, resulting in quick reconfiguration of the path when the existing path fails due to node failure.
- Since the infrastructure built is in the form of small radio relaying devices, the investment required in wireless mesh networks is much less than what is required for the cellular network counterpart.
- The possible deployment scenarios of wireless mesh networks include: residential zones, highways, business zones, important civilian regions and university campuses
- Wireless mesh networks should be capable of self-organization and maintenance.
- It operates at license-free ISM band around 2.4 GHz & 5 GHz.

- It is scaled well to provide support to large number of points.
- Major advantage is the support for a high data rate, quick & low cost of deployment, enhanced services, high scalability, easy extendibility, high availability & low cost per bit.

1.4.5 Wireless Sensor Networks

- The Wireless Sensor Networks (WSN) are special category of Adhoc wireless network that are used to provide a wireless communication infrastructure among the sensors deployed in a specific application domain.(Figure 1.5)
- Sensor nodes are tiny devices that have capability of sensing physical parameters processing the data gathered, & communication to the monitoring system.

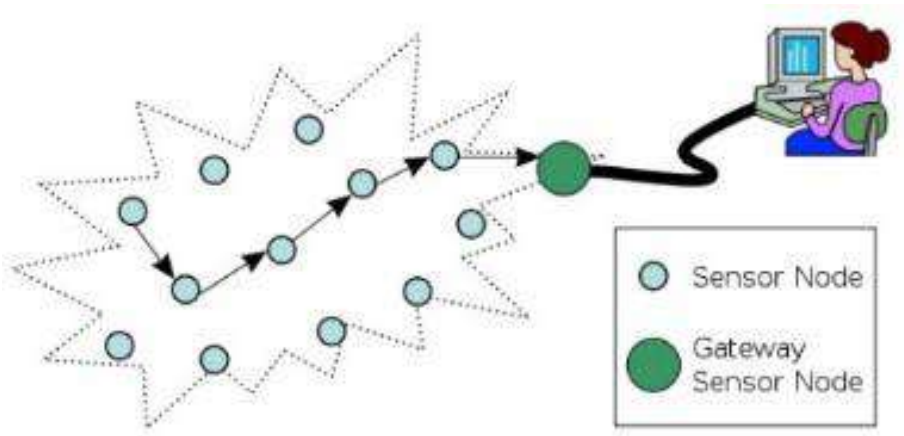


Figure 1.5 Wireless Sensor Networks

Source **:** Ad Hoc Wireless Networks Architectures and Protocol by C. Siva Ram Murthy and B. S. Manoj

- The issue that make sensor network a distinct category of adhoc wireless network are the following:

Mobility of nodes

- Mobility of nodes is not a mandatory requirement in sensor networks.
- For example, the nodes used for periodic monitoring of soil properties are not required to be mobile & the nodes that are fitted on the bodies of patients in a post-surgery ward of a hospital are designed to support limited or partial mobility.
- In general, sensor networks need not in all cases be designed to support mobility of sensor nodes.

Size of the network

- The number of nodes in sensor network can be much larger than that in a typical ad hoc wireless network.

Density of deployment

- The density of nodes in a sensor network varies with the domain of application.
- For example, Military applications require high availability of the network, making redundancy a high priority.

Power constraints

- The power constraints in sensor networks are much more stringent than those in ad hoc wireless networks. This is mainly because the sensor nodes are expected to operate in harsh environmental or geographical conditions, with minimum or no human supervision and maintenance.
- In certain case, the recharging of the energy source is impossible.
- Running such a network, with nodes powered by a battery source with limited energy, demands very efficient protocol at network, data link, and physical layer.
- The power sources used in sensor networks can be classified into the following 3 categories:
  - Replenishable Power source: The power source can be replaced when the existing source is fully drained.
  - Non-replenishable Power source: The power source cannot be replenished once the network has been deployed. The replacement of sensor node is the only solution.
  - Regenerative Power source: Here, Power source employed in sensor network have the capability of regenerating power from the physical parameter under measurement.

Data / Information fusion

- Data fusion refers to the aggregation of multiple packets into one before relaying it.

- Data fusion mainly aims at reducing the bandwidth consumed by redundant headers of the packets and reducing the media access delay involved in transmitting multiple packets.

- Information fusion aims at processing the sensed data at the intermediate nodes and relaying the outcome to the monitor node.

Traffic Distribution

- The communication traffic pattern varies with the domain of application in sensor networks.
- For example, the environmental sensing application generates short periodic packets indicating the status of the environmental parameter under observation to a central monitoring station.
- This kind of traffic requires low bandwidth.
- Ad hoc wireless networks generally carry user traffic such as digitized & packetized voice stream or data traffic, which demands higher bandwidth.

1.4.6 Hybrid Wireless Networks

- One of the major application area of ad hoc wireless network is in the hybrid wireless architecture such as Multi-hop Cellular Network [MCN] & Integrated Cellular Adhoc Relay [iCAR].
- The primary concept behind cellular networks is geographical channel reuse.
- Several techniques like cell sectoring, cell resizing and multi-tier cells increase the capacity of cellular networks.
- MCNs combine the reliability & support of fixed base station of cellular network with flexibility & multi - hop relaying adhoc wireless networks.
- Major advantages are:
  - Higher capacity than cellular networks due to the better channel reuse.
  - Increased flexibility & reliability in routing.
  - Better coverage & connectivity in holes of a cell can be provided by means of multiple hops through intermediate nodes in a cell.

1.5 Ad hoc wireless Internet

- Ad hoc wireless internet extends the services of the internet to the end users over an ad hoc wireless network. It shows in figure 1.6.
- Some of the applications of ad hoc wireless internet are :
  - Wireless mesh network.
  - Provisioning of temporary internet services to major conference venues.
  - Sports venues.
  - Temporary military settlements.
  - Battlefields
  - Broadband internet services in rural regions.
- The major issues to be considered for a successful ad hoc wireless internet are the following :
  - **Gateway**
    - They are the entry points to the wired internet.
    - Generally owned & operated by a service provider.
    - They perform following tasks ,
      - ✓ Keeping track of end users.
      - ✓ Bandwidth management.
      - ✓ Load balancing.
      - ✓ Traffic shaping.
      - ✓ Packet filtering.
      - ✓ Width fairness &
      - ✓ Address, service & location discovery.
  - **Address mobility**
    - This problem is worse here as the nodes operate over multiple wireless hops.
    - Solution such as Mobile IP can provide temporary alternative.

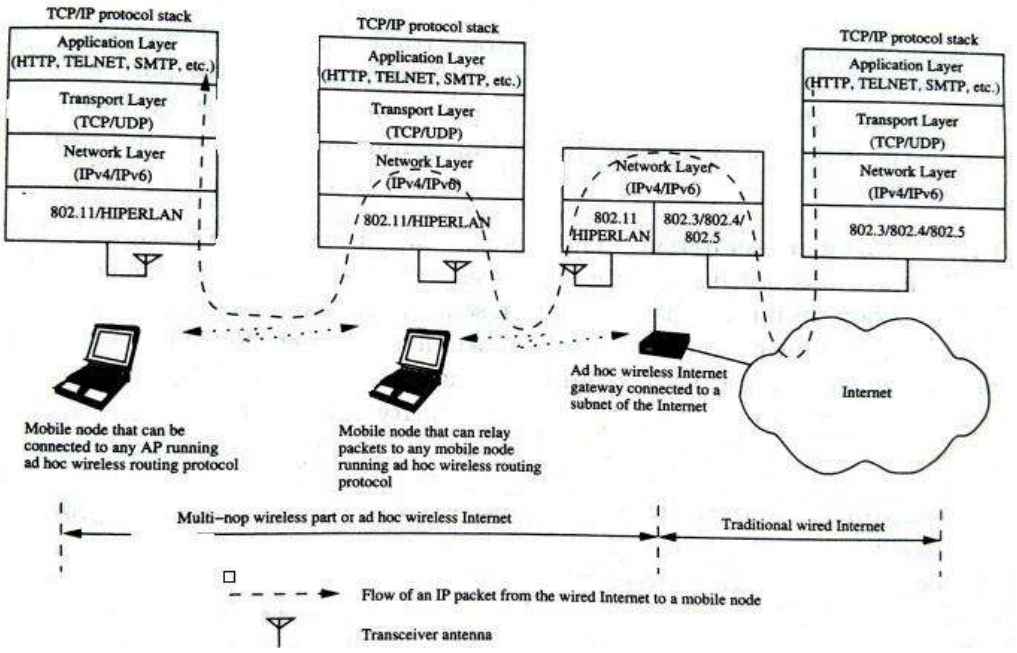


Figure 1. 6 Ad Hoc Wireless Internet

Source **:** Ad Hoc Wireless Networks Architectures and Protocol by C. Siva Ram Murthy and B. S. Manoj

- **Routing**

  - It is a major problem in ad hoc wireless internet, due to dynamic topological changes, the presence of gateways, multi-hop relaying, & the hybrid character of the network.
  - Possible solution is to use separate routing protocol for the wireless part of ad hoc wireless internet.

- **Transport layer protocol**

  - Several factors are to be considered here, the major one being the state maintenance overhead at the gateway nodes.

- **Load balancing**

  - They are essential to distribute the load so as to avoid the situation where the gateway nodes become bottleneck nodes.

- **Pricing / Billing**

  - Since internet bandwidth is expensive, it becomes very important to introduce pricing/billing strategies for the ad hoc wireless internet.

- **Provisioning of security**
  - Security is a prime concern since the end users can utilize the ad hoc wireless internet infrastructure to make e-commerce transaction.
- **QoS support**
  - With the widespread use of Voice Over IP (VOIP) & growing multimedia applications over the internet, provisioning of QoS support in the ad hoc wireless internet becomes a very important issue.
- **Service, address & location discovery**
  - Service discovery refers to the activity of discovering or identifying the party which provides service or resource.
  - Address discovery refers to the services such as those provided by Address Resolution Protocol (ARP) or Domain Name Service (DNS) operating within the wireless domain.
  - Location discovery refers to different activities such as detecting the location of a particular mobile node in the network or detecting the geographical location of nodes.

1.6 Routing Protocols for Ad Hoc Wireless Networks

- Routing is the exchange of information from one station of networks to other and Protocol is the set of standard or rules to exchange data between two devices.
- An ad hoc routing protocol is a convention, or standard, that controls how nodes decide which way to route packets between computing devices in a mobile ad hoc network.
- An ad hoc wireless network consists of a set of mobile nodes (hosts) that are connected by wireless links. The network topology (the physical connectivity of the communication network) in such a network may keep changing randomly.
- Routing protocols that find a path to be followed by data packets from a source node to a destination node used in traditional wired networks cannot be directly applied in ad hoc wireless networks due to their highly dynamic topology absence of established infrastructure for centralized administration (e.g., base stations or access points), bandwidth-constrained wireless links, and resource (energy)-constrained nodes.

1.7 Issues in Designing a Routing Protocol for Ad Hoc Wireless Networks

- The major challenges that a routing protocol designed for ad hoc wireless networks faces are:
  - Mobility of nodes
  - Bandwidth Constraints
  - Error-Prone channel state
  - Hidden Terminal Problem
  - Exposed Terminal Problems
  - Resource Constraints

1.7.1 Mobility

- Network topology is highly dynamic due to movement of nodes. Hence, an ongoing session suffers frequent path breaks.
- Disruption occurs due to the movement of either intermediate nodes in the path or end nodes.
- Wired network routing protocols cannot be used in adhoc wireless networks because the nodes are here are not stationary and the convergence is very slow in wired networks.
- Mobility of nodes results in frequently changing network topologies
- Routing protocols for ad hoc wireless networks must be able to perform efficient and effective mobility management.

1.7.2 Bandwidth Constraint

- Abundant bandwidth is available in wired networks due to the advent of fiber optics and due to the exploitation of wavelength division multiplexing (WDM) technologies.
- In a wireless network, the radio band is limited, and hence the data rates it can offer are much less than what a wired network can offer.
- This requires that the routing protocols use the bandwidth optimally by keeping the overhead as low as possible.

- The limited bandwidth availability also imposes a constraint on routing protocols in maintaining the topological information.

1.7.3 Error-prone shared broadcast radio channel

- The broadcast nature of the radio channel poses a unique challenge in ad hoc wireless networks.
- The wireless links have time-varying characteristics in terms of link capacity and link-error probability.
- This requires that the adhoc wireless network routing protocol interact with the MAC layer to find alternate routes through better-quality links.
- Transmissions in ad hoc wireless networks result in collisions of data and control packets.
- Therefore, it is required that ad hoc wireless network routing protocols find paths with less congestion.

1.7.4 Hidden Terminal Problem

- The hidden terminal problem refers to the collision of packets at a receiving node due to the simultaneous transmission of those nodes that are not within the direct transmission range of the receiver, but are within the transmission range of the receiver.
- Collision occurs when both nodes transmit packets at the same time without knowing about the transmission of each other.

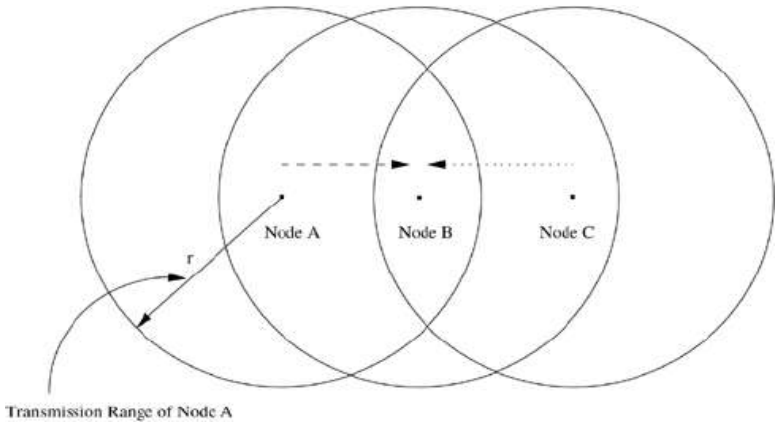


Figure 1. 7 Hidden Terminal Problem

Source **:** Ad Hoc Wireless Networks Architectures and Protocol by C. Siva Ram Murthy and B. S. Manoj

- For example, consider figure 1.7. Here, if both node A and node C transmit to node B at the same time, their packets collide at node B. This is due to the fact that both node A and C are hidden from each other, as they are not within the direct transmission range of each other and hence do not know about the presence of each other.
- Solution for this problem (figure 1.8), include medium access collision avoidance (MACA)

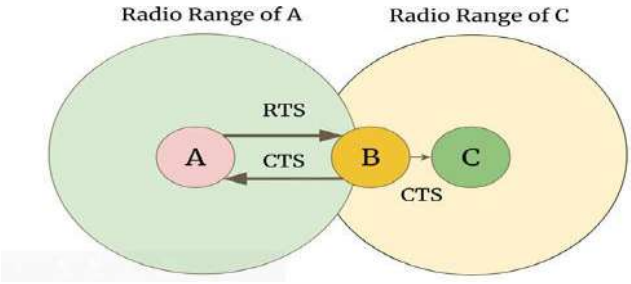


Figure 1.8 Solution for Hidden Terminal Problem

- Transmitting node first explicitly notifies all potential hidden nodes about the forthcoming transmission by means of a two way handshake control protocol called RTS-CTS protocol exchange. This may not solve the problem completely but it reduces the probability of collisions.

1.7.5 Exposed Terminal Problem

- The exposed terminal problem refers to the inability of a node which is blocked due to transmission by a nearby transmitting node to transmit to another node.
- For example, consider the figure 1.9, Here, if a transmission from node B to another node A is already in progress, node C cannot transmit to node D, as it concludes that its neighbor node B, is in transmitting mode and hence should not interfere with the on-going transmission. Thus, reusability of the radio spectrum is affected.

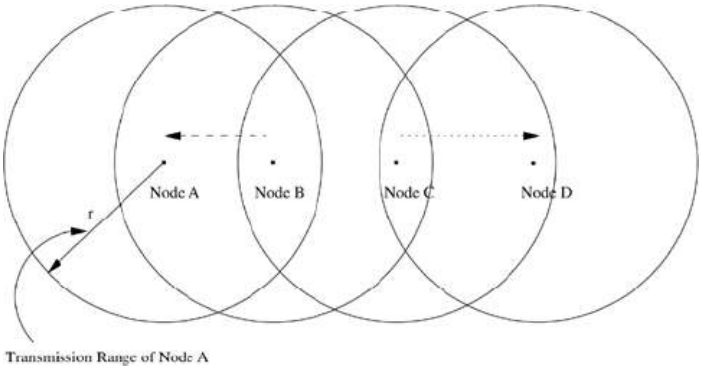


Figure 1. 9 Exposed Terminal Problem

Source **:** Ad Hoc Wireless Networks Architectures and Protocol by C. Siva Ram Murthy and B. S. Manoj

- Solution for this problem, illustrated in figure 1.10. In this case, node A did not successfully receive the CTS originated by node R and hence assumes that there is no on-going transmission in the neighborhood. Since node A is hidden from node T, any attempt to originate its own RTS would result in collision of the on-going transmission between nodes T and R.

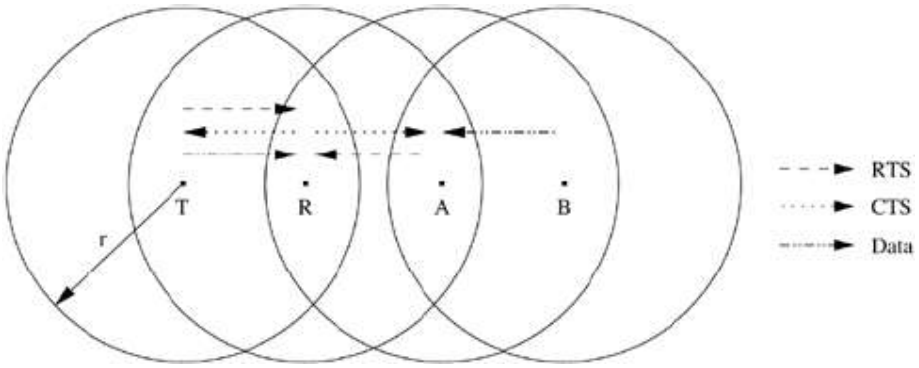


Figure 1. 10 Solution for Exposed Terminal Problem

Source **:** Ad Hoc Wireless Networks Architectures and Protocol by C. Siva Ram Murthy and B. S. Manoj

1.7.6 Resource Constraints

- Two essential and limited resources are battery life and processing power.
- Devices used in adhoc wireless networks require portability, and hence they also have size and weight constraints along with the restrictions on the power source.
- Increasing the battery power and processing ability makes the nodes bulky and less portable.

1.8 Characteristics of an Ideal Routing Protocol for Ad Hoc Wireless Networks

- A routing protocol for ad hoc wireless networks should have the following characteristics:
  - It must be fully distributed as centralized routing involves high control overhead and hence is not scalable.
  - It must be adaptive to frequent topology changes caused by the mobility of nodes.
  - Route computation and maintenance must involve a minimum number of nodes. Each node in the network must have quick access to routes, that is, minimum connection setup time is desired.
  - It must be localized, as global state maintenance involves a huge state propagation control overhead.
  - It must be loop-free and free from state routes.

- The number of packet collisions must be kept to a minimum by limiting the number of broadcasts made by each node. The transmissions should be reliable to reduce message loss and to prevent the occurrence of state routes.
- It must converge to optimal routes once the network topology becomes stable. The convergence must be quick.
- It must optimally use scarce resources such as bandwidth, computing power, memory, and battery power.
- Every node in the network should try to store information regarding the stable local topology only. Changes in remote parts of the network must not cause updates in the topology information maintained by the node.
- It should be able to provide a certain level of quality of service (QoS) as demanded by the applications, and should also offer support for time-sensitive traffic.

1.9 Classifications of Routing Protocols

➢ Routing protocols for ad hoc wireless networks can be classified into several types based on different criteria. A classification tree is shown in Figure 1.11.

➢ The routing protocol for adhoc wireless networks can be broadly classified into 4 categories based on

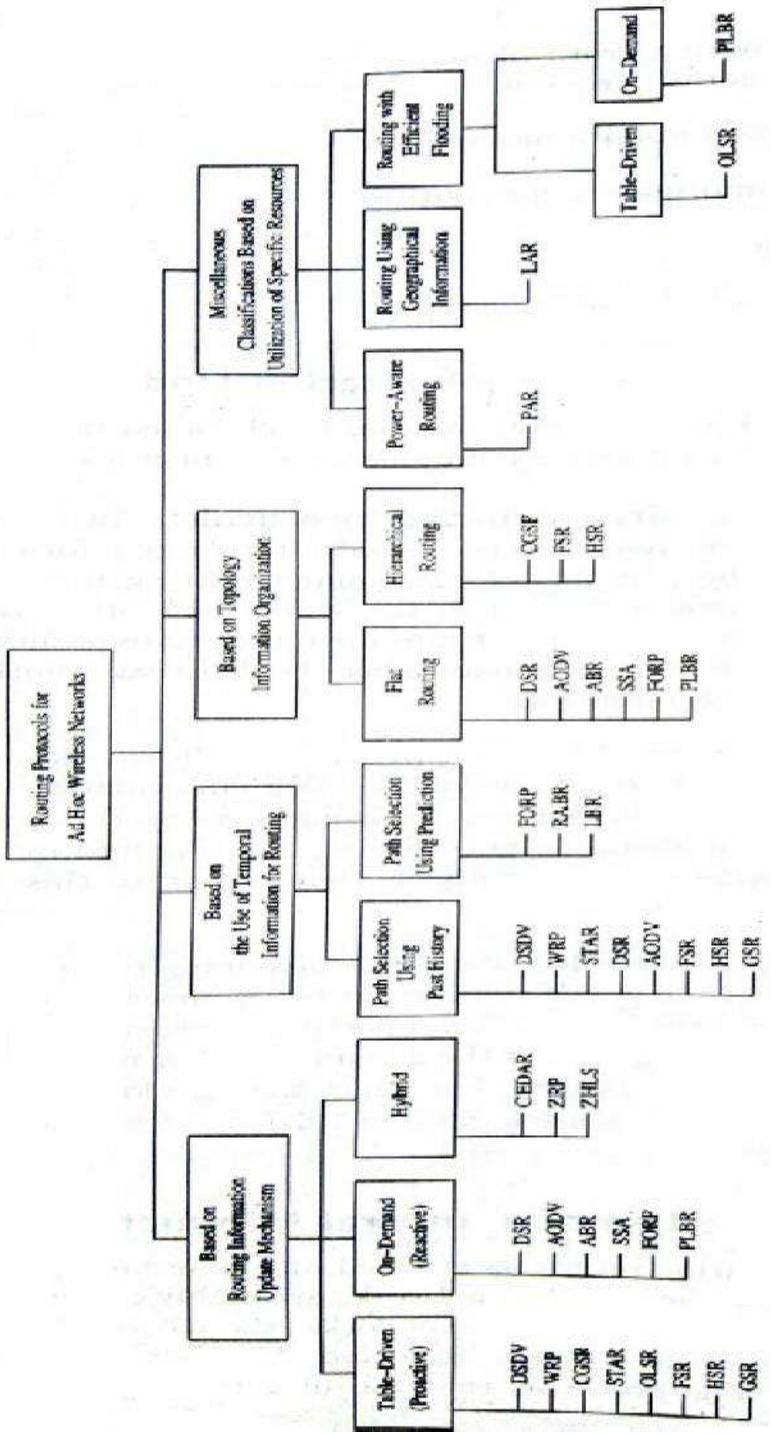
- Routing information update mechanism.
- Use of temporal information for routing
- Routing topology
- Utilization of specific resources.

1.9.1 Based on the routing information update mechanism

➢ Ad hoc wireless network routing protocols can be classified into 3 major categories based on the routing information update mechanism. They are:

- **Proactive or table-driven routing protocols**
  - Every node maintains the network topology information in the form of routing tables by periodically exchanging routing information.
  - Routing information is generally flooded in the whole network.
  - Whenever a node requires a path to a destination, it runs an appropriate path-finding algorithm on the topology information it maintains.
- **Reactive or on-demand routing protocols**
  - Do not maintain the network topology information.
  - Obtain the necessary path when it is required, by using a connection establishment process.

Figure 1.11 Classification of Sensor Network Protocols



Source : Ad Hoc Wireless Networks Architectures and Protocol by C. Siva Ram Murthy and B. S. Manoj

- **Hybrid routing protocols**
  - Combine the best features of the above two categories.
  - Nodes within a certain distance from the node concerned, or within a particular geographical region, are said to be within the routing zone of the given node.
  - For routing within this zone, a table-driven approach is used.
  - For nodes that are located beyond this zone, an on-demand approach is used.

1.9.2 Based on the use of temporal information for routing

➢ The protocols that fall under this category can be further classified into two types

- **Routing protocols using past temporal information**
  - Use information about the past status of the links or the status of links at the time of routing to make routing decisions.
- **Routing protocols that use future temporal information**
  - Use information about the about the expected future status of the wireless links to make approximate routing decisions.
  - Apart from the lifetime of wireless links, the future status information also includes information regarding the lifetime of the node, prediction of location, and prediction of link availability.

1.9.3 Based on the Routing Topology

➢ Ad hoc wireless networks, due to their relatively smaller number of nodes, can make use of either a flat topology or a hierarchical topology for routing.

- **Flat topology routing protocols**
  - Make use of a flat addressing scheme similar to the one used in IEEE 802.3 LANs.
  - It assumes the presence of a globally unique addressing mechanism for nodes in an ad hoc wireless network.

- **Hierarchical topology routing protocols**
  - Make use of a logical hierarchy in the network and an associated addressing scheme.
  - The hierarchy could be based on geographical information or it could be based on hop distance.

1.9.4 Based on the utilization of specific resources

- **Power-aware routing**
  - Aims at minimizing the consumption of a very important resource in the ad hoc wireless networks: the battery power.
  - The routing decisions are based on minimizing the power consumption either logically or globally in the network.
- **Geographical information assisted routing**
  - Improves the performance of routing and reduces the control overhead by effectively utilizing the geographical information available.

1.10 Table Driven Routing Protocols

- These protocols are extensions of the wired network routing protocols.
- They maintain the global topology information in the form of tables at every node.
- Tables are updated frequently in order to maintain consistent and accurate network state information.
- Example:
  - **Destination Sequenced Distance Vector Routing Protocol (DSDV)**
  - **Wireless Routing Protocol (WRP)**
  - **Source-Tree Adaptive Routing Protocol (STAR)**
  - **Cluster-head Gateway Switch Routing Protocol (CGSR)**

1.10.1 Destination Sequenced Distance Vector Routing Protocol (DSDV)

- Destination Sequenced Distance Vector (DSDV) is a hop-by-hop vector routing protocol requiring each node to periodically broadcast routing updates. Destination Sequenced Distance Vector Routing protocol is a modified version of Bellman Ford Algorithm and is based upon the concepts of Distance Vector Routing.
- In Distance Vector Routing (DVR), each node broadcasts a table containing its distance from nodes which are directly connected and based upon this, other nodes broadcasts the updated routing. Those nodes which are unreachable directly are labelled as "infinite".
- But, this updation of routing tables keeps on happening and an infinite loop is generated which is commonly known as Count-To-Infinityproblem.
- To overcome this problem of count to infinity by generating sequence number in the routing table, every time the routing table is updated. The process of DSDV is same as that of Distance Vector Routing but an extra attribute of sequence number is added.

Destination Sequenced Distance Vector Routing: Concept

- DSDV protocol uses and maintains a single table only, for every node individually. The table contains the following attributes.
  - Routing Table: It contains the distance of a node from all the neighbouring nodes along with the sequence number (SEQ No means the time at which table is updated).

Node	Destination	Next Hop	Distance	SEQ No

Table: 1.1 DSDV Table Format

- This table is updated on every step and ensures that each node broadcast as well as receives correct information about all the nodes including their distance and sequence number.

Destination Sequenced Distance Vector Routing Protocol: Working

- In DSDV, nodes broadcasts their routing tables to immediate neighbors with the sequence number. Every time any broadcasting occurs, the sequence number is also updated along with distances of nodes.

- Consider a network (Figure 1.12) of 3 nodes having distances of "1" on each of the edges respectively. Below mentioned steps will let you know how DSDV works and routing tables are updated.

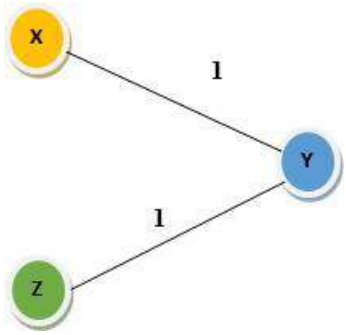


Figure 1.12 Sample Network of DSDV

Step-1: Draw separate tables for all the nodes "X", "Y" & " Z" along with the distance and sequence number.

For X:

Source	Destination	Next Hop	Cost	SEQ No
X	X	X	0	100-X
X	Y	Y	1	200-Y
X	Z	Y	2	300-Z

For Y:

Source	Destination	Next Hop	Cost	SEQ No
Y	X	X	1	100-X
Y	Y	Y	0	200-Y
Y	Z	Y	1	300-Z

For Z:

Source	Destination	Next Hop	Cost	SEQ No
Z	X	Y	2	100-X
Z	Y	Y	1	200-Y
Z	Z	Z	0	300-Z

- If "Y" wants to broadcast the routing table. Then updated routing tables of all the nodes in the network will look like as depicted in the below tables where Bold marked cell denotes the change in sequence number.

For X:

Source	Destination	Next Hop	Cost	SEQ No
X	X	X	0	100-X
X	Y	Y	1	**210-Y**
X	Z	Y	2	300-Z

For Y:

Source	Destination	Next Hop	Cost	SEQ No
Y	X	X	1	100-X
Y	Y	Y	0	**210-Y**
Y	Z	Z	1	300-Z

For Z:

Source	Destination	Next Hop	Cost	SEQ No
Z	X	Y	2	100-X
Z	Y	Y	1	**210-Y**
Z	Z	Z	0	300-Z

Advantages

- Less delay involved in the route setup process.
- Mechanism of incremental update with sequence number tags makes the existing wired network protocols adaptable to ad hoc wireless networks.
- The updates are propagated throughout the network in order to maintain an up-to-date view of the network topology at all nodes.

Disadvantages

- The updates due to broken links lead to a heavy control overhead during high mobility.
- Even a small network with high mobility or a large network with low mobility can completely choke the available bandwidth.
- Suffers from excessive control overhead.
- In order to obtain information about a particular destination node, a node has to wait for a table update message initiated by the same destination node.

1.11 On–Demand Routing protocols

➢ In table-driven protocols, each node maintain up-to-date routing information to all the nodes in the network where in **on-demand protocols** a node finds the route to a destination when it desires to send packets to the destination.

1.11.1 Ad hoc On–Demand Distance Vector Routing (AODV)

- This protocol is an example of reactive routing protocol which does not maintain routes but build the routes as per requirements. That means, Route is established only when it is required by a source node for transmitting data packets.
- AODV is used to overcome the drawbacks of Dynamic Source Routing Protocol and Distance Vector Routing Protocol i.e. Dynamic Source Routing is capable of maintaining information of the routes between source and destination which makes it slow. If the network is very large containing a number of routes from source to destination, it is difficult for the data packets header to hold whole information of the routes.
- In case of Dynamic Source Routing, multiple routes are present for sending a packet from source to destination but AODV overcomes this disadvantage too.
- In AODV, along with routing tables of every node, two counters including Sequence Number (SEQ NO) and broadcast ID are maintained also.
- The destination IP is already known to which data is to be transferred from source. Thus, the destination Sequence Number (SEQ NO) helps to determine an updated path from source to destination.
- Along with these counters, Route Request (RREQ) and Route Response (RRESP) packets are used in which RREQ is responsible for discovering of route from source to destination and RRESP sends back the route information response to its source.

Ad-Hoc On - Demand Distance Vector Routing Protocol: Working

- In Ad-Hoc On-Demand Distance Vector Routing, the source node and destination nodes IP addresses are already known.
- The goal is to identify, discover and maintain the optimal route between source and destination node in order to send/receive data packets and informative.
- Each node comprises of a routing table along with below mentioned format of Route Request (RREQ) packet.
- RREQ {Destination IP, Destination Sequence Number, Source IP, Source Sequence Number, Hop Count}.

- Consider a network (Figure 1.13) containing 5 nodes that are "X", "Y", "Z","T","D" present at unit distance from each other, where "X" being the source node and "D" being the destination node.

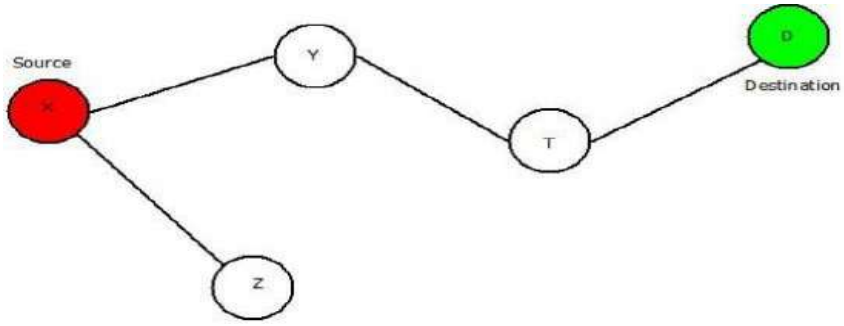


Figure 1.13 Sample Network of AODV

Source **:** Ad Hoc Wireless Networks Architectures and Protocol by C. Siva Ram Murthy and B. S. Manoj

- The IP addresses of source node "X" and destination node "D" is already known. Below mentioned steps will let you know how AODV works and concept of Route Request (REREQ) and Route Response (RRESP) is used.(Figure 1.14)

  - **Step 1:** Source node "X" will send Route Request i.e. RREQ packet to its neighbours "Y" and "Z".

  - **Step 2:** Node "Y" & "Z" will check for route and will respond using RRESP packet back to source "X". Here in this case "Z" is the last node but the destination. It will send the RREQ packet to "X" stating "Route Not Found". But node "Y" will send RRESP packet stating "Route Found" and it will further broadcast the RRESP to node "T".

  - **Step 3:** Now the field of net hop in the RREQ format will be updated, Node "T" will send back the "Route Found" message to Node "Y" and will update the next hop field further.

  - **Step 4:** Then Node "T" will broadcast and RREQ packet to Node "D", which is the destination and the next hop field is further updated. Then it will send RRES packet to "T" which will further be sent back to the source node "X" via node "Y" and Node "T" resulting in generation of an optimal path. The updated network would be:

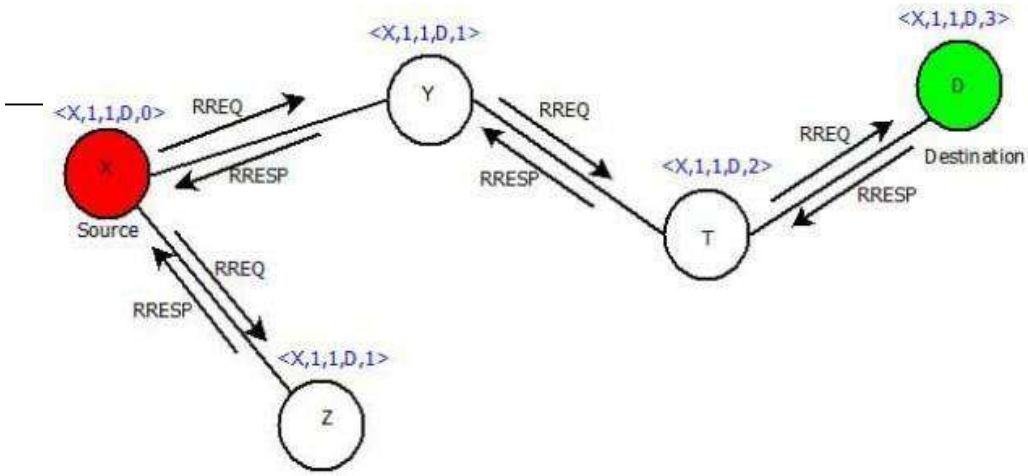


Figure 1.14 Example of AODV Network

Source **:** Ad Hoc Wireless Networks Architectures and Protocol by C. Siva Ram Murthy and B. S. Manoj

Advantages

- Dynamic networks can be handled easily.
- No loop generation.

Disadvantages

- A delayed protocol because of its route discovery process.
- High bandwidth requirement.

UNIT – 2

SENSOR NETWORKS – INTRODUCTION & ARCHITECTURES

Challenges for Wireless Sensor Networks, Enabling Technologies for Wireless Sensor Networks, WSN application examples, Single-Node Architecture – Hardware Components, Energy Consumption of Sensor Nodes, Network Architecture – Sensor Network Scenarios, Transceiver Design Considerations, Optimization Goals and Figures of Merit.

TABLE OF CONTENTS

2.1 *Introduction – Wireless Sensor Networks*

2.1.1 Sensor

- A Sensor is a device that responds and detects some type of input from both the physical or environmental conditions, such as pressure, heat, light, etc.
- The output of the sensor is generally an electrical signal that is transmitted to a controller for further processing.
- All types of sensors can be basically classified into analog sensors and digital sensors. But, there are a few types of sensors such as temperature sensors, IR sensors, ultrasonic sensors, pressure sensors, proximity sensors, and touch sensors are frequently used in most of the electronics applications.

2.1.2 Wireless Sensor Networks

- Wireless Sensor Network (WSN) is an infrastructure-less wireless network that is deployed in a large number of wireless sensors in an ad-hoc manner that is used to monitor the system, physical or environmental conditions (Figure 2.1).
- A sink or base station acts like an interface between users and the network. One can retrieve required information from the network by injecting queries and gathering results from the sink. Typically a wireless sensor network contains hundreds of thousands of sensor nodes.

- The sensor nodes can communicate among themselves using radio signals. Base Station in a WSN System is connected through the Internet to share data. WSN can be used for processing, analysis, storage, and mining of the data.

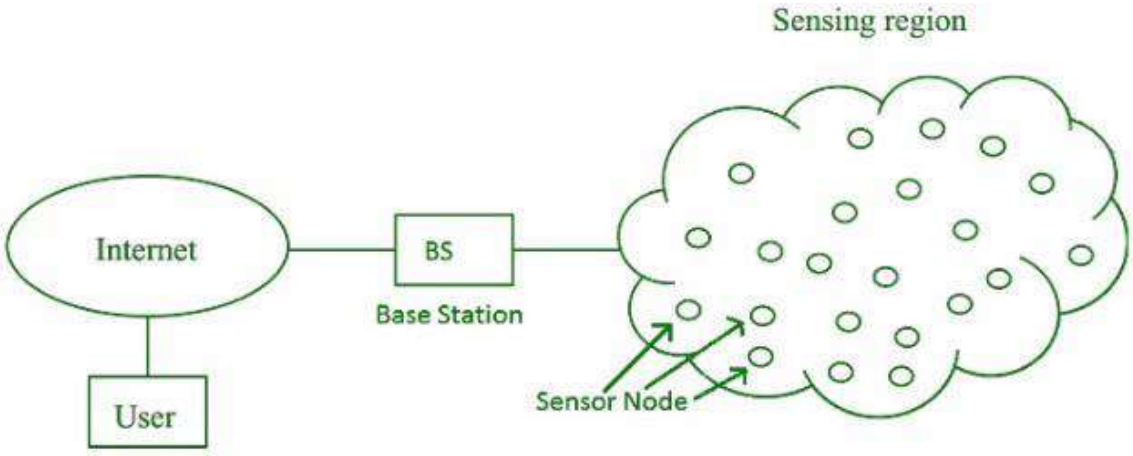


Figure 2.1 Wireless Sensor Networks

Source **:** Ad Hoc Wireless Networks Architectures and Protocol by C. Siva Ram Murthy and B. S. Manoj

Components of WSN Sensor Networks – Introduction & Architectures

- **Sensors** : Sensors in WSN are used to capture the environmental variables and which is used for data acquisition. Sensor signals are converted into electrical signals.
- **Radio Nodes** : It is used to receive the data produced by the Sensors and sends it to the WLAN access point. It consists of a microcontroller, transceiver, external memory, and power source.
- **WLAN Access Point** : It receives the data which is sent by the Radio nodes wirelessly, generally through the internet.
- **Evaluation Software** : The data received by the WLAN Access Point is processed by a software called as Evaluation Software for presenting the report to the users for further processing of the data which can be used for processing, analysis, storage, and mining of the data.

Characteristic of Wireless Sensor Network

- Power consumption constraints for nodes using batteries or energy harvesting
- Chance to cope with node failures (resilience)
- Mobility of nodes
- Heterogeneity of nodes
- Scalability to large scale of deployment

- Capability to withstand harsh environmental conditions
- Simplicity of use
- Cross layer design

Advantages

- Network setups can be carried out without fixed infrastructure.
- Suitable for the non-reachable places such as over the sea, mountains, rural areas or deep forests
- Suitable for the non-reachable places such as over the sea, mountains, battlefield surveillance, rural areas or deep forests
- Flexible if there is random situation when additional workstation is needed
- Implementation pricing is cheap.
- It avoids plenty of wiring
- It might accommodate new devices at any time
- It's flexible to undergo physical partitions.

Disadvantages

- WSN is it is not fully secure. Hackers hack the network easily. It is easy for hackers to hack it we couldn't control propagation of waves.
- Works in short communication range – consumes a lot of power
- Short battery life. Sensor nodes need to be charged at after few times at intervals.
- Communication speed is very poor, on the other hand, wired networks have good speed of communication.
- More complicated to configure compared to a wired network

Applications of WSN

- Internet of Things (IOT)
- Surveillance and Monitoring for security, threat detection
- Environmental temperature, humidity, and air pressure

- Noise Level of the surrounding
- Medical applications like patient monitoring
- Agriculture
- Landslide Detection

Difference between Cellular and Ad-hoc Networks

Cellular Networks	Ad-hoc Networks
Fixed, pre-located cell sites and base stations	No fixed base stations
Slow Deployment	Very rapid deployment
Static backbone network topology	Highly dynamic network topologies
Single Hop	Single and Multi-hop Communication
Relatively favourable environment	Hostile environment
Stable connectivity	Irregular connectivity
Detailed planning before base stations can be installed	Ad-hoc network automatically forms and conforms to change

Difference between Ad hoc Network and Wireless Sensor Networks

Features	Wireless Sensor Networks	Ad hoc Networks
Number of Sensor Nodes	Large in Quantity	Medium in quantity
Deployment Type	Very much dense	Scattered
Rate of failure	More	Very rare
Battery	Not Rechargeable/ Hard to recharge	Rechargeable
Redundancy	High	Low
Data rate	Low	High
Change in network topology	Frequency	Rare

2.2 Challenges for Wireless Sensor Networks

- Major issues that affect the design and performance of a
- wireless sensor network are as follows
- Major issues and challenges that affect the design and performance of a wireless sensor network are as follows
  - **Energy Efficiency**
  - **Quality of Service**
  - **Security Issue**
  - **Deployment**
  - **Node Costs**
  - **Limited Bandwidth**
  - **Scalability to large scale of deployment**
  - **Fault-Tolerance**
  - **Multi-hop communication**
  - **Design Constraints**

Energy Efficiency

- The first and often most important design challenge for a WSN is energy efficiency. Power consumption can be allocated to three functional domains: sensing, communication, and data processing, each of which requires optimization.
- The sensor node lifetime typically exhibits a strong dependency on battery life. The constraint most often associated with sensor network design is that sensor nodes operate with limited energy budgets.
- Typically, sensors are powered through batteries, which must be either replaced or recharged when depleted. For non-rechargeable batteries, a sensor node should be able to operate until either its mission time has passed or the battery can be replaced. The length of the mission time depends on the type of application.

Quality of Service (QOS)

- Some real time sensor application are very time critical which means the data should be delivered within a certain period of time from the moment it is sensed, otherwise the data will be unusable .So this must be a QOS parameter for some applications

Security Issue

- Many wireless sensor networks collect sensitive information, especially for military applications which carry sensitive data. The remote and unattended operation of sensor nodes increases their exposure to malicious intrusions and attacks.
- Further, wireless communications make it easy for an adversary to eavesdrop on sensor transmissions. For example, one of the most challenging security threats is a denial-of service (DoS) attack, whose goal is to disrupt the correct operation of a sensor network.

Deployment

- Node deployment is a fundamental issue to be solved in Wireless Sensor Networks. A proper node deployment scheme can reduce the complexity of problems.
- Deploying and managing a high number of nodes in a relatively bounded environment requires special techniques. Hundreds to thousands of sensors may be deployed in a sensor region.
- There are two deployment models at present: (i) static deployment (ii) dynamic deployment. The static deployment chooses the best location according to the optimization strategy, and the location of the sensor nodes has no change in the lifetime of the WSN. The dynamic deployment throws the nodes randomly for optimization

Node Costs

- A sensor network consists of a large set of sensor nodes. It follows that the cost of an individual node is critical to the overall financial metric of the sensor network.
- Clearly, the cost of each sensor node has to be kept low for the global metrics to be acceptable. Depending on the application of sensor network, large number sensors might be scattered randomly over an environment, such as weather monitoring.
- If the overall cost was appropriate for sensor networks and it will be more acceptable and successful to users which need careful consideration.

Limited Bandwidth

- Bandwidth limitation directly affects message exchanges among sensors, and synchronization is impossible without message exchanges.

- Sensor networks often operate in a bandwidth and performance constrained multi-hop wireless communications medium. Presently, wireless communication is limited to a data rate in the order of 10–100 Kbits/second.

Scalability to large scale of deployment

- In some applications, tens of thousands of sensors might be deployed. At any time numbers of nods can be increased or decreased. A synchronization scheme should scale well with increasing number of nodes and/or high density in the network.

Fault-Tolerance

- In a hostile environment, a sensor node may fail due to physical damage or lack of energy (power). If some nodes fail, the protocols that are working upon must accommodate these changes in the network. As an example, for routing or aggregation protocol, they must find suitable paths or aggregation point in case of these kinds of failures.

Multi-hop communication

- The need for multi-hop communication arises due to the increase in the size of wireless sensor networks. In such settings, sensors in one domain communicate with sensors in another domain via an intermediate sensor that can relate to both domains. Communication can also occur as a sequence of hops through a chain of pair-wise adjacent sensors.

Design Constraints

- The primary goal of wireless sensor design is to create smaller, cheaper, and more efficient devices.

- A variety of additional challenges can affect the design of sensor nodes and wireless sensor networks. WSN have challenges on both software and hardware design models with restricted constraints.

2.3 *Enabling Technologies for Wireless Sensor Networks*

- The building of wireless sensor networks has only become possible with some fundamental advances in enabling technologies.

2.3.1 Hardware Design

- First and foremost among these technologies is the miniaturization of hardware. Smaller feature sizes in chips have driven down the power consumption of the basic

components of a sensor node to a level that the constructions of WSNs can be contemplated.

- This is particularly relevant to microcontrollers and memory chips as such, but also, the radio modems, responsible for wireless communication, have become much more energy efficient.

- Reduced chip size and improved energy efficiency is accompanied by reduced cost, which is necessary to make redundant deployment of nodes affordable.

2.3.2 Energy Consumption

- The sensor nodes consume the power for important three functional domains such as sensing, communication, and data processing. The sensor node lifetime typically exhibits a strong dependency on battery life. The constraint most often associated with sensor network design is that sensor nodes operate with limited energy budgets.

- Typically, sensors are powered through batteries, which must be either replaced or recharged when depleted. For non-rechargeable batteries, a sensor node should be able to operate until either its mission time has passed or the battery can be replaced. The length of the mission time depends on the type of application.

- A sensor node also has a device for energy scavenging. Energy Scavenging (also known as power harvesting or energy harvesting or ambient power) is the process by which energy is derived from external sources (e.g., solar power, thermal energy, wind energy, salinity gradients, and kinetic energy, also known as ambient energy), captured, and stored for small, wireless autonomous devices, like those used in wearable electronics and wireless sensor networks. Such a concept requires the battery to be efficiently chargeable with small amounts of current, which is not a standard ability.

Classical energy consumption model

- An energy consumption model for sensors based on the observation that the energy consumption would likely be dominated by the data communications subsystem. Table 2.1 reproduces their model. The energy level is mentioned by Joules.

Table 2.1 Radio Characteristics, Classical model

Radio mode	Energy Consumption
Transmitter Electronics ($E_{Tx-elec}$) Receiver Electronics ($E_{Rx-elec}$) ($E_{Tx-elec} = E_{Rx-elec} = E_{elec}$)	$50nJ / bit$
Transmit Amplifier ($\mathcal{E}_{amp}$)	$100\,pJ / bit / m^2$
Idle (E_{idle})	$40nJ / bit$
Sleep	0

Modelling energy consumption during transmission

- The energy consumed by a transmitter is due to two sources. One part is due to RF signal generation, which mostly depends on chosen modulation and target distance and hence on the transmission power P_{tx}, that is, the power radiated by the antenna.
- A second part is due to electronic components necessary for frequency synthesis, frequency conversion, filters, and so on. These costs are basically constant. One of the most crucial decisions when transmitting a packet is thus the choice of P_{tx}.
- Let us assume that the desired transmission power P_{tx} is known. The transmitted power is generated by the amplifier of a transmitter. Its own power consumption P_{amp} depends on its architecture, but for most of them, their consumed power depends on the power they are to generate. In the most simplistic models, these two values are proportional to each other, but this is an oversimplification.
- A more realistic model assumes that a certain constant power level is always required irrespective of radiated power, plus a proportional offset:

$$P_{amp} = \alpha_{amp} + \beta_{amp} P_{tx}$$

where α_{amp} and β_{amp} are constants depending on process technology and amplifier architecture.

2.3.3 Software

- Energy is the scarcest resource of WSN nodes, and it determines the lifetime of WSNs. WSNs may be deployed in large numbers in various environments, including remote and hostile regions, where ad hoc communications are a key component.

- For this reason, algorithms and protocols need to address the following issues:
  - Increased lifespan
  - Robustness and fault tolerance
  - Self-configuration
- Lifetime maximization: Energy/Power Consumption of the sensing device should be minimized and sensor nodes should be energy efficient since their limited energy resource determines their lifetime. To conserve power, wireless sensor nodes normally power off both the radio transmitter and the radio receiver when not in use

2.3.4 Routing protocols

- Wireless sensor networks are composed of low-energy, small-size, and low-range unattended sensor nodes. Recently, it has been observed that by periodically turning on and off the sensing and communication capabilities of sensor nodes, we can significantly reduce the active time and thus prolong network lifetime.
- However, this duty cycling may result in high network latency, routing overhead, and neighbour discovery delays due to asynchronous sleep and wake-up scheduling. These limitations call for a countermeasure for duty-cycled wireless sensor networks which should minimize routing information, routing traffic load, and energy consumption.

2.3.5 Operating systems

- Operating systems for wireless sensor network nodes are typically less complex than general-purpose operating systems.
- They more strongly resemble embedded systems, for two reasons. First, wireless sensor networks are typically deployed with a particular application in mind, rather than as a general platform. Second, a need for low costs and low power leads most wireless sensor nodes to have low-power microcontrollers ensuring that mechanisms such as virtual memory are either unnecessary or too expensive to implement.

2.4 WSN Application Examples

- Wireless sensor network (WSN) refers to a group of spatially dispersed and dedicated sensors for monitoring and recording the physical conditions of the environment and organizing the collected data at a central location.
- There are numerous applications of WSNs. Some of applications are:

- Military or Border Surveillance Applications
- Environmental Applications
  - Air Pollution Monitoring
  - Forest Fire Detection
  - Landslide Detection
  - Water Quality Monitoring
  - Natural Disaster Prevention
- Health Care Applications
- Home Intelligence
- Industrial Process Control
- Agriculture
- Structural Monitoring

2.4.1 Military or Border Surveillance Applications

- WSNs are becoming an integral part of military command, control, communication and intelligence systems.
- Sensors can be deployed in a battle field to monitor the presence of forces and vehicles, and track their movements, enabling close surveillance of opposing forces.
- The chemical, nuclear and biological attacks can also be detected through the sensor nodes.
- An example of this is the 'sniper detection system' which can detect the incoming fire through acoustic sensors and the position of the shooter can also be estimated by processing the detected audio from the microphone.

2.4.2 Environmental Applications

- These sensor networks have a huge number of applications in the environment. They can be used to track movement of animals, birds and record them.
- Monitoring of earth, soil, atmosphere context, irrigation and precision agriculture can be done through these sensors.

- A common example is of 'Zebra Net'. The purpose of this system is to track and monitor the movements and interactions of zebras within themselves and with other species also.
- There are many applications in monitoring environmental parameters. Some examples of which are given below:

  - **Air Pollution Monitoring**
    - Wireless sensor networks have been deployed in several cities to monitor the concentration of dangerous gases for citizens. These can take advantage of the ad hoc wireless links rather than wired installations, which also make them more mobile for testing readings in different areas.
  - **Forest Fire Detection**
    - A network of Sensor Nodes can be installed in a forest to detect when a fire has started. The nodes can be equipped with sensors to measure temperature, humidity and gases which are produced by fire in the trees or vegetation. The early detection is crucial for a successful action of the fire fighters.
  - **Landslide Detection**
    - A landslide detection system makes use of a wireless sensor network to detect the slight movements of soil and changes in various parameters that may occur before or during a landslide. Through the data gathered it may be possible to know the impending occurrence of landslides long before it actually happens.
  - **Water Quality Monitoring**
    - Water quality monitoring involves analyzing water properties in dams, rivers, lakes and oceans, as well as underground water reserves. The use of many wireless distributed sensors enables the creation of a more accurate map of the water status, and allows the permanent deployment of monitoring stations in locations of difficult access, without the need of manual data retrieval.
  - **Natural Disaster Prevention**
    - Wireless sensor networks can be effective in preventing adverse consequences of natural disasters, like floods. Wireless nodes have been deployed successfully in rivers, where changes in water levels must be monitored in real time.

2.4.3 Health Care Applications

- Wireless sensor networks can be used to monitor and track elders and patients for health care purposes, which can significantly relieve the severe shortage of health care personnel and reduce the health care expenditures in the current health care systems.
- For example sensors can be deployed in a patient's home to monitor the behaviours of the patient. It can alert doctors when the patient falls and requires immediate medical attention.
- An example of this is 'artificial retina' which helps the patient in detecting the presence of light and the movement of objects.

2.4.4 Home Intelligence

- Wireless sensor networks can be used to provide more convenient and intelligent living environments for human beings.
- For example, wireless sensors can be used to remotely read utility meters in a home like water, gas, electricity and then send the readings to a remote centre through wireless communication

2.4.5 Industrial Process Control

- In industry, WSNs can be used to monitor manufacturing process or the condition of manufacturing equipment.
- For example, chemical plants or oil refiners can use sensors to monitor the condition of their miles of pipelines. These sensors are used to alert in case of any failures occurred.

2.4.6 Agriculture

- Sensors used in smart farming are known as agriculture sensors.
- These sensors provide data which assist farmers to monitor and optimize crops by adapting to changes in the environmental conditions. These sensors are installed on weather stations, drones and robots used in the agriculture industry.

2.4.7 Structural Monitoring

- Wireless sensors can be used to monitor the movement within buildings and infrastructure such as bridges, flyovers, embankments, tunnels etc... enabling Engineering practices to monitor assets remotely without the need for costly site visits, as well as having the advantage of daily data, whereas traditionally this data was

collected weekly or monthly, using physical site visits, involving either road or rail closure in some cases.

- It is also far more accurate than any visual inspection that would be carried out.

2.5 Single-Node Architecture

2.5.1 Hardware Components

- Choosing the hardware components for a wireless sensor node, obviously the applications has to consider size, costs, and energy consumption of the nodes.
- A basic sensor node comprises five main components such as (Figure 2.2)
  - **Controller**
  - **Memory**
  - **Sensors and Actuators**
  - **Communication Devices**
  - **Power Supply Unit**

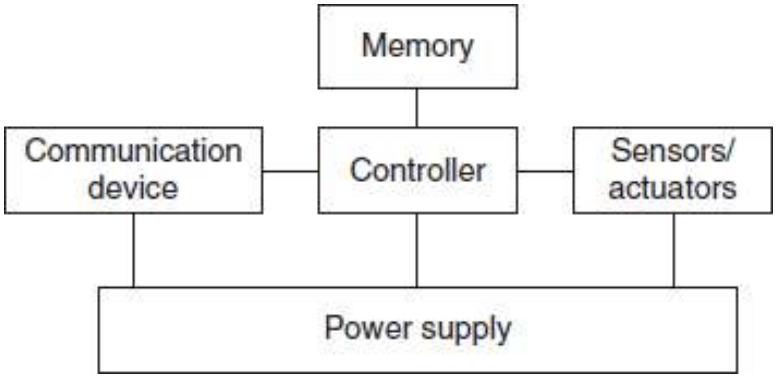


Figure 2.2 Hardware Components of Sensor Node

Source **:** Ad Hoc Wireless Networks Architectures and Protocol by C. Siva Ram Murthy and B. S. Manoj

2.5.1.1 Controller

- A controller to process all the relevant data, capable of executing arbitrary code. The controller is the core of a wireless sensor node.
- It collects data from the sensors, processes this data, decides when and where to send it, receives data from other sensor nodes, and decides on the actuator's behavior.

- It has to execute various programs, ranging from time critical signal processing and communication protocols to application programs; it is the Central Processing Unit (CPU) of the node.
- For General-purpose processors applications microcontrollers are used. These are highly overpowered, and their energy consumption is excessive. These are used in embedded systems.
- Some of the key characteristics of microcontrollers are particularly suited to embedded systems are their flexibility in connecting with other devices like sensors and they are also convenient in that they often have memory built in.
- A specialized case of programmable processors are Digital Signal Processors (DSPs). They are specifically geared, with respect to their architecture and their instruction set, for processing large amounts of vectorial data, as is typically the case in signal processing applications.
- In a wireless sensor node, such a DSP could be used to process data coming from a simple analog, wireless communication device to extract a digital data stream. In broadband wireless communication, DSPs are an appropriate and successfully used platform.
- An FPGA can be reprogrammed (or rather reconfigured) "in the field" to adapt to a changing set of requirements; however, this can take time and energy – it is not practical to reprogram an FPGA at the same frequency as a microcontroller could change between different programs.
- An ASIC is a specialized processor, custom designed for a given application such as, for example, high-speed routers and switches. The typical trade-off here is loss of flexibility in return for a considerably better energy efficiency and performance.
- On the other hand, where a microcontroller requires software development, ASICs provide the same functionality in hardware, resulting in potentially more costly hardware development. Examples: Intel Strong ARM, Texas Instruments MSP 430, Atmel ATmega.

2.5.1.2 Memory

- Some memory to store programs and intermediate data; usually, different types of memory are used for programs and data.

- In WSN there is a need for Random Access Memory (RAM) to store intermediate sensor readings, packets from other nodes, and so on. While RAM is fast, its main disadvantage is that it loses its content if power supply is interrupted.
- Program code can be stored in Read-Only Memory (ROM) or, more typically, in Electrically Erasable Programmable Read-Only Memory (EEPROM) or flash memory.
- Flash memory can also serve as intermediate storage of data in case RAM is insufficient or when the power supply of RAM should be shut down for some time.

2.5.1.3 Sensors and Actuators

- The actual interface to the physical world: devices that can observe or control physical parameters of the environment.

Sensors

- Sensors can be roughly categorized into three categories as
  - **Passive, omnidirectional sensors:** These sensors can measure a physical quantity at the point of the sensor node without actually manipulating the environment by active probing – in this sense, they are passive. Moreover, some of these sensors actually are self-powered in the sense that they obtain the energy they need from the environment – energy is only needed to amplify their analog signal.
  - **Passive, narrow-beam sensors:** These sensors are passive as well, but have a well-defined notion of direction of measurement.
  - **Active sensors:** This last group of sensors actively probes the environment, for example, a sonar or radar sensor or some types of seismic sensors, which generate shock waves by small explosions. These are quite specific – triggering an explosion is certainly not a lightly undertaken action – and require quite special attention.

Actuators

- Actuators are just about as diverse as sensors, yet for the purposes of designing a WSN that converts electrical signals into physical phenomenon.

2.5.1.4 Communication Device

- Turning nodes into a network requires a device for sending and receiving information over a wireless channel.

- **Choice of transmission medium:** The communication device is used to exchange data between individual nodes. In some cases, wired communication can actually be the method of choice and is frequently applied in many sensor networks. The case of wireless communication is considerably more interesting because it include radio frequencies. Radio Frequency (RF) - based communication is by far the most relevant one as it best fits the requirements of most WSN applications.

- **Transceivers:** For Communication, both transmitter and receiver are required in a sensor node to convert a bit stream coming from a microcontroller and convert them to and from radio waves. For two tasks a combined device called transceiver is used. Transceiver structure has two parts as Radio Frequency (RF) front end (Figure 2.3) and the baseband part.

  - A) The radio frequency front end performs analog signal processing in the actual radio frequency Band.

  - B) The baseband processor performs all signal processing in the digital domain and communicates with a sensor node's processor or other digital circuitry.

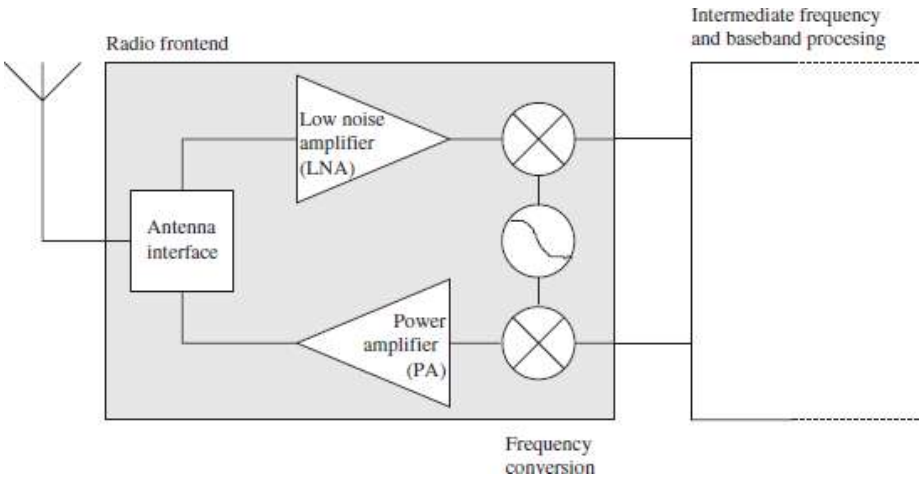


Figure 2.3 RF Front end

Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

- **The Power Amplifier (PA):** It accepts upconverted signals from the IF or baseband part and amplifies them for transmission over the antenna.

- **The Low Noise Amplifier (LNA):** It amplifies incoming signals up to levels suitable for further processing without significantly reducing the SNR. The range of powers of

the incoming signals varies from very weak signals from nodes close to the reception boundary to strong signals from nearby nodes; this range can be up to 100 dB.

- **Elements like local oscillators or voltage-controlled oscillators and mixers**, are used for frequency conversion from the RF spectrum to intermediate frequencies or to the baseband. The incoming signal at RF frequencies f_{RF} is multiplied in a mixer with a fixed frequency signal from the local oscillator (frequency f_{LO}). The resulting intermediate frequency signal has frequency $f_{LO} - f_{RF}$. Depending on the RF front end architecture, other elements like filters are also present.

  - **Transceiver tasks and characteristics**

    - **Service to upper layer:** A receiver has to offer certain services to the upper layers, most notably to the Medium Access Control (MAC) layer. Sometimes, this service is packet oriented; sometimes, a transceiver only provides a byte interface or even only a bit interface to the microcontroller.

    - **Power consumption and energy efficiency:** The simplest interpretation of energy efficiency is the energy required to transmit and receive a single bit.

    - **Carrier frequency and multiple channels:** Transceivers are available for different carrier frequencies; evidently, it must match application requirements and regulatory restrictions.

    - **State change times and energy:** A transceiver can operate in different modes: sending or receiving, use different channels, or be in different power-safe states.

    - **Data rates:** Carrier frequency and used bandwidth together with modulation and coding determine the gross data rate.

    - **Modulations:** The transceivers typically support one or several of on/off-keying, ASK, FSK, or similar modulations.

    - **Coding:** Some transceivers allow various coding schemes to be selected.

    - **Transmission power control:** Some transceivers can directly provide control over the transmission power to be used; some require some external circuitry for that purpose. Usually, only a discrete number of power levels are available from which the actual transmission power can be chosen. Maximum output power is usually determined by regulations.

- **Noise figure:** The noise figure NF of an element is defined as the ratio of the Signal-to-Noise Ratio (SNR) ratio SNR_I at the input of the element to the SNR ratio SNR_O at the element's output:

$$\mathbf{NF = SNR_I / SNR_O}$$

It describes the degradation of SNR due to the element's operation and is typically given in dB: NF dB= SNR_I dB − SNR_O dB.

- **Gain:** The gain is the ratio of the output signal power to the input signal power and is typically given in dB. Amplifiers with high gain are desirable to achieve good energy efficiency.
- **Power efficiency:** The efficiency of the radio front end is given as the ratio of the radiated power to the overall power consumed by the front end; for a power amplifier, the efficiency describes the ratio of the output signal's power to the power consumed by the overall power amplifier.
- **Receiver sensitivity:** The receiver sensitivity (given in dBm) specifies the minimum signal power at the receiver needed to achieve a prescribed E_b/N_0 or a prescribed bit/packet error rate.
- **Range:** The range of a transmitter is clear. The range is considered in absence of interference; it evidently depends on the maximum transmission power, on the antenna characteristics.
- **Blocking performance:** The blocking performance of a receiver is its achieved bit error rate in the presence of an interferer.
- **Out of band emission:** The inverse to adjacent channel suppression is the out of band emission of a transmitter. To limit disturbance of other systems, or of the WSN itself in a multichannel setup, the transmitter should produce as little as possible of transmission power outside of its prescribed bandwidth, centered around the carrier frequency.
- **Carrier sense and RSSI:** In many medium access control protocols, sensing whether the wireless channel, the carrier, is busy (another node is transmitting) is a critical information. The receiver has to be able to provide that information. The signal strength at which an incoming data packet has been received can provide useful information a receiver has to provide this information in the Received Signal Strength Indicator (RSSI).

- **Frequency stability:** The frequency stability denotes the degree of variation from nominal center frequencies when environmental conditions of oscillators like temperature or pressure change.
- **Voltage range:** Transceivers should operate reliably over a range of supply voltages. Otherwise, inefficient voltage stabilization circuitry is required.

2.5.1.5 Power Supply Unit

➢ The batteries are necessary to provide energy. Sometimes, some form of recharging by obtaining energy from the environment is available as well (e.g. solar cells). There are essentially two aspects:

- **Storing Energy**
- **Energy Scavenging**

Storing Energy: Batteries

➢ **Traditional batteries:** The power source of a sensor node is a battery, either non-rechargeable ("primary batteries") or, if an energy scavenging device is present on the node, also rechargeable ("secondary batteries").

Table 2.2 Energy densities for various primary and secondary battery types

Primary batteries			
Chemistry	Zinc-air	Lithium	Alkaline
Energy (J/cm^3)	3780	2880	1200
Secondary batteries			
Chemistry	Lithium	NiMHd	NiCd
Energy (J/cm^3)	1080	860	650

Upon these batteries the requirements are

- **Capacity:** They should have high capacity at a small weight, small volume, and low price. The main metric is energy per volume, J/cm^3.

- **Capacity under load:** They should withstand various usage patterns as a sensor node can consume quite different levels of power over time and actually draw high current in certain operation modes.

- **Self-discharge:** Their self-discharge should be low. Zinc-air batteries, for example, have only a very short lifetime (on the order of weeks).

- **Efficient recharging:** Recharging should be efficient even at low and intermittently available recharge power.

- **Relaxation:** Their relaxation effect – the seeming self-recharging of an empty or almost empty battery when no current is drawn from it, based on chemical diffusion processes within the cell – should be clearly understood. Battery lifetime and usable capacity is considerably extended if this effect is leveraged.

- **DC–DC Conversion:** Unfortunately, batteries alone are not sufficient as a direct power source for a sensor node. One typical problem is the reduction of a battery's voltage as its capacity drops. A DC – DC converter can be used to overcome this problem by regulating the voltage delivered to the node's circuitry. To ensure a constant voltage even though the battery's supply voltage drops, the DC – DC converter has to draw increasingly higher current from the battery when the battery is already becoming weak, speeding up battery death. The DC – DC converter does consume energy for its own operation, reducing overall efficiency.

Energy Scavenging

➢ Depending on application, high capacity batteries that last for long times, that is, have only a negligible self-discharge rate, and that can efficiently provide small amounts of current. Ideally, a sensor node also has a device for energy scavenging, recharging the battery with energy gathered from the environment – solar cells or vibration-based power generation are conceivable options.

- **Photovoltaics:** The well-known solar cells can be used to power sensor nodes. The available power depends on whether nodes are used outdoors or indoors, and on time of day and whether for outdoor usage. The resulting power is somewhere between 10 $\mu W/cm^2$ indoors and 15 mW/cm^2 outdoors. Single cells achieve a fairly stable output voltage of about 0.6 V (and have therefore to be used in series) as long as the drawn current does not exceed a critical threshold, which depends on the light intensity. Hence, solar cells are usually used to recharge secondary batteries.

- **Temperature gradients:** Differences in temperature can be directly converted to electrical energy.
- **Vibrations:** One almost pervasive form of mechanical energy is vibrations: walls or windows in buildings are resonating with cars or trucks passing in the streets, machinery often has low frequency vibrations. Both amplitude and frequency of the vibration and ranges from about 0.1 $\mu W/cm^3$ up to 10, 000 $\mu W/cm^3$ for some extreme cases.
- **Pressure variations:** Somewhat akin to vibrations, a variation of pressure can also be used as a power source.
- **Flow of air/liquid:** Another often-used power source is the flow of air or liquid in wind mills or turbines. The challenge here is again the miniaturization, but some of the work on millimeter scale MEMS gas turbines might be reusable.

Table 2.3 Comparison of Energy Sources

Energy source	Energy density
Batteries (zinc-air)	1050–1560 mWh/cm^3
Batteries (rechargeable lithium)	300 mWh/cm^3 (at 3–4 V)
Energy source	**Power density**
Solar (outdoors)	15 mW/cm^2 (direct sun)
	0.15 mW/cm^2 (cloudy day)
Solar (indoors)	0.006 mW/cm^2 (standard office desk)
	0.57 mW/cm^2 (<60 W desk lamp)
Vibrations	0.01–0.1 mW/cm^3
Acoustic noise	$3 \cdot 10^{-6}$ mW/cm^2 at 75 dB
	$9,6 \cdot 10^{-4}$ mW/cm^2 at 100 dB
Passive human-powered systems	1.8 mW (shoe inserts)
Nuclear reaction	80 mW/cm^3, 10^6 mWh/cm^3

2.5.2 Energy Consumption of Sensor Nodes

- Energy efficiency is the key requirement to maximize sensor node lifetime. Sensor nodes are typically powered by a battery source that has finite lifetime.
- Hence, the energy consumption of a sensor node must be tightly controlled. The main consumers of energy are the controller, the radio front ends, the memory, and type of the sensors.
- One method to reduce power consumption of these components is designing low-power chips, it is the best starting point for an energy-efficient sensor node. But any advantages gained by such designs can easily be squandered/ wasted when the components are improperly operated.
- Second method for energy efficiency in wireless sensor node is reduced functionality by using multiple states of operation with reduced energy consumption. These modes can be introduced for all components of a sensor node, in particular, for controller, radio front end, memory, and sensors.

2.5.2.1 Microcontroller Energy Consumption

- For a controller, typical states are "active", "idle", and "sleep". A radio modem could turn transmitter, receiver, or both on or off. At time t_1, the microcontroller is to be put into sleep mode should be taken to reduce power consumption from P_{active} to P_{sleep}.
- If it remains active and the next event occurs at time t_{event}, then a total energy is $E_{active} = P_{active} (t_{event} - t_1)$. On the other hand, requires a time τ_{down} until sleep mode has been reached.
- Let the average power consumption during this phase is $(P_{active} + P_{sleep})/2$. Then, P_{sleep} is consumed until t_{event}. The energy saving is given by

$$E_{saved} = (t_{event} - t_1)P_{active} - (\tau_{down}(P_{active} + P_{sleep})/2 + (t_{event} - t_1 - \tau_{down})P_{sleep})$$

- Once the event to be processed occurs, however, an additional overhead of

$$E_{overhead} = \tau_{Up} (P_{active} + P_{sleep})/2$$

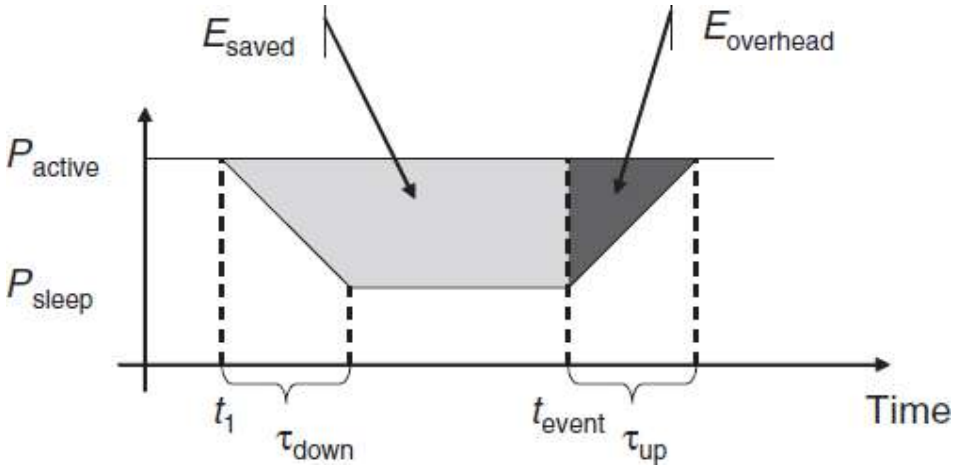


Figure 2.4 Energy Savings and Overheads for Sleep Modes
Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

- Switching to a sleep mode is only beneficial if $E_{overhead} < E_{saved}$ or, equivalently, if the time to the next event is sufficiently large:

$$(t_{event} - t_1) > \frac{1}{2}\left(\tau_{down} + \frac{P_{active} + P_{sleep}}{P_{active} - P_{sleep}}\tau_{up}\right)$$

Examples:

1. **Intel StrongARM** : The Intel StrongARM provides three sleep modes:

   - **In normal mode,** all parts of the processor are fully powered. Power consumption is up to 400 mW.
   - **In idle mode,** clocks to the CPU are stopped; clocks that pertain to peripherals are active. Any interrupt will cause return to normal mode. Power consumption is up to 100 mW.
   - **In sleep mode,** only the real-time clock remains active. Wakeup occurs after a timer interrupt and takes up to 160 ms. Power consumption is up to 50 µW.

2. **Texas Instruments MSP 430** : The MSP430 family features a wider range of operation modes: One fully operational mode, which consumes about 1.2 mW (all power values given at 1 MHz and 3 V). There are four sleep modes in total. The deepest sleep mode, LPM4, only consumes 0.3 µW, but the controller is only woken

up by external interrupts in this mode. In the next higher mode, LPM3, a clock is also till running, which can be used for scheduled wake ups, and still consumes only about 6 µW.

3. **Atmel ATmega** : The Atmel ATmega 128L has six different modes of power consumption, which are in principle similar to the MSP 430 but differ in some details. Its power consumption varies between 6 mW and 15 mW in idle and active modes and is about 75 µW in power-down modes.

2.5.2.2 Memory Energy Consumption

- The most relevant kinds of memory are on-chip memory and FLASH memory. Off-chip RAM is rarely used. In fact, the power needed to drive on-chip memory is usually included in the power consumption numbers given for the controllers.

- Hence, the most relevant part is FLASH memory. In fact, the construction and usage of FLASH memory can heavily influence node lifetime. The relevant metrics are the read and write times and energy consumption. Read times and read energy consumption tend to be quite similar between different types of FLASH memory. Energy consumption necessary for reading and writing to the Flash memory is used on the Mica nodes. Hence, writing to FLASH memory can be a time- and energy-consuming task that is best avoided if somehow possible.

2.5.2.3 Radio Transceivers Energy Consumption

- A radio transceiver has essentially two tasks: transmitting and receiving data between a pair of nodes. Similar to microcontrollers, radio transceivers can operate in different modes, the simplest ones are being turned on or turned off.

- To accommodate the necessary low total energy consumption, the transceivers should be turned off most of the time and only be activated when necessary – they work at a low duty cycle.

- The energy consumed by a transmitter is due to two sources one part is due to RF signal generation, which mostly depends on chosen modulation and target distance. Second part is due to electronic components necessary for frequency synthesis, frequency conversion, filters, and so on.

- The transmitted power is generated by the amplifier of a transmitter. Its own power consumption P_{amp} depends on its architecture $P_{amp} = \alpha_{amp} + \beta_{amp}P_{tx}$, where α_{amp} and β_{amp} are constants depending on process technology and amplifier architecture. The energy to transmit a packet n-bits long (including all headers) then depends on how

long it takes to send the packet, determined by the nominal bit rate R and the coding rate R_{code}, and on the total consumed power during transmission.

$$E_{tx}(n, R_{code}, P_{amp}) = T_{start} P_{start} + \frac{n}{R R_{code}} (P_{txElec} + P_{amp})$$

- Similar to the transmitter, the receiver can be either turned off or turned on. While being turned on, it can either actively receive a packet or can be idle, observing the channel and ready to receive. Evidently, the power consumption while it is turned off is negligible.
- Even the difference between idling and actually receiving is very small and can, for most purposes, be assumed to be zero. To elucidate, the energy E_{rcvd} required to receive a packet has a startup component T_{start} P_{start} similar to the transmission case when the receiver had been turned off (startup times are considered equal for transmission and receiving here); it also has a component that is proportional to the packet time n / RR_{code}.
- During this time of actual reception, receiver circuitry has to be powered up, requiring a (more or less constant) power of P_{rxElec}.

$$E_{rcvd} = T_{start} P_{start} + \frac{n}{R R_{code}} P_{rxElec} + n E_{decBit}$$

2.5.2.4 Power Consumption of Sensor and Actuators

- Providing any guidelines about the power consumption of the actual sensors and actuators is impossible because of the wide variety of these devices.
- For example, passive light or temperature sensors – the power consumption can possibly be ignored in comparison to other devices on a wireless node. For others, active devices like sonar, power consumption can be quite considerable in the dimensioning of power sources on the sensor node, not to overstress batteries.

2.6 *Network Architecture*

- It introduces the basic principles of turning individual sensor nodes into a wireless sensor network.

2.6.1 Sensor Network Scenarios

2.6.1.1 Types of Sources and Sinks

- Source is any unit in the network that can provide information (sensor node). A sink is the unit where information is required, it could belong to the sensor network or outside this network to interact with another network or a gateway to another larger Internet**.** Sinks are illustrated by Figure 2.5, showing sources and sinks in direct communication.

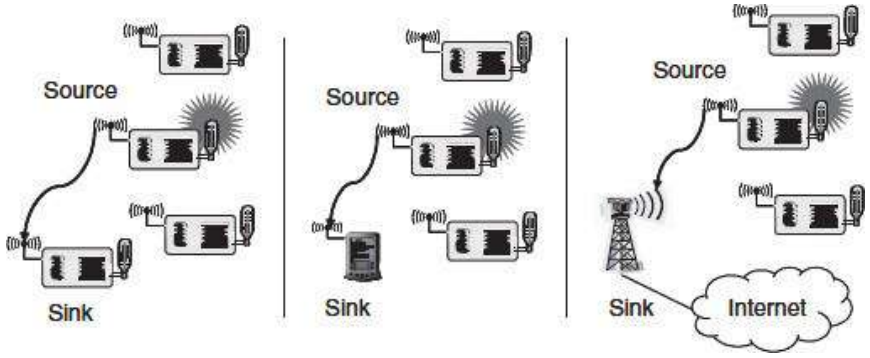


Figure 2.5 Three types of sinks in a very simple, single-hop sensor network
Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

2.6.1.2 Single-Hop versus Multi-Hop Networks

- Because of limited distance the direct communication between source and sink is not always possible.
- In WSNs, to cover a lot of environment the data packets taking multi hops from source to the sink. To overcome such limited distances it better to use relay stations. The data packets taking multi hops from source to the sink as shown in Figure 2.6.
- Depending on the particular application of having an intermediate sensor node at the right place is high.

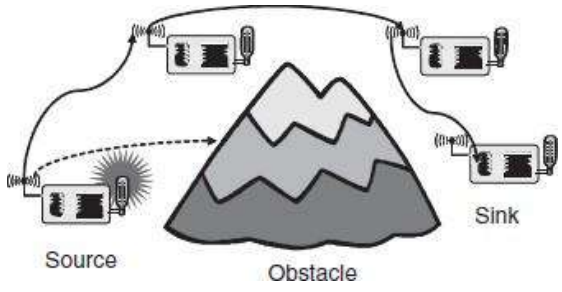


Figure 2.6 Multi-hop networks: As direct communication is impossible because of distance and/or obstacles

Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

- Multi-hopping also to improves the energy efficiency of communication as it consumes less energy to use relays instead of direct communication, the radiated energy required for direct communication over a distance d is cd^{α} (c some constant, $\alpha \geq 2$ the path loss coefficient) and using a relay at distance d/2 reduces this energy to $2c(d/2)^{\alpha}$

- This calculation considers only the radiated energy. It should be pointed out that only multi-hop networks operating in a store and forward fashion are considered here. In such a network, a node has to correctly receive a packet before it can forward it somewhere.

2.6.1.3 Multiple Sinks and Sources

- In many cases, multiple sources and multiple sinks present. Multiple sources should send information to multiple sinks. Either all or some of the information has to reach all or some of the sinks. This is illustrated in figure 2.7.

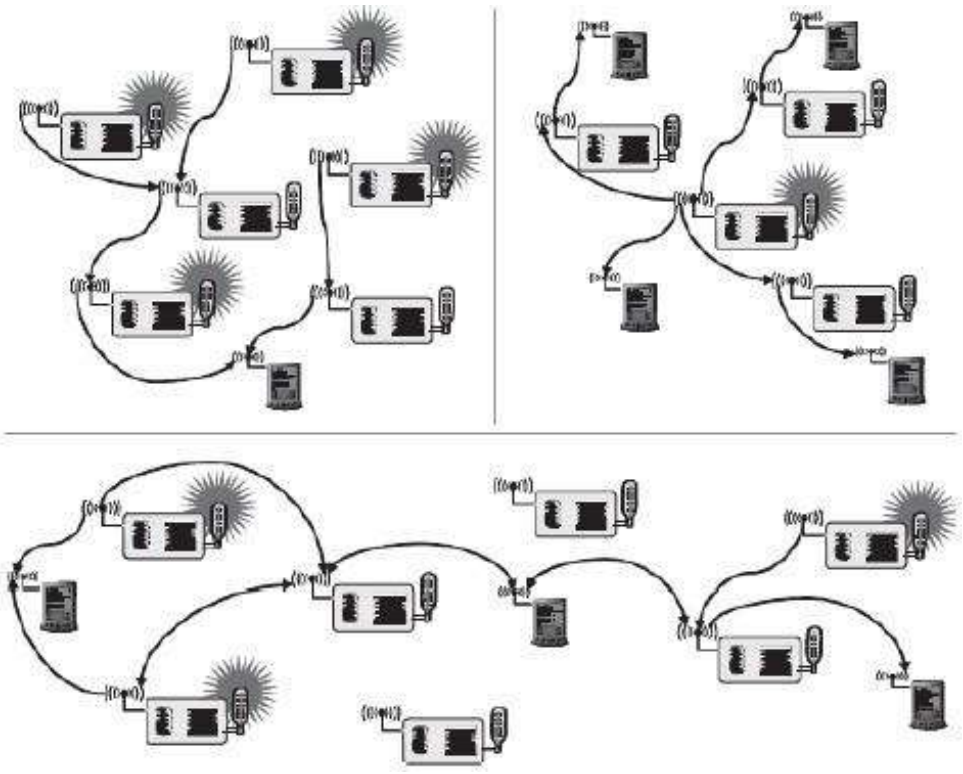


Figure 2.7 Multiple sources and/or multiple sinks.

Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

2.6.1.4 Three types of mobility

- In the scenarios discussed above, all participants were stationary. But one of the main virtues of wireless communication is its ability to support mobile participants.
- In wireless sensor networks, mobility can appear in three main forms
  - Node mobility
  - Sink mobility
  - Event mobility

Node Mobility

- The wireless sensor nodes themselves can be mobile. The meaning of such mobility is highly application dependent. In examples like environmental control, node mobility should not happen; in livestock surveillance (sensor nodes attached to cattle, for example), it is the common rule. In the face of node mobility, the network has to reorganize to function correctly.

Sink Mobility

- The information sinks can be mobile. For example, a human user requested information via a PDA while walking in an intelligent building (Figure 2.8)
- In a simple case, such a requester can interact with the WSN at one point and complete its interactions before moving on. In many cases, consecutive interactions can be treated as separate, unrelated requests.

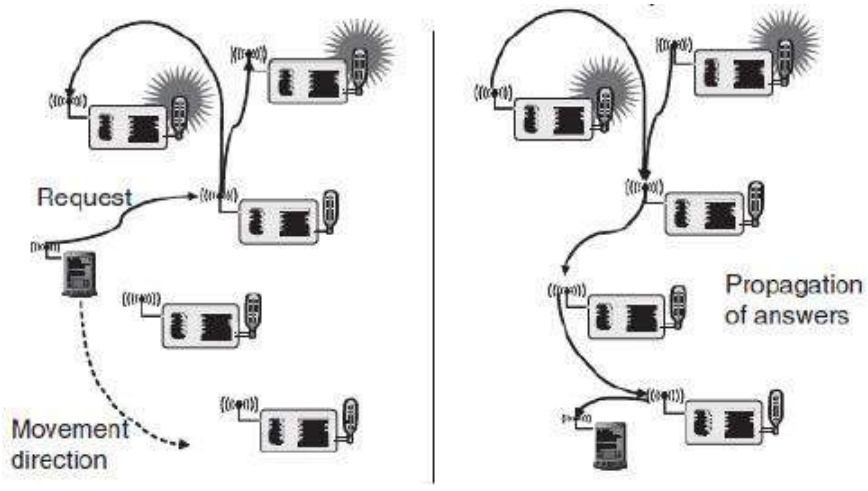
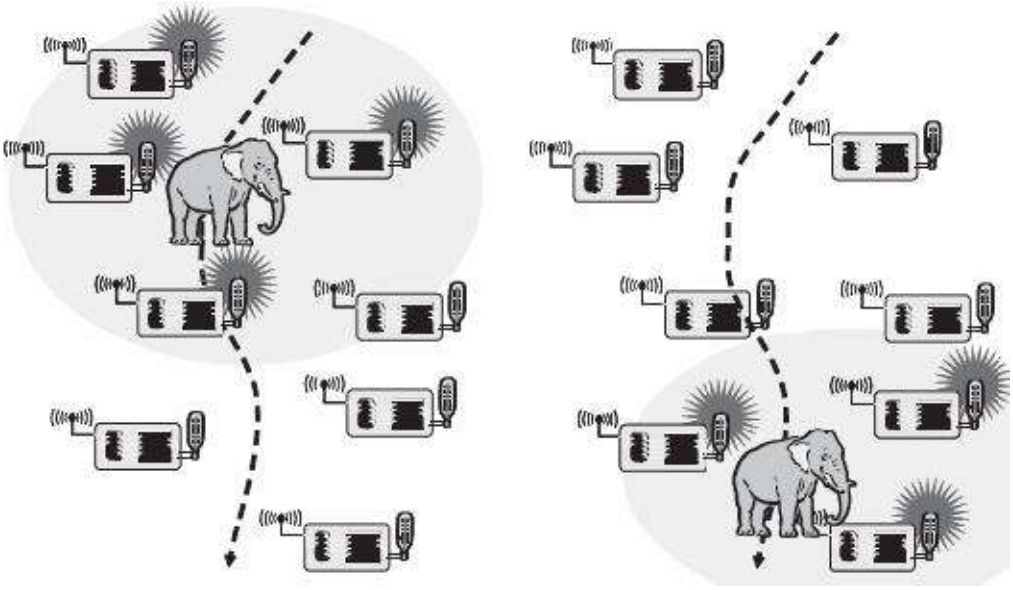


Figure2.8 Sink mobility: A mobile sink moves through a sensor network as information is being retrieved *on its behalf*

Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

Event Mobility

- In tracking applications, the cause of the events or the objects to be tracked can be mobile. In such scenarios, it is (usually) important that the observed event is covered by a sufficient number of sensors at all time. As the event source moves through the network, it is accompanied by an area of activity within the network – this has been called the frisbee model. This notion is described by Figure 2.9, where the task is to detect a moving elephant and to observe it as it moves around



Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

Figure 2.9 Area of sensor nodes detecting an event – an elephant– that moves through the network along with the event source (dashed line indicate the elephant's trajectory; shaded ellipse the activity area following or even preceding the elephant)

2.6.2 Transceiver Design Considerations

- Some of the most crucial points influencing PHY design in wireless sensor networks are:

  - Low power consumption.
  - As one consequence: small transmit power and thus a small transmission range.
  - As a further consequence: low duty cycle. Most hardware should be switched off or operated in a low-power standby mode most of the time.

- Comparably low data rates, on the order of tens to hundreds kilobits per second, required.
- Low implementation complexity and costs.
- Low degree of mobility.
- A small form factor for the overall node.

2.6.3 Optimization Goals and Figures of Merit

➢ For all WSN scenarios and application types have to face the challenges such as

- How to optimize a network and How to compare these solutions?
- How to decide which approach is better?
- How to turn relatively inaccurate optimization goals into measurable figures of merit?

➢ For all the above questions the general answer is obtained from

- Quality of service
- Energy efficiency
- Scalability
- Robustness

2.6.3.1 Quality of Service

➢ WSNs differ from other conventional communication networks in the type of service they offer. These networks essentially only move bits from one place to another. Some generic possibilities are

- **Event detection/reporting probability:** The probability that an event that actually occurred is not detected or not reported to an information sink that is interested in such an event For example, not reporting a fire alarm to a surveillance station would be a severe shortcoming.
- **Event classification error:** If events are not only to be detected but also to be classified, the error in classification must be small

- **Event detection delay:** It is the delay between detecting an event and reporting it to any/all interested sinks.
- **Missing reports:** In applications that require periodic reporting, the probability of undelivered reports should be small.
- **Approximation accuracy:** For function approximation applications, the average/maximum absolute or relative error with respect to the actual function.
- **Tracking accuracy:** Tracking applications must not miss an object to be tracked, the reported position should be as close to the real position as possible, and the error should be small.

2.6.3.2 Energy Efficiency

➢ **E**nergy efficiency should be optimization goal. The most commonly considered aspects are:

- **Energy per correctly received bit:** How much energy is spent on average to transport one bit of information (payload) from the transmitter to the receiver.
- **Energy per reported (unique) event:** What is the average energy spent to report one event.
- **Delay/energy trade-offs:** "urgent" events increases energy investment for a speedy reporting events. Here, the trade-off between delay and energy overhead is interesting.
- **Network lifetime:** The time for which the network is operational.
- **Time to first node death:** When does the first node in the network run out of energy or fail and stop operating?
- **Network half-life:** When have 50 % of the nodes run out of energy and stopped operating.
- **Time to partition:** When does the first partition of the network in two (or more) disconnected parts occur?
- **Time to loss of coverage:** The time when for the first time any spot in the deployment region is no longer covered by any node's observations.

- **Time to failure of first event notification:** A network partition can be seen as irrelevant if the unreachable part of the network does not want to report any events in the first place.

2.6.3.3 Scalability

- The ability to maintain performance characteristics irrespective of the size of the network is referred to as scalability. With WSN potentially consisting of thousands of nodes, scalability is an obviously essential requirement.
- The need for extreme scalability has direct consequences for the protocol design. Often, a penalty in performance or complexity has to be paid for small networks.
- The architectures and protocols should implement appropriate scalability support rather than trying to be as scalable as possible. Applications with a few dozen nodes might admit more-efficient solutions than applications with thousands of nodes.

2.6.3.4 Robustness

- Wireless sensor networks should also exhibit an appropriate robustness.

They should not fail just because a limited number of nodes run out of energy, or because their environment changes and severs existing radio links between two nodes. If possible, these failures have to be compensated by finding other routes

UNIT III
WSN NETWORKING CONCEPTS AND PROTOCOLS

MAC Protocols for Wireless Sensor Networks, Low Duty Cycle Protocols And Wakeup Concepts - S-MAC, The Mediation Device Protocol, Contention based protocols - PAMAS, Schedule based protocols – LEACH, IEEE 802.15.4 MAC protocol, Routing Protocols- Energy Efficient Routing, Challenges and Issues in Transport layer protocol.

MAC protocols for wireless sensor networks

1. Explain the design considerations for MAC protocols in wireless sensor networks.

Balance of requirements

- The importance of energy efficiency for the design of MAC protocols is relatively new and many of the "classical" protocols like ALOHA and CSMA contain no provisions toward this goal.
- Other typical performance figures like fairness, throughput, or delay tend to play a minor role in sensor networks.
- Further important requirements for MAC protocols are scalability and robustness against frequent topology changes.
- It is caused by nodes powering down temporarily to replenish their batteries by energy scavenging, mobility, deployment of new nodes, or death of existing nodes.

Energy problems on the MAC layer

- A nodes transceiver consumes a significant share of energy.
- The transceiver has four main states: transmitting, receiving, idling, or sleeping.
- Transmitting is costly, receive costs often have the same order of magnitude as transmit costs, idling can be significantly cheaper but also about as expensive as receiving, and sleeping costs almost nothing but results in a "deaf" node.
- Some **energy problems** and design goals are mentioned below:

Collisions

- Collisions incur useless receive costs at the destination node, useless transmit costs at the source node, and the prospect to expend further energy upon packet retransmission.

- Hence, collisions should be avoided, either by design (fixed assignment/TDMA or demand assignment protocols) or by appropriate collision avoidance/hidden-terminal procedures in CSMA protocols.

Overhearing

- Unicast frames have one source and one destination node.
- However, the wireless medium is a broadcast medium and all the source's neighbors that are in receive state hear a packet and drop it when it is not destined to them; these nodes overhear the packet.

Protocol overhead

- Protocol overhead is induced by MAC-related control frames like, RTS and CTS packets or request packets in demand assignment protocols.

Idle listening

- A node being in idle state is ready to receive a packet but is not currently receiving anything.
- This readiness is costly and useless in case of low network loads; the idle state still consumes significant energy.
- Switching off the transceiver is a solution
- A design constraint somewhat related to energy concerns is the requirement for **low complexity operation**.
- Sensor nodes shall be simple and cheap and cannot offer plentiful resources in terms of processing power, memory, or energy.
- Therefore, computationally expensive operations like complex scheduling algorithms should be avoided.

****

Low duty cycle protocols and wakeup concepts

2. Explain about Low duty protocols in WSN with neat diagram.

- **Low duty cycle protocols** try to avoid spending time in the idle state and to reduce the communication activities of a sensor node to a minimum.
- In an ideal case, the sleep state is left only when a node is about to transmit or receive packets.
- A concept for achieving this is called wakeup radio.
- In several protocols, a **periodic wakeup** scheme is used. Such schemes exist in different flavors. One is the **cycled receiver** approach is illustrated in below Figure.

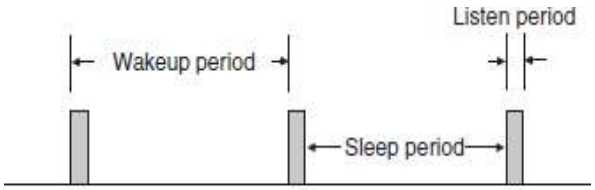


Figure 5.4 Periodic wakeup scheme

- In this approach, nodes spend most of their time in the sleep mode and wake up periodically to *receive* packets from other nodes.
- Specifically, a node *A* listens onto the channel during its **listen period** and goes back into sleep mode when no other node takes the opportunity to direct a packet to *A.*
- A potential transmitter *B* must acquire knowledge about *A*'s listen periods to send its packet at the right time – this task corresponds to a *rendezvous.*
- This rendezvous can be accomplished by letting node *A* transmit a short beacon at the beginning of its listen period to indicate its willingness to receive packets.
- Another method is to let node *B* send frequent request packets until one of them hits *A*'s listen period and is really answered by *A*.
- However, in either case, node *A* only *receives* packets during its listen period.
- If node *A* itself wants to transmit packets, it must acquire the target's listen period.
- A whole cycle consisting of sleep period and listen period is also called a **wakeup period**.
- The ratio of the listen period length to the wakeup period length is also called the node's **duty cycle**.

- By choosing a small duty cycle, the transceiver is in sleep mode most of the time, avoiding idle listening and conserving energy.
- By choosing a small duty cycle, the traffic directed from neighboring nodes to a given node concentrates on a small time window (the listen period) and in heavy load situations significant competition can occur.
- Choosing a long sleep period induces significant **per-hop latency.** In the multihop case, the per-hop latencies add up and create significant end-to-end latencies.
- Sleep phases should not be too short lest the start-up costs outweigh the benefits.
- In other protocols like S-MAC, there is also a periodic wakeup but nodes can both *transmit and receive* during their wakeup phases.
- When nodes have their wakeup phases at the same time, there is no necessity for a node wanting to transmit a packet to be awake *outside* these phases to rendezvous its receiver.

****

S-MAC

3. Explain about S-MAC protocol in WSN with neat diagram.

- The S-MAC (Sensor-MAC) protocol provides mechanisms to circumvent idle listening, collisions, and overhearing.
- S-MAC adopts a periodic wakeup scheme, that is, each node alternates between a fixed-length listen period and a fixed-length sleep period according to its **schedule.**
- The listen period of S-MAC can be used to receive *and transmit* packets.
- S-MAC attempts to coordinate the schedules of neighboring nodes such that their listen periods start at the same time.

Phases in listen period:

- A node *x*'s listen period is subdivided into three different phases:

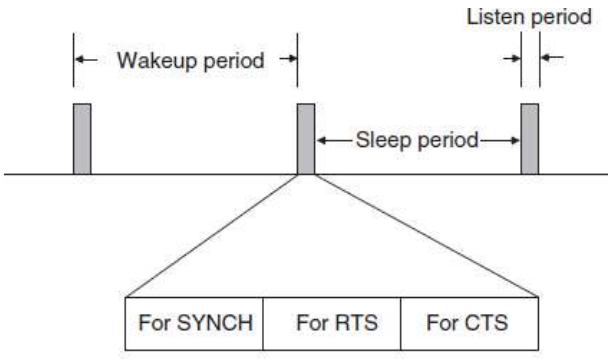


Figure 5.6 S-MAC principle

1. ***first phase***
   - In the first phase (**SYNCH phase**), node *x* accepts SYNCH packets from its neighbors.
   - In these packets, the neighbors describe their own schedule and *x* stores their schedule in a table (the **schedule table**).
   - Node *x*'s SYNCH phase is subdivided into time slots and *x*'s neighbors contend according to a CSMA scheme with additional backoff.
   - Each neighbor *y* wishing to transmit a SYNCH packet picks one of the time slots randomly and starts to transmit if no signal was received in any of the previous slots.
   - In the other case, *y* goes back into sleep mode and waits for *x*'s next wakeup.
   - In the other direction, since *x* knows a neighbor *y*'s schedule, *x* can wake at appropriate times and send its own SYNCH packet to *y* (in broadcast mode).
   - It is not required that *x* broadcasts its schedule in every of *y*'s wakeup periods.
   - However, for reasons of time synchronization and to allow new nodes to learn their local network topology, *x* should send SYNCH packets periodically. The according period is called **synchronization period**.

2. ***Second phase***
   - In the second phase (**RTS phase**), *x* listens for RTS packets from neighboring nodes.
   - In S-MAC, the RTS/CTS handshake is used to reduce collisions of data packets due to hidden-terminal situations.
   - Again, interested neighbors contend in this phase according to a CSMA scheme with additional backoff.

3. ***Third Phase***
   - In the third phase (**CTS phase**), node *x* transmits a CTS packet if an RTS packet was received in the previous phase. After this, the packet exchange continues, extending into *x*'s nominal sleep time.

<u>Working of S-MAC Protocol</u>

- When competing for the medium, the nodes use the RTS/CTS handshake, including the virtual carrier-sense mechanism.
- When transmitting in a broadcast mode (for example SYNCH packets), the RTS and CTS packets are dropped and the nodes use CSMA with backoff.

- If we can arrange that the schedules of node *x* and its neighbors are synchronized, node *x* and all its neighbors wake up at the same time and *x* can reach all of them with a single SYNCH packet.
- The S-MAC protocol allows neighboring nodes to agree on the same schedule and to create **virtual clusters**.
- The clustering structure refers solely to the exchange of schedules; the transfer of data packets is not influenced by virtual clustering.
- The S-MAC protocol proceeds as follows to form the virtual clusters:
  - A node *x*, newly switched on, listens for a time of at least the synchronization period.
  - If *x* receives any SYNCH packet from a neighbor, it adopts the announced schedule and broadcasts it in one of the neighbors' next listen periods.
  - In the other case, node *x* picks a schedule and broadcasts it.
  - If *x* receives another node's schedule during the broadcast packet's contention period, it drops its own schedule and follows the other one.
  - It might also happen that a node *x* receives a different schedule after it already has chosen one, for example, because bit errors destroyed previous SYNCH packets.
  - If node *x* already knows about the existence of neighbors who adopted its own schedule, it keeps its schedule and in the future has to transmit its SYNCH and data packets according to both schedules.
  - On the other hand, if *x* has no neighbor sharing its schedule, it drops its own and adopts the other one.
  - Since there is always a chance to receive SYNCH packets in error, node *x* periodically listens for a whole synchronization period.

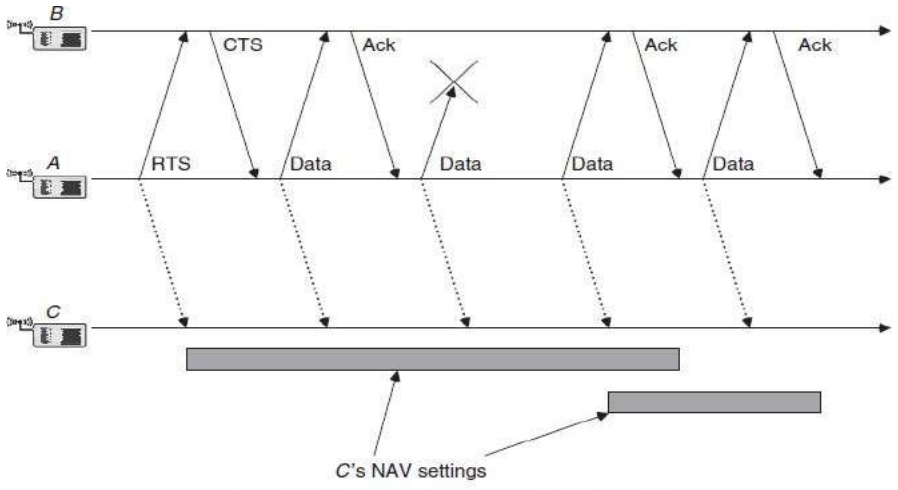


Figure 5.7 S-MAC fragmentation and NAV setting

S-MAC includes a fragmentation scheme

- ✓ A series of fragments is transmitted with only one RTS/CTS exchange between the transmitting node *A* and receiving node *B*.
- ✓ After each fragment, *B* has to answer with an acknowledgment packet.
- ✓ All the packets (data, ack, RTS, CTS) have a duration field and a neighboring node *C* is required to set its NAV field accordingly.
- ✓ In S-MAC, the duration field of all packets carries the remaining length of the whole transaction, including all fragments and their acknowledgments. Therefore, the whole message shall be passed at once.
- ✓ If one fragment needs to be retransmitted, the remaining duration is incremented by the length of a data plus ack packet, and the medium is reserved for this prolonged time.
- ✓ However, there is the problem of how a nonparticipating node shall learn about the elongation of the transaction when he has only heard the initial RTS or CTS packets.

Drawbacks:

- ✓ It is hard to adapt the length of the wakeup period to changing load situations, since this length is essentially fixed, as is the length of the listen period.

****

The mediation device protocol

4. **Explain the mediation device protocol with neat diagram.**

- ✓ The mediation device protocol is compatible with the peer-to-peer communication mode of the IEEE 802.15.4 low-rate WPAN standard.
- ✓ It allows each node in a WSN to go into sleep mode periodically and to wake up only for short times to receive packets from neighbor nodes.
- ✓ There is no global time reference, each node has its own sleeping schedule, and does not take care of its neighbors sleep schedules.
- ✓ Upon each periodic wakeup, a node transmits a short **query beacon**, indicating its node address and its willingness to accept packets from other nodes.
- ✓ The node stays awake for some short time following the query beacon, to open up a window for incoming packets.
- ✓ If no packet is received during this window, the node goes back into sleep mode.

- ✓ When a node wants to transmit a packet to a neighbor, it has to synchronize with it.
- ✓ The **dynamic synchronization** approach achieves this synchronization without requiring the transmitter to be awake permanently to detect the destinations query beacon.
- ✓ To achieve this, a **mediation device** (MD) is used.

Working of Mediation Device:

- ✓ Consider the scenario, mediation device is not energy constrained and can be active all the time.
- ✓ Because of its full duty cycle, the mediation device can receive the query beacons from all nodes in its vicinity and learn their wakeup periods.

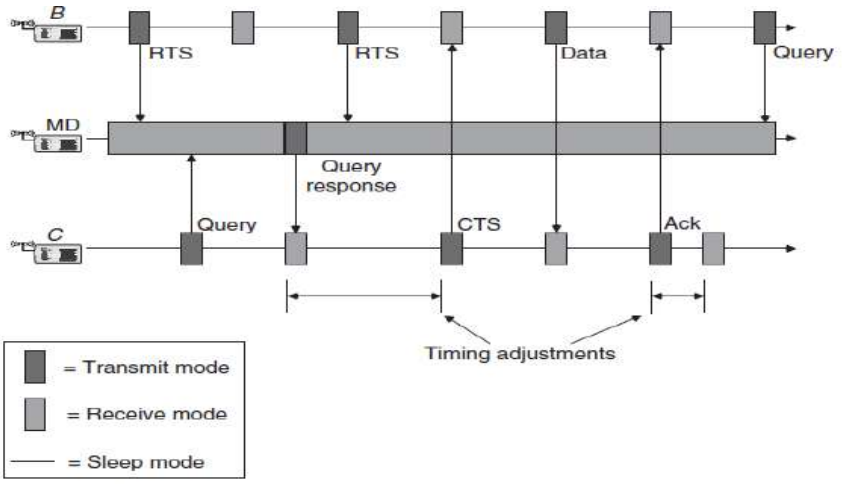


Figure 5.8 Mediation device protocol with unconstrained mediators [115, Chap. 4, Fig. 3]

- ❖ Suppose that node *A* wants to transmit a packet to node *B*.
- ❖ Node *A* announces this to the mediation device by sending periodically **request to send** (RTS) packets, which the MD captures.
- ❖ Node *A* sends its RTS packets instead of its query beacons and thus they have the same period.
- ❖ Again, there is a short answer window after the RTS packets, where *A* listens for answers.
- ❖ After the MD has received *A*'s RTS packet, it waits for *B*'s next query beacon.
- ❖ The MD answers this with a **query response** packet, indicating *A*'s address and a timing offset, which lets *B* know when to send the answering **clear to send** (CTS) to *A* such that the CTS packet hits the short answer window after *A*'s next RTS packet.
- ❖ Therefore, *B* has learned *A*'s period. After *A* has received the CTS packet, it can send its data packet and wait for *B*'s immediate acknowledgment.

- After the transaction has finished, *A* restores its periodic wakeup cycle and starts to emit query beacons again.
- Node *B* also restores its own periodic cycle and thus *decouples* from *A*'s period.

Advantages:

- It does not require any time synchronization between the nodes, only the mediation device has to learn the periods of the nodes.
- The protocol is asymmetric in the sense that most of the energy burden is shifted to the mediation device, which so far is assumed to be power unconstrained.
- The other nodes can be in the sleep state most of the time and have to spend energy only for the periodic beacons.

Drawbacks:

- The nodes transmit their query beacons without checking for ongoing transmissions and, thus, the beacons of different nodes may collide repeatedly when nodes have the same period and their wakeup periods overlap.
- However, in case of higher node densities or unwanted synchronization between the nodes, the number of collisions can be significant.

****

Wakeup radio concepts

5. **Explain the wakeup radio concepts in WSN.**

- If a node were always in the receiving state when a packet is transmitted to it, in the transmitting state when it transmits a packet, and in the sleep state at all other times; the idle state should be avoided.
- The **wakeup radio** concept strives to achieve this goal by a simple, "powerless" receiver that can trigger a main receiver if necessary.
- One proposed wakeup MAC protocol assumes the presence of several parallel data channels, separated either in frequency (FDMA) or by choosing different codes in a CDMA schemes.
- A node wishing to transmit a data packet randomly picks one of the channels and performs a carrier sensing operation.
- If the channel is busy, the node makes another random channel choice and repeats the carrier-sensing operation.

- After a certain number of unsuccessful trials, the node backs off for a random time and starts again.
- If the channel is idle, the node sends a wakeup signal to the intended receiver, indicating both the receiver identification and the channel to use.
- The receiver wakes up its data transceiver, tunes to the indicated channel, and the data packet transmission can proceed. Afterward, the receiver can switch its data transceiver back into sleep mode.
- It has the significant advantage that only the low-power wakeup transceiver has to be switched on all the time while the much more energy consuming data transceiver is nonsleeping if and only if the node is involved in data transmissions.
- Furthermore, this scheme is naturally **traffic adaptive**, that is, the MAC becomes more and more active as the traffic load increases.
- Periodic wakeup schemes do not have this property. However, there are also some drawbacks.
- *First*, there is no real hardware yet for such an ultralow power wakeup transceiver.
- *Second*, the range of the wakeup radio and the data radio should be the same.
- If the range of the wakeup radio is smaller than the range of the data radio, possibly not all neighbor nodes can be woken up.
- On the other hand, if the range of the wakeup radio is significantly larger, there can be a problem with local addressing schemes.
- These schemes do not use globally or network wide-unique addresses but only locally unique addresses, such that no node has two or more one-hop neighbors with the same address.
- Since the packets exchanged in the neighbor discovery phase have to use the data channel, the two hop neighborhood as seen on the data channel might be different from the two-hop neighborhood on the wakeup channel.
- *Third*, this scheme critically relies on the wakeup channel's ability to transport useful information like node addresses and channel identifications;
- This might not always be feasible for transceiver complexity reasons and additionally requires methods to handle collisions or transmission errors on the wakeup channel.
- If the wakeup channel does not support this feature, the transmitter wakes up *all* its neighbors when it emits a wakeup signal, creating an overhearing situation for most of them.

- If the transmitting node is about to transmit a long data packet, it might be worthwhile to prepend the data packet with a short **filter packet** announcing the receiving node's address.
- All the other nodes can go back to sleep mode after receiving the filter packet. Instead of using an extra packet, all nodes can read the bits of the data packet until the destination address appeared.
- If the packet's address is not identical to its own address, the node can go back into sleep mode.

****

Contention-based protocols

6. **Explain the contention based protocol PAMAS with neat diagram.**

- In contention-based protocols, a given transmit opportunity toward a receiver node can in principle be taken by any of its neighbors.
- If only one neighbor tries its luck, the packet goes through the channel.
- If two or more neighbors try their luck, these have to compete with each other and in unlucky cases due to hidden-terminal situations, a collision might occur, wasting energy for both transmitter and receiver.

PAMAS

- The PAMAS protocol (Power Aware Multiaccess with Signaling) originally designed for ad hoc networks.
- It provides a detailed overhearing avoidance mechanism while it does not consider the idle listening problem.
- The protocol combines the busy-tone solution and RTS/CTS handshake similar to the MACA protocol

Features of PAMAS:

- It uses two channels: a **data channel** and a **control channel**.
- All the signaling packets (RTS, CTS, busy tones) are transmitted on the control channel, while the data channel is reserved for data packets.

Protocol operation of PAMAS:

- Let us consider an idle node *x* to which a new packet destined to a neighboring node *y* arrives.
- First, *x* sends an RTS packet on the control channel without doing any carrier sensing. This packet carries both *x*'s and *y*'s MAC addresses.
- If *y* receives this packet, it answers with a CTS packet if *y* does not know of any ongoing transmission in its vicinity.
- Upon receiving the CTS, *x* starts to transmit the packet to *y* on the data channel. When *y* starts to receive the data, it sends out a **busy-tone** packet on the control channel.
- If *x* fails to receive a CTS packet within some time window, it enters the backoff mode, where a binary exponential backoff scheme is used.
- The backoff time is uniformly chosen from a time interval that is doubled after each failure to receive a CTS.
- Now, let us look at the nodes receiving *x*'s RTS packet on the control channel. There is the intended receiver *y* and there are other nodes; let *z* be one of them.
- If *z* is currently receiving a packet, it reacts by sending a busy-tone packet, which overlaps with *y*'s CTS at node *x* and effectively destroys the CTS.
- Therefore, *x* cannot start transmission and *z*'s packet reception is not disturbed. Since the busy-tone packet is longer than the CTS, we can be sure that the CTS is really destroyed.
- Next, we consider the intended receiver *y*. If *y* knows about an ongoing transmission in its vicinity, it suppresses its CTS, causing *x* to back off.
- Node *y* can obtain this knowledge by either sensing the data channel or by checking whether there was some noise on the control channel immediately after receiving the RTS.
- This noise can be an RTS or CTS of another node colliding at *y*.
- In the other case, *y* answers with a CTS packet and starts to send out a busy-tone packet as soon as *x*'s transmission has started.
- Furthermore, *y* sends out busy-tone packets each time it receives some noise or a valid packet on the control channel, to prevent its neighborhood from any activities.

****

Schedule-based protocols

***7.* Write short notes on advantages and disadvantages of scheduled based protocols.**

Advantages:

- Schedule-based protocols that do not explicitly address idle listening avoidance but do so implicitly, for example, by employing TDMA schemes, which explicitly assign transmission and reception opportunities to nodes and let them sleep at all other times.
- In schedule-based protocols is that transmission schedules can be computed such that no collisions occur at receivers and hence no special mechanisms are needed to avoid hidden-terminal situations.

Disadvantages:

- First, the setup and maintenance of schedules involves signaling traffic, especially when faced to variable topologies.
- Second, if a TDMA variant is employed, time is divided into comparably small slots, and both transmitter and receiver have to agree to slot boundaries to actually meet each other and to avoid overlaps with other slots, which would lead to collisions.
- However, maintaining time synchronization involves some extra signaling traffic.
- Third drawback is that such schedules are not easily adapted to different load situations on small timescales. Specifically, in TDMA, it is difficult for a node to give up unused time slots to its neighbors.
- Fourth drawback is that the schedule of a node may require a significant amount of memory, which is a scarce resource in several sensor node designs.
- Finally, distributed assignment of conflict-free TDMA schedules is a difficult problem in itself.

LEACH

8. Explain the operation of LEACH protocol.

- The LEACH protocol (Low-energy Adaptive Clustering Hierarchy) assumes a dense sensor network of homogeneous, energy-constrained nodes, which shall report their data to a sink node.
- In LEACH, a TDMA based MAC protocol is integrated with clustering and a simple "routing" protocol.

- LEACH partitions the nodes into **clusters** and in each cluster a dedicated node, the **clusterhead**, is responsible for creating and maintaining a TDMA schedule; all the other nodes of a cluster are **member nodes**.
- To all member nodes, TDMA slots are assigned, which can be used to exchange data between the member and the clusterhead; there is no peer-to-peer communication.
- With the exception of their time slots, the members can spend their time in sleep state.
- The clusterhead aggregates the data of its members and transmits it to the sink node or to other nodes for further relaying.
- Since the sink is often far away, the clusterhead must spend significant energy for this transmission.
- For a member, it is typically much cheaper to reach the clusterhead than to transmit directly to the sink.
- The clusterheads role is energy consuming since it is always switched on and is responsible for the long-range transmissions.
- If a fixed node has this role, it would burn its energy quickly, and after it died, all its members would be "headless" and therefore useless.
- Therefore, this burden is rotated among the nodes. Specifically, each node decides independent of other nodes whether it becomes a clusterhead, and therefore there is no signaling traffic related to clusterhead election.
- This decision takes into account when the node served as clusterhead the last time, such that a node that has not been a clusterhead for a long time is more likely to elect itself than a node serving just recently.
- The protocol is round based, that is, all nodes make their decisions whether to become a clusterhead at the same time and the nonclusterhead nodes have to associate to a clusterhead subsequently.
- The nonclusterheads choose their clusterhead based on received signal strengths.
- The network partitioning into clusters is time variable and the protocol assumes global time synchronization.
- After the clusters have been formed, each clusterhead picks a random CDMA code for its cluster, which it broadcasts and which its member nodes have to use subsequently.
- This avoids a situation where a border node belonging to clusterhead *A* distorts transmissions directed to clusterhead *B*.

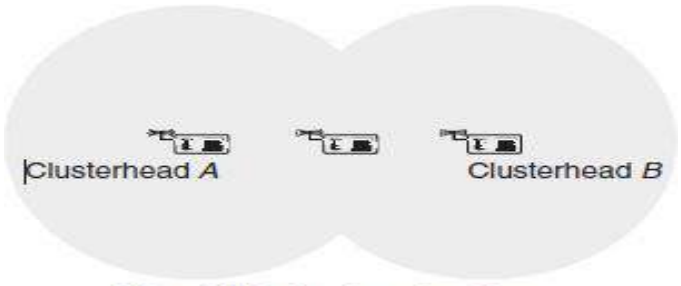


Figure 5.10 Intercluster interference

Stages of LEACH protocol:

- The protocol is organized in **rounds** and each round is subdivided into a setup phase and a steady-state phase.

Setup Phase:

- The **setup phase** starts with the self-election of nodes to clusterheads.
- In the following **advertisement phase**, the clusterheads inform their neighborhood with an advertisement packet.
- The clusterheads contend for the medium using a CSMA protocol with no further provision against the hidden-terminal problem.
- The nonclusterhead nodes pick the advertisement packet with the strongest received signal strength.
- In the following cluster-setup phase, the members inform their clusterhead ("join"), again using a CSMA protocol.
- After the cluster setup-phase, the clusterhead knows the number of members and their identifiers.
- It constructs a TDMA schedule, picks a CDMA code randomly, and broadcasts this information in the broadcast schedule subphase.

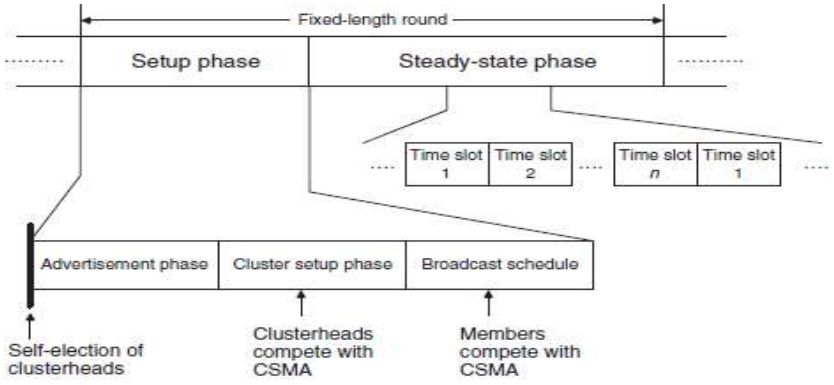


Figure 5.11 Organization of LEACH rounds

Steady state phase:

- After this, the TDMA steady-state phase begins. Because of collisions of advertisement or join packets, the protocol cannot guarantee that each non clusterhead node belongs to a cluster.
- However, it can guarantee that nodes belong to at most one cluster.
- The clusterhead is switched on during the whole round and the member nodes have to be switched on during the setup phase and occasionally in the steady-state phase, according to their position in the cluster's TDMA schedule.

Drawback:

- unable to cover large geographical areas because a clusterhead two miles away from the sink likely does not have enough energy to reach the sink at all, not to mention achieving a low BER.

Solution:

- If it can be arranged that a clusterhead can use other clusterheads for forwarding, this limitation can be mitigated.

****

IEEE 802.15.4

(Low-rate WPANs)

9. Explain about the MAC protocol in WSN (April/May 2018)(Dec 2019)

- IEEE 802.15.4: The fourth working group goes in the opposite direction for data rates.
- This group standardizes low-rate wireless personal area networks (LR-WPAN.
- The ZigBee consortium tries to standardize the higher layers of 802.15.4 similar to the activities of the Bluetooth consortium for 802.15.1 (ZigBee, 2002).
- IEEE 802.15.4 – Low-rate WPANs The reason for having low data rates is the focus of the working group on extremely low power consumption enabling multi-year battery life.
- Compared to 802.11 or Bluetooth, the new system should have a much lower complexity making it suitable for low-cost wireless communication.
- Example applications include industrial control and monitoring, smart badges, interconnection of environmental sensors, interconnection of peripherals, remote controls etc.
- The new standard should offer data rates between 20 and 250 Kbit/s as maximum and latencies down to 15 ms.

- This is enough for many home automation and consumer electronics applications.
- IEEE 802.15.4 offers two different PHY options using DSSS.
- The 868/915 MHz PHY operates in Europe at 868.0–868.6 MHz and in the US at 902–928 MHz. At 868 MHz one channel is available offering a data rate of 20 kbit/s.
- At 915 MHz 10 channels with 40 kbit/s per channel are available (in Europe GSM uses these frequencies).
- The advantages of the lower frequencies are better propagation conditions.
- However, there is also interference in these bands as many analog transmission systems use them. The 2.4 GHz PHY operates at 2.4–2.4835 GHz and offers 16 channels with 250 kbit/s per channel.
- This PHY offers worldwide operation but suffers from interference in the 2.4 GHz ISM band and higher propagation loss.
- Typical devices with 1 mW output power are expected to cover a 10–20 m range. All PHY PDUs start with a 32 bit preamble for synchronization.
- After a start-of-packet delimiter, the PHY header indicates the length of the payload (maximum 127 bytes).
- Compared to Bluetooth the MAC layer of 802.15.4 is much simpler. \

Network architecture and types/roles of nodes

- The standard distinguishes on the MAC layer two types of nodes:
  - ✓ A Full Function Device (FFD) can operate in three different roles: it can be a PAN coordinator (PAN = Personal Area Network), a simple coordinator or a device.
  - ✓ A Reduced Function Device (RFD) can operate only as a device.

Superframe structure

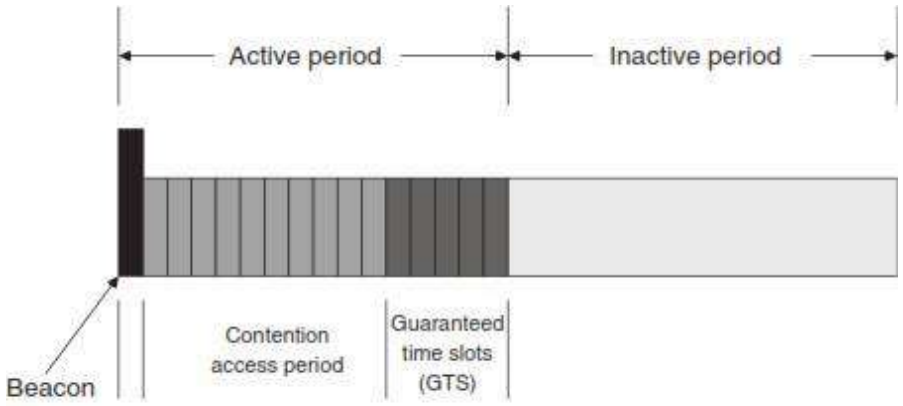
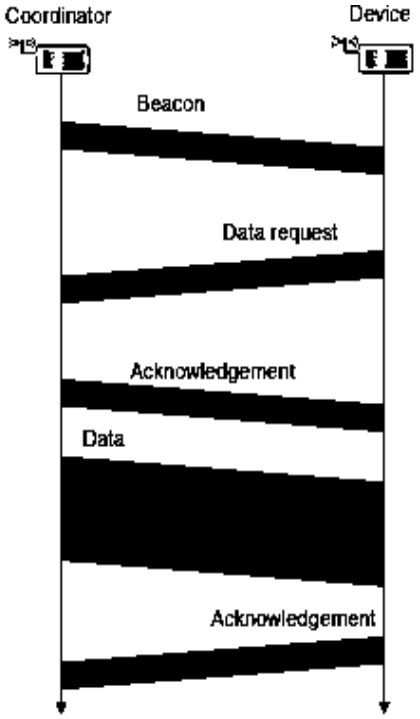


Figure 5.14 Superframe structure of IEEE 802.15.4

- ✓ The coordinator of a star network operating in the beaconed mode organizes channel access and
  data transmission with the help of a superframe structure displayed in Figure 5.14.
- ✓ All superframes have the same length. The coordinator starts each superframe by sending a frame beacon packet. The frame beacon includes a superframe specification describing the length of the
  various components of the following superframe:
- ✓ The superframe is subdivided into an active period and an inactive period. During the inactive period, all nodes including the coordinator can switch off their transceivers and go into sleep state. The nodes have to wake up immediately before the inactive period ends to receive the next beacon. The inactive period may be void.
- ✓ The active period is subdivided into 16 time slots. The first time slot is occupied by the beacon frame and the remaining time slots are partitioned into a Contention Access Period (CAP) followed by a number (maximal seven) of contiguous Guaranteed Time Slots (GTSs).

Slotted CSMA-CA protocol

- ❖ No synchronous voice links are supported. MAC frames start with a 2-byte frame control field, which specifies how the rest of the frame looks and what it contains.



Handshake between coordinator and device when the device retrieves a packet

- The following 1-byte sequence number is needed to match acknowledgements with a previous data transmission. The variable address field (0–20 bytes) may contain source and/or destination addresses in various formats.
- The payload is variable in length; however, the whole MAC frame may not exceed 127 bytes in length.
- A 16-bit FCS protects the frame. Four different MAC frames have been defined: beacon, data, acknowledgement, and MAC command.
- The time slots making up the CAP are subdivided into smaller time slots, called backoff periods.
- Optionally, this LR-WPAN offers a superframe mode. In this mode, a PAN coordinator transmits beacons in predetermined intervals (15ms–245s).
- With the help of beacons, the medium access scheme can have a period when contention is possible and a period which is contention free.
- Furthermore, with beacons a slotted CSMA/CA is available. Without beacons standard CSMA/CA is used for medium access.
- Acknowledgement frames confirming a previous transmission do not use the CSMA mechanism. These frames are sent immediately following the previous packet.
- IEEE 802.15.4 specifies three levels of security: no security, access control lists, and symmetric encryption using AES-128. Key distribution is not specified further.
- Security is a must for home automation or industry control applications. Up to now, the success of this standard is unclear as it is squeezed between Bluetooth, which also aims at cable replacement, and enhanced RFIDs/RF controllers.

****

Energy-Efficient Routing

10. **Explain the Energy efficient unicast routing protocol with an example.**

- Energy-efficient unicast routing appears to be a simple problem: take the network graph, assign to each link a cost value that reflects the energy consumption across this link, and pick any algorithm that computes least-cost paths in a graph.
- Figure shows an example scenario for a communication between nodes A and H including link energy costs and available battery capacity per node.
- The minimum energy route is A-B-E-H, requiring 3 units of energy.
- The minimum hop count route would be A-D-H, requiring 6 units of energy.

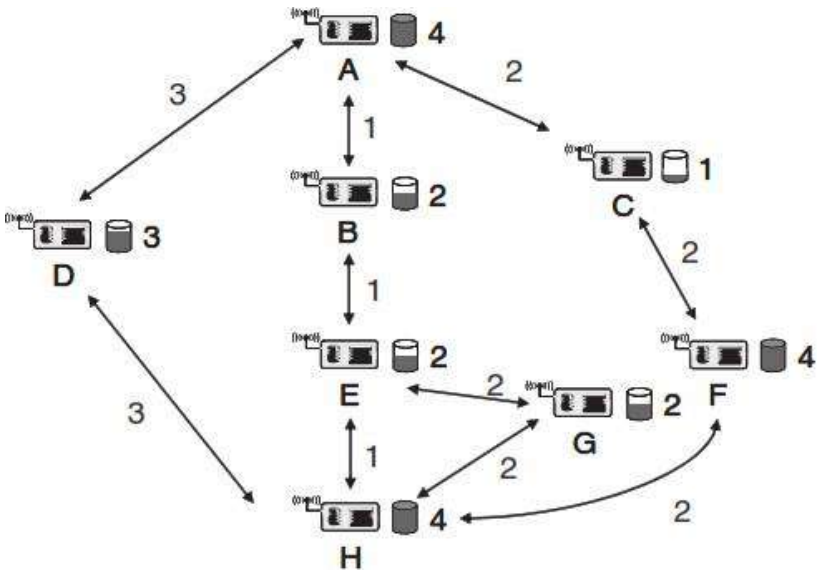


Fig: Communication between A and H through different routes.

Minimize energy per packet (or per bit):

- The total energy required to transport a packet over a multihop path from source to destination.
- The goal is then to minimize, for each packet, this total amount of energy by selecting a good route.

Maximize network lifetime:

- The network should be able to fulfill its duty for as long as possible.

Routing considering available battery energy:

- As the finite energy supply in nodes' batteries is the limiting factor to network lifetime, it stands to reason to use information about battery status in routing decisions.

Maximum Total Available Battery Capacity:

- Choose that route where the sum of the available battery capacity is maximized, without taking "maximum available power".
- Looking only at the intermediate nodes in above Figure, route A-B-E-G-H has a total available capacity of 6 units, but that is only because of the extra node G that is not really needed – such detours can of course arbitrarily increase this metric.

- Hence, AB-E-G-H should be discarded as it contains A-B-E-H as a proper subset. Eventually, route A-C-F-H is selected.

Minimum Battery Cost Routing (MBCR):

- MBCR looks at the "reluctance" of a node to route traffic instead of looking directly into the sum available battery capacities.
- This reluctance increases as its battery is drained; for example, reluctance or routing cost can be measured as the reciprocal of the battery capacity.
- Then, the cost of a path is the sum of this reciprocals and the rule is to pick that path with the smallest cost.
- Since the reciprocal function assigns high costs to nodes with low battery capacity, this will automatically shift traffic away from routes with nodes about to run out of energy.
- In the example of Figure, route A-C-F-H is assigned a cost of 1/1 + 1/4 = 1.25, but route A-D-H only has cost 1/3.
- Consequently, this route is chosen, protecting node C from needless effort.

Min–Max Battery Cost Routing (MMBCR)

- The main idea behind this routing is to protect nodes with low energy battery resources.
- Instead of using the sum of reciprocal battery levels, simply the largest reciprocal level of all nodes along a path is used as the cost for this path.
- Then, again the path with the smallest cost is used.
- In the example of Figure, route A-D-H will be selected.

Conditional Max–Min Battery Capacity Routing (CMMBCR):

- If there are routes along which all nodes have a battery level exceeding a given threshold, then select the route that requires the lowest energy per bit.
- If there is no such route, then pick that route which maximizes the minimum battery level.

Minimize variance in power levels:

- To ensure a long network lifetime, one strategy is to use up all the batteries uniformly to avoid some nodes prematurely running out of energy and disrupting the network.

- Hence, routes should be chosen such that the variance in battery levels between different routes is reduced.

Minimum Total Transmission Power Routing (MTPR) :

- A given transmission is successful if its SINR exceeds a given threshold.
- The goal is to find an assignment of transmission power values for each transmitter such that all transmissions are successful and that the sum of all power values is minimized.

****

Issues in Designing a Transport Layer Protocol For Ad Hoc Wireless Networks

11. Explain the issues in designing a transport layer protocol for adhoc wireless networks.(May/ June 2013 (NOV/DEC 2018))

1. ***Induced Traffic:***
   - In a path having multiple link, the traffic at any given link (or path) due to the traffic through neighbouring links (or paths) is referred to as induced traffic.
   - This is due to the broadcast nature of the channel and the location-dependent contention on the channel.
   - Induced Traffic affects the throughput achieved by the transport layer protocol.
2. ***Induced throughput unfairness:***
   - It refers to the throughput unfairness at the transport layer due to the throughput/delay unfairness existing at the lower layer such as the n/w and MAC layers.
   - A transport layer should consider these in order to provide a fair share of throughput across contending flows.
3. ***Separation of congestion control, reliability and flow control:***
   - A transport layer protocol can provide better performance if end-to-end reliability, flow control and congestion control are handled separately.
   - Reliability and flow control are end-to-end activities, whereas congestion can at times be a local activity.
   - The Objective is minimisation of the additional control overhead generated by them.

4. ***Power and Band width constraints:***
    - Nodes in ad hoc wireless networks face resource constraints including the two most important resources: (i) power source and (ii) bandwidth.
    - The performance of a Transport layer protocol is significantly affected by these resource constraints.
5. ***Interpretation of congestion:***
    - Interpretation of network congestion as used in traditional networks is not appropriate in ad hoc networks.
    - This is because the high error rates of wireless channel, location-dependent contention, hidden terminal problem, packet collisions in the network, path breaks due to mobility of nodes, and node failure due to drained battery can also lead to packet loss in ad hoc wireless networks.
6. ***Completely decoupled transport layer:***
    - Another challenge faced by Transport layer protocol is the interaction with the lower layers.
    - Cross-layer interaction between the transport layer and lower layers is important to adapt to the changing network environment
7. ***Dynamic topology:***
    - Experience rapidly changing network topology due to mobility of nodes.
    - Leads to frequent path breaks, partitioning and remerging of networks & high delay in re-establishment of paths.
    - Performance is affected by rapid changes in network topology.

****

Design Goals Of A Transport Layer Protocol For Ad Hoc Wireless Networks

Explain the significance and design goals of transport layer protocol for adhoc network.

- The protocol should maximize the throughput per connection.
- It should provide throughput fairness across contending flows.
- It should incur minimum connection set up and connection maintenance overheads.
- It should have mechanisms for congestion control and flow control in the network.
- It should be able to provide both reliable and unreliable connections as per the requirements of the application layer.
- It should be able to adapt to the dynamics of the network such as rapid changes in topology.

- Bandwidth must be used efficiently.
- It should be aware of resource constraints such as battery power and buffer sizes and make efficient use of them.
- It should make use of information from the lower layers for improving network throughput.
- It should have a well-defined cross-layer interaction framework.
- It should maintain End-to-End Semantics.

****

UNIT- 4

SENSOR NETWORK SECURITY

Network Security Requirements, Issues and Challenges in Security Provisioning, Network Security Attacks, Layer wise attacks in wireless sensor networks, possible solutions for jamming, tampering, black hole attack, flooding attack. Key Distribution and Management, Secure Routing – SPINS, reliability requirements in sensor networks.

TABLE OF CONTENTS

4.1 Introduction: Security in Wireless Sensor Networks

- WSN is a special type of network. The sensor networks, based on an inherently broadcast wireless medium, are vulnerable to a variety of attacks.
- Security is of prime importance in sensor networks because the absence of central authority, random deployment of nodes in the network and nodes assume a large amount of trust among themselves during data aggregation and event detection.
- From a set of sensor nodes in a given locality, only one final aggregated message may be sent to the BS, so it is necessary to ensure that communication links are secure for data exchange.
- Cryptographic solutions based on symmetric or public key cryptography are not suitable for sensor networks, due to the high processing requirements of the algorithms. So, need special type of protocol to ensure the security in sensor networks.

4.2 Network Security Requirements

- The security services in a WSN should protect the information communicated over the network and the resources from attacks and misbehaviour of nodes.
- The most important security requirements in WSN are listed below:
  - **Data Confidentiality**
  - **Authentication**
  - **Data Integrity**
  - **Data Freshness**

- **Availability**
- **Self-Organization**
- **Time synchronization**
- **Source Localization**
- **Scalability**

Data Confidentiality

- Data Confidentiality requirement is required to ensure that sensitive information is well protected and not revealed to unauthorized third parties.
- The confidentiality objective helps to protect information traveling between the sensor nodes of the network or between the sensors and the base station from disclosure, since an adversary having the appropriate equipment may eavesdrop on the communication.
- By eavesdropping, the adversary could overhear critical information such as sensing data and routing information. Based on the sensitivity of the data stolen, an adversary may cause severe damage since he can use the sensing data for many illegal purposes i.e. sabotage, blackmail.
- Furthermore, by stealing routing information the adversary could introduce his own malicious nodes into the network in an attempt to overhear the entire communication.
- If we consider eavesdropping to be a network level threat, then a local level threat could be a compromised node that an adversary has in his possession. Compromised nodes are a big threat to confidentiality objective since the adversary could steal critical data stored on nodes such as cryptographic keys that are used to encrypt the communication.

Authentication

- It ensures that the communicating node is the one that it claims to be. An adversary can not only modify data packets but also can change a packet stream by injecting fabricated packets.
- It is, therefore, essential for a receiver to have a mechanism to verify that the received packets have indeed come from the actual sender node.
- In case of communication between two nodes, data authentication can be achieved through a message authentication code (MAC) computed from the shared secret key.

Data Integrity

- The mechanism should ensure that no message can be altered by an entity as it traverses from the sender to the recipient.

Data Freshness

- It implies that the data is recent and ensures that no adversary can replay old messages.
- This requirement is especially important when the WSN nodes use shared-keys for message communication, where a potential adversary can launch a replay attack using the old key as the new key is being refreshed and propagated to all the nodes in the WSN.
- A nonce or time-specific counter may be added to each packet to check the freshness of the packet.

Availability

- Availability ensures that services and information can be accessed at the time that they are required.
- In sensor networks, there are many risks that could result in loss of availability such as sensor node capturing and denial of service attacks.
- Lack of availability may affect the operation of many critical real-time applications like those in the healthcare sector that require a 24/7 operation that could even result in the loss of life.
- Therefore, it is critical to ensure resilience to attacks targeting the availability of the system and find ways to fill in the gap created by the capturing or disablement of a specific node by assigning its duties to some other nodes in the network.

Self-Organization

- In WSN no fixed infrastructure exists, hence, every node is independent having properties of adaptation to the different situations and maintains self-organizing and self-healing properties. This is a great challenge for security in WSN.

Time synchronization

- Most of the applications in sensor networks require time synchronization. Any security mechanism for WSN should also be time-synchronized. A collaborative WSN may require synchronization among a group of sensors.

Source Localization

- For data transmission some applications use location information of the sink node. It is important to give security to the location information.
- Non-secured data can be controlled by the malicious node by sending false signal strengths or replaying signals.

Scalability

- Hundreds of thousands of nodes are deployed in a network carrying out distributed operations. Because of this explosive proliferation of sensor nodes, scalability is becoming an important requirement in WSN.
- WSN must be scalable to provide capacity for additional nodes. New nodes insertion and old nodes removal should be easy with no bad impact over the network operations.

4.3 Issues and Challenges in Security Provisioning

- A strong routing protocol can only protect the network from various malicious activities. Designing a strong security routing protocol for wireless sensor network is a very challenging task.
- WSN must have the richest set of different protocols to carryout application requirements; a WSN protocol must handle a hostile environment.
- Routing protocol should provide a high throughput, and a decrease packet loss ratio. Routing algorithm should handle mobility and dynamic changing behavior in WSNs.
- Unreliable wireless media can drop packets; routing protocols should prevent packet loss. Designing a new routing protocol for WSN should consider the following security and privacy issues.
    - **Node Mobility**
    - **Coverage Problem**
    - **Shared Broadcast Radio Channel**
    - **Insecure Operational Environment**
    - **Lack of Central Authority**
    - **Lack of Association**

- **Limited Resource Availability**
- **Physical Vulnerability**
- **Quality of Service**
- **Programming Wireless Sensor Networks**

Node Mobility

- The mobility sink node is used to collect data from all sensors. A static sink node collects data from all sensors without changing its constant position. A mobile sink node has its own effects on the network, e.g., performance and dynamic change behavior. Routing protocols must provide better connectivity, an efficient energy consumption, a controlled flooding mechanism, etc.

Coverage Problem

- Coverage is an important performance metric in WSNs; it reflects how well the environment is monitored. The surrounding vicinity should be monitored all times to collect data; a dead node cannot forward any packets; consequently, it degrades network services.

Shared Broadcast Radio Channel

- Unlike in wired networks where a separate dedicated transmission line can be provided between a pair of end users, the radio channel used for communication in wireless sensor networks is broadcast in nature and is shared by all nodes in the network.
- Data transmitted by a node is received by all nodes within its direct transmission range. So a malicious node could easily obtain data being transmitted in the network. This problem can be minimized to a certain extent by using directional antennas.

Insecure Operational Environment

- The operating environments where wireless sensor networks are used may not always be secure.
- One important application of such networks is in battlefields. In such applications, nodes may move in and out of hostile and insecure enemy territory, where they would be highly vulnerable to security attacks.

Lack of Central Authority

- In wired networks and infrastructure-based wireless networks, it would be possible to monitor the traffic on the network through certain important central points (such as routers, base stations, and access points) and implement security mechanisms at such points. Since wireless networks do not have any such central points, these mechanisms cannot be applied in wireless sensor networks.

Lack of Association

- Since these networks are dynamic in nature, a node can join or leave the network at any point of the time. If no proper authentication mechanism is used for associating nodes with a network, an intruder would be able to join into the network quite easily and carry out his/her attacks.

Limited Resource Availability

- Resources such as bandwidth, battery power, and computational power are scarce in wireless sensor networks. Hence, it is difficult to implement complex cryptography-based security mechanisms in such networks.

Physical vulnerability

- Nodes in these networks are usually compact and hand-held in nature. They could get damaged easily and are also vulnerable to theft.

Quality of Service

- QoS is the function of its application. The proper congestion control provides better QoS. In WSNs, there is a minimum chance of congestion outside the base station area. Congestion near the base station results into: channel occupancy, buffer overflow, packet collision, channel contention, high data rate, and minimum node's life.

- For better services, minimum congestion in the network is necessary. Congestion avoidance ensures high throughput, better link utilization, minimum delay, energy efficiency, and minimum data rate error. Control packets are used to prevent congestion.

Programming Wireless Sensor Networks

- Programming a large network of highly resource-constraint devices that are self-organized and globally consistent, with a robust behavior and a dynamically changing environment, is a big challenge.

- Programming in a hostile or un-secure environment, to monitor the surroundings, is a daunting task. Programming WSNs must be equipped with proper software engineering principles; it must be well coded, tested, debugged, and should provide a flawed free design.

4.4 Network Security Attacks

- Wireless networks are vulnerable to security attacks due to the broadcast nature of the transmission medium. Furthermore, WSNs have an additional vulnerability because nodes are often placed in a hostile or dangerous environment where they are not physically protected.

- For a large-scale sensor network, it is impractical to monitor and protect each individual sensor from physical or logical attack. Attackers may device different types of security attacks to make the WSN system unstable.

4.4.1 Based On the Capability of the Attacker

Outsider versus insider (node compromise) attacks

- Outside attacks are defined as attacks from nodes, which do not belong to a WSN; insider attacks occur when legitimate nodes of a WSN behave in unintended or unauthorized ways.

Passive versus Active attacks

- Passive attacks include eavesdropping on or monitoring packets exchanged within a WSN; active attacks involve some modifications of the data steam or the creation of a false stream.

Mote-class versus laptop-class attacks

- In mote-class attacks, an adversary attacks a WSN by using a few nodes with similar capabilities to the network nodes; in laptop-class attacks, an adversary can use more powerful devices (e.g., a laptop) to attack a WSN. These devices have greater transmission range, processing power, and energy reserves than the network nodes.

4.4.2 Attacks on Information in Transit

- In a sensor network, sensors monitor the changes of specific parameters or values and report to the sink according to the requirement. While sending the report, the information in transit may be attacked to provide wrong information to the base stations or sinks. The attacks are:

- **Interruption:** Communication link in sensor networks becomes lost or unavailable. This operation threatens service availability. The main purpose is to launch denial-of service (DoS) attacks. From the layer-specific perspective, this is aimed at all layers.

- **Interception:** Sensor network has been compromised by an adversary where the attacker gains unauthorized access to sensor node or data in it. Example of this type of attacks is node capture attacks. This threatens message confidentiality. The main purpose is to eavesdrop on the information carried in the messages.

- **Modification:** Unauthorized party not only accesses the data but also tampers with it. This threatens message integrity. The main purpose is to confuse or mislead the parties involved in the communication protocol. This is usually aimed at the network layer and the application layer, because of the richer semantics of these layers.

- **Fabrication:** An adversary injects false data and compromises the trustworthiness of information. This threatens message authenticity. The main purpose is to confuse or mislead the parties involved in the communication protocol. This operation can also facilitate DOS attacks, by flooding the network.

- **Replaying existing messages:** This operation threatens message freshness. The main purpose of this operation is to confuse or mislead the parties involved in the communication protocol that is not time- aware.

4.4.3 Host Based Vs Network Based

- **Host-based attacks:** It is further broken down in to User compromise: This involves compromising the users of a WSN, e.g. by cheating the users into revealing information such as passwords or keys about the sensor nodes. Hardware compromise: This involves tampering with the hardware to extract the program code, data and keys stored within a sensor node. The attacker might also attempt to load its program in the compromised node. Software compromise: This involves breaking the software running on the sensor nodes. Chances are the operating system and/or the applications running in a sensor node are vulnerable to popular exploits such as buffer overflows.

- **Network-based attacks:** It has two orthogonal perspectives layer-specific compromises, and protocol-specific compromises. This includes all the attacks on information in transit. Apart from that it also includes Deviating from protocol: When the attacker is, or becomes an insider of the network, and the attacker's

purpose is not to threaten the service availability, message confidentiality, integrity and authenticity of the network, but to gain an unfair advantage for itself in the usage of the network, the attacker manifests selfish behaviours, behaviours that deviate from the intended functioning of the protocol.

4.5 Layer wise Attacks in Wireless Sensor Networks

- This section discusses about the WSN layer wise attack.

4.5.1 Physical Layer Attacks

4.5.1.1 Jamming

- This is one of the Denial of Service Attacks in which the adversary attempts to disrupt the operation of the network by broadcasting a high-energy signal.
- Jamming attacks in WSNs, classifying them as constant (corrupts packets as they are transmitted), deceptive (sends a constant stream of bytes into the network to make it look like legitimate traffic), random (randomly alternates between sleep and jamming to save energy), and reactive (transmits a jam signal when it senses traffic).
- To defense against this attack, use spread-spectrum techniques for radio communication. Handling jamming over the MAC layer requires Admission Control Mechanisms.

4.5.1.2 Radio Interference

- Here, adversary either produces large amounts of interference intermittently or persistently. To handle this issue, use of symmetric key algorithms in which the disclosure of the keys is delayed by some time interval.

4.5.1.3 Tampering or Destruction

- Given physical access to a node, an attacker can extract sensitive information such as cryptographic keys or other data on the node.
- One defense to this attack involves tamper-proofing the node's physical package.
- Self-Destruction (tamper-proofing packages) – whenever somebody accesses the sensor nodes physically the nodes vaporize their memory contents and this prevents any leakage of information.

4.5.2 Data Link Layer Attacks

4.5.2.1 Continuous Channel Access (Exhaustion)

- A malicious node disrupts the Media Access Control protocol, by continuously requesting or transmitting over the channel. This eventually leads a starvation for other nodes in the network with respect to channel access.
- One of the countermeasures to such an attack is Rate Limiting to the MAC admission control such that the network can ignore excessive requests, thus preventing the energy drain caused by repeated transmissions.
- A second technique is to use time division multiplexing where each node is allotted a time slot in which it can transmit.

4.5.2.2 Collision

- This is very much similar to the continuous channel attack. A collision occurs when two nodes attempt to transmit on the same frequency simultaneously. When packets collide, a change will likely occur in the data portion, causing a checksum mismatch at the receiving end. The packet will then be discarded as invalid. A typical defense against collisions is the use of error-correcting codes.

4.5.2.3 Unfairness

- Repeated application of these exhaustion or collision based MAC layer attacks or an abusive use of cooperative MAC layer priority mechanisms, can lead into unfairness.
- This kind of attack is a partial DOS attack, but results in marginal performance degradation.
- One major defensive measure against such attacks is the usage of small frames, so that any individual node seizes the channel for a smaller duration only.

4.5.2.4 Interrogation

- Exploits the two-way request-to-send/clear-to-send (RTS/CTS) handshake that many MAC protocols use to mitigate the hidden-node problem.
- An attacker can exhaust a node's resources by repeatedly sending RTS messages to elicit CTS responses from a targeted neighbour node.

- To put a defense against such type of attacks a node can limit itself in accepting connections from same identity or use Anti replay protection and strong link-layer authentication.

4.5.2.5 Sybil Attack

- In this attack, a single node presents multiple identities to all other nodes in the WSN. This may mislead other nodes, and hence routes believed to be disjoint with respect to node can have the same adversary node.

- A countermeasure to Sybil Attack is by using a unique shared symmetric key for each node with the base station.

4.5.3 Network Layer Attacks

4.5.3.1 Sinkhole Attack

- Sinkhole attacks normally occur when compromised node send fake routing information to other nodes in the network with aim of attracting as many traffic as possible.

4.5.3.2 Hello Flood

- This attack exploits Hello packets that are required in many protocols to announce nodes to their neighbors. A node receiving such packets may assume that it is in radio range of the sender.

- A laptop class adversary can send this kind of packet to all sensor nodes in the network so that they believe the compromised node belongs to their neighbors. This causes a large number of nodes sending packets to this imaginary neighbour and thus into oblivion. Authentication is the key solution to such attacks. Such attacks can easily be avoided by verify bi-directionality of a link before taking action based on the information received over that link.

4.5.3.3 Node Capture

- Node capture attack is a serious attack through which an intruder can performs various operations on the network and can easily compromise the entire network. It is one of the hazardous attack in WSNs.

- A single node capture is sufficient for an attacker to take over the entire network.

4.5.3.4 Selective Forwarding/ Black Hole Attack

- In Black Hole attack, a malicious node falsely advertises good paths (e.g., shortest path or most stable path) to the destination node during the path-finding process (in on-demand routing protocols) or in the route update messages (in table-driven routing protocols). The intention of the malicious node could be to hinder the path-finding process or to intercept all data packets being sent to the destination node concerned. Malicious or attacking nodes can however refuse to route certain messages and drop them. If they drop all the packets through them, then it is called a Black Hole Attack.
- However if they selectively forward the packets, then it is called selective forwarding.
- To overcome this, Multi path routing can be used in combination with random selection of paths to destination, or braided paths can be used which represent paths which have no common link or which do not have two consecutive common nodes, or use implicit acknowledgments, which ensure that packets are forwarded as they were sent.

4.5.3.5 Wormhole Attacks

- An adversary can tunnel messages received in one part of the network over a low latency link and replay them in another part of the network. This is usually done with the coordination of two adversary nodes, where the nodes try to understate their distance from each other, by broadcasting packets along an out-of-bound channel available only to the attacker.
- To overcome this, the traffic is routed to the base station along a path, which is always geographically shortest or use very tight time synchronization among the nodes, which is infeasible in practical environments.

4.5.3.6 Spoofed, Altered, or Replayed Routing Information

- The most direct attack against a routing protocol in any network is to target the routing information itself while it is being exchanged between nodes. An attacker may spoof, alter, or replay routing information in order to disrupt traffic in the network. These disruptions include the creation of routing loops, attracting or repelling network traffic from select nodes, extending and shortening source routes, generating fake error messages, partitioning the network, and increasing end-to-end latency.
- A countermeasure against spoofing and alteration is to append a message authentication code (MAC) after the message. Efficient encryption and authentication techniques can defend spoofing attacks.

4.5.3.7 Misdirection

- This is a more active attack in which a malicious node present in the routing path can send the packets in wrong direction through which the destination is unreachable. In place of sending the packets in correct direction the attacker misdirects those and that too towards one node and thus this node may be victimized.

4.5.3.8 Homing

- In a homing attack, the attacker looks at network traffic to deduce the geographic location of critical nodes, such as cluster heads or neighbors of the base station. The attacker can then physically disable these nodes. This leads to another type of black hole attack.

4.5.4 Transport layer Attacks

4.5.4.1 Flooding

- Sometime, the malicious node can cause immense traffic of useless messages on the network. This is known as the flooding. Sometimes, malicious nodes replay some actual broadcast messages, and hence generating useless traffic on the network. This can cause congestion, and may eventually lead to the exhaustion of complete nodes. This is a form of Denial of Service attack.

4.5.4.2 De-synchronization Attacks

- In this attack, the adversary repeatedly forges messages to one or both end points which request transmission of missed frames. Hence, these messages are again transmitted and if the adversary maintains a proper timing, it can prevent the end points from exchanging any useful information.

4.5.5 Application layer Attacks

4.5.5.1 Overwhelm Attack

- An attacker might attempt to overwhelm network nodes with sensor stimuli, causing the network to forward large volumes of traffic to a base station. This attack consumes network bandwidth and drains node energy.

4.5.5.2 Path-based DOS Attack

- It involves injecting spurious or replayed packets into the network at leaf nodes. This attack can starve the network of legitimate traffic, because it consumes resources on the path to the base station, thus preventing other nodes from sending data to the base station.

4.5.5.3 Deluge (reprogram) Attack

- Network programming system let you remotely reprogram nodes in deployed networks. If the reprogramming process isn't secure, an intruder can hijack this process and take control of large portions of a network. It can use authentication streams to secure the reprogramming process.

4.6 Possible Solutions for Jamming

- Jamming in wireless networks is defined as the disruption of existing wireless communications by decreasing the signal-to-noise ratio at receiver sides through the transmission of interfering wireless signals.

- Jamming can be done at different levels, from hindering transmission to distorting packets in legitimate communications.

- Jamming makes use of intentional radio interferences to harm wireless communications by keeping communicating medium busy, causing a transmitter to back-off whenever it senses busy wireless medium, or corrupted signal received at receivers. Jamming mostly targets attacks at the physical layer but sometimes cross-layer attacks are possible too.

4.6.1 Types of Jammers

- Jammers are malicious wireless nodes planted by an attacker to cause intentional interference in a wireless network. Depending upon the attack strategy, a jammer can either have the same or different capabilities from legitimate nodes in the network which they are attacking.

- The jamming effect of a jammer depends on its radio transmitter power, location and influence on the network or the targeted node. A jammer may jams a network in various ways to make the jamming as effective as possible. Basically, a jammer can be either **Proactive** and **Reactive**

Proactive jammer

- Proactive jammer transmits jamming (interfering) signals whether or not there is data communication in a network. It sends packets or random bits on the channel it is operating on, putting all the others nodes on that channel in non-operating modes. However, it does not switch channels and operates on only one channel until its energy is exhausted. There are three basic types of proactive jammers: constant, deceptive and random

- **Constant jammer,** emits continuous, random bits without following the CSMA protocol. A constant jammer prevents legitimate nodes from communicating with each other by causing the wireless media to be constantly busy. This type of attack is energy inefficient and easy to detect but is very easy to launch and can damage network communications.
- **Deceptive jammer,** sends a constant stream of bytes into the network to make it look like legitimate traffic.
- **Random jammer,** intermittently transmits either random bits or regular packets into networks. It continuously switches between two states: sleep phase and jamming phase. It sleeps for a certain time of period and then becomes active for jamming before returning back to a sleep state.

Reactive Jammer

- Reactive jammer starts jamming only when it observes a network activity occurs on a certain channel. As a result, a reactive jammer targets on compromising the reception of a message. It can disrupt both small and large sized packets. Since it has to constantly monitor the network, reactive jammer is less energy efficient than random jammer. However, it is much more difficult to detect a reactive jammer than a proactive jammer because the Packet Delivery Ratio (PDR) cannot be determined accurately in practice. There are two different ways to implement a reactive jammer
- **Reactive RTS/CTS jammer**, jams the network when it senses a request-to-send (RTS) message is being transmitted from a sender. It starts jamming the channel as soon as the RTS is sent. In this way, the receiver will not send back clear-to-send (CTS) reply because the RTS packet sent from a sender is distorted. Then, the sender will not send data because it believes the receiver is busy with another on-going transmission.
- **Reactive Data/ACK jammer**, jams the network by corrupting the transmissions of data or acknowledgement (ACK) packets. This type of jammer can corrupt data packets, or it waits until the data packets reach the receiver and then corrupts the ACK packets. The corruptions of both data packets and ACK messages will lead to re-transmissions at the sender end.

4.6.2 Countermeasures for Proactive Jammer

- In proactive jamming, the jammer chokes the bandwidth so that a transmitter is unable to transmit. Therefore, carrier-sensing thresholds can be used to detect such type of jammers. When jamming is detected, nodes in the network can map the jammed area and re-route traffic, switch channel, or perform spatial retreat to counteract this jamming act.

4.6.3 Countermeasures for Reactive Jammer

- Reactive Jamming detection using BER. It is used to detect jamming using the bit error rate (BER) for reactive jammers that keep the received signal strength (RSS) low while introducing disruption in a packet.
- By looking at the RSS of each bit during the reception, it identifies the cause of bit errors for individual packet using predetermined knowledge, error correcting codes (ECC), or wired node chain systems. If the error is due to weak signal, the RSS should be low. .
- If the RSS value is high for a bit error, there are external interference or jamming. Assuming nodes can assess the expected local interference, the sequential jamming probability test calculates the marginal likelihood of errors due to 10 unintentional collisions. If this value is less than the log of the ratio of targeted probability for a missed alarm to the targeted probability, then there is jamming and an alarm is raised.
- If the marginal likelihood is less than the ratio, there is no jamming and the sequence is reset. There is also a possibility that no conclusion is made until there are more conclusive evidences for jamming.

4.7 Tampering Attack and its Countermeasures

- An attacker can damage or replace sensor and computation hardware and the program codes or remove sensitive materials like cryptographic keys to allow unrestricted access to higher levels of communication (Figure4.1). Thereby these tampering nodes interfere in the physical access of sensor nodes.

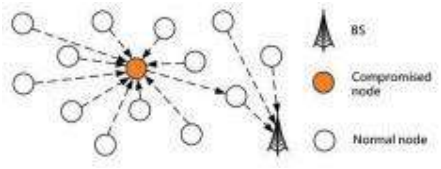


Figure 4.1 Tampering Attack

Countermeasures

- Some attacks in the physical layer are quite hard to cope with. For example, after sensors are deployed in the field, it is difficult or infeasible to prevent every single sensor from device tampering. Therefore, although there are some mechanisms that attempt to reduce the occurrences of attacks, more of them focus on protecting information from divulgence.

Access Restriction

- Obviously, restricting adversaries from physically accessing or getting close to sensors is effective on tampering attacks. It is good to have such restrictions if we can, but unfortunately, they are either difficult or infeasible in most cases. Therefore, we usually have to fall back on another type of restrictions: communication media access restriction.
- A few techniques exist nowadays that prevent attackers from accessing the wireless medium in use, including sleeping/hibernating and spread spectrum communication.
- This technique uses either analog schemes where the frequency variation is continuous, or digital schemes (e.g. frequency hopping) where the frequency variation is abrupt.
- By this way, attackers cannot easily locate the communication channel, and are thus restrained from attacking. The spread spectrum communications are not yet feasible for WSNs that are usually constrained in resources. Directional antenna is another technique for access restriction. By confining the directions of the signal propagation, it reduces the chances of adversaries accessing the communication channel.

Encryption

- In general, cryptography is the all-purpose solution to achieve security goals in WSNs. To protect data confidentiality, cryptography is indispensable.
- Cryptography can be applied to the data stored on sensors. Once data are encrypted, even if the sensors are captured, it is difficult for the adversaries to obtain useful information. A more costly encryption can yield higher strength, but it also drains the limited precious energy faster and needs more memory. More often, cryptography is applied to the data in transmission.
- There are basically two categories of cryptographic mechanisms: asymmetric and symmetric. In asymmetric mechanisms (e.g. RSA), the keys used for encryption and decryption are different, allowing for easier key distribution. It usually requires a third trusted party called Certificate Authority (CA) to distribute and check certificates so that the identity of the users using a certain key can be verified. However, due to the lack of a priori trust relationship and infrastructure support, it is infeasible to have CAs in WSNs.
- Furthermore, asymmetric cryptography usually consumes more resources such as computation and memory.

- In comparison, symmetric mechanisms are more economical in terms of resource consumption. As long as two nodes share a key, they can use this key to encrypt and decrypt data and securely communicate with each other.

4.8 Block Hole Attack and its Countermeasures

- Black Hole attack occurs under Dos (Denial of service) attack in the network layer of OSI Model. In this kind of attacks the malicious node forgery other nodes by announcing a shortest false route to the destination then attracts additional traffic and drops continually the packets.

- During data transmission the source node sends a Route REQuest (RREQ) message to all the nodes including malicious node. Given that a malicious node may become active by receiving RREQ message and replies using Route REPly (RREP) message.

- It attracts additional traffic by falsely claiming the shortest route to the destination. This causes blocking and increasing the energy consumption in each node, leading to the formation of routing holes which disturb or stop the network functionality.

- The Fig. 4.3 illustrates the Black hole attack: while the source node A broadcasts an RREQ messages to discover the route for sending packets to destination node C. An RREQ broadcast from node A is received by neighbouring nodes B, D and the malicious node E. The RREP message sent by the malicious attacker node E is the first message reaching the source node. This last updates its routing table for the new route to the intended node destination, discarding any RREP message from other neighbouring nodes including the actual node destination and starts sending the buffered data packets immediately. In the same time the Black hole node drops all coming data packets rather than forwarding.

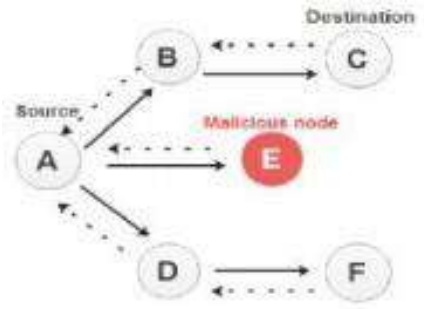


Figure 4.2 Black hole Attack schematic illustration using RREQ and RREP Packet

Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

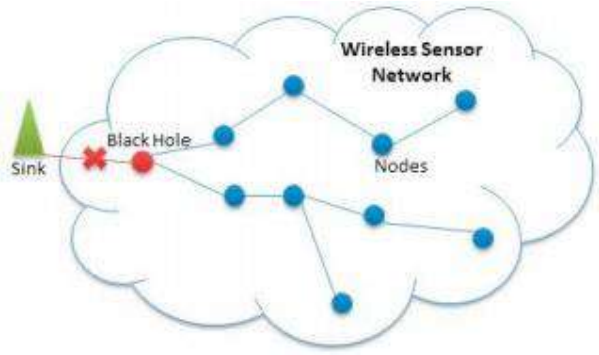


Figure 4.3 Black Hole Attaack
Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

Countermeasures

Routing Access Restriction

- Routing may be one of the most attractive attack targets in WSNs. If we can exclude attackers from participating in the routing process, i.e. restrict them from accessing routing, a large number of attacks in the network layer will be prevented or alleviated.
- Multi-path routing is one of the methods to reduce the effectiveness of attacks launched by attackers on routing paths. In these schemes, packets are routed through multiple paths. Even if the attacker on one of the paths breaks down the path, the routing is not necessarily broken as other paths still exist.
- This alleviates the impact of routing attacks, although does not prevent these attacks. A general way is to use authentication methods. With authentication, it can be easily determined whether a sensor can participate in routing or not.
- Authentication can be either end-to-end or hop-to-hop. In end-to-end authentication, the source and destination share some secret and can thus verify each other. When a node receives a routing update, it always verify the sender of the update before accepting the update.
- In hop-to-hop authentication, each message in transmission is authenticated hop by hop. Therefore, the trust between the source and the destination is built upon the trust on all the intermediate nodes in the path.
- Data are authenticated hop by hop between associated nodes until they reach the base station. Hop-to-hop authentication can be combined with multi-path routing. This paths can be physical, meaning that messages are routed through multiple physically different communication paths.

False Routing Information Detection

- Sometimes attackers do have chances to send false routing information into the network, e.g. during route discovery stages. If the false information does not lead to network failure such as broken routes, we really cannot do much about it. Otherwise, we can apply the idea of misbehaviour detection method.
- For example, watchdog or IDS (Intrusion Detection System) may find that some node fails to route messages along the routing path due to the wrong information it keeps. This anomaly of route failure may trigger out an alarm.
- Nodes can start to trace the source of false routing information. The Reputation can also be maintained, depending on whether nodes are providing valid routing information.

4.9 Flooding Attack and its Countermeasures

- Many protocols require nodes to broadcast HELLO packets to announce themselves to their neighbors, and a node receiving such a packet may assume that it is within (normal) radio range of the sender (Figure 4.4).
- This assumption may be false: a laptop-class attacker broadcasting routing or other information with large enough transmission power could convince every node in the network that the adversary is its neighbour.
- For example, an adversary advertising a very high-quality route to the base station to every node in the network could cause a large number of nodes to attempt to use this route, but those nodes sufficiently far away from the adversary would be sending packets into oblivion. The network is left in a state of confusion.

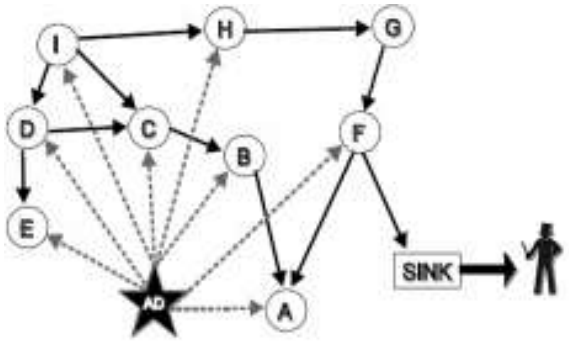


Figure 4.4 Flooding Attacks

Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

Countermeasures

Using Secret Keys Method

- In multi-path multi-base station data forwarding technique, each sensor node maintains number of different secrets (keys) in a multiple tree.
- Sensor node can forward its sensed data to multiple routes by using these secrets. There are multiple base stations in the network that have control over specific number of nodes and also, there are common means of communication among base stations.
- Each base station has all the secrets that are shared by all the sensor nodes, covered by it, according to the key assignment protocol.

Using Threshold Method

- A threshold based solution is used to defend against flooding attacks in WSN.
- The mobile nodes use a threshold value to check whether its neighbors are intruders or not.
- When the number of route request packets broadcasted by a node exceeds the predefined threshold value, it is treated as an intruder and the node stops providing its services to the intruder.

4.10 Key Distribution and Management

4.10.1 Key Management

- Cryptography is one of the most common and reliable means to ensure security. It is the study of the principles, techniques, and algorithms by which information is transformed into a disguised version which no unauthorized person can read, but which can be recovered in its original form by an intended recipient.
- In cryptography, the original information to be sent from one person to another is called plaintext. This plaintext is converted into ciphertext by the process of encryption, that is, the application of certain algorithms or functions.
- An authentic receiver can decrypt/decode the ciphertext back into plaintext by the process of decryption. The processes of encryption and decryption are governed by keys, which are small amounts of information used by the cryptographic algorithms.
- When the key is to be kept secret to ensure the security of the system, it is called a secret key. The secure administration of cryptographic keys is called key management.

- The four main goals of cryptography are confidentiality, integrity, authentication (the receiver should be able to identify the sender and verify that the message actually came from that sender), and non-repudiation.
- There are two major kinds of cryptographic algorithms: symmetric key algorithms, which use the same key for encryption and decryption, and asymmetric key algorithms, which use two different keys for encryption and decryption.
- Symmetric key algorithms are usually faster to execute electronically, but require a secret key to be shared between the sender and receiver. If the same key is used among more than two parties, a breach of security at any one point makes the whole system vulnerable.
- The asymmetric key algorithms are based on some mathematical principles which make it infeasible or impossible to obtain one key from another; therefore, one of the keys can be made public while the other is kept secret (private). This is called public key cryptography.

Symmetric Key Algorithms

- Symmetric key algorithms rely on the presence of the shared key at both the sender and receiver, which has been exchanged by some previous arrangement.
- There are two kinds of symmetric key algorithms, one involving block ciphers and the other stream ciphers. A block cipher is an encryption scheme in which the plaintext is broken into fixed-length segments called blocks, and the blocks are encrypted one at a time.
- The simplest examples include substitution and transposition. In substitution, each alphabet of the plaintext is substituted by another in the ciphertext, and this table mapping the original and the substituted alphabet is available at both the sender and receiver.
- A transposition cipher permutes the alphabet in the plaintext to produce the ciphertext. Figure 4.5 **(**a) illustrates the encryption using substitution, and Figure 4.5 (b) shows a transposition cipher. The block length used is five.

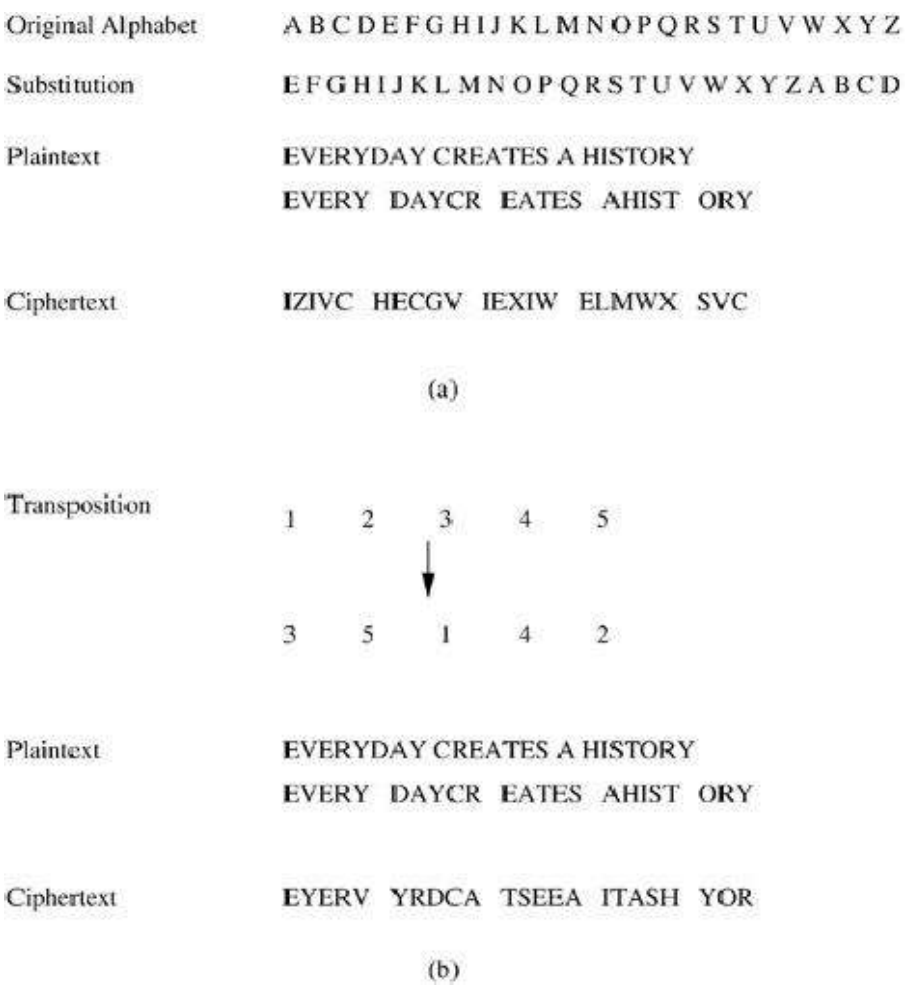


Figure 4.5 Substitution and Transposition

Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

- A stream cipher is a symmetric key cipher where plaintext digits are combined with a pseudorandom cipher digit stream. In a stream cipher, each plaintext digit is encrypted one at a time with the corresponding digit of the keystream, to give a digit of the ciphertext stream

- One of the simplest stream ciphers is the Vernam cipher, which uses a key of the same length as the plaintext for encryption. For example, if the plaintext is the binary string 10010100, and the key is 01011001, then the encrypted string is given by the XOR of the plaintext and key, to be 11001101. The plaintext is again recovered by XOR the ciphertext with the same key. If the key is randomly chosen, transported securely to the receiver, and used for only one communication, this forms the one-time pad which has proven to be the most secure of all cryptographic systems.

Asymmetric Key Algorithms

- Asymmetric key (or public key) algorithms use different keys at the sender and receiver ends for encryption and decryption, respectively.
- Let the encryption process be represented by a function E, and decryption by D. Then the plaintext M is transformed into the ciphertext C as C = E (M).
- The receiver then decodes C by applying D. Hence, D is such that M = D(C) = D (E (M)). When this asymmetric key concept is used in public key algorithms, the key E is made public, while D is private, known only to the intended receiver. Anyone who wishes to send a message to this receiver encrypts it using E. Though C can be overheard by adversaries, the function E is based on a computationally difficult mathematical problem, such as the factorization of large prime numbers.
- Hence, it is not possible for adversaries to derive D given E. Only the receiver can decrypt C using the private key D.
- A very popular example of public key cryptography is the RSA system developed by Rivest, Shamir, and Adleman, which is based on the integer factorization problem.
- Digital signatures schemes are also based on public key encryption. In these schemes, the functions E and D are chosen such that D (E (M)) = E (D (M)) = M for any message M. These are called reversible public key systems.
- In this case, the person who wishes to sign a document encrypts it using his/her private key D, which is known only to him/her. Anybody who has his/her public key E can decrypt it and obtain the original document, if it has been signed by the corresponding sender.

4.10.2 Key Distribution (Management) Approaches

- The primary goal of key distribution is to share a secret among a specified set of participants. There are several methods that can be employed to perform this operation, all of them requiring varying amounts of initial configuration, communication, and computation. The main approaches to key distribution are
  - Key Pre-distribution
  - Pairwise Key Generation
  - Key Transport
  - Key Agreement

Key Pre-distribution

- Key pre-distribution, as the name suggests, involves distributing keys to all interested parties before the start of communication. This method involves much less communication and computation, but all participants must be known a priori, during the initial configuration.
- Once deployed, there is no mechanism to include new members in the group or to change the key. As an improvement over the basic pre-distribution scheme, sub-groups may be formed within the group, and some communication can be restricted to a subgroup. However, the formation of sub-groups is also an a priori decision with no flexibility during the operation.

Pairwise Key Generation

- In WSN, if it is known which nodes will be in the same neighbourhood before deployment, pairwise keys can be established between these nodes a priori. Any pair of nodes can use this master secret key to achieve key agreement

Key Transport

- In key transport systems, one of the communicating entities generates keys and transports them to the other members. The simplest scheme assumes that a shared key already exists among the participating members.
- This prior shared key is used to encrypt a new key and is transmitted to all corresponding nodes. Only those nodes which have the prior shared key can decrypt it. This is called the key Encrypting Key (KEK) method. However, the existence of a prior key cannot always be assumed. If the public key infrastructure (PKI) is present, the key can be encrypted with each participant's public key and transported to it.

Key Agreement

- Most key agreement schemes are based on asymmetric key algorithms. They are used when two or more people want to agree upon a secret key, which will then be used for further communication.
- Key agreement protocols are used to establish a secure context over which a session can be run, starting with many parties who wish to communicate and an insecure channel.
- In group key agreement schemes, each participant contributes a part to the secret key. These need the least amount of pre-configuration, but such schemes have high computational complexity. The most popular key agreement schemes use the Diffie-Hellman exchange, an asymmetric key algorithm based on discrete logarithms.

4.11 Secure Routing in Wireless Sensor Networks

- Routing is one of the most important operations in wireless sensor networks (WSNs) as it deals with data delivery to base stations.
- Routing attacks can cripple it easily and degrade the operation of WSNs significantly. Hence, providing security becomes a challenging task in the networks.
- Various other factors which make the task of ensuring secure communication in wireless sensor networks difficult include the mobility of nodes, a promiscuous mode of operation, limited processing power, and limited availability of resources such as battery power, bandwidth, and memory.
- The secure routing protocol should be resilient in the presence of malicious nodes that may launch various types of attacks. Some of the mechanisms proposed for secure routing.

4.11.1 Requirements of a Secure Routing Protocol for Wireless Sensor Networks

The fundamental requisites of a secure routing protocol for wireless sensor networks are listed as follows:

Detection of malicious nodes

- A secure routing protocol should be able to detect the presence of malicious nodes in the network and should avoid the participation of such nodes in the routing process. Even if such malicious nodes participate in the route discovery process, the routing protocol should choose paths that do not include such nodes.

Guarantee of correct route discovery

- If a route between the source and the destination nodes exists, the routing protocol should be able to find the route, and should also ensure the correctness of the selected route.

Confidentiality of network topology

- Information disclosure attack may lead to the discovery of the network topology by the malicious nodes. Once the network topology is known, the attacker may try to study the traffic pattern in the network.
- If some of the nodes are found to be more active compared to others, the attacker may try to mount (e.g., DoS) attacks on such bottleneck nodes. This may ultimately affect the on-going routing process.

- Hence, the confidentiality of the network topology is an important requirement to be met by the secure routing protocols.

Stability against attacks

- The routing protocol must be self-stable in the sense that it must be able to revert to its normal operating state within a finite amount of time after a passive or an active attack. The routing protocol should take care that these attacks do not permanently disrupt the routing process.

- The protocol must also ensure Byzantine robustness, that is, the protocol should work properly even if some of the nodes, which were earlier participating in the routing process, turn out to become malicious at a later point of time or are intentionally damaged.

4.12 Security Protocols for Sensor Networks (SPINS)

- Security protocols for sensor networks (SPINS) consists of a suite of security protocols that are optimized for highly resource-constrained sensor networks. SPINS consists of two main modules:

  - Sensor Network Encryption Protocol (SNEP)
  - Micro-version of Timed Efficient Stream Loss-Tolerant Authentication protocol (μTESLA)

4.12.1 Sensor Network Encryption Protocol (SNEP)

- SPIN is abbreviation of sensor protocol for information via negotiation. This protocol is defined to use to remove the deficiency like flooding and gossiping that occurs in other protocols.

- The main idea is that the sharing of data, which is sensed by the node, might take more resources as compare to the meta-data, which is just a descriptor about the data sensed, by the node.

- The resource manager in each node monitors its resources and adapts their functionality accordingly.

- SNEP is sensor network encryption protocol. The SNEP protocol offers the following nice properties:

- **Semantic security:** Since the counter value is incremented after each message, the same message is encrypted differently each time. The counter value is long enough that it never repeats within the lifetime of the node.
- **Data authentication:** If the MAC verifies correctly, a receiver can be assured that the message originated from the claimed sender.
- **Replay protection:** The counter value in the MAC prevents replaying old messages. Note that if the counter were not present in the MAC, an adversary could easily replay messages.
- **Weak freshness:** If the message verified correctly, a receiver knows that the message must have been sent after the previous message it received correctly (that had a lower counter value). This enforces a message ordering and yields weak freshness.
- **Low communication overhead:** SNEP has low communication overhead since it only adds 8 bytes per message. The counter state is kept at each end point and does not need to be sent in each message

4.12.1.1 Key Generation /Setup

- Nodes and base station share a master key pre-deployment
- Other keys are bootstrapped from the master key:
  - Encryption key
  - Message Authentication code key
  - Random number generator key

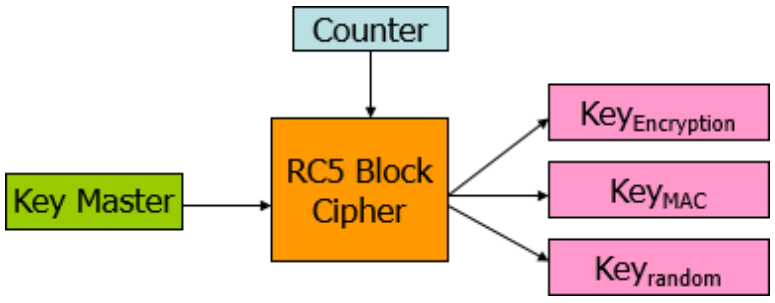


Figure 4.6 SNEP Key Generation

Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

4.12.1.2 Authentication, Confidentiality

- The two communicating parties A and B share a master secret key X_{AB}, and they derive independent keys using the pseudorandom function F: encryption keys $K_{AB} = F_X(1)$ and $K_{BA} = F_X(3)$ for each direction of communication, and MAC keys $K'_{AB} = F_X(2)$ and $K'_{BA} = F_X(4)$ for each direction of communication.
- The combination of these mechanisms form our Sensor Network Encryption Protocol SNEP.
- The encrypted data has the following format: $E = \{M\}_{(K, C)}$, where M is the data, the encryption key is K, and the counter is C. The MAC is M = MAC (K', C||E). The complete message that A sends to B is

$$A \rightarrow B: \{M\}_{(K_{AB}, C_A)}, MAC(K'_{AB} C_A \,||\, \{M\}_{(K_{AB}, C_A)})$$

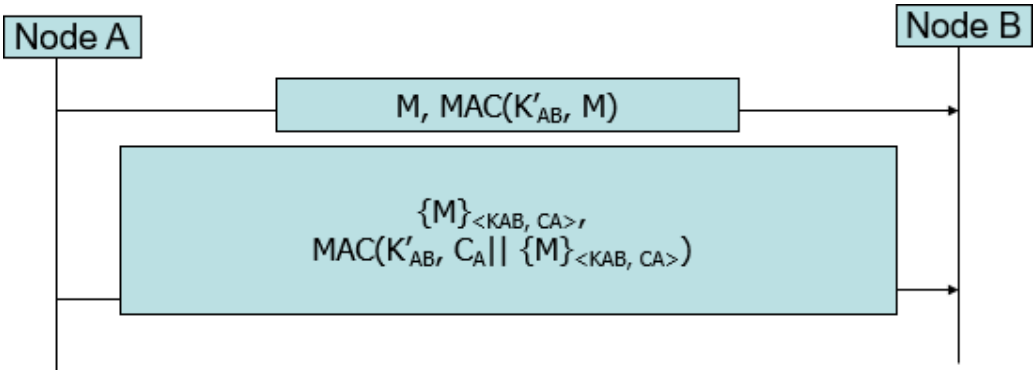


Figure 4.7 SNEP Authentication, Confidentiality

Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

4.12.1.3 Strong Freshness

- Node A achieves strong data freshness for a response from node B through a nonce NA. Node A generates N_A randomly and sends it along with a request message R_A to node B. The simplest way to achieve strong freshness is for B to return the nonce with the response message R_B in an authenticated protocol.

$$A \rightarrow B: \quad N_A, R_A,$$
$$B \rightarrow A:$$
$$\{R_B\}_{(K_{BA}, C_B)}, \; MAC(K'_{BA}, N_A \,||\, C_B \,||\, \{R_B\}_{(K_{BA}, C_B)})$$

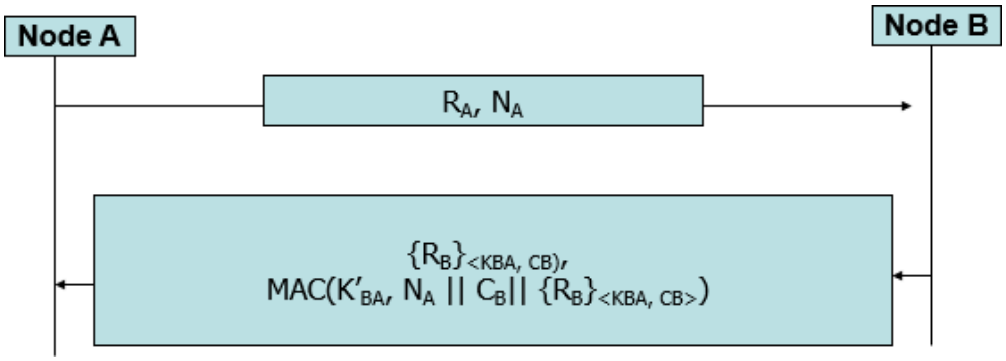


Figure 4.8 SNEP Strong Freshnes

Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

4.12.1.4 Counter exchange protocol

- To achieve small SNEP messages, we assume that the communicating parties A and B know each other's counter values C_A and C_B and so the counter does not need to be added to each encrypted message.

$$\begin{aligned} A \rightarrow B: &\quad C_A, \\ B \rightarrow A: &\quad C_B,\ \mathrm{MAC}\left(K'_{BA} C_A \,||\, C_B\right). \\ A \rightarrow B: &\quad \mathrm{MAC}\left(K'_{AB}, C_A \,||\, C_B\right). \end{aligned}$$

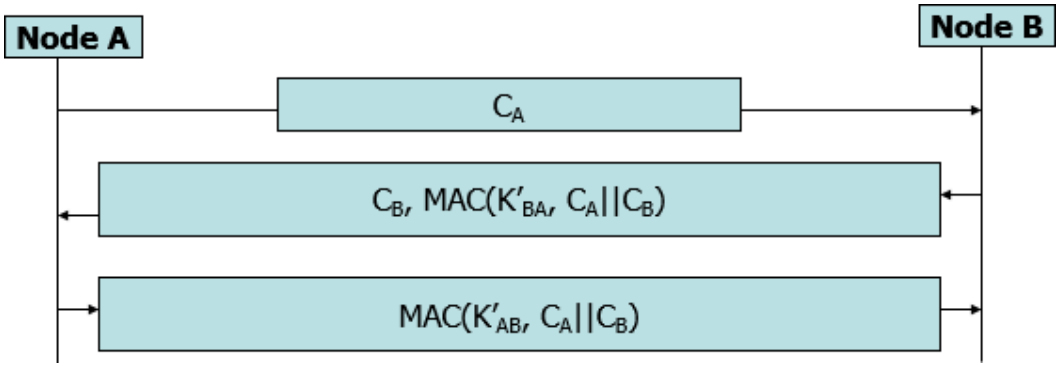


Figure 4.9 SNEP Counter Exchange

Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

- If party A realizes that the counter CB of party B is not synchronized any more, A can request the current counter of B using a nonce NA to ensure strong freshness of the reply:

$$\begin{aligned} A \rightarrow B: &\quad N_A, \\ B \rightarrow A: &\quad C_B,\ \mathrm{MAC}(K'_{BA}, N_A \,||\, C_B). \end{aligned}$$

4.12.2 Micro Timed Efficient Stream Loss-tolerant Authentication (µTESLA)

➢ Micro Timed Efficient Stream Loss-tolerant Authentication delivers broadcast authentication. The pure TESLA is not practical for a node to broadcast authenticated data.

Problems with TESLA

- Digital Signature for initial packet authentication
  - µTESLA uses only symmetric mechanism
- Overhead of 24 bytes per packet
  - µTESLA discloses key once per epoch
- One way key chain is too big
  - µTESLA restricts number of authenticated senders

4.12.2.1 Authentication

➢ To send an authenticated packet, the base station simply computes a MAC on the packet with a key that is secret at that point in time.

➢ When a node gets a packet, it can verify that the corresponding MAC key was not yet disclosed by the base. Since a receiving node is assured that the MAC key is known only by the base station, the receiving node is assured that no adversary could have altered the packet in transit.

➢ The node stores the packet in a buffer. At the time of key disclosure, the base station broadcasts the verification key to all receivers. When a node receives the disclosed key, it can easily verify the correctness of the key. If the key is correct, the node can now use it to authenticate the packet stored in its buffer

4.12.2.2 Key Setup

➢ Each MAC key is a key of a key chain, generated by a public one-way function F. To generate the one-way key chain, the sender chooses the last key K_n of the chain randomly, and repeatedly applies F to compute all other keys: $K_i = F(K_{i+1})$.

➢ Each node can easily perform time synchronization and retrieve an authenticated key of the key chain for the commitment in a secure and authenticated manner.

- For example, Figure 4.10 shows an example of µTESLA. Each key of the key chain corresponds to a time interval and all packets sent within one time interval are authenticated with the same key. The time until keys of a particular interval are disclosed is 2 time intervals in this example.

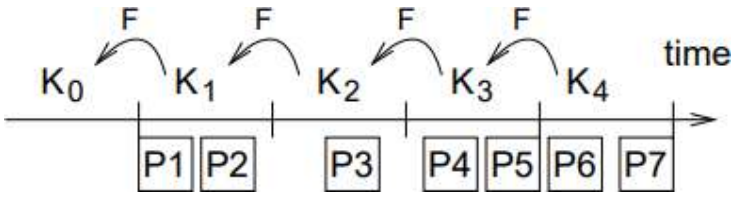


Figure 4.10 Key setup

- Assume that the receiver node is loosely time synchronized and knows K_0 (a commitment to the key chain) in an authenticated way. Packets P1 and P2 sent in interval 1 contain a MAC with key K_1. Packet P3 has a MAC using key K_2. So far, the receiver cannot authenticate any packets yet. Let us assume that packets P4, P5, and P6 are all lost, as well as the packet that discloses key K_1, so the receiver can still not authenticate P1, P2, or P3. In interval 4 the base station broadcasts key K_2, which the node authenticates by verifying $K_0 = F(F(K_2))$, and hence knows also $K_1 = F(K_2)$, so it can authenticate packets P1, P2 with K_1, and P3 with K_2.

- Instead of adding a disclosed key to each data packet, the key disclosure is independent from the packets broadcast, and is tied to time intervals. Within the context of µTESLA, the sender broadcasts the current key periodically in a special packet.

4.12.2.3 µTESLA Detailed Description

- µTESLA has multiple phases:

  - **Sender setup:** The sender first generates a sequence of secret keys (or key chain). To generate the one-way key chain of length n, the sender chooses the last key K_n randomly, and generates the remaining values by successively applying a one-way function F.

  - **Sending authenticated packets:** Time is divided into time intervals and the sender associates each key of the one-way key chain with one time interval. In time interval t, the sender uses the key of the current interval, K_t, to compute the message authentication code (MAC) of packets in that interval.

- **Bootstrapping new receivers:** The important property of the one-way key chain is that once the receiver has an authenticated key of the chain, subsequent keys of the chain are self-authenticating, which means that the receiver can easily and efficiently authenticate subsequent keys of the one-way key chain using the one authenticated key. For example, if a receiver has an authenticated value K_i of the key chain, it can easily authenticate K_{i+1}, by verifying $K_i = F(K_{i+1})$. Therefore to bootstrap µTESLA, each receiver needs to have one authentic key of the one-way key chain as a commitment to the entire chain

- **Authenticating packets:** When a receiver receives the packets with the MAC, it needs to ensure that the packet could not have been spoofed by an adversary. The threat is that the adversary already knows the disclosed key of a time interval and so it could forge the packet since it knows the key used to compute the MAC. Hence the receiver needs to be sure that the sender did not disclose the key yet which corresponds to an incoming packet, which implies that no adversary could have forged the contents. This is called the security condition, which receivers check for all incoming packets.

4.13 Reliability Requirements in Sensor Networks

➢ The sensor networks are not designed with the goal of transporting multiple independent data streams. Sensor networks are data-centric and rely on in-network processing. The reliability requirements are pretty much application specific and the protocols can take advantage of this;

4.13.1 Single packet versus block versus stream delivery

➢ The cases of delivering only a single packet on the one hand and of delivering a number or even an infinite stream of packets on the other hand differ substantially in the protocol mechanisms usable in either case.

➢ In the single packet delivery problem, a single packet must be reliably transported between two nodes.

➢ In the block delivery problem, a finite data block comprising multiple packets must be delivered to a sensor or a set of sensors.

➢ In the stream delivery problem, a theoretically unbounded number of packets has to be transported between two nodes.

4.13.2 Sink-to-sensors versus sensors-to-sink versus localsensor-to-sensor

- It can be assumed that most communications in sensor networks are not between arbitrary peer nodes, but information flows either from sensor nodes towards a single or a few sink/gateway nodes or in critical environments such as military applications, it is necessary that the sink is able to transmit the data to the sensors in the least possible time**.**
- In the case of sensor to sensor communications, the sensors monitor a region and transmit the collected data packets through routes (intermediate sensor nodes) to the sinks.

4.13.3 Guaranteed versus stochastic delivery

- In the case of guaranteed delivery, it is expected that all transmitted packets reach the destination; anything else is considered a failure. In general, guaranteed delivery is challenging and costly in terms of energy and bandwidth expenditure, specifically over links with sometimes high error rates like wireless ones.
- The concept of stochastic delivery guarantees allows a limited amount of losses. There are several ways to specify stochastic guarantees. For example, one might specify that for periodic data delivery within every k subsequent packets at least m packets must reach the destination; any number below m is considered a failur

UNIT – 5 SENSOR NETWORK PLATFORMS AND TOOLS

Sensor Node Hardware – Berkeley Motes, Programming Challenges, Node-level software platforms – TinyOS, nesC, CONTIKIOS, Node-level Simulators – NS2 and its extension to sensor networks, COOJA, TOSSIM, Programming beyond individual nodes – State centric programming.

TABLE OF CONTENTS

5.1 Sensor Node Hardware

- Sensor node hardware can be grouped into three categories, each of which entails a different trade-offs in the design choices.
  - Augmented general-purpose computers
  - Dedicated embedded sensor nodes
  - System on-chip (SoC) nodes

5.1.1 Augmented general-purpose computers

- These nodes typically run off-the-shelf operating systems such as WinCE, Linux, or real-time operating systems and use standard wireless communication protocols such as IEEE 802.11, Bluetooth, Zigbee etc.

- Because of their relatively higher processing capability, they can accommodate wide variety of sensors, ranging from simple microphones to more sophisticated video cameras. It is fully supported for popular programming languages.
- Examples include low-power PCs, embedded PCs (e.g. PC104), custom-designed PCs, (e.g. Sensoria WINS NG nodes), and various personal digital assistants (PDA).

5.1.2 Dedicated embedded sensor nodes

- These platforms typically use commercial off-the-shelf (COTS) chip sets with emphasis on small form factor, low power processing and communication, and simple sensor interfaces.

- Because of their COTS CPU, these platforms typically support at least one programming language, such as C. However, in order to keep the program footprint small to accommodate their small memory size, programmers of these platforms are given full access to hardware but rarely any operating system support.
- Examples include the Berkeley mote family, the UCLA Medusa family, Ember nodes and MIT µAMP. A classical example is the TinyOS platform and its companion programming language, nesC, mica.

5.1.3 System on-chip (SoC) nodes

- These platforms try to push the hardware limits by fundamentally rethinking the hardware architecture trade-offs for a sensor node at the chip design level.
- The goal is to find new ways of integrating CMOS, MEMS, and RF technologies to build extremely low power and small footprint sensor nodes that still provide certain sensing, computation, and communication capabilities.
- Examples of SoC hardware include smart dust the BWRC picoradio node, and the PASTA node.

5.2 Berkeley Motes

- Berkeley Mote platform as it is an open hardware/software, smart-sensing platform with a large user community.
- The Berkeley Mote platform was developed under the Networked Embedded Systems Technology (NEST) program with the quantitative target of building dependable, real-time, distributed, embedded applications comprising 100 to 100 000 simple computing nodes.
- Berkeley motes tiny, self-contained, battery powered computers with radio links, which enable to communicate and exchange data with one other, and to self- organize into ad hoc networks.
- Motes form the building blocks of wireless sensor networks.
- The platform consists of four basic components: Power, sensors, computation, and communication. These motes are autonomous and connectable to other motes.
- The main advantages are small physical size, low cost, modest power consumption, and diversity in design and usage. The latest versions of the Berkeley Mote include the MicaZ, Mica2, and Mica2dot processor boards (Fig. 5.1). The Motes have

improvements in memory and radio over predecessors and specifications are summarised in Table 5.1. The same sensor board can be also used for the MicaZ and modified for use with the Mica2dot

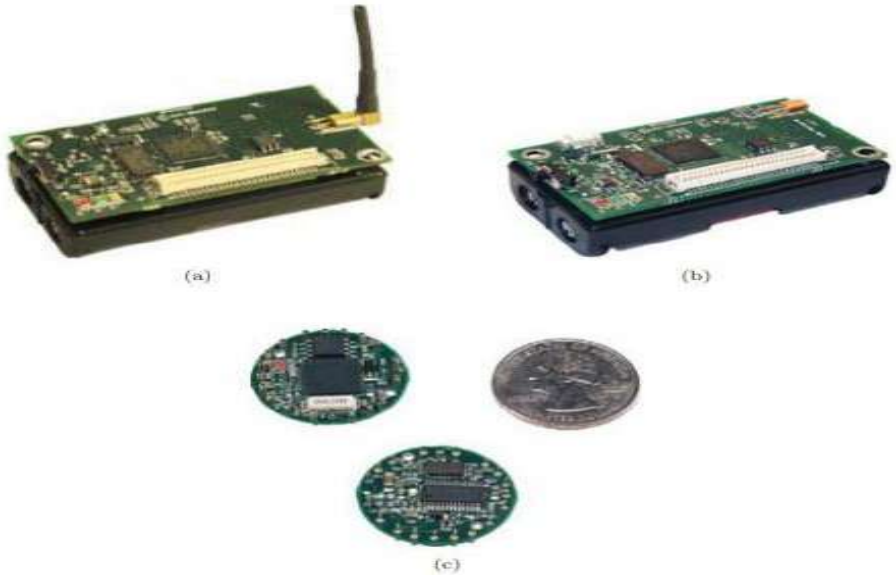


Figure 5.1 Berkeley Mote processor boards: (a) MicaZ, (b) Mica2, and (c) Mica2dot.
Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

Table 5.1 Characteristics of the MicaZ, Mica2, and Mica2dot processor boards

	MicaZ	Mica2	Mica2dot
Flash memory	128 K bytes	128 K bytes	128 K bytes
Measurement memory	512 K bytes	512 K bytes	512 K bytes
EEPROM	4 K bytes	4 K bytes	4 K bytes
A/D (Channels)	10 bits (8)	10 bits (8)	10 bits (6)
Frequency	1400 MHz–2483.5 MHz	433/868/916 MHz	433/868/916 MHz
Data rate	250 K bps	19.2 K bps	19.2 K bps
Outdoor range	100 m	300 m	300 m
Size	6×3×1 cm	6×3×1 cm	2.5×0.6 cm

- The Motes have the versatility to connect different Printed Circuit Boards (PCB); that is, the Motes have the modularity to support different types of sensors. Users can switch the sensor board or customise it, independent of the other hardware components. Table 5. 2 provides a description of the sensors that are used on the respective boards.

- The current sensor board designs for the Mica2 platform includes: MTS101CA, MTS300CA, MTS310CA, MTS400CA, and MTS420CA

Table 5. 2. Description of the available sensor boards

Name	Sensors
MTS101CA	Photo resistor and thermistor.
MTS300CA	Photo resistor, thermistor, acoustic sensor and acoustic actuator.
MTS310CA	Photo resistor, thermistor, acoustic sensor, acoustic actuator, biaxial accelerometer and magnetometer.
MTS400/420CA	Thermistor, hygrometer, barometer, Photo resister, accelerometer, GPS(only on MTS420CA)

5.3 Sensor Network Programming Challenges

- Traditional programming technologies rely on operating systems to provide abstraction for processing, I/O, networking, and user interaction hardware. When applying such a model to programming networked embedded systems, such as sensor networks, the application programmers need to explicitly deal with message passing, event synchronization, interrupt handling, and sensor reading.
- As a result, an application is typically implemented as a finite state machine (FSM) that covers all extreme cases: unreliable communication channels, long delays, irregular arrival of messages, simultaneous events etc.
- For resource-constrained embedded systems with real-time requirements, several mechanisms are used in embedded operating systems to reduce code size, improve response time, and reduce energy consumption.
- The microkernel technologies modularize the operating system so that only the necessary parts are deployed with the application. Real-time scheduling allocates resources to more urgent tasks so that they can be finished early.
- Event-driven execution allows the system to fall into low-power sleep mode when no interesting events need to be processed.
- At the extreme, embedded operating systems tend to expose more hardware controls to the programmers, who now have to directly face device drivers and scheduling algorithms, and optimize code at the assembly level.
- Although these techniques may work well for small, stand-alone embedded systems, they do not scale up for the programming of sensor networks for two reasons:

- **Sensor networks are large-scale distributed systems**, where global properties are derivable from program execution in a massive number of distributed nodes. Distributed algorithms themselves are hard to implement, especially when infrastructure support is limited due to the ad hoc formation of the system and constrained power, memory, and bandwidth resources.

- **As sensor nodes deeply embed into the physical world,** a sensor network should be able to respond to multiple concurrent stimuli at the speed of changes of the physical phenomena of interest.

➢ There no single universal design methodology for all applications. Depending on the specific tasks of a sensor network and the way the sensor nodes are organized, certain methodologies and platforms may be better choices than others.

➢ For example, if the network is used for monitoring a small set of phenomena and the sensor nodes are organized in a simple star topology, then a client-server software model would be sufficient.

➢ If the network is used for monitoring a large area from a single access point (i.e., the base station), and if user queries can be decoupled into aggregations of sensor readings from a subset of nodes, then a tree structure that is rooted at the base station is a better choice. However, if the phenomena to be monitored are moving targets, as in the target tracking, then neither the simple client-server model nor the tree organization is optimal. More sophisticated design and methodologies and platforms are required.

5.4 Node-Level Software Platforms

➢ Most design methodologies for sensor network software are node-centric, where programmers think in terms of how a node should behave in the environment.

➢ A node level platform can be node-centric operating system, which provides hardware and networking abstractions of a sensor node to programmers, or it can be a language platform, which provides a library of components to programmers.

➢ A typical operating system abstracts the hardware platform by providing a set of services for applications, including file management, memory allocation, task scheduling, peripheral device drivers, and networking.

➢ For embedded systems, due to their highly specialized applications and limited resources, their operating systems make different trade-offs when providing these services.

- For example, if there is no file management requirement, then a file system is obviously not needed. If there is no dynamic memory allocation, then memory management can be simplified. If prioritization among tasks is critical, then a more elaborate priority scheduling mechanism may be added.

5.5 Operating System Design Issues

- Traditional operating systems are system software, including programs that manage computing resources, control peripheral devices, and provide software abstraction to the application software.

- Traditional OS functions are therefore to manage processes, memory, CPU time, file system, and devices. This is often implemented in a modular and layered fashion, including a lower layer of kernels and a higher layer of system libraries.

- Traditional OSs are not suitable for wireless sensor networks because WSNs have constrained resources and diverse data-centric applications, in addition to a variable topology.

- Hence, WSNs need a new type of operating system, considering their special characteristics. There are several issues to consider when designing operating systems for wireless sensor networks.

- The **first issue** is process management and scheduling. The traditional OS provides process protection by allocating a separate memory space (stack) for each process. Each process maintains data and information in its own space. But this approach usually causes multiple data copying and context switching between processes. This is obviously not energy efficient for WSNs. For some real-time applications in WSNs, a real-time scheduler such as earliest deadline first (EDF) or its variants may be a good choice, but the number of processes should be confined since that would determine the time complexity of the EDF scheduler.

- The **second issue** is memory management. Memory is often allocated exclusively for each process/task in traditional operating systems, which is helpful for protection and security of the tasks. Since sensor nodes have small memory, another approach, sharing, can reduce memory requirements.

- The **third issue** is the kernel model. The event-driven and finite state machine(FSM) models have been used to design microkernels for WSNs. The event-driven model may serve WSNs well because they look like event-driven systems. An event may comprise receiving a packet, transmitting a packet, detection of an event of interest,

alarms about energy depletion of a sensor node, and so on. The FSM-based model is convenient to realize concurrency, reactivity, and synchronization.

- The **fourth issue** is the application program interface (API). Sensor nodes need to provide modular and general APIs for their applications. The APIs should enable applications access the underlying hardware.

- The **fifth issue** is code upgrade and reprogramming. Since the behavior of sensor nodes and their algorithms may need to be adjusted either for their functionality or for energy conservation, the operating system should be able to reprogram and upgrade.

- **Finally,** because sensor nodes generally have no external disk, the operating system for WSNs cannot have a file system. These issues should be considered carefully in the design of WSN OSs and to meet their constrained resources, network behavior, and data-centric application requirements.

- Sensor operating systems (SOS) should represent the following functions, bearing in mind the limited resource of sensor nodes:

  - Should be compact and small in size since the sensor nodes have very small memory. The sensor nodes often have memories of only tens or hundreds of kilobytes.

  - Should provide real-time support, since there are real-time applications, especially when actuators are involved. The information received may become outdated rather quickly. Therefore, information should be collected and reported as quickly as possible.

  - Should provide efficient resource management mechanisms in order to allocate microprocessor time and limited memory. The CPU time and limited memory must be scheduled and allocated for processes carefully to guarantee fairness (or priority if required).

  - Should support reliable and efficient code distribution since the functionality performed by the sensor nodes may need to be changed after deployment. The code distribution must keep WSNs running normally and use as little wireless bandwidth as possible.

  - Should support power management, which helps to extend the system lifetime and improve its performance. For example, the operating system may schedule the process to sleep when the system is idle, and to wake up with the advent of an incoming event or an interrupt from the hardware.

- Should provide a generic programming interface up to sensor middleware or application software. This may allow access and control of hardware directly, to optimize system performance.

5.6 Operating System: TinyOS

- The design of TinyOS allows application software to access hardware directly when required. TinyOS is a tiny micro threaded OS that attempts to address two issues:
  - How to guarantee concurrent data flows among hardware devices, and
  - How to provide modularized components with little processing and storage overhead.
- These issues are important since TinyOS is required to manage hardware capabilities and resources effectively while supporting concurrent operation in an efficient manner.
- TinyOS uses an event-based model to support high levels of concurrent application in a very small amount of memory. Compared with a stack-based threaded approach, which would require that stack space be reserved for each execution context, and because the switching rate of execution context is slower than in an event-based approach, TinyOS achieves higher throughput.
- It can rapidly create tasks associated with an event, with no blocking or polling. When CPU is idle, the process is maintained in a sleep state to conserve energy. TinyOS includes a tiny scheduler and a set of components. The scheduler schedules operation of those components.
- Each component consists of four parts: command handlers, event handlers, an encapsulated fixed-size frame, and a group of tasks
- Commands and tasks are executed in the context of the frame and operate on its state. Each component will declare its commands and events to enable modularity and easy interaction with other components.
- The current task scheduler in TinyOS is a simple FIFO mechanism whose scheduling data structure is very small, but it is power efficient since it allows a processor to sleep when the task queue is empty and while the peripheral devices are still running. The frame is fixed in size and is assigned statically. It specifies the memory requirements of a component at compile time and removes the overhead from dynamic assignment. Commands are non-blocking requests made to the low-level components. Therefore, commands do not have to wait a long time to be executed.

- A command provides feedback by returning status indicating whether it was successful (e.g., in the case of buffer overrun or of timeout). A command often stores request parameters into its frame and conditionally assigns a task for later execution.
- The occurrence of a hardware event will invoke event handlers. An event handler can store information in its frame, assign tasks, and issue high-level events or call low-level commands. Both commands and events can be used to perform a small and usually fixed amount of work as well as to pre-empt tasks.
- Tasks are a major part of components. Like events, tasks can call low-level commands, issue high-level events, and assign other tasks. Through groups of tasks, TinyOS can realize arbitrary computation in an event-based model.
- The design of components makes it easy to connect various components in the form of function calls. The architecture of TinyOS shown in Figure 5.2.
- This WNS operating system defines three type of components:
  - Hardware abstractions
  - Synthetic hardware
  - High-level software components
- **Hardware abstraction components** are the lowest-level components. They are actually the mapping of physical hardware such as I/O devices, a radio transceiver, and sensors. Each component is mapped to a certain hardware abstraction.
- **Synthetic hardware components** are used to map the behavior of advanced hardware and often sit on the hardware abstraction components. TinyOS designs a hardware abstract component called the radio-frequency module (RFM) for the radio transceiver, and a synthetic hardware component called radio byte, which handles data into or out of the underlying RFM.
- **Higher-level components** encapsulate software functionality, but with a similar abstraction. They provide commands, signal events, and have internal handlers, task threads, and state variables

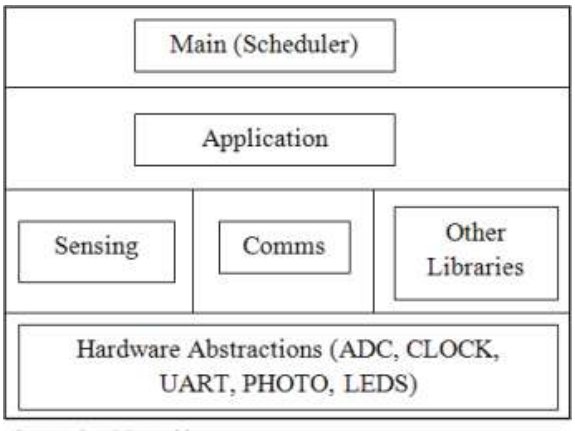


Figure 5.2 TinyOS Architecture

Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

- An evaluation of TinyOS shows that it achieves the following performance gains or advantages:
  - It requires very little code and a small amount of data.
  - Events are propagated quickly and the rate of posting a task and switching the corresponding context is very high.
  - It enjoys efficient modularity.

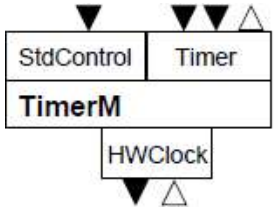
5.7 nesC

- nesC is a component-based, event-driven programming language used to build applications for the TinyOS platform. TinyOS is an operating environment designed to run on embedded devices used in distributed wireless sensor networks.
- The name nesC is an abbreviation of "network embedded systems C". nesC is an extension of C.
- nesC programs are subject to whole program analysis (for safety) and optimization (for performance). Therefore we do not consider separate compilation in nesC's design. The limited program size on motes makes this approach tractable.
- nesC is a "static language". There is no dynamic memory allocation and the call-graph is fully known at compile-time. These restrictions make whole program analysis and optimization significantly simpler and more accurate. nesC's component model and parameterized interfaces eliminate many needs for dynamic memory allocation and dynamic dispatch.

- nesC is based on the concept of components, and directly supports TinyOS's event based concurrency model. Additionally, nesC explicitly addresses the issue of concurrent access to shared data. In practice, nesC resolved many ambiguities in the TinyOS concepts of components and concurrency.

Component Specification

- nesC applications are built by writing and assembling components. A component provides and uses interfaces. These interfaces are the only point of access to the component. An interface generally models some service (e.g., sending a message) and is specified by an interface type. Figure 5.3 shows the TimerM component, part of the TinyOS timer service that provides the StdControl and Timer interfaces and uses a Clock interface (all shown in Figure 5.4).




```
module TimerM {
  provides {
    interface StdControl;
    interface Timer;
  }
  uses interface Clock as Clk;
} ...
```

Figure 5.3 Specification and graphical depiction of the TimerM component

Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

```
interface StdControl {
  command result_t init();
}

interface Timer {
  command result_t start(char type, uint32_t interval);
  command result_t stop();
  event result_t fired();
}

interface Clock {
  command result_t setRate(char interval, char scale);
  event result_t fire();
}

interface Send {
  command result_t send(TOS_Msg *msg, uint16_t length);
  event result_t sendDone(TOS_Msg *msg, result_t success);
}

interface ADC {
  command result_t getData();
  event result_t dataReady(uint16_t data);
}
```

Figure 5.4 Some Interface Types

- TimerM provides the logic that maps from a hardware clock (Clock) into TinyOS's timer abstraction (Timer).
- Interfaces in nesC are bidirectional. They contain commands and events, both of which are essentially functions. The providers or an interface implement the commands, while the users implements the events. For instance, the Timer interface (Figure 5.4) defines start and stop commands and a fired event.
- In Figure 5.3 provided interfaces are shown above the TimerM component and used interfaces are below; downward-pointing arrows depict commands and upward-pointing arrows depict events. Although this same interaction between the timer and its client could have been provided via two separate interfaces (one for start and stop, and one for fired), grouping these commands and events in the same interface makes the specification much clearer and helps prevent bugs when wiring components together.
- Split-phase operations are cleanly modelled by placing the command request and event response in the same interface. Figure 5.4 shows two examples of this.
- The Send interface has the send command and sendDone event of the split-phased packet send. The ADC interface is similarly used to model split-phase sensor value reads. The separation of interface type definitions from their use in components promotes the definition of standard interfaces, making components more reusable and flexible.
- A component can provide and use the same interface type (e.g., when interposing a component between a client and service), or provide the same interface multiple times. In these cases, the component must give each interface instance a separate name using the as notation shown for Clk in Figure 5.3.
- The components are also a clean way to abstract the boundary between hardware and software. For instance, on one sensor board, the temperature sensor (accessed via a component named Temp) is mostly in hardware; Temp is a thin layer of software accessing on-chip hardware registers.

Component Implementation

- There are two types of components in nesC: modules and configurations. Modules provide application code, implementing one or more interfaces. Configurations are used to wire other components together, connecting interfaces used by components to interfaces provided by others.

Concurrency and Atomicity

- nesC detects the data races at compile time. Data races occur due to concurrent updates to shared state. In order to prevent them, a compiler must
  - Understand the concurrency model,
  - Determine the target of every update

5.8 ContikiOS

- ContikiOS is open source operating system for resource constraint hardware devices with low power and less memory. It was developed by Adam Dunkels in 2002. This OS is fully GUI based system requires only 30 KB ROM and 10 KB RAM. It also provide multitasking feature and have the built in TCP/IP suit.
- The working environments of the WSNs are often energy-limited. This is one of the most important constraint for WSNs. Likewise, tiny and simple designs of the nodes are the other constraints. For this reason, WSNs should have some important hardware and software features to cope with these constraints.
- Contiki OS is one of the convenient solutions to cope with mentioned constraints to its flexibility and support of lightweight and low-powered networks.
- Contiki can provide communication over IPv4, IPv6 and Rime Network Stack. Contiki Network Stack shown in Figure 5.5 gives more details for its structure.

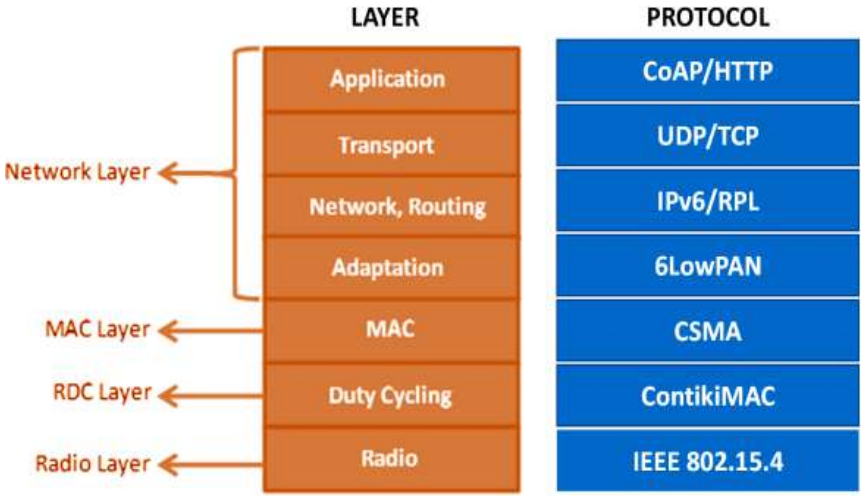


Figure 5.5 Contiki Network Stack

Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

- Many Contiki systems are severely power-constrained. Battery operated wireless sensors may need to provide years of unattended operation and with little means to recharge or replace batteries.

- Contiki provides a set of mechanisms to reduce the power consumption of systems on which it runs. The default mechanism for attaining low-power operation of the radio is called ContikiMAC. With ContikiMAC, nodes can be running in low-power mode and still be able to receive and relay radio messages.

- The Contiki programming model is based on protothreads. A protothread is a memory-efficient programming abstraction that shares features of both multithreading and event-driven programming to attain a low memory overhead of each protothread.

- The kernel invokes the protothread of a process in response to an internal or external event. Examples of internal events are timers that fire or messages being posted from other processes. Examples of external events are sensors that trigger or incoming packets from a radio neighbour.

5.9 Node-Level Simulators

- Node-level design methodologies are usually associated with simulators that simulate the behavior of a sensor network on a per-node basis. Using simulation, designers can quickly study the performance (in terms of timing, power, bandwidth, and scalability) of potential algorithms without implementing them on actual hardware and dealing with the vagaries of actual physical phenomena. A node-level simulator typically has the following components:

  - **Sensor node model:** A node in a simulator acts as a software execution platform, a sensor host, as well as a communication terminal. In order for designers to focus on the application-level code, a node model typically provides or simulates a communication protocol stack, sensor behaviours (e.g., sensing noise), and operating system services. If the nodes are mobile, then the positions and motion properties of the nodes need to be modelled. If energy characteristics are part of the design considerations, then the power consumption of the nodes needs to be modelled.

  - **Communication model:** Depending on the details of modelling, communication may be captured at different layers. The most elaborate simulators model the communication media at the physical layer, simulating the RF propagation delay and collision of simultaneous transmissions. Alternately, the communication may

be simulated at the MAC layer or network layer, using, for example, stochastic processes to represent low-level behaviours.

- **Physical environment model:** A key element of the environment within a sensor network operates is the physical phenomenon of interest. The environment can also be simulated at various levels of details. For example, a moving object in the physical world may be abstracted into a point signal source. The motion of the point signal source may be modelled by differential equations or interpolated from a trajectory profile.

- **Statistics and visualization:** The simulation results need to be collected for analysis. Since the goal of a simulation is typically to derive global properties from the execution of individual nodes, visualizing global behaviours is extremely important. An ideal visualization tool should allow users to easily observe on demand the spatial distribution and mobility of the nodes, the connectivity among nodes, link qualities, end-to-end communication routes and delays, phenomena and their spatio-temporal dynamics, sensor readings on each node, sensor nodes states, and node lifetime parameters (e.g., battery power).

➢ A sensor network simulator simulates the behavior of a subset of the sensor nodes with respect to time. Depending on how the time is advanced in the simulation, there are two types of execution models:

  - Cycle-Driven Simulation
  - Discrete-Event Simulation

5.9.1 Cycle-Driven Simulation

- A cycle-driven (CD) simulation discretizes the continuous notion of real time into (typically regularly spaced) ticks and simulates the system behavior at these ticks. At each tick, the physical phenomena are first simulated, and then all nodes are checked to see if they have anything to sense, process, or communicate.

- Sensing and computation are assumed to be finished before the next tick. Sending a packet is also assumed to be completed by then. However, the packet will not be available for the destination node until next tick. This split-phase communication is a key mechanism to reduce cyclic dependencies that may occur in cycle-driven simulations.

5.9.2 Discrete-Event Simulation

- A Discrete-Event (DE) simulator assumes that the time is continuous and an event may occur at any time. As event is 2-tuple with a value and a time stamp indicating when the event is supposed to be handled. Components in a DE simulation react to input events and produce output events. In node-level simulators, a component can be a sensor node, and the events can be communication packets; or a component can be software module within and the events can be message passing among these nodes.

- Typically, components are causal, in the sense that if an output event is computed from an input event, then the time stamp of the output should not be earlier than that of the input event. Non-causal components require the simulators to be able to roll back in time, and worse, they may not define a deterministic behavior of a system.

- A DE simulator typically requires a global event queue. All events passing between nodes or modules are put in the event queue and sorted according to their chronological order. At each iteration of the simulation, the simulator removes the first event (the one with earliest time stamp) from the queue and triggers the component that reacts to that event.

➢ In terms of timing behaviour, a DE simulator is more accurate than a CD simulator, and as a consequence, DE simulators run slower. The overhead of ordering all events and computation, in addition to the values and time stamps of events, usually dominates the computation time.

➢ CD simulations usually require less complex components and give faster simulations. DE simulations are sometimes considered as good as actual implementations, because of their continuous notion of time and discrete notion of events.

➢ There are several open source or commercial simulators available. One class of these simulators comprises extensions of classical network simulators, such as ns-2, J-Sim (previously known as JavaSim), and GloMoSim/ Qualnet. The focus of these simulators is on network modelling, protocol stacks, and simulation performance.

➢ Another class of simulators, sometimes called software-in-the-loop simulators, incorporate the actual node software into the simulation. For this reason, they are typically attached to particular hardware platforms and are less portable. Example include TOSSIM for Berkeley motes.

5.10 NS2 and its Extension to Sensor Networks

- The NS-2 (Network Simulator-2) is a well-known network simulator for discrete event simulation. Simulations are based on a combination of C++ and OTcl.
- NS-2 includes a large number of simulated network protocols and tools used for simulating transport control protocol (TCP), routing algorithm, multicast protocol over the wired or wireless (local connection or via satellite connection) networks.
- NS-2 is committed to OSI model simulation, including the behaviour of physical layer and it is a free open source software and available for free download.

Limitations of NS-2

- It puts some restrictions on the customisation of packet formats, energy models, MAC protocols, and the sensing hardware models, which limits its flexibility.
- The lack of an application model makes it ineffective in environments that require interaction between applications and the network protocols.
- It does not run real hardware code.
- It has been built by many developers and contains several inherent known and unknown bugs.
- It does not scale well for WSNs due to its object-oriented design.
- Using C++ code and OTcl scripts make it difficult to use.

- Actually, NS-2 was not initially designed to simulate wireless sensor network, but a few research groups had extended NS-2 in order to enable it to support wireless sensor network simulation, including sensor model, battery model, a small stack, and hybrid simulation tools.
- It is extensible, but not very scalable because of the split programming model and object-oriented structure. In addition, because NS-2 can simulate very detailed data packet close to the exact number of running packets, it is unable to carry out large-scale network simulation.
- To overcome the above drawbacks the improved NS-3 simulator was developed. NS-3 supports simulation and emulation. It is totally written in C++, while users can use python scripts to define simulations.

- Hence, transferring NS-2 implementation to NS-3 require manual intervention. Besides the scalability and performance improvements, simulation nodes have the ability to support multiple radio interfaces and multiple channels.
- Furthermore, NS-3 supports a real-time schedule that makes it possible to interact with real systems. For example, a real network device can emit and receive NS-3 generated packets.

5.11 COOJA

- Cooja simulator is the efficient simulate wireless sensor networks. Cooja is the default simulator of Contiki operating system that helps to simulate the wireless sensor networks in addition it helps to do the performance evolution.
- Contiki is a light weight operating system that is developed mainly for wireless nodes. The motes that are developed by the contiki offers many advantages.
- Contiki offers a java based simulator called as cooja which is used to simulate the wireless sensors. Cooja simulator is more flexible so that many parts of the simulator is replaceable and extendable. The parts of the simulator like simulated node hardware, plug-ins and radio medium can be replaceable.

Characteristics of Cooja

- Scalability
- Efficiency
- Extensibility
- Flexibility

- Wireless sensor network has the powerful tool called tool in which it can be simulate the idea before it is implementing in real time. Contiki Cooja WSN Simulator mainly used to simulate many wireless scenario.

Contiki Cooja WSN simulator

- Contiki cooja is the best simulator to simulate any wireless sensors with its own property. For example, if we are designing a wireless sensor network that detects the earth quake, the sensor has its own property like lifetime, withstand ability, capacity, etc.

- We can design this wireless sensors with the same property in contiki cooja. When compared to other simulators cooja is developed purely for wireless sensor networks.

- In addition cooja is more flexible to change the properties of a node so that we could implement our own idea exactly. Wireless sensors play important role in IOT (Internet of Thins), where contiki Operating system was developed mainly for IOT devices, cooja is a simulator comes with the Contiki. So we can use the Cooja simulator for simulating any wireless sensor networks.

5.12 TOSSIM

- TOSSIM (TinyOS Mote Simulator) is an open-source operating system specially developed for the wireless embedded sensor networks. There are few hardware platforms available for TinyOS, some commercial and some non-commercial.

- TinyOS release includes a simulator called TOSSIM. It is built especially for Berkeley Mica Mote platform. TOSSIM is an emulator rather than a simulator, as it runs actual application code. Simulated application code can be transferred directly to the platform, but it might not run in a mote as it runs in a simulation due to the simplifying assumptions in TOSSIM.

- Figure 5.6 shows the working flow of TOSSIM. The TOSSIM architecture is consisted of five segments: Frames, Components, Models, Services and Events.

- TOSSIM is a very simple but powerful emulator for WSN. Each node can be evaluated under perfect transmission conditions, and using this emulator can capture the hidden terminal problems.

- As a specific network emulator, TOSSIM can support thousands of nodes simulation. This is a very good feature, because it can more accurately simulate the real world situation. Besides network, TOSSIM can emulate radio models and code executions. This emulator may be provided more precise simulation result at component levels because of compiling directly to native codes.

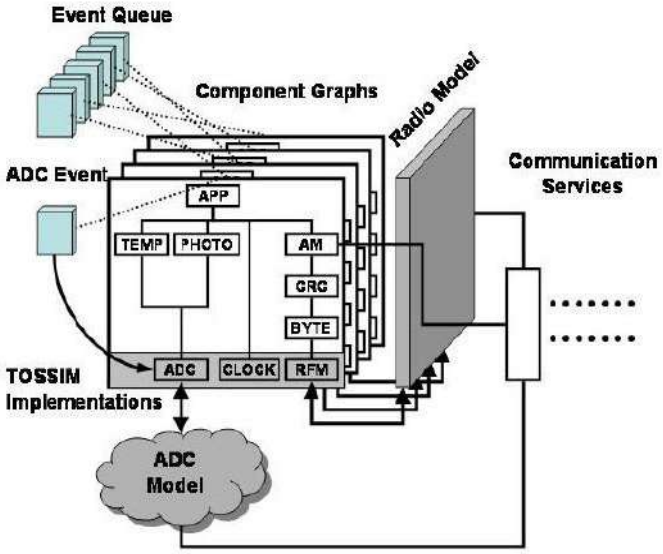


Figure 5.6 TOSSIM Architecture

Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

- TOSSIM is a bit-level discrete event network emulator built in Python, a high-level programming language emphasizing code readability, and C++. It can run TOSSIM on Linux Operating Systems or on Cygwin on Windows.
- TOSSIM also provides open sources and online documents. Developers had set four requirements for TOSSIM: scalability, completeness, fidelity and bridging.
- To be scalable, a simulator should manage networks of thousands of nodes in a wide variety of configurations. To achieve this, each node in TOSSIM is connected in a directed graph where each edge has a probabilistic biterror.
- For completeness, a simulator must capture behavior and interactions of a system at a wide variety of levels. And for fidelity, a simulator must capture behavior of a network with a subtle timing of interactions on a mote and between motes. Requirement for bridging is met as the simulated code runs directly in a real mote.
- The goal of TOSSIM is to study the behavior of TinyOS and its applications rather than performance metrics of some new protocol. Hence, it has some limitations, for instance, it does not capture energy consumption. Another drawback of this framework is that every node must run the same code. Therefore, TOSSIM cannot be used to evaluate some types of heterogeneous applications.

5.13 Programming Beyond Individual Nodes

- Sensor-actuator network systems offer some unique advantages. Dense networks of distributed sensors can improve perceived signal-to-noise ratio by reducing average distances from sensor to physical phenomena.
- In-network processing and actuation shorten the feedback chain and improve the timeliness of observation and response. Untethered network nodes and infrastructure less mesh network topologies reduce deployment costs. However, the greatest advantages of networked systems are improved robustness and scalability.
- A decentralized system is inherently more robust against individual node or link failures because of network redundancy. Decentralized algorithms are also far more scalable in practical deployment; they might be the only way to achieve the large scales needed for some applications. Because of decentralized systems spatial coverage and multiplicity in sensing aspect and modality, the detection, classification, and tracking of moving, nonlocal, or low-observable events require cross-node collaboration among sensors.

Target tracking as a motivating example

- Tracking is a canonical problem for sensor networks and essential for many commercial and military applications such as traffic monitoring, facility security, and battlefield situational awareness.
- Given a moving point signal source or target in a 2D sensor field, a tracking system's goal is to estimate target state histories, such as spatial trajectory, on the basis of sensor measurements.
- From a tracking expert's point of view, each sensor node provides a local measurement useful in estimating the target state. However, in most cases, only a relatively small subset of sensors contribute significantly to the estimation, owing to sensing-range limitations. In this case, a good solution is a leader-based tracking scheme, such as Information-Driven Sensor Querying (IDSQ), to fuse information from only the sensors that provide high-quality measurements.
- As Figure 5.7 illustrates, at any time instant t, IDSQ designates a single node, located close to the target, as leader. The leader node fuses these high signal-to-noise ratio measurements and updates its current target location estimate, referred to as the belief.
- For most sensor types, owing to the physical properties of signal propagation, the sensors with high signal-to-noise ratio will be within a limited range of the leader node. So, we can minimize the communication cost and latency for gathering sensor data.

- As the target traverses the sensor field and the belief evolves to follow its motion, the most "informative" sensors might no longer be those closest to the current leader. A nearby sensor might then be selected to replace this leader on the basis of the updated belief and a criterion combining resource constraints with some measure of sensing utility (such as mutual information). The current leader then hands off the belief to this sensor, which becomes the next leader at time $t + \delta$, where δ is the communication delay. The process of sensing, estimation, and leader selection repeats.

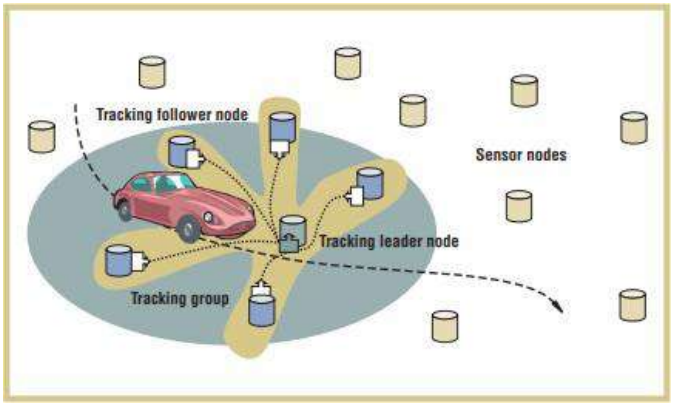


Figure 5.7. Collaborative processing in a leader-based object-tracking scenario.
Source : Protocol and Architecture for Wireless Sensor Networks by Holger Karl , Andreas willig

- As a vehicle moves through a sensor field, nearby sensors detect it. An elected leader node aggregates data from the active sensors and migrates the information from node to node as the vehicle moves.
- The sensor nodes collaborate primarily to improve sensing accuracy, and acceptable estimation quality might be achieved using only a subset of the sensors.
- One node, the leader, plays a key role in fusing others' sensor measurements. If no leader is present, all sensors that form the contour are equally important. Each node might locally update and repair its observation of a contour section, but the global state can only be assembled from observations of many nodes along the entire contour. Hence, System designers must explicitly write code to
  - Maintain sensor connectivities in a neighbourhood
  - Discover the best node for handoff
  - Invite neighbour nodes into the group
  - Handle communication delays and failures

5.14 State Centric Programming

- Many sensor network applications, such as target tracking (Discussed in previous section 5.13), are not simply generic distributed programs over an ad hoc network of energy-constrained nodes.
- Deeply rooted in these applications is the notion of states of physical phenomena and models of their evolution over space and time. Some of these states may be represented on a small number of nodes and evolve over time, as in the target tracking problem, while others may be represented over a large and spatially distributed number of nodes, as in tracking a temperature contour.
- A distinctive property of physical states, such as location, shape, and motion of objects, is their continuity in space and time. Their sensing and control is typically done through sequential state updates. System theories, the basis for most signal and information processing algorithms, provide abstractions for state updates, such as:

  $x_{k+1} = f(x_k, u_k)$

  $y_k = g(x_k, u_k)$
- Where **x** is the state of a system, u is the system input, **y** is the output and k is an integer update index over space and/or time, f is the state update function, and g is the output or observation function.
- This formulation is broad enough to capture a wide variety of algorithms in sensor fusion, signal processing, and control (e.g., Kalman filtering, Bayesian estimation, system identification, feedback control laws, and finite-state automata).
- However, in distributed real-time embedded systems such as sensor networks, the formulation is not as clean as represented in the above equations. The relationships among subsystems can be highly complex and dynamic over space and time.
- The following issues must be properly addressed during the design to ensure the correctness and efficiency of the system.
  - Where are the state variables stored?
  - Where do the inputs come from?
  - Where do the outputs go?
  - Where are the functions f and g evaluated?
  - How long does the acquisition of input take?
  - Are the inputs in uk collected synchronously?
  - Do the inputs arrive in the correct order through communication?
  - What is the time duration between indices k and k +1? Is it a constant?

- These issues, addressing where and when, rather than how, to perform sensing, computation, and communication, play a central role in the overall system performance.
- However, these 'non-functional" aspects of computation, related to concurrency, responsiveness, networking, and resource management, are not well supported by traditional programming models and languages.
- **State-centric programming aims** at providing design methodologies and frameworks that give meaningful abstractions for these issues, so that system designers can continue to write algorithms on top of an intuitive understanding of where and when the operations are performed.
- A collaborative group is such an abstraction. A collaborative group is a set of entities that contribute to a state update. These entities can be physical sensor nodes, or they can be more abstract system components such as virtual sensors or mobile agents hopping among sensors. These are all referred to as agents.
- Intuitively, a collaboration groups provides two abstractions: its scope to encapsulate network topologies and its structure to encapsulate communication protocols. The scope of a group defines the membership of the nodes with respect to the group.
- A software agent that hops among the sensor nodes to track a target is a virtual node, while a real node is physical sensor. Limiting the scope of a group to a subset of the entire space of all agents improves scalability.
- Grouping nodes according to some physical attributes rather than node addresses is an important and distinguishing characteristic of sensor networks. The structure of a group defines the "roles" each member plays in the group, and thus the flow of data.
  - Are all members in the group equal peers?
  - Is there a "leader" member in the group that consumes data?
  - Do members in the group form a tree with parent and children relations?
- For example, a group may have a leader node that collects certain sensor readings from all followers. By mapping the leader and the followers onto concrete sensor nodes, one can effectively define the flow of data from the hosts of followers to the host of the leader. The notion of roles also shields programmers from addressing individual nodes either by name or address.
- Furthermore, having multiple members with the same role provides some degree of redundancy and improves robustness of the application in the presence of node and link failures.

EC 8702-AD HOC AND WIRELESS SENSOR NETWORKS

TWO MARKS QUESTIONS WITH ANSWERS & QUESTION BANK

UNIT-1 Ad-Hoc Networks- Introductions and Routing Protocol

Elements of Ad hoc Wireless Networks, Issues in Ad hoc wireless networks, Example commercial applications of Ad hoc networking, Ad hoc wireless Internet, Issues in Designing a Routing Protocol for Ad Hoc Wireless Networks, Classifications of Routing Protocols, Table Driven Routing Protocols - Destination Sequenced Distance Vector (DSDV), On–Demand Routing protocols –Ad hoc On–Demand Distance Vector Routing (AODV).

PART-A

1. What is an Ad-Hoc Network?

- The term 'ad hoc' implies spontaneous construction of temporary wireless network that composed of individual devices communicating with one another directly with no centralized administration.
- Due to above ad hoc is an infrastructure less network otherwise Multi-hop wireless network (MHWN) in that computers or other devices are enabled to send data packets on to each other instead of browsing a centralized access point
- In other words Ad hoc networks are temporary network composed of mobile nodes without pre-existing communication infrastructure such as AP (Access Point) and BSS (Basic service set). Each node plays the role of router for multi-hop routing.

2. Define MHWN?

Multi-hop wireless network (MHWN) is defined as a set of nodes that communicate with one another wirelessly by using radio signals with a shared common channel. There are several names for MHWN; it might be called packet radio network, Ad-Hoc network or temporary mobile network.

3. List Characteristics/Features of Ad-Hoc network?

- Dynamic topology
- Bandwidth constraints and variable links
- Energy controlled nodes
- Multi-hop communications
- Limited security
- Determining/detecting the sudden changes in network topology
- Maintaining network topology/connectivity
- Scheduling of packet transmission
- Finding the shortest path to reach required destination by proper routing protocols

- Maintaining network connectivity even under changing radio conditions and mobility
- Proper transmission scheduling and channel assignment.

4. List the advantages of Ad-Hoc networks over existing traditional networks?

- Ad-hoc network is more flexible than traditional networks.
- It supports even the nodes under the mobility
- It's having special capability like self-organization and self-reconfiguration.
- It can be "Turn Up" and "Turn Down" in a very short time.
- It can be more economical.
- No need of router and switches during construction of network.
- It supports a robust network because of its non-hierarchical distributed control and management mechanisms.

5. Difference between traditional (cellular) network and ad-hoc network?

Sno.	Cellular Network	Ad-Hoc Network
1.	Based on fixed infrastructure	Based on independent infrastructure or infrastructureless network
2.	It is a single hop wireless links	It is multi hop wireless links
3.	Guaranteed Bandwidth	Not Guaranteed Bandwidth
4.	It is circuit switched	Packet switched
5.	More time deployment required	Less time deployment
6.	Maintenance cost is high	Maintenance cost is low
7.	It belongs seamless connectivity	Frequent Path breaks take places
8.	Static frequency reuse spectrum	Dynamic frequency reuse spectrum
9.	Centralized routing based network	Distributed routing based networks
10.	Easier to achieve time synchronization	Time synchronization is difficult and consumes bandwidth
11.	Easier to employ bandwidth reservation	Bandwidth reservations requires complex medium access control protocols
12.	Application domains includes mainly civilian and commercial sectors	Application domains includes battlefields, emergency search and rescue operations and collaborative computing
13.	There is no self-organizing property	Self-organization, Self reconfiguration and maintenance properties are built into the network

14.	Mobile hosts are relatively less complexity	Mobile host requires more intelligence
15.	Major goals of routing and call admissions are to maximize the call acceptance ratio and minimize the call drop	Goal of routing is to find paths with minimum overheads and also quick reconfiguration of broken paths.
16.	Widely deployed and currently in the third generation of evolution.	Several issues are to be addresses for successful commercial deployment even though widespread use exists in defense.

6. Why Ad-Hoc network is needed?

Ad-hoc networking is often justified by scenario where you cannot deploy and manage an infrastructure based network; on that area we can construct a temporary wireless network without presence of any access point or base station for exchanging the required information in form of data packets between clients (nodes).

7. What are all the challenging issues that affect the performance of Ad-hoc wireless network maintenance?

The major issues that affect the design, deployment and performance of an ad-hoc wireless network system during

- Medium access scheme
- Routing
- Multicasting
- Transport layer protocol
- Pricing scheme
- Quality of service provisioning
- Self-Organization
- Security
- Energy management
- Addressing and Service discovery
- Scalability and deployment considerations

8. List out the various applications of ad-hoc network?

Due to their quick and economically less demanding deployment it finds applications in several areas like

- Military applications like remote sensing area(battle fields)
- Collaborative and distributed computing
- Environmental applications (during different weather conditions, forest fire detection and etc.,)

- Medical applications (monitoring medical diagnosing equipment and patients)
- Educational applications (video conferencing, virtual class rooms)
- During crisis conditions (Flood, earthquake, Tsunami locations)

9. What is Ad-Hoc Wireless Internet?

It is extend the service to end user in ad-hoc network, for provisioning the temporary internet service to

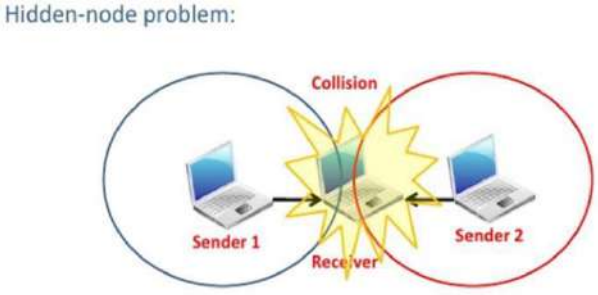
- Major conference venues
- Sports venues
- Temporary military settlements
- Battlefields
- Broadband internet service in rural regions

10. What is Gateway node? What are all the roles of Gateway node in ad-hoc wireless internet?

Gateway is one among the node in ad-hoc network which act as entry point to wired internet, and major part of internet service provisioning lies through this gateway node, generally this node is owned and operated by service provider, and its majors roles are

- Keep tracking the end user
- Bandwidth management
- Load balancing
- Traffic shaping
- Packet filtering
- Bandwidth fairness
- Address service and location discovery.

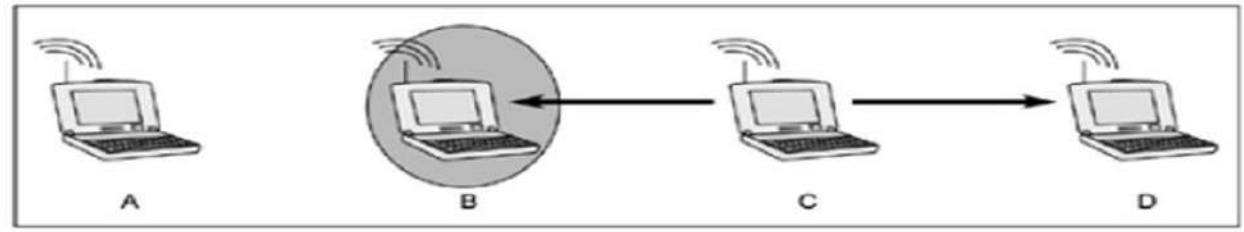
11. What is Hidden terminal problem?



- It refers to collision of packets at a receiving node due to simultaneous transmission of data packets by both senders (1 &2), because both senders are not in the radio range of each other.

- In other words hidden terminal problem is data packet transmission problem that arises when two sender nodes are out of range to each other, they transmits data packets simultaneously to a common receiver node, in that situation there is a chance for collision at receiver.

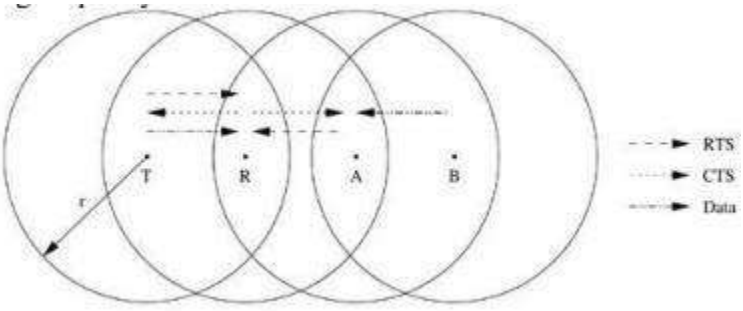
12. What is Exposed terminal Problem?



- C is transmitting to D.
- B overhears this, and is blocked.
- B wants to transmit to A, but is being blocked by C.
- Wasted bandwidth!

- It refers to inability of node for packet data transmission to its required destination node because the corresponding node is blocked by nearby transmitting node, such problems is called a exposed terminal problem.
- In other words exposed terminal problem is a transmission problem that arises when transmitting node is prevented from sending data packets due to interference with another transmitting node.

13. Give the solution for hidden terminal and exposed terminal problem?



- The solution of both problems is the transmitting node first explicitly notifies all potential hidden nodes about the forthcoming transmission by means of “two-way handshake control” called RTS (Request to send) and CTS (Clear to send). This may not solve the problem completely, but it reduces the probability of collisions.

- For reducing the probability of collisions an improved version of protocol has been proposed named as MACAW (Medium Access Collision Avoidance for Wireless).
- This MACAW protocol requires that the receiver acknowledges each successful reception of data packet. Hence successful avoidance of prevented transmissions and collisions to certain level by RTS, CTS, Data transmission and Data Acknowledgement.

14. List the four categories based that the routing protocols are classified in Ad-hoc Networks?

The routing protocols for ad hoc wireless networks can be broadly classified into four categories such as

- Routing information update mechanism
- Use of temporal information for routing
- Routing topology
- Utilization of specific resources.

15. What is proactive routing protocol or Table-driven routing protocol?

- In this protocol every node maintains the network global topology information in form of routing tables.
- Each routing tables consist information like destination node, next node (or) next hop, distance to reach destination, time-in, time-out along with sequence number.
- Routing information's is generally flooded in whole network by periodical exchanging.
- Whenever a node requires a path to destination, it runs an appropriate path-finding algorithm.
- Example-DSDV protocol.

16. What is Reactive (or) On-Demand Routing protocol?

- Protocols that fall under this category do not maintain the network topology information, also not maintaining any table information like proactive.
- They obtain the necessary path when it is required, by using a connection establishment process; hence these protocols do not exchange routing information periodically.
- Example- AODV protocol.

17. What are Hybrid Routing protocols?

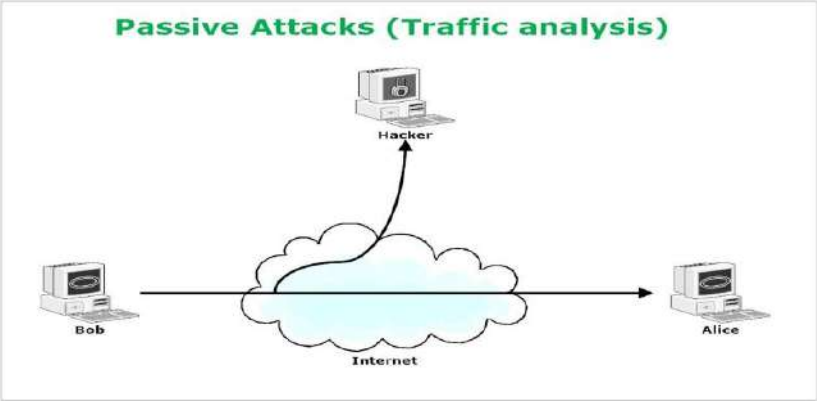
- It combines the best features of both proactive and reactive protocol, Nodes(TX/RX) with in reachable distance, or within a particular geographical

region, or within in same routing zone, a table-drive approach is used for data packet transmission.

- Nodes are in beyond the routing zone means on-demand approach is used for packet transmission

18. What are active and passive attacks during security issues in Ad-hoc wireless network?

Passive Attacks:

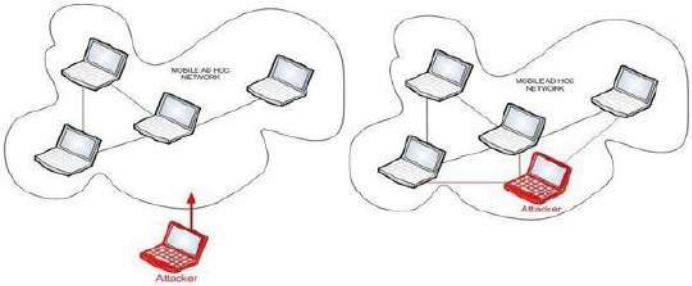


An interruption made by malicious node during data packet transmission between nodes and tries to observe or copy the message without disturbing the network operations called passive attacks

Active Attacks:

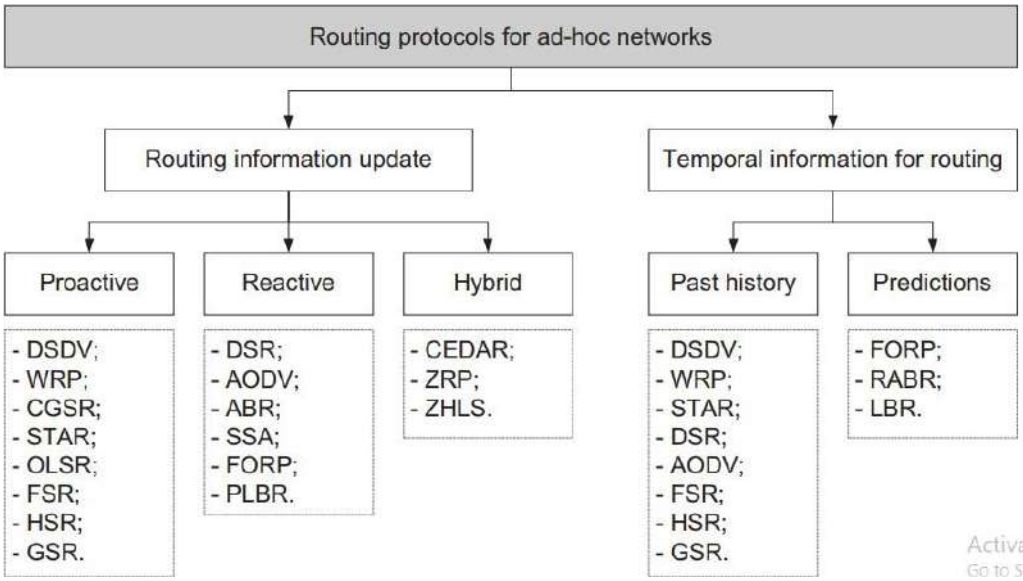
An attempt made by a node from outside the network, and tries to copy and modify the message in all disturbing the network operations is called active attacks.

19. What are internal and external attacks in ad-hoc wireless network?

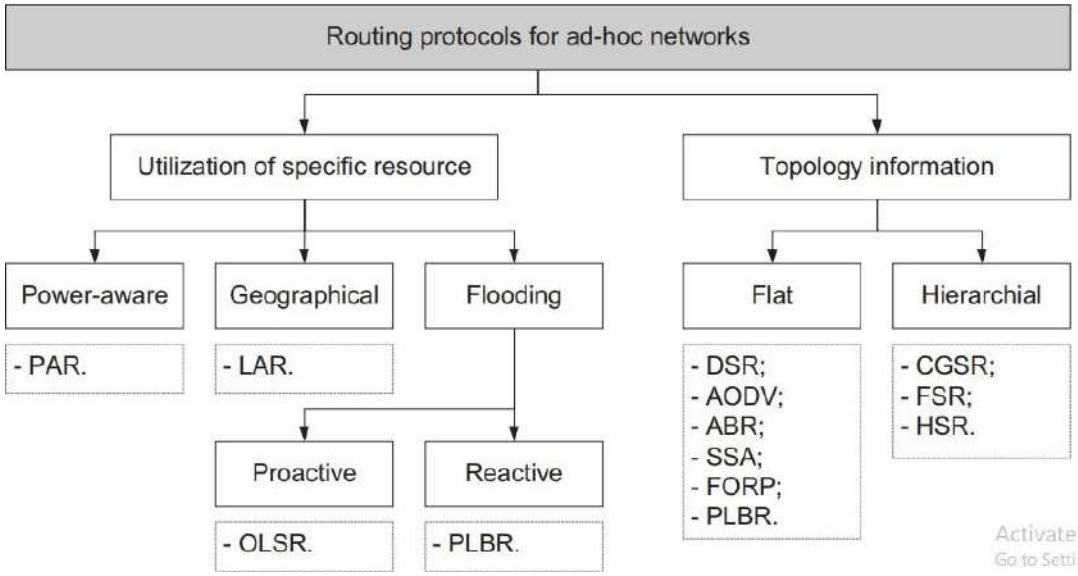


Copying or observing the data packets, modifying the users data packets during TX/RX by node with in network called *Internal Attacks*, if suppose by a node from outside the network belongs *External Attacks*

20. Write down the classification of routing protocol based on routing information update and temporal information in Ad-hoc network?



21. Write down the classification of routing protocol based on utilization of specific resource and topology information in ad-hoc network?



22. What are major activities in self-organization system in Ad-hoc wireless network?

The major activities of self-organization in ad hoc wireless network is following performances

a. Neighbor discovery
b. Topology organizations and
c. Topology reorganization.

During neighbor discovery phase every node in the network gathers information about its neighbor and maintain that information in appropriate data structures.

During topology organization phase, every node in network gathers information about the entire network or part of network in order to maintain topological information's

During topology reorganization phase, updating the topology information by incorporating the topology changes due to the following reasons

- Due to mobility of nodes
- Due to failure of nodes
- By complete depletion of power sources of the nodes.

23. What is meant self-configuring mechanism in ad-hoc network?

- This types mechanism based on typical approaches known as *route discovery approach* and *route update approach*
- In route discovery which can be "*done proactively* or *On-Demand basis*"
- In route update, single or multiple routes are maintained between a pair of nodes which helps a lot during unexpected path breaks takes place during packet transmission or reception duration.

24. What is meant by self-optimizing?

Self-Optimizing which helps to improve the routes with respect to route length (Path aware) or energy consumption (Energy aware) in ad-hoc network.

25. List the advantages of DSDV protocol?

- Less delay involved in the route setup process.
- Mechanism of incremental update with sequence number tags makes the existing wired network protocols adaptable to ad hoc wireless networks.
- The updates are propagated throughout the network in order to maintain an up-to-date view of the network topology at all nodes.

26. List the advantages of AODV protocol?

- Routes are established on demand
- Destination sequence numbers are used to find the latest route to the destination.
- The connection setup delay is less.

27. What is replay attack? How can it be prevented?

- A replay attack is a form of network attack in which a valid data transmission is maliciously or fraudulently repeated or delayed. This is carried out either by the originator or by an adversary who intercepts the data and re-transmits it, possibly as part of a masquerade attack by IP packet substitution.
- Prevention Methods:
- Passwords:
  By implementing one-time passwords for sensitive communications, we can prevent the Reply attack. One-time passwords expire either after they've been used or after a short period of time. Either way they are useful for ensuring that important transactions or communications are only taking place between the intended parties.

Digital Signatures:

- Next, you could use digital signatures. A digital signature isn't your name, but rather a complex process that involves algorithms and "keys." Each computer has its own private key, one that only those machines know, to encrypt information on one end and decrypt it on the other. Think of it like sending a coded message to a friend - the only person who knows how to break the code.

28. Is a table-driven routing protocol suitable for high-mobility environments?

No. Table driven routing protocols are not suitable for high mobility environment.

UNIT-1 -PART-B & PART-C -IMPORTANT QUESTIONS

1. Explain the important issues in Adhoc wireless networks. [13 Marks]

2. Explain the commercial applications of Adhoc wireless networks. [13 Marks]

3. Write short notes on Adhoc-wireless internet. [8 Marks]

4. Explain the issues in designing a Routing Protocol for Adhoc wireless networks. [13 Marks]

5. Explain the classifications of Routing Protocols. [8 Marks]

6. Explain Table Driven Destination Sequence Distance Vector Routing Protocols [DSDV] with necessary diagrams. [15 Marks]

7. Explain Adhoc on Demand Distance Vector [AODV] Routing Protocol with necessary Diagrams. [15 Marks]

8. List the major advantages and disadvantages of the ad hoc wireless Internet. [6 Marks]

UNIT II: SENSOR NETWORKS – INTRODUCTION & ARCHITECTURES

Challenges for Wireless Sensor Networks, Enabling Technologies for Wireless Sensor Networks, WSN application examples, Single-Node Architecture - Hardware Components, Energy Consumption of Sensor Nodes, Network Architecture - Sensor Network Scenarios, Transceiver Design Considerations, Optimization Goals and Figures of Merit.

PART A

1. What are the types of node architecture?

(i) Single node architecture means only one sensor will placed on the system architecture.

(ii) Multiple node architecture means more than one sensor will be placed on the system architecture.

2. Mention the components used in the wireless sensor nodes.

1. Controller
2. Sensor/actuators
3. Memory
4. Communication devices
5. Power supply

3. What is meant by controller?

A controller is a processor that process all the relevant data to the task and capable of executing arbitrary code.

4. What is meant by flash memory?

Flash memory is a fastest memory it is used as immediate storage of data in case RAM is insufficient or when power supply of RAM should be shut down.

5. Give the application states of transceivers

1. Transmit state
2. Receive state
3. Idle state
4. Sleep state

6. What are the types of sensors?

1. Passive, Omni directional sensors
2. Passive, narrow beam sensors
3. Active sensors

7. Give the operational states of controller.

1. Active
2. Idle
3. Sleep

8. What is dynamic voltage scaling?

Dynamic voltage scaling is a technique used to reduce the energy consumption by applying such a technique the special care has to be taken to operate the controller within its specification.

9. What are the memories used to reduce energy consumption?

1. On- chip
2. Flash memory

10. What are the advantages of event base programming over process based programming?

1. Event based programming model on the same hardware and the performance is improved by a factor of 8.
2. In event based the instruction/data memory requirements were reduced by a factor of 2 to 30
3. Power consumption was reduced by a factor of 12.

11. What are types of mobility?

1. Node mobility
2. Sink mobility
3. Event mobility

12. What is need for gateway concepts?

The sensor network has a capability to interact with itself only, so the gateway is needed to enable the sensor network over the interact and other information.

13. What is robustness?

Robustness means network should fail when limited number of nodes runs out of energy in the network. Failure of nodes will be compensated using other router.

14. Give some examples of sensor nodes.

1. The mica mote nodes
2. EYES nodes
3. BT nodes
4. Scatter web

15. What is source and types of sources?

A source is an entity in the network that can provide information, i.e., typically a sensor node.

16. Give some examples of radio transceivers?

1. RFM TR100 family
2. Mica motes
3. Chipcon.CC100 and CC2420
4. IEEE 802.15.4/ Ember EM2420 RF transceiver.

17. What is the need of gateway concepts in WSN?

- Gateway is an entry and exit point for wireless data sensing and processing networks.
- It is considered to be a static node with no energy issues, with high calculation capability.
- Provides various connectivity options between sensor nodes (wireless sensing field devices) and task manager (central monitoring station).
- Aggregates sensed data.
  Ex: 9791_WSN Ethernet gateway, 9792 gateway.
- The gateway devices are used as protocol converters, command data forwarders and it can be used as security manager and synchronizer in network

18. Why microcontroller is prepared than other controllers like microprocessor, DSP, FPGA and ASIC in WSN sensor nodes?

- Microcontroller is Suitable for all commercial and specific applications
- Low power consumption (i.e., Reduce the power consumption by going sleep states)
- Instructions sets are amenable to time critical signal processing

- Flexible to connect other devices
- It offers sufficient inbuilt memory's hence no need of external memory unit
- Easy of programming
- Economically Low cost

19. Explain the term 'Auto configuration' in WSN.

- WSN should configure most of its operation parameters automatically without any interruption of external configuration tools
- Nodes should be able to identify their own geographical location
- Nodes should be able to tolerate failure nodes
- Nodes should be able to integrate with new nodes in the network

20. Write down the categories of sensors.

Ability of the sensor device which receives and to measure the natural emission like vibrations, heat, light or other phenomena from its environment, is referred as **Passive sensors.**

A device which provides their own source of energy and observing information about targeted objects in their environment referred as ***Active Sensors*** (or) A sensor which emits its own radiations towards the directions of targeted objects to be investigated.

Sensors are roughly categorized into three categories

- **Passive Omnidirectional**
  Ex: light, thermometer, microphones, hygrometer
- **Passive Narrow Beam**
  Ex: Camera
- **Active Sensor**
  Ex: Radar, Sonar

21. Differentiate Single Hop and Multi Hop networks.

SINGLE HOP	MULTI HOP
Packet transmission - direct path to reach destination	Packet transmission - Multipath to reach destination
There is no intermediate nodes between source to destination	One or more than one nodes act as intermediate nodes
Transmission get failure if any one of node gets shutdown	Transmission may occurs even in any one of intermediate node get failure by finding alternate path
Radio coverage area is less	Radio coverage area is large
High power consumption	Low power consumption

Channel should be Line of sight for event execution	Event execution takes place even in poor channel quality

22. State the differences between ad-hoc and sensor network.

Adhoc networks	Wireless Sensor networks
The medium used in wireless adhoc network is radio waves	The medium used in wireless sensor network are radio waves, infrared, optical media.
Application independent based network	Application dependent based network
Point to point traffic pattern	Traffic pattern is any to any, one to many, many to few.
Wireless router is used as inter connecting device	Application level gateway is used as an inter-connecting devices
Have Global id	Does not have global id
Address centric	Data centric
Topology based	Not topology based
Supports common services	Supports specific applications.

23. Define figure of merit in WSN transmission control.

- Figure of merit or Noise figure of an element is defined as the ratio of the signal to noise ratio at the input of the element to the signal to noise ratio at the output of element.

$$NF = SNR_I \ / \ SNR_O$$
$$NF\ dB = SNR_I\ dB - SNR_O\ dB$$

24. List two even driven application of sensor network.

- Forest fire detection
- Precession agriculture farming applications

25. Write the goal of sensor networks.

- Reliable event detection
- Finding accurate geographic location where event is detected
- Performing specific task without delay
- Distributive and collaborative organizations
- Auto configurations during different environment conditions.

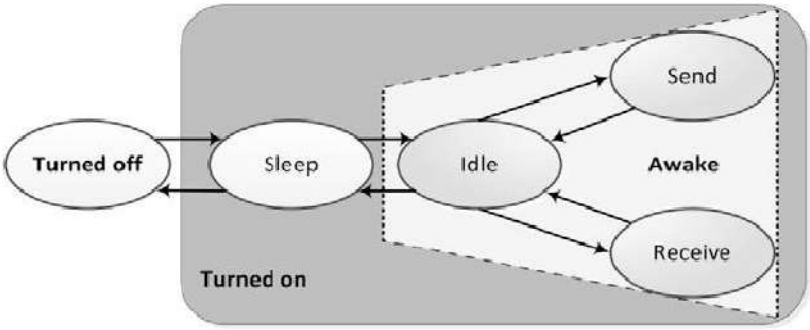
26. List out the various modes of a sensor node.

- ***Transmit mode***: Transmitting data
- ***Receive mode***: Receiving data
- ***Idle mode***: - Ready to receive, but not doing so

  - Some functions in hardware can be switched off

  - Reducing energy consumption a little

- ***Sleep mode***: -Significant parts of the transceiver nodes in network are switched off for reducing energy consumption much more



UNIT-2 -PART-B AND PART-C QUESTIONS

1. Summarize the challenges and the required mechanisms of a wireless sensor network.
2. What are the applications of wireless sensor networks and explain any two with an example each.
3. Explain how the sensor networks are deployed for Military and SAR application.
4. Sketch the RF front end of a transceiver and outline the behavior of operational states.
5. Discuss about the transceiver tasks and characteristics in a sensor node in a wireless sensor network.
6. Describe the enabling technologies and characteristic requirements of the wireless sensor networks.
7. Explain the transceiver characteristics and structure used in the sensor node.
8. Analyze how energy scavenging is realized in wireless sensor network.
9. Distinguish sensor networks from the mobile ad hoc network.
10. Write the detailed notes on energy consumption during the transmission and reception of a signal in WSN with the supporting equations.

11. Derive the expression for energy consumption in a sensor node with an appropriate diagram.
12. Draw the sensor network architecture and describe the components in detail.
13. Categorize the sensor network scenario with diagrams and also explain how mobility can appear in WSN?
14. Explain how optimization goals and figure of merits achieved in WSN with list of factors used to optimize the wireless sensor network.
15. Explain Single node architecture and Hardware components of WSN in detail.

UNIT III

WSN NETWORKING CONCEPTS AND PROTOCOLS

MAC Protocols for Wireless Sensor Networks, Low Duty Cycle Protocols and Wakeup Concepts - S-MAC, The Mediation Device Protocol, Contention based protocols - PAMAS, Schedule based protocols – LEACH, IEEE 802.15.4 MAC protocol, Routing Protocols Energy Efficient Routing, Challenges and Issues in Transport layer protocol.

PART A

1. What is MAC protocol for WSN?

The MAC layer is responsible for the establishment of a reliable and efficient communication link between WSN nodes and is responsible for energy waste. This technique enables dividing collisions from weak signals and takes appropriate decisions to reduce energy consumption.

2. What is meant by path loss and attenuation?

Wireless waveforms are propagating through free space and it is subjected to a distance dependent loss of power called path loss and different kinds of path loss called attenuation

3. Define interference.

Interference refers to the presence of any unwanted signals from external resources which mask a signal. Interference has three types

1. Multiple access interference
2. Co-channel interference
3. Adjacent channel interference

4. Write down the types of synchronization in WSN.

1. Carrier synchronization
2. Bits/Symbol synchronization
3. Frame Synchronization

5. Write the importance responsibilities of data link layer.

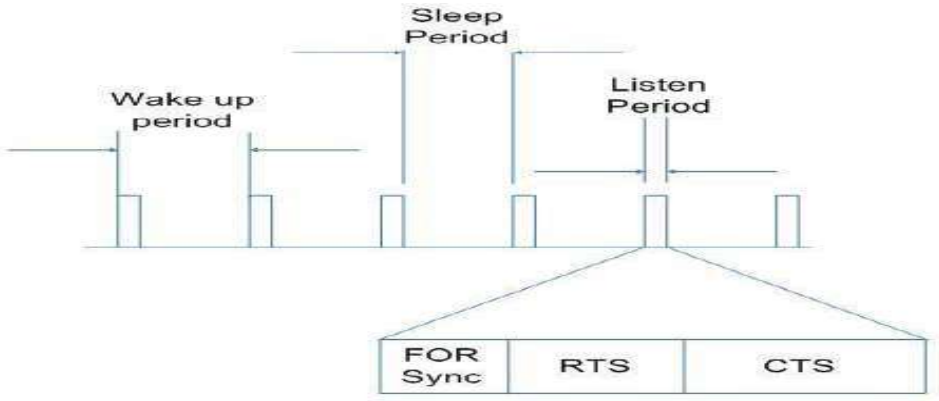
1. Error control
2. Flow control

Error control is used to ensure correctness of transmission and to take appropriate actions in case of transmission errors.

Flow control regulates the rate of transmission to protect a slow receiver from being overwhelmed with data.

6. What is S-MAC?

This protocol tries to reduce energy consumption due to overhearing, idle listening and collision. In this protocol also every node has two states, sleep state and active state. SMAC adopts a periodic wake up scheme. SMAC tries to synchronize the listen periods of neighboring nodes. The listen period of a node is divided into three phases as shown below. The listen period is the time, during which a node is awake rest of the time node is sleeping. The listen and sleep periods in the S-MAC are fixed intervals.



7. Define routing protocol

The transmission of packets from source to destination will be taken by router. Then the protocols used in the routing mechanisms are known as routing.

Types

1. Energy efficient routing
2. Geographic routing

8. Write down the names of address types used in the sensor networks

1. Unique node identifier
2. MAC address
3. Network address
4. Network identifiers
5. Resource identifiers
6. Address management tasks
7. Address allocation
8. Address representation

9. Write the usage of wakeup radio concepts

Wakeup radio concept used to avoid idle state by a simple and powerful receiver that can trigger a main receiver if necessary. Proposed wake up MAC protocol assumes the presence of several parallel data channels separated using FDMA or CDMA schemes.

10. Define frequency band

The frequency allocation is based on frequency band, for communication purposes always a finite portion of electromagnetic spectrum provide a single frequency capacity called frequency band.

11. Define antenna efficiency

An important parameter in a transmission system is the antenna efficiency which is defined as the ratio of the radiated power to the total input power to the antenna and remaining power to the antenna and remaining power is dissipated as heat.

12. Define symbol rate.

The symbol rate is the inverse of the symbol duration for binary modulation. It is also called bit rate.

13. Define demodulation

Modulation is carried out at the transmitter side. The receiver waves to recover the transmitted symbols from a received waveform. The mapping from a received waveform the symbols are called demodulation.

14. How many classes are there in the MAC protocols?

MAC protocols are having three kinds of classes

1. Fixed assignment protocols-TDMA, FDMA, CDMA and SDMA
2. Demands assignment protocols-HIPERLAN12
3. Random access protocol-ALOHA protocol

15. Define low duty cycle

Low duty cycle protocols try to avoid spending much time in the idle state to reduce the communication activities of a sensor node is a minimum level.

16. How many states in nodes?

Each node having four states

1. Transmitting state
2. Receiving state
3. Idling state
4. Sleeping state

17. What are the major problems in wireless transmission?

1. Bit error rate
2. Frequencies
3. Depending on the modulation schemes
4. Thermal noise
5. Time variable
6. Path loss

UNIT-3 -PART-B AND PART-C QUESTIONS

1. Explain MAC protocol for Wireless sensor network with neat diagrams.
2. Explain S-MAC protocol for Wireless sensor network with neat diagrams.
3. Explain Mediation Device protocol for Wireless sensor network with neat diagrams.
4. Explain Contention Based protocol for Wireless sensor network with neat diagrams.
5. Explain Schedule Based protocols for WSN.

6. Explain IEEE 802.14 MAC protocol with neat diagram.
7. Explain Energy Efficient Routing Protocols for wireless sensor networks.
8. Explain the Challenges and Issues in Transport layer protocol.

UNIT IV

SENSOR NETWORK SECURITY

Network Security Requirements, Issues and Challenges in Security Provisioning, Network Security Attacks, Layer wise attacks in wireless sensor networks, possible solutions for jamming, tampering, black hole attack, flooding attack. Key Distribution and Management, Secure Routing – SPINS, reliability requirements in sensor networks

PART-A

1. **Define Confidentiality?**

   The data sent by the sender (source node) must be comprehensible only to the intended receiver (destination node). Though an intruder might get hold of the data being sent, he/she must not be able to derive any useful information out of the data. One of the popular techniques used for ensuring confidentiality is data encryption.

2. **Define Integrity ?**

   The data sent by the source node should reach the destination node as it was sent: unaltered. In other words, it should not be possible for any malicious node in the network to tamper with the data during transmission.

3. **Define Availability?**

   The network should remain operational all the time. It must be robust enough to tolerate link failures and also be capable of surviving various attacks mounted on it. It should be able to provide the guaranteed services whenever an authorized user requires them.

4. **Define Non-repudiation?**

   Non-repudiation is a mechanism to guarantee that the sender of a message cannot later deny having sent the message and that the recipient cannot deny having received the message. Digital signatures, which function as unique identifiers for each user, much like a written signature, are used commonly for this purpose.

5. **Define Shared broadcast radio channel?**

Unlike in wired networks where a separate dedicated transmission line can be provided between a pair of end users, the radio channel used for communication in ad hoc wireless networks is broadcast in nature and is shared by all nodes in the network. Data transmitted by a node is received by all nodes within its direct transmission range. So a malicious node could easily obtain data being transmitted in the network. This problem can be minimized to a certain extent by using directional antennas.

6. **Define Insecure operational environment?**

The operating environments where ad hoc wireless networks are used may not always be secure. One important application of such networks is in battlefields. In such applications, nodes may move in and out of hostile and insecure enemy territory, where they would be highly vulnerable to security attacks.

7. **Define Lack of central authority?**

In wired networks and infrastructure-based wireless networks, it would be possible to monitor the traffic on the network through certain important central points (such as routers, base stations, and access points) and implement security mechanisms at such points. Since ad hoc wireless networks do not have any such central points, these mechanisms cannot be applied in ad hoc wireless networks.

8. **Define Lack of association ?**

Since these networks are dynamic in nature, a node can join or leave the network at any point of the time. If no proper authentication mechanism is used for associating nodes with a network, an intruder would be able to join into the network quite easily and carry out his/her attacks.

9. **Define Limited resource availability ?**

Resources such as bandwidth, battery power, and computational power (to a certain extent) are scarce in ad hoc wireless networks. Hence, it is difficult to implement complex cryptography-based security mechanisms in such networks.

10. **Define Physical vulnerability ?**

Nodes in these networks are usually compact and hand-held in nature. They could get damaged easily and are also vulnerable to theft.

11. What are the types of Network Security Attacks?

Attacks on ad hoc wireless networks can be classified into two broad categories, namely, passive and active attacks.

12. Define Passive attack?

A passive attack does not disrupt the operation of the network; the adversary snoops the data exchanged in the network without altering it. Here, the requirement of confidentiality can be violated if an adversary is also able to interpret the data gathered through snooping.

13. Define active attack?

An active attack attempts to alter or destroy the data being exchanged in the network, thereby disrupting the normal functioning of the network.

14. What are the types of active attack?

Active attacks can be classified further into two categories, namely,

External attack and internal attacks.

15. Define External attacks ?

External attacks are carried out by nodes that do not belong to the network. These attacks can be prevented by using standard security mechanisms such as encryption techniques and firewalls.

16. Define Internal attacks ?

Internal attacks are from compromised nodes that are actually part of the network. Since the adversaries are already part of the network as authorized nodes, internal attacks are more severe and difficult to detect when compared to external attacks.

17. Define firewall ?

A firewall is used to separate a local network from the outside world. It is software which works closely with a router program and filters all packets entering the network to determine whether or not to forward those packets toward their intended destinations.

18. Define Wormhole attack?

In this attack, an attacker receives packets at one location in the network and tunnels them (possibly selectively) to another location in the network, where the packets are resent into the network.

19. Define Black hole attack?

In this attack, a malicious node falsely advertises good paths (e.g., shortest path or most stable path) to the destination node during the path-finding process (in on-demand routing protocols) or in the route update messages.

20. Define Byzantine attack ?

Here, a compromised intermediate node or a set of compromised intermediate nodes works in collusion and carries out attacks such as creating routing loops, routing packets on non-optimal paths, and selectively dropping packets.

21. Define Information disclosure?

A compromised node may leak confidential or important information to unauthorized nodes in the network. Such information may include information regarding the network topology, geographic location of nodes, or optimal routes to authorized nodes in the network.

22. Define Routing attacks?

There are several types attacks mounted on the routing protocol which are aimed at disrupting the operation of the network.

23. Define Routing table overflow?

In this type of attack, an adversary node advertises routes to non-existent nodes, to the authorized nodes present in the network.

24. Define Routing table poisoning?

Here, the compromised nodes in the networks send fictitious routing updates or modify genuine route update packets sent to other uncompromised nodes.

25. Define Packet replication?

In this attack, an adversary node replicates stale packets. This consumes additional bandwidth and battery power resources available to the nodes and also causes unnecessary confusion in the routing process.

26. Define Session hijacking?

Here, an adversary takes control over a session between two nodes. Since most authentication processes are carried out only at the start of a session, once the session between two nodes gets established, the adversary node masquerades as one of the end nodes of the session and hijacks the session.

27. Define Repudiation?

In simple terms, repudiation refers to the denial or attempted denial by a node involved in a communication of having participated in all or part of the communication.

UNIT-4 -PART-B AND PART-C QUESTIONS

1. Write short notes on Network Security Requirements, Issues and Challenges in Security Provisioning.

2. Explain Network Security Attacks, Layer wise attacks in wireless sensor networks.

3. What are the possible solutions for jamming, tampering, black hole attack, flooding attack?

4. Explain Key Distribution and Management in sensor network security.

5. Explain Secure Routing in SPINS and Reliability requirements in sensor networks.

UNIT V

SENSOR NETWORK PLATFORMS AND TOOLS

Sensor Node Hardware – Berkeley Motes, Programming Challenges, Node-level software platforms – TinyOS, nesC, CONTIKIOS, Node-level Simulators – NS2 and its extension to sensor networks, COOJA, TOSSIM, Programming beyond individual nodes – State centric programming.

PART A

1. **Write sensor nodes Hardware**
   - (i) Augmented general purpose computers-Embedded PC's
   - (ii) Dedicated embedded sensor nodes-Berkley mote
   - (iii) System-on-chip nodes –PASTA
2. **Write down the types of programming for sensor networks?**

   Two types of programming method is available for sensor networks

   (i) These which carried out by end users
   (ii) Those which performed by application developers
3. **What is meant by Moto?**

   A Sensor node on a network is called as mote. This node capable of performing some processing, gathering information and communicate with other is connected nodes in network.

4. **Write the types of hardware in sensor node.**

   There are two types of hardware granted for sensor nodes

   - (i) Augmented general purpose computers
   - (ii) Dedicated embedded sensor nodes
   - (iii) System-on-chip
5. **Give some examples for augmented general purpose computers.**
   - (i) Personal digital assistants
   - (ii) Embedded PC's
   - (iii) Linux
   - (iv) Real time operating system
   - (v) Win CE

6. **Define Berkeley Motes?**

A Berkeley mote is a wireless sensor module manufactured by Berkeley. The Berkeley motes are a family of embedded sensor nodes. This node composed of sensing capabilities communication radio, computation unit and power source.

7. **What are the things to be followed in traditional programming in sensor network?**

To apply the traditional programming in sensor networks the following things to be followed

(i) Message passing
(ii) Event synchronization
(iii) Interrupt handling
(iv) Sensor reading

8. **What are the services provided by operating system?**
    1. File management
    2. Memory location
    3. Task scheduling
    4. Peripheral device drivers
    5. Networking
9. **Say some example programming for node level.**

There are two examples available for node level programming such as

1. Tiny OS

2. Tiny GALS

10. **Define tasks in TinyOS.**

Task are providing source for concurrency. Tasks are created by components to a task scheduler. The default implementation of the tinyOS scheduler maintains a task queue and task queue maintain information according the task order posted.

11. **Define nesC.**

The nesC is an imperative language it an extension of C to support and reflect the design of TinyOS V1.0 and above version. It provides a set of language constructs and restrictions to implement Tiny OS components and application.

12. Write the types of interface in nesC components interface.

There are two types of component interface in nesC

1. Provide interface
2. Uses interface

13. Write the types of component in nesC component implementation.

There are two types of component in nesC component implementation

1. Modules
2. Configurations

Modules are implemented by application code and configurations are implemented by connecting interfaces existing components.

14. How many codes in nesC?

The nesC code can be classified into two types.

1. Asynchronous code(AC)
2. Synchronous Code(SC)

15. Define dataflow style language.

Dataflow style languages are more reliable for expressing computation on interrelated data units by specifying data dependencies among those data. Tiny GALS is the example for this language.

16. What is meant by actor?

In a dataflow style language processing units called actors. Actors have ports to receive and produce data.

17. How applications built in Tiny GALS.

An application in Tiny GALS is built in two steps

Step 1: Constructing asynchronous actors from synchronous components

Step2: Constructing an application by connecting the asynchronous components through FIFO queues.

18. State advantages of Tiny GALS applications

1. It has highly structured architecture
2. Efficient scheduling
3. It has event handling code

19. Write the types of node level simulator components.

1. Sensor node model
2. Communication model
3. Physical environmental model
4. Statistics and visualization

20. Write the types of execution model in simulation.

Depending on how the time is advanced in the simulation, there are two types of execution models.

1. Cycle driven simulation
2. Discrete event simulation

21. What is meant by causal component?

Causal component means the output event is computed from an input event. The time stamps of the output event always the reference of input event at present only.

22. What is meant by non-causal component?

Non-causal component means the output event is computed from an input event. The time stamp of the output event contains past and present value of input event.

23. State some functions of processing algorithm.

1. Kalman filtering
2. Bayesian estimation
3. System identification
4. Feedback control laws
5. Finite state automate

24. How many groups available in state centric programming?

1. Collaboration groups
2. Geographically constrained group
3. N-hop neighborhood group
4. Publish/Subscribe
5. Acquaintance group

UNIT-5 -PART-B AND PART-C QUESTIONS

1. Write detailed notes on any one node-level software platform.
2. Discuss on the sensor network programming challenges.
3. Explain the system architecture of Berkeley motes with neat diagram
4. Write short notes on TinyOS and contikios?compare both.
5. Explain the concept of state centric programming in target tracking application carried out using WSN?
6. Discuss about nesC programming and write program for field monitor application?
7. Discuss about NS2 simulator for WSN.
8. Write short note on the following:
   a) TOSSIM
   b) COOJA

MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

EC8701 ANTENNA AND MICROWAVE ENGINEERING

Semester - 07

Question Bank

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

Vision

To excel in providing value based education in the field of Electronics and Communication Engineering, keeping in pace with the latest technical developments through commendable research, to raise the intellectual competence to match global standards and to make significant contributions to the society upholding the ethical standards.

Mission

- ✓ To deliver Quality Technical Education, with an equal emphasis on theoretical and practical aspects.
- ✓ To provide state of the art infrastructure for the students and faculty to upgrade their skills and knowledge.
- ✓ To create an open and conducive environment for faculty and students to carry out research and excel in their field of specialization.
- ✓ To focus especially on innovation and development of technologies that is sustainable and inclusive, and thus benefits all sections of the society.
- ✓ To establish a strong Industry Academic Collaboration for teaching and research, that could foster entrepreneurship and innovation in knowledge exchange.
- ✓ To produce quality Engineers who uphold and advance the integrity, honour and dignity of the engineering.

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

1. To provide the students with a strong foundation in the required sciences in order to pursue studies in Electronics and Communication Engineering.
2. To gain adequate knowledge to become good professional in electronic and communication engineering associated industries, higher education and research.
3. To develop attitude in lifelong learning, applying and adapting new ideas and technologies as their field evolves.
4. To prepare students to critically analyze existing literature in an area of specialization and ethically develop innovative and research oriented methodologies to solve the problems identified.
5. To inculcate in the students a professional and ethical attitude and an ability to visualize the engineering issues in a broader social context.

PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO1: Design, develop and analyze electronic systems through application of relevant electronics, mathematics and engineering principles.

PSO2: Design, develop and analyze communication systems through application of fundamentals from communication principles, signal processing, and RF System Design & Electromagnetics.

PSO3: Adapt to emerging electronics and communication technologies and develop innovative solutions for existing and newer problems.

EC8701	ANTENNAS AND MICROWAVE ENGINEERING	L	T	P	C
		3	0	0	3

OBJECTIVES:

- To enable the student to understand the basic principles in antenna and microwave system design
- To enhance the student knowledge in the area of various antenna designs.
- To enhance the student knowledge in the area of microwave components and antenna for practical applications.

UNIT I INTRODUCTION TO MICROWAVE SYSTEMS AND ANTENNAS 9

Microwave frequency bands, Physical concept of radiation, Near- and far-field regions, Fields and Power Radiated by an Antenna, Antenna Pattern Characteristics, Antenna Gain and Efficiency, Aperture Efficiency and Effective Area, Antenna Noise Temperature and G/T, Impedance matching, Friis transmission equation, Link budget and link margin, Noise Characterization of a microwave receiver.

UNIT II RADIATION MECHANISMS AND DESIGN ASPECTS 9

Radiation Mechanisms of Linear Wire and Loop antennas, Aperture antennas, Reflector antennas, Microstrip antennas and Frequency independent antennas, Design considerations and applications.

UNIT III ANTENNA ARRAYS AND APPLICATIONS 9

Two-element array, Array factor, Pattern multiplication, Uniformly spaced arrays with uniform and non-uniform excitation amplitudes, Smart antennas.

UNIT IV PASSIVE AND ACTIVE MICROWAVE DEVICES 9

Microwave Passive components: Directional Coupler, Power Divider, Magic Tee, attenuator, resonator, Principles of Microwave Semiconductor Devices: Gunn Diodes, IMPATT diodes, Schottky Barrier diodes, PIN diodes, Microwave tubes: Klystron, TWT, Magnetron.

UNIT V MICROWAVE DESIGN PRINCIPLES 9

Impedance transformation, Impedance Matching, Microwave Filter Design, RF and Microwave Amplifier Design, Microwave Power amplifier Design, Low Noise Amplifier Design, Microwave Mixer Design, Microwave Oscillator Design

TOTAL: 45 PERIODS

UNIT I -INTRODUCTION MICROWAVE SYSTEMS AND ANTENNAS
PART-A

1. **Define antenna**

   An antenna is defined as a metallic device for radiating or receiving electromagnetic waves or radio waves. It is a transitional structure between freespace wave and guided waves.

2. **Give important design parameters for antennas.**

   Important design parameters of an antenna are
   a. Desired Frequency.
   b. Gain.
   c. Bandwidth.
   d. Impedance.
   e. Polarization.

3. **What are properties of antenna?**

   The properties of antenna are,
   - Antenna has identical impedance in spite of being used as transmitter or receiver.
   - It exhibits directional characteristics and pattern.
   - It exhibits effective height in spite of being used as either transmitter or receiver.

4. **Define isotropic radiator.**

   Anisotropic radiator is a fictitious or hypothetical radiator which radiates electromagnetic energy in all directions uniformly. It is also called isotropic source or omnidirectional radiator or unipole.

5. **How does an antenna radiate?**

   When an alternating voltage is applied to an antenna, it pushes and pulls the charge backward and forward in the 'wire'. This movement of charge creates a changing electric and magnetic field which can create an electromagnetic wave capable of radiating energy from the antenna.

6. **What are Antenna parameters?**

   The antenna parameters are, Gain, Directivity, Effective aperture, Radiation Resistance, Band width, Beam width, Input Impedance. Matching – Baluns, Polarization mismatch and Antenna noise temperature.

7. **What is meant by radiation pattern?**

   Radiation pattern of an antenna is a graphical representation of the radiation properties of the antenna as a function of space coordinates.

8. **Define radian and steradian.(or) Differentiate radian and steradian.**
   - The measure of a plane angle is radian. One *radian* is defined as the plane angle with its vertex at the center of a circle of radius *r.* It is subtended by an arc whose length is *r*.
   - The measure of a solid angle is a steradian. One *steradian* is defined as the solid angle with its vertex at the center of a sphere of radius *r* .It is subtendedby a spherical surface area equal to that of a square with each side of length*r*.

9. **What is beam solid angle?**

   Beam solid angle (is the angle through which all the power is radiated to the free space. It is a three dimensional angle formed by the major lobe. It is measured by a unit called *steradian.* (-S_r ||)

10. Define gain.

The ratio of maximum radiation intensity in a given direction to the maximum radiation intensity from reference antenna produced in the same direction with same input power.

$$\text{Gain (G)} = \frac{\text{Maximum radiation intensity from test antenna}}{\text{Maximum radiation intensity from the reference antenna with same input power}}$$

11. Define absolute gain.

Absolute gain of an antenna is defined as the ration of the intensity in a given direction to the radiation intensity that could be obtained if the power accepted by the antenna were radiated isotropically.

12. What is the Significance of gain of an antenna?

a. Gain of an antenna is a relative measure of antenna's ability to direct radio frequency energy in a particular direction.
b. Higher the gain and efficiency is more.
c. By means of gain, the total power radiated by an antenna can be obtained which is otherwise difficult to find.

13. Write about power gain and directive gain.

a. Directive gain and power gain of an antenna represent the ability of theantenna to focus its beam in a particular direction.
b. Directive gain is a parameter dependent only on the shape of radiation patternwhile power gain takes ohmic and other losses.
c. The power gain of an Antenna is an actual or realized quantity which is lessthan directive gain due to ohmic losses in the antenna.

14. Define Directivity.

- Directivity is a measure of the concentration of radiated power in a particular direction.
- The ratio of the maximum power density to the average power radiated is called maximum directive gain (or) directivity of an antenna.

15. Define effective aperture (A_e).

Area over which the power is extracted from the incident wave and delivered to the load is called effective aperture.

16. What is the significance of aperture of an antenna?

- The Gain of an antenna is directly proportional to its aperture.
- Larger the aperture, higher the gain and narrower the Beam width

17. Define Radiation Resistance.

Radiation Resistance is defined as the fictitious resistance which when inserted in series with the antenna will consume the same amount of power as it is actually radiated. The antenna appears to the transmission line as a resistive component and this is known as the radiation resistance.

18. What is the significance of Radiation resistance of an antenna?

a. Radiation resistance accounts for the power radiated by the antenna into space.
b. It is equal to the radiated power in watts divided by the square of theeffective current in amperes at the point of power supply ($R_r = P / I^2$).
c. Thus the radiation resistance of an antenna is a **good indicator** of the strength of the electromagnetic field radiated by a transmitting antenna (received by a receiving antenna,) since its value is directly proportional to the power radiated.

19. Define Half power Beam width.

In pattern maximum, the angle between the two directions in which the radiation intensity is one half of the maximum value is called half Power Beam width.

20. Define Beam efficiency.

The ratio of the main beam area to the total beam area is called beam efficiency.

Beam efficiency $S_M = W_M / W_A$.

The total beam area (W_A) consists of the main beam area (W_M) plus the minor lobe area (W_m) .

Thus $W_A = W_M + W_m$.

21. Define Antenna Temperature.(or)Define Brightness temperature of an antenna.

Antenna temperature is defined as the temperature of far field region (space and near surroundings) which are coupled to the antenna through radiation resistance.

22. Define induction field. (or)What do you mean by induction field?

The induction field is the field which predominates at distance close to the current element, r<<wavelength. It represents the energy stored in the magnetic field surrounding the current element or conductor. This field is also known as nearfield.

23. Define Radiation field. (Or) What do you mean by radiation field

The radiation field will be produced at a larger distance from the current element, r>> wavelength. It is also called as distant field or far field.

24. What is the field zone?

The field around an antenna is called field zone. It may be divided into two principal regions.

- Near field (Induction field) zone called as Fresnel zone
- Far field (Radiation field) zone called as Fraunhofer zone

25. Define antenna efficiency.

The efficiency of an antenna is defined as the ratio of power radiated tothe total input power supplied to the antenna.

Antenna efficiency = Power radiated / Total input power.

PART-B

1. Explain the radiation concept of an antenna with a diagram.
2. What are antenna field zones? Explain the types with the necessary equation.
3. Discuss the fields and power radiated by an antenna.
4. Explain the different types of antenna parameters with their characteristics.
5. Explain about Antenna gain and efficiency.
6. Explain the terms (i) Aperture efficiency (ii) Effective area.
7. Explain the concept of antenna noise temperature and derive the expression for G/T.
8. Derive the expression for the Friis transmission formula.
9. Explain the concept of link budget and link margin with the necessary equation.
10. Explain noise characterization of a microwave receiver and derivethe expression for S/N ratio.
11. Discuss the concept of impedance matching.

UNIT-II RADIATION MECHANISMS AND DESIGN ASPECTS

PART-A

1. Define Hertzian dipole or oscillating dipole.

A Hertzian dipole is a short linear antenna which when radiating is assumed to carry constant current along its length. It is also defined as an infinitesimal current element Idl which does not exist in real life.

2. What is a short dipole?

A short dipole is the one in which the field is oscillating because of the oscillating voltage and current. It is named so because the length of the dipole is short and the current is almost constant throughout the entire length of the dipole.

3. How radiations are created from a short dipole.

The dipole has two equal charges of opposite sign oscillating up and down in harmonic motion. The charges will move towards each other and electric field lines were created. When the charges meet the midpoint, the field lines cut each other, and new fields are created. This process is spontaneous, and hence more fields are created around the antenna.

4. Why a short dipole is also called an elemental dipole?

A short dipole that does have a uniform current will be called as an elemental dipole. This dipole will be shorter than one-tenth of a wavelength.

5. What is infinitesimal dipole?

When the length of the short dipole is vanishingly small, then that dipole is called an infinitesimal dipole. If dl be the infinitesimal small length and I will be current, then Idl is called the current element.

6. Define half-wave dipole.

Half wavelength dipole or simply half-wave dipole or $\lambda/2$ antenna is one of the simplest antennas and is frequently employed as an element of a more complex directional system. Ex: Antenna arrays. It is used above 2 MHz.

7. What are the salient features of a folded dipole antenna?

- It is a single antenna consisting of 2 or 3 elements.
- The input impedance of a folded dipole is 4 times that of a straight dipole.

8. List the advantages and applications of a folded dipole antenna.

Advantages:

- It has a high impedance.
- It has greater bandwidth
- It has wideband in frequency
- Construction is simple and cheap.

Applications:

- Used in wideband operation such as television
- Used as a feed element in Yagi uda antennas

9. What is a loop antenna?

A loop antenna is a radiating coil of any convenient cross-section of one or more turns carrying radiofrequency current. It may assume any shape (e.g. rectangular, square, triangular, and hexagonal).

10. What are electrically small loop antennas?

Electrically small loop antennas are one in which the overall length of the loop is less than one-tenth of the wavelength. Electrically small loop antennas hare small radiation resistances that are usually smaller than the loop resistances They are very poor radiators and seldom employed for transmission in radio communication.

11. List the uses of a loop antenna.

Various uses of the loop antenna are:

- It is used as a receiving antenna in portable radio and pagers.
- It is used as probes for field measurements and as directional antennas for radio wave navigation.
- It is used to estimate the direction of radio wave propagation.

12. What is the difference between the slot antenna and its complementary dipole antenna?

- Polarization is different. i.e., The electric fields associated with the slot antenna is identical with the magnetic field of the complementary dipole antenna.
- The electric field will be vertically polarized for the slot and horizontally polarized for the dipole.
- The radiation from the backside of the conducting plane of the slot antenna has the opposite polarity from that of the complementary antenna.

13. What is a horn antenna?

It is a flared-out waveguide. It is a transition (or) matching section from the guided mode inside the waveguide to the unguided (free space) mode outside the waveguide.

14. What are the different types of horn antennas?

1. Sectoral horn antenna
2. Pyramidal horn antenna
3. Conical horn antenna
4. Biconical horn antenna

15. What are the various feeds used in reflector antennas?

1. Dipole antenna
2. Horn antenna
3. End fire feed
4. Cassegrain feed

16. What is the reflector type of antenna?

The antenna which is used to eliminate the backward radiations from an antenna and to modify the radiation pattern in the desired manner to the desired direction is called a reflector type of antenna.

17. What are the most widely used types of reflectors?

(i) Plane sheet reflector
(ii) Comer reflector
(iii) Parabolic reflector
(iv) Hyperbolic reflector
(v) Elliptical reflector
(vi) Circular reflector

18. What is a parabolic reflector?

It is a parabola shaped reflective device used to collect or distribute energy entering the reflector at an angle.

19. What is a frequency-independent antenna?

If the structure of the antenna is defined in terms of angles only, then it comes under the category of a frequency-independent antenna.

e.g., Log periodic antenna, spiral antenna.

20. Define pitch angle concerning the helical antenna. What happens when $\alpha=0^0$ and $\alpha=90^0$

It is the angle between a line tangent to the helix wire and the plane normal to the helix axis.

Pitch angle, $= \tan^{-1}\left(\frac{S}{\pi D}\right)$

where, S- helix turn spacing D-diameter of helix If

$\alpha=0^0$, then helix becomes a loop

If $\alpha=90^0$, then helix becomes a linear conductor

21. What are the limitations of the normal mode operation of a helical antenna?

Bandwidth is very narrow.

The efficiency of radiation (η) is low.

22. Define the log periodic antenna.

It is a broadband, multi-element narrow beam, a frequency-independent antenna that has impedance and radiation characteristics that are regularly repetitive as a logarithmic function of frequency.

23. List out the applications of a helical antenna.

A helical antenna is used in

- VHF transmission such as satellite communication.
- Space telemetry link with ballistic missiles, satellites, etc.
-

24. Define Rumsey's principle for frequency-independent antennas.

Rumsey's principle: It states that the impedance and pattern properties of an antenna will be frequency independent if the antenna shape is specified only in terms of angles.

25. State Huygens's principle.

Huygens's principle states that -each point on a primary wave front can be considered to be a new source of a secondary spherical wave and that a secondary wave front can be constructed as the envelope of these secondary spherical waves‖.

PART-B

1. Derive the radiation resistance of an oscillating electric dipole.
2. Derive magnetic field components of dipole having dimension $l<<\lambda/2$.
3. Deduce the field quantities and draw the radiation pattern of a half-wavedipole.
4. Obtain the expression for power radiated by half-wave dipole and find its radiation resistance.
5. What is a loop antenna? Explain in detail.
6. What is a reflector antenna? Explain the principle of operation and application ofparabolic reflector and various types of feed used?
7. Compare flat reflector and corner reflector antennas. Explain how a paraboloidal antenna gives a highly directional pattern.
8. Explain in detail about different types of horn antenna with relevant diagrams and equations.
9. Discuss the various feed techniques for Rectangular patch antenna with neat diagrams.
10. Explain the radiation mechanism of the slot antenna with a diagram. Explain different feed methods of slot antenna?
11. Explain in detail about Microstrip patch antennas.
12. Explain the construction and characteristics features of frequency independent antennas.
13. With a neat diagram explain the helical antenna and briefly describe its operation inaxial mode. How does it differ from other antennas?
14. With necessary illustrations explain the radiation characteristics of the multi- element log periodic antenna and mention its possible applications.
15. Discuss in detail about Spiral antenna.

UNIT-III ANTENNA ARRAYS AND APPLICATIONS

PART-A

1. What is an antenna array?

An antenna array is a radiating system of similar antennas spaced properly to get greater directivity in a desired direction.

2. What is the need of antenna array?

In the point to point communication, it is desired to have most of theenergy radiated in one particular direction. This means it is desired to have greater directivity in a desired direction particularly which is not possible with single dipole antenna. Hence to increase field strength in the desired direction antenna array is used which consists of a group of similar antennas properly spaced and orientedin desired direction.

3. List the uses of antenna arrays?

The uses of antenna arrays are

- To achieve high gain in one particular direction.
- To provide diversity reception.
- To cancel interference from a particular set of direction.
- To steer the array so that it is most sensitive in one particular direction.
- To maximize the signal to Interference plus noise ratio (SINR).

4. What are the advantages of antenna arrays?

The advantages of antenna arrays are

- ✓ Greater directivity in a desired direction.
- ✓ Diversity reception.
- ✓ Interference is cancelled from a particular set of direction.

Maximum signal to Interference plus noise ratio(SINR)

5. List the types of Arrays by positioning.

The types of arrays are,

1. Broadside Array
2. End-fire Array:
3. Phased Array:
4. Parasitic Array:

6. What is broadside Array?

An array with equally spaced elements which are fed with a current of equal amplitude and phase is known as broadside array. In this array, maximum radiation occurs at right angles to the axis of antenna array.

7. What is end-fire Array?

An array with equally spaced elements which are fed with a current of equal amplitude and opposite phase is called as end-fire array. In this array, the maximum radiation occurs along the axis of antenna array.

8. What is phased Array?

An array of many elements with variable phase elements providing control of beam direction and pattern shape including side lobes is called phased array.

9. What is linear array?

The antenna array is said to be linear if the elements of the antenna arrays are equally spaced along a straight line.

10. What is meant by uniform linear array?

The linear antenna array is said to be uniform linear array if all the elements are fed with current of equal magnitude with progressive uniform phase-shift along the line.

11. What are the conditions to obtain end fire array pattern?

End fire array is defined as an arrangement in which the principle direction of radiation coincides with the array axis.

For end fire array, $\alpha = -\beta d$

Where, α = Phase difference of the current fed between the sources of the end fire array.

d = Distance between the elements

12. Calculate the directivity of a given linear end fire array of 10 elements with a separation of λ/4 between the elements.

Given data:

No. of isotropic radiators, n = 10.

Distance between 2 elements (Antenna) = λ/4

Solution: Directivity of end fire array $4(nd/\lambda)$

$$= \frac{4(10x\,\lambda/4)}{\lambda} = 10\,degrees$$

13. Define beam width of major lobe.

Beam width of major lobe is defined as

1. The angle between first nulls (or)
2. Double the angle between first null and major lobe maxima directions.

14. What is pattern multiplication?

The total field pattern of an array of non-isotopic but similar sources is the multiplication of the individual source patterns and the pattern of array of isotopic point sources each located at the phase centre of individual source and has the relative amplitude and phase, whereas the total phase pattern is the addition of the phase pattern of the individual sources and that of the array of isotopic point sources.

15. Mention the features of radiation pattern multiplication principle.

The features of radiation pattern multiplication principle are

1. Useful tool in designing antenna.

It approximates the pattern of a complicated array without making lengthy computations

16. State the disadvantage of pattern multiplication.

The disadvantages of pattern multiplication are,

1. It is the technique which is useful only for arrays containing identical elements.
2. It is not useful for very larger arrays.

17. Write the advantages of pattern multiplication.

The advantages of pattern multiplication are,

- Pattern multiplication provides a speedy method for sketching the radiation patterns of complicated arrays just by inspection. .
- It is a useful tool in design of antenna arrays.
- This method provides the exact pattern of the resultant.
- The secondary lobes are determined from the number of nulls in the resultant pattern.

18. Distinguish between active and passive arrays.

- In active array, all the elements are driven by a physical feed.Example:phased array
- In passive array, one element (driven element) is fed and other elementsare coupled to it electromagnetically.
  Example: parasitic array

19. State the features of Binomial array.

Features of Binomial array

- In Binomial array, radiating sources at the centre radiates more strongly than the sources at the edges. Minor lobes can be eliminated (But at the cost of directivity).
- HPBW is more than that of uniform array for the same array length.

20. List the Advantages and disadvantages of Binomial array.

Advantage of Binomial array:

There are no side lobes in the resultant pattern. Disadvantages of Binomial array**:**

1. Small directivity
2. Undesirable large beam width of main lobe.
3. For the design of a large array, larger amplitude ratio of sources is required.

21. What is Hansen – Wood yard array? (or)

What is the condition on phase for the end fire array with increased directivity?

Hansen – Wood yard array is an end fire array with increased directivity. It has been shown by Hansen and Wood yard that a maximum directivity is achieved by increasing the phase change of the current between the sources so that $\beta = -(\beta d + \pi)$ where β = phase difference of the current in adjacent point sources.

22. What is tapering of arrays? Why we go for non-uniform amplitude distribution of current?

In uniform linear array as the array length is increased to increase the directivity, the secondary lobes also occurs. It is tapering of arrays. To reduce the side lobe level, we go for non-uniform distribution of current.

Example: Binomial array. Tapering is done from center to end.

Example: Binomial Array: Tapering follows the coefficient of binomial series.

23. What are the disadvantages of Binomial array?

The disadvantages of binomial array are

- Increased beam width, hence the directivity decreases.
- Maintaining the large ratio of current amplitude in large arrays is difficult.

24. What is role of array element in smart antenna?

The antenna elements play an important role in shaping and scanning the radiation pattern and constraining the adaptive algorithm used by the digital signal processor. Smart antenna uses an array of printed elements. There are a number of printed element geometries, like patches or microstrips.

25. What are the parameters that determine the overall pattern of an antenna array?

Design parameters of Arrays are

- ✓ General array shape (linear, circular, planar)Element spacing.
- ✓ Element excitation amplitude.
- ✓ Element excitation phase.
- ✓ Patterns of array elements

PART-B

1) Derive the expression for an array factor of n - element linear array.
2) Explain in detail about BSA and EFA.
3) Compare BSA and EFA.
4) Derive the expression for pattern maxima, minima, and half-power beam width for broadside array.
5) Derive the expression for pattern maxima, minima, and half-power beam width for an end-fire array.
6) Explain in detail the principle of pattern multiplication.
7) Explain in detail the concept ofphased arrays. Describe the working principle of an adaptive array.
9) Explain in detail about binomial array.
10) In a linear array of 4 isotropic elements spaced $\lambda/2$ apart and with equal currents fed in phase, plot the radiation pattern.
11) In a linear array of 4 isotropic elements spaced $\lambda/2$ apart and with equal currents fed out of phase, plot the radiation pattern.
12) Explain in detail the concept of smart antennas with an example.

UNIT IV – PASSIVE AND ACTIVE MICROWAVE DEVICES

PART-A

1. Define Microwave.

Microwaves are electro waves whose frequencies range from $1GH_z$ to 300GHz. Signals at these frequencies have wavelength that range from 30cm to 1mm.

2. What are the various bands available in Microwave? Give their frequency range. (or) Give IEEE microwave frequency bands.

Various bands available in microwave are

Bands	Frequency
L	1 - 2 GH_Z
S	2 – 4 GH_Z
C	4 – 8 GH_Z
X	8– 12 GH_Z
KU	12–18GH_Z
K	18–26GH_Z

3. List the applications of RF and Microwave.

The applications of RF and microwave are Communication, Radar , Radio Astronomy, Navigation, Heating and Power Applications and Spectroscopy.

4. Define S-matrix.

S matrix is a square matrix which relates all possible combinations of input ports andoutput ports in a microwave junction.

5. What are power dividers?

Power dividers are used to divide the input power into a number of smaller amounts of power for exciting the radiating elements in an array antenna.

6. What is E-plane tee junction?

Waveguide in which the axis of its side arm is parallel to the E-field of the main arm. Wave fed to side arm (port 3) produce waves of opposite phase and same magnitude in its collinear arms.

$S_{13} = - S_{23}$

7. What is H- plane tee junction?

Waveguide in which the axis of the side arm is shunting the E-field or parallel to H-field of main guide. Input at port3, produces in phase and same magnitude waves at port1 and port2.

Inputs at port1 and port2, produces output at port3 which is in phase and additive.

$S_{13} = S_{23}$

8. Mention the some characteristics of reflex klystrons.

- Frequency range: 1to 25GHz
- Power output: It is a low-power generator of 10 to 500mW
- Efficiency: About 20 to 30%

9. What is Faraday's rotation law?

If a circular polarized wave is made to pass through a ferrite rod which has been influenced by an axial magnetic field B ,then the axis of polarization gets tilted in clockwise direction and amount of tilt depends upon the strength of magnetic field and geometry of the ferrite.

10. What is hybrid ring?

Hybrid ring consists of an annular line of proper electrical length to sustain standing waves, to which four arms are connected at proper intervals by means of series or parallel junctions.

11. Give some coupling parameters of directional coupler?

Some of the coupling parameters of directional coupler are Coupling coefficient, Directivity, Insertion loss, Isolation.

12. Define coupling factor.

The characteristics of a directional coupler can be expressed in terms of its coupling factor and its directivity.

13. Define directivity.

The directivity is a measure of how well the forward traveling wave in the primary waveguide couples only to a specific port of the secondary waveguide.

14. Write the applications of Magic tee.

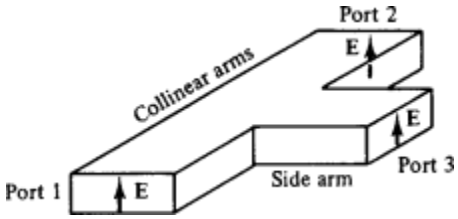
The applications of magic tee are mixing, duplexing, and impedance measurements.

A particular application requires twice more input power to an antenna than either transmitter can deliver. A magic tee may be used to couple the two transmitters to the antenna in such a way that the transmitters do not load each other.

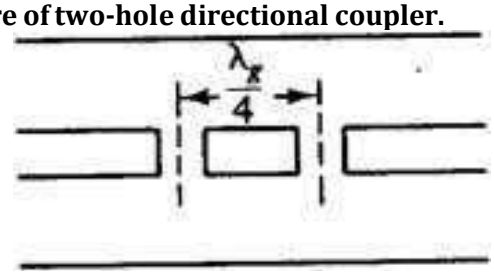
15. What is meant by hybrid rings (or) rat-race circuits? Give the significance of rat race junctions.

A hybrid ring consists of an annular line of proper electrical length to sustain standing waves, to which four arms are connected at proper intervals by means of series or parallel junctions. When a wave is fed into port 1, it will not appear at port 3 because the difference of phase shifts for the waves traveling in the clockwise and counterclockwise directions is 180°.

16. Draw the diagram of H-plane Tee junctions.



17. Draw the structure of two-hole directional coupler.



18. Find the resonant frequency of TE101 mode of an air filled rectangular cavity resonator with dimensions 5cmx4cmx2.5cm.

Sol: Given: a=5cm = $5x10^{-2}$m, b=4cm=$4x10^{-2}$m,
d=2.5cm=$2.5x10^{-2}$m TE101 mode, m=1, n=0, p=1

$= (3x10^{8})x44.72 = 1.34x10^{10}$ m/sec

$$f_0 = \frac{v}{2}\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2 + (d)^2}$$

19. Obtain the coupling coefficient of directional coupler if the magnitude of its scattering parameter is 0.707.

Sol: Given: S_{41}= 0.707, To find Coupling factor = - 10 log S_{41}

= 1.50

20. Give some examples of reciprocal devices.

Circulator and directional coupler are examples of reciprocal devices.

21. What is transferred electron effect?

Some materials like GaAs exhibit a negative differential mobility when biased above a threshold value of the electric field. The electrons in the lower – energy band will be transferred into the higher energy band. The behavior is called transferred electron effect and the device is called transferred electron device or Gunn diode.

22. What is negative resistance in Gunn diode?

The carrier drift velocity is linearly increased from zero to a maximum when the electric field is varied from zero to a threshold value. When the electric field is beyond the threshold value of 3000V/cm, the drift velocity is decreased and the diode exhibits negative resistance.

23. What are the various modes of operation of Gunn diode?

i. Gunn oscillation mode.
ii. Stable amplification mode.
iii. LSA oscillation mode.

24. What are the elements that exhibit Gunn Effect?

The elements are

1. Gallium arsenide
2. Indium phosephide
3. Cadmium telluride
4. Indium arsenide

25. Define GUNN EFFECT.

Gunn effect was first observed by GUNN in n-type GaAs bulk diode. According to GUNN, above some critical voltage corresponding to an electric field of 2000-4000v/cm, the current in every specimen became a fluctuating function of time. The frequency of oscillation was determined mainly by the specimen and not by the external circuit.

26. Mention the disadvantage of IMPATT diodes.

The major disadvantages of the IMPATT diodes are

(1) Dc power is drawn due to induced electron current in the external circuit, IMPATT diodes have low efficiency.

(2) Tend to be noisy due to the avalanche process and to the high level of operating current.

(3) A typical noise figure is 30dB which is worse than that of Gunn diodes.

27. What are the factors reducing efficiency of IMPATT diode?

The factors reducing efficiency of IMPATT diode are

1) Space charge effect
2) Reverse saturation current effect
3) High frequency skin effect
4) Ionization saturation effect.

28. List the types of microwave tubes.

- O–type microwave tube or linear beam
- M –type microwave tube

29. What are the applications of reflex klystron?

Reflex klystron is widely used in the laboratory for microwave measurements and in microwave receivers as local oscillators in commercial, military, and airborne Doppler radars as well as missiles.

30. What are hybrid couplers?

Hybrid couplers are inter-digitated microstrip couplers consisting of four parallel Strip lines with alternate lines tied together, It has four ports. This type of coupler is called Lange hybrid coupler.

31. Write the types of directional couplers.

Several types of directional couplers exist, such as a two—hole directional coupler, four-hole directional coupler, reverse-coupling directional coupler (Schwinger coupler), and Bethe-hole directional.

32. A directional coupler is having coupling factor of 20dB and directivity of 40dB. If the incident power is 100mW, what is the coupled power?

Sol: Given:

coupling factor = 20 dB

Directivity= 40 dB

Incident power =100mW

To find:

Coupling factor = 10 log P1/P4

Therefore coupled power P4 = $100 \times 10^{-3} / 10^{2} = 1$mW.

PART-B

1. What do you understand by microwave components? List variouscomponents using microwave frequency.
2. Explain the operation of Directional coupler with neat diagram. Discuss the various types of directional coupler.
3. Explain the working of flap and vane type attenuator.
4. What is a power divider? Explain its uses in microwave engineering.
5. What is a Magic Tee? Derive the scattering matrix of it and list theadvantages of Magic Tee.
6. Write a note on Microwave resonators.
7. What are Avalanche transit time device? Explain the operation, construction andapplications of IMPATT diode.
8. Briefly explain the working principle of PIN diode.
9. Explain in detail about Schottky barrier diode.
10. What are the limitations of conventional tubes at microwave frequencies? Explain how these limitations can be overcome.
11. What are the performance characteristics of a klystron amplifier?
12. By means of an Applegate diagram explain the operation of a reflex klystron.
13. What are cross field devices? How does a magnetron sustain its oscillations using this cross field?
14. How is bunching achieved in a cavity magnetron? Explain the phasefocusing effect.
15. Derive an expression for the cut off magnetic flux density with reference to a cylindrical cavity magnetron.
16. What are slow wave structures? Explain how a helical TWT achieves amplification.
17. Differentiate Klystrons and TWT.
18. Explain the terms frequency pulling and frequency pushing with reference to a magnetron.

UNIT-V MICROWAVE DESIGN PRINCIPLES

PART-A

1. What is the impedance transformation?

The ability to change impedance by adding a length of a transmission line is known as impedance transformation. When operated at a frequency corresponding to a standing wave of 1/4-wavelength along the transmission line, the line's characteristic impedance necessary for impedance transformation must be equal to the square root of the product of the source impedance and the load's impedance.

2. What is the lossless and lossy line?

If the attenuation coefficient =0 the line is called a lossless line. A lossless line is defined as a transmission line that has no line resistance and no dielectric loss. This would imply that the conductors act like perfect conductors and the dielectric acts as a perfect dielectric. (Here the attenuation coefficient =0). A lossy transmission line includes a termˆ to represent the resistance of the signal flowing. down the line and a conductanceˆ to represent the possibility of a leakage current between the conductors through the insulator. (Here the attenuation coefficient ≠0)

3. What is impedance matching?

Impedance matching is the practice of designing the input impedance of an electrical load or the output impedance of its corresponding signal source to maximize the power transfer or minimize signal reflection from the load. Here ($Z_L=Z_O$)

4. Why impedance matching or tuning is important?

Impedance matching or tuning is important for the following reasons:

1. Maximum power is delivered when the load is matched to the line (assuming the generator is matched), and power loss in the feed line is minimized.
2. Impedance matching sensitive receiver components (antenna, low-noise amplifier, etc.) may improve the signal-to-noise ratio of the system.
3. Impedance matching in a power distribution network (such as an antenna array feed network) may reduce amplitude and phase errors.

5. What are the different methods of impedance matching?

The various methods are using:

1. L section matching network
2. Single stub matching
3. Double stub matching
4. Quarter wave transformer

6. What is double stub matching?

A double-stub matching network matches a complex load impedance (Z_{load}) to a desired complex input impedance (Z_{in}) using two shunt stubs and a connecting line.

7. What are the applications of the Smith chart?

1. It is used to calculate impedance and admittance on any load.
2. It is used to find V_{max}, V_{min}, SWR, and reflection coefficient K.
3. It is used to find the length and position of the stub.

8. What is a microwave filter?

Microwave filters are two-port, reciprocal, passive, linear devices that heavily attenuate the unwanted signal frequencies while permitting transmission of wanted frequencies.

9. List the important filter parameters.

In designing a filter, the following important parameters are generally considered.

1. Pass-band width
2. Stop-band attenuation and frequencies
3. Input and output impedances
4. Return loss
5. Insertion loss
6. Group delay.

10. What is insertion loss?

The insertion loss is defined as the ratio of incident power to the load power.

It is given by $IL(dB) = 10 \log \frac{\text{incident power}(P_i)}{\text{Load power } (P_L)}$

11. Define return loss.

It is defined as the ratio of incident power to the reflected power, which tells about the amount of impedance matching at the input port.

12. Define the group delay.

The group delay is important for the multi-frequency or pulsed signals to determine the frequency dispersion or deviation from constant group delay over a given frequency band.

13. What are the commonly used filters?

The commonly used filters are

1. Low pass filter
2. High pass filter
3. Bandpass filter
4. Bandstop filter

14. What are the types of filter design?

Two methods are normally used. They are

1. Image parameter method
2. Insertion loss method

15. What are the steps involved in the insertion loss method?

Insertion loss method consists of the following steps:

1. Design of a prototype low-pass filter with the desired passband characteristics.
2. Transformation of this prototype network to the required type (low-pass, high- pass, band-pass, or band-stop) filter with the specified center and band-edge frequencies.
3. Realization of the network in microwave form by using sections of microwave transmission lines whose reactance are corresponded to those of distributed circuit elements.

16. What is a microwave power amplifier?

Power amplifiers are used in the final stages of radar and radio transmitters to increase the radiated power level. Typical output powers may be on the order of 100–500 mW for mobile voice or data communications systems, or in the range of 1– 100 W for radar or fixed-point radio systems. Important considerations for RF and microwavepower amplifiers are efficiency, gain, inter modulation distortion, and thermal effects.

17. What are the two mixer characteristics?

1. Frequency up conversion
2. Frequency down-conversion.

18. List the applications of microwave mixers.

Mixer circuits can be used to shift the frequency of an input signal like as in a receiver. They can also be used as a product detector, modulator, frequency multiplier, or phase detector.

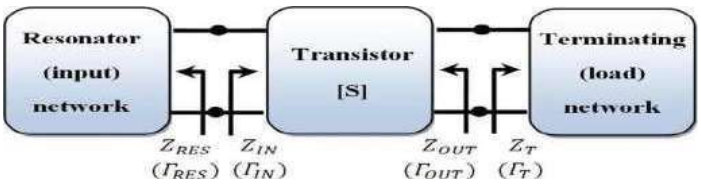
19. List the applications of an oscillator.

They provide a critical clocking function for high-speed digital systems, generate electromagnetic energy for radiation, enable frequency up and down conversion when used as local oscillators, and are used as a reference source for system synchronization.

20. What is a microwave mixer?

A mixer is a three-port device that uses a nonlinear or time-varying element to achieve frequency conversion an ideal mixer produces an output consisting of the sum and difference frequencies of its two input signals. The operation of practical RF and microwave mixers is usually based on the nonlinearity provided by either a diode or a transistor.

21. Draw the model of a transistor oscillator.



22. What is a microwave oscillator?

RF and microwave oscillators are found in all modern wireless communications, radar, and remote sensing systems to provide signal sources for frequency conversion and carrier generation. A solid-state oscillator uses an active nonlinear device, such as a diode or transistor, in conjunction with a passive circuit to convert DC to a sinusoidal steady-state RF signal. Basic transistor oscillator circuits can generally be used at low frequencies, often with crystal resonators to provide improved frequency stability and low noise performance.

23. List the important consideration of oscillators in the microwave system.

Important considerations for oscillators used in RF and microwave systems include the following:

1. Tuning range (specified in MHz/V for voltage-tuned oscillators)
2. Frequency stability (specified in PPM/◦ C)

AM and FM noise (specified in dBc/Hz below the carrier, offset from the carrier)
Harmonics(specified indBc below carrier)

24. What are the drawbacks of single stub matching?

The single-stub tuner can match any load impedance (having a positive real part) to a transmission line but suffers from the disadvantage of requiring a variable length of line between the load and the stub. This may not be a problem for a fixed matching circuit but would probably pose some difficulty if an adjustable tuner was desired. In this case, the double-stub tuner, which uses two tuning stubs in fixed positions, can be used. Such tuners are often fabricated in coaxial lines with adjustable stubs connected in shunt to the main coaxial line.

25. List the various types of mixers.

1. Single-ended diode mixer
2. Single-ended FET mixer
3. Balanced mixer
4. Image reject mixer
5. Differential FET Mixer and Gilbert cell mixer.

PART-B

1. Explain the impedance of impedance transformation in transmission lines.
2. With one example, explain how the L section is used in impedance matching.
3. Explain the concept of single stub impedance matching with one example.
4. List the drawbacks of single stub matching. How double stub matching overcomes? Explain.
5. Explain how QWT is used in impedance matching?
6. Explain the design of the microwave filter using the insertion loss method.
7. Derive the design equation of microwave low pass filter.
8. Derive the equation for unilateral transducer power gain.
9. Explain the concept of stability circles with neat diagrams.
10. Explain the design of a single-stage transistor amplifier design.
11. Explain the various types of a broadband transistor amplifiers.
12. With a neat diagram, explain the operation of low noise amplifier.
13. Discuss the design of microwave oscillators using i) Transistors ii) Dielectric resonator.

14. Explain the operation of the microwave power amplifier with neat sketches.
15. Explain the design of class A power amplifier with diagrams.
16. Explain the operation of the single-ended diode mixer.
17. Discuss the working principle of i) Single-ended FET mixer ii) Balanced mixer.
18. Explain Differential FET Mixer and Gilbert cell mixer.

MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

EC8701 ANTENNA AND MICROWAVE ENGINEERING

Semester - 07

Notes

MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

Vision

To excel in providing value based education in the field of Electronics and Communication Engineering, keeping in pace with the latest technical developments through commendable research, to raise the intellectual competence to match global standards and to make significant contributions to the society upholding the ethical standards.

Mission

- ✓ To deliver Quality Technical Education, with an equal emphasis on theoretical and practical aspects.
- ✓ To provide state of the art infrastructure for the students and faculty to upgrade their skills and knowledge.
- ✓ To create an open and conducive environment for faculty and students to carry out research and excel in their field of specialization.
- ✓ To focus especially on innovation and development of technologies that is sustainable and inclusive, and thus benefits all sections of the society.
- ✓ To establish a strong Industry Academic Collaboration for teaching and research, that could foster entrepreneurship and innovation in knowledge exchange.
- ✓ To produce quality Engineers who uphold and advance the integrity, honour and dignity of the engineering.

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

1. To provide the students with a strong foundation in the required sciences in order to pursue studies in Electronics and Communication Engineering.
2. To gain adequate knowledge to become good professional in electronic and communication engineering associated industries, higher education and research.
3. To develop attitude in lifelong learning, applying and adapting new ideas and technologies as their field evolves.
4. To prepare students to critically analyze existing literature in an area of specialization and ethically develop innovative and research oriented methodologies to solve the problems identified.
5. To inculcate in the students a professional and ethical attitude and an ability to visualize the engineering issues in a broader social context.

PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO1: Design, develop and analyze electronic systems through application of relevant electronics, mathematics and engineering principles.

PSO2: Design, develop and analyze communication systems through application of fundamentals from communication principles, signal processing, and RF System Design & Electromagnetics.

PSO3: Adapt to emerging electronics and communication technologies and develop innovative solutions for existing and newer problems.

EC8701	ANTENNAS AND MICROWAVE ENGINEERING	L	T	P	C
		3	0	0	3

OBJECTIVES:

- To enable the student to understand the basic principles in antenna and microwave system design
- To enhance the student knowledge in the area of various antenna designs.
- To enhance the student knowledge in the area of microwave components and antenna for practical applications.

UNIT I INTRODUCTION TO MICROWAVE SYSTEMS AND ANTENNAS 9

Microwave frequency bands, Physical concept of radiation, Near- and far-field regions, Fields and Power Radiated by an Antenna, Antenna Pattern Characteristics, Antenna Gain and Efficiency, Aperture Efficiency and Effective Area, Antenna Noise Temperature and G/T, Impedance matching, Friis transmission equation, Link budget and link margin, Noise Characterization of a microwave receiver.

UNIT II RADIATION MECHANISMS AND DESIGN ASPECTS 9

Radiation Mechanisms of Linear Wire and Loop antennas, Aperture antennas, Reflector antennas, Microstrip antennas and Frequency independent antennas, Design considerations and applications.

UNIT III ANTENNA ARRAYS AND APPLICATIONS 9

Two-element array, Array factor, Pattern multiplication, Uniformly spaced arrays with uniform and non-uniform excitation amplitudes, Smart antennas.

UNIT IV PASSIVE AND ACTIVE MICROWAVE DEVICES 9

Microwave Passive components: Directional Coupler, Power Divider, Magic Tee, attenuator, resonator, Principles of Microwave Semiconductor Devices: Gunn Diodes, IMPATT diodes, Schottky Barrier diodes, PIN diodes, Microwave tubes: Klystron, TWT, Magnetron.

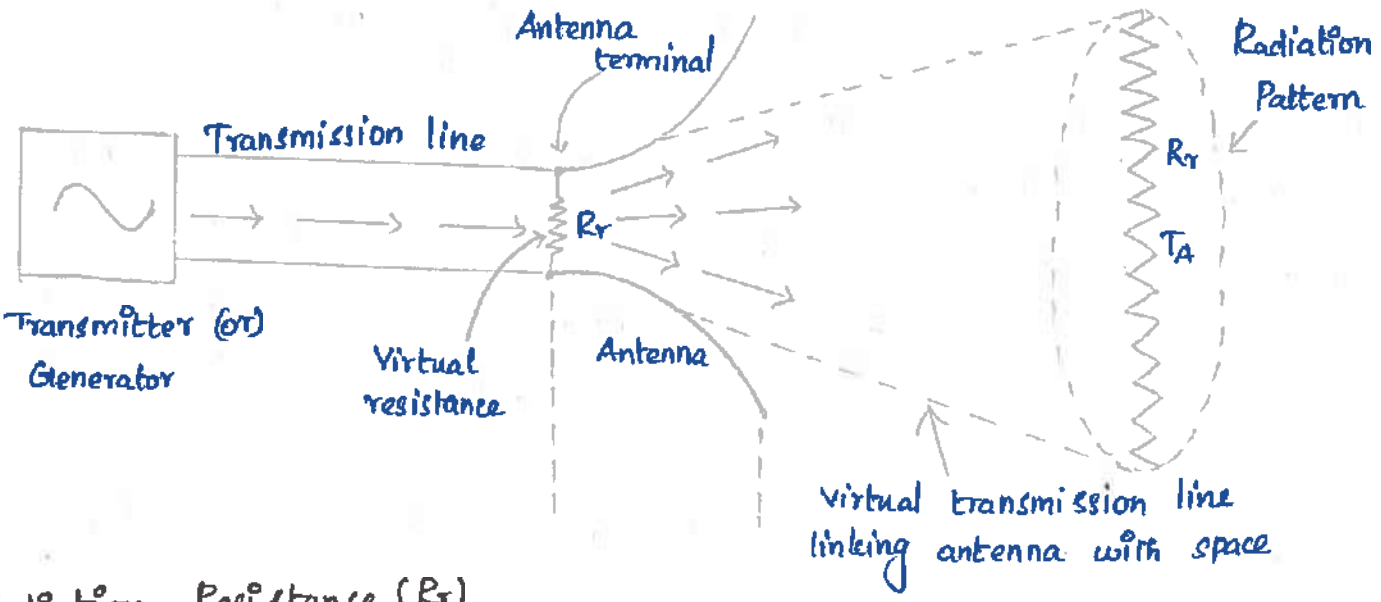
UNIT V MICROWAVE DESIGN PRINCIPLES 9

Impedance transformation, Impedance Matching, Microwave Filter Design, RF and Microwave Amplifier Design, Microwave Power amplifier Design, Low Noise Amplifier Design, Microwave Mixer Design, Microwave Oscillator Design

TOTAL: 45 PERIODS

PHYSICAL CONCEPT OF RADIATION

An antenna is an important basic component in the communication system. Basically antennas are the metallic structures designed for radiating and receiving an Electromagnetic (EM) energy in an effective manner which is used for conveying the information.



(i) Radiation Resistance (R_r)

The antenna appears as a terminal resistance to the transmission line which is commonly called as radiation resistance and denoted by 'R_r'. It is a resistance coupled from space to the antenna terminals.

(ii) Antenna Temperature (T_A)

The receiving antenna receives two types of radiations namely, passive radiations which are the reflections from any obstacle distant objects, while active radiations from other antennas. These radiation increases the apparent temperature of the radiation resistance, which is related to the temperature of the distant objects.

The basic equation of the radiation can be simply expressed as,

$$IL = Q_v \ (A - m/s)$$

Where, I - Time changing current (As^{-1})

l - Length of the current element (m)

Q - charge (c)

v - time change of velocity which equals the acceleration of the charge (ms^{-2}).

Radiation Pattern : Antenna Pattern.

* Any antenna is characterized by its radiation pattern which is a mathematical or graphical representation of the radiation properties of an antenna as a function of space coordinates in a desired direction. This is called as the radiation pattern.

* The total radiation field strength is expressed as,

$$E = \sqrt{E_\theta^2 + E_\phi^2}$$

where, E_θ - amplitude of θ component

E_ϕ - amplitude of ϕ component

* There are two basic types of radiation pattern :

(i) If the radiation of an antenna is expressed in terms of field strength (E) in V/m, then the graphical representation is called field strength pattern or field radiation pattern.

(ii) Similarly, if the radiation of an antenna is expressed in terms of the power per unit solid angle, then the graphical representation is called Power radiation pattern or simply power pattern.

* Field pattern typically represents a plot of the magnitude of an electric or magnetic field as a function of the angular space.

* Power pattern (in linear scale) typically represents a plot of the square of the magnitude of an electric or magnetic field as a function of the angular space.

* Power pattern (in dB) represents the magnitude of an electric or magnetic field in decibels, as a function of the angular space.

Normalized Field Pattern:

It is obtained, when dividing a field component of radiation pattern by its maximum value. It is a dimensionless number with a maximum value of unity.

$$E_\theta(\theta,\phi)_n = \frac{E_\theta(\theta,\phi)}{E_\theta(\theta,\phi)_{max}} \text{ (dimensionless)}$$

Normalized Power pattern:

It is obtained, when dividing a power component of radiation pattern by its maximum value as a function of angle. It is a dimensionless number with the maximum value of unity.

$$P_n(\theta,\phi)_n = \frac{S_\theta(\theta,\phi)}{S_\theta(\theta,\phi)_{max}} \text{ (dimensionless)}$$

where

$S(\theta,\phi)$ = Poynting vector

$$= \frac{E_\theta^2(\theta,\phi) + E_\phi^2(\theta,\phi)}{Z_0} \ W m^{-2}$$

Z_0 = Intrinsic impedence of space = 376.7 Ω

Antenna Beamwidth.

* It is the measure of the directivity of an antenna and it is defined as, "the angular seperation, that is, angular width in degrees between the two identical points on the opposite side of the main radiation pattern".

* In an antenna pattern, there are a number of beam widths possible, but two of the most widely used beam-widths are:

(i) Half-Power Beam width (HPBW)

(ii) First-Null Beam Width (FNBW)

(1) Half-Power Beam Width (HPBW)

* HPBW is an angular width in degrees, measured on the major lobe radiation pattern between points where the radiated power has fallen to half of its maximum value, which is called half power points.

* HPBW is also known as "3-dB beamwidth" because at half power points, the power is 3-dB down the maximum power value of the major lobe.

(2) First-Null Beam-width (FNBW)

* FNBW is defined as "the angular width between the first nulls or first side lobes, which has a beamwidth of 10 dB down from the power maximum of the main lobe".

* FNBW is also known as 10-dB beamwidth and it is usually used to approximate the HPBW as,

$$HPBW = \frac{FNBW}{2}$$

Radiation Pattern Lobes :-

* Different parts of radiation pattern are referred to as "lobes". A radiation lobe is a 3 dimensional portion of strong fields which is surrounded by a weak field. It is the portion of significant field strength in a particular direction.

* Depends on the field strength, the radiation lobes of an antenna may be classified into four types:

(i) Major lobe
(ii) Minor Lobe
(iii) Side Lobe
(iv) Back Lobe

(i) Major Lobe :

* This is the radiation lobe containing the maximum radiation in a desired direction, which is also referred to as main lobe or main beam.

* The maximum power is transmitted to the free space from an antenna only by the major lobes.

(ii) Minor Lobe :-

* All the lobes other than the main lobe are called minor Lobes. It represents the radiation in an undesired direction. The level of a minor lobe is usually expressed as "a ratio of power density in that lobe to that of the major lobe".

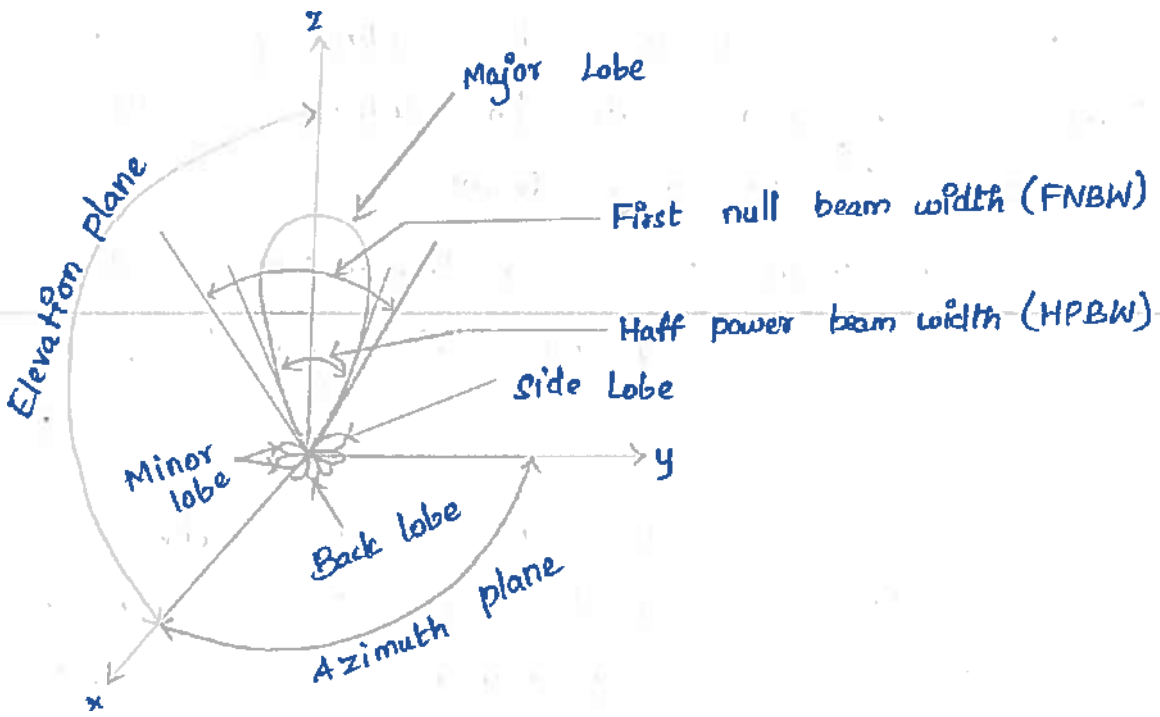
* In most of the wireless systems, minor lobes are undesired. Hence, a good antenna design should minimize the minor lobes.

(3) Side lobe :

These are the minor lobes adjacent to the main lobe and which are seperated by the various nulls. The side lobes are the largest among the minor lobes.

(4) Back Lobe :

This is the minor Lobe diametrically opposite to the main lobe. It is the radiation lobe whose axis makes an angle of approximately 180° with the major lobe direction.



Lobes and beamwidths of an antenna radiation pattern.

NEAR AND FAR-FIELD REGIONS: ANTENNA FIELD ZONES.

The space surrounding an antenna is usually subdivided into three regions.

(i) Reactive near-field region (or) Antenna region.

(ii) Radiating near-field (or) Near field (or) Fresnel region.

(iii) Far-field regions (or) Fraunhofer region.

Near-Field Region:

* Near-field region (or) Fresnel region is defined as, "region of the field of an antenna between the reactive near-field region and the far-field region wherein radiation fields predominate and wherein the angular field distribution is dependent upon the distance from the antenna.

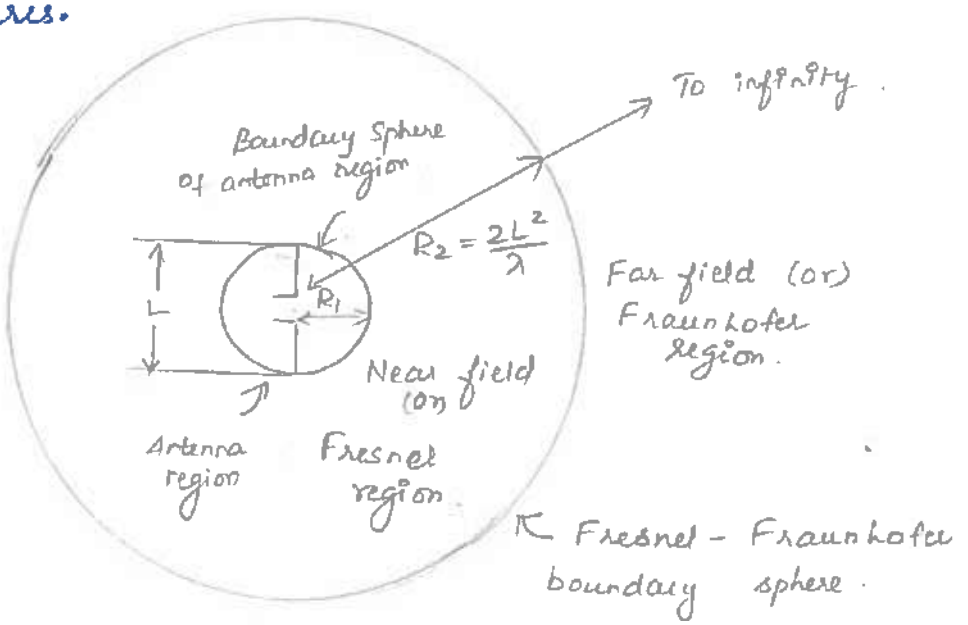
* The inner boundary is taken to be the distance $R_1 \geq 0.62 \sqrt{L^3/\lambda}$ (m) and the outer boundary distance is $R_2 \geq 2L^2/\lambda$ (m).

Far-field Region:

* The far field region is a region which is commonly taken to exists at a distances greater than

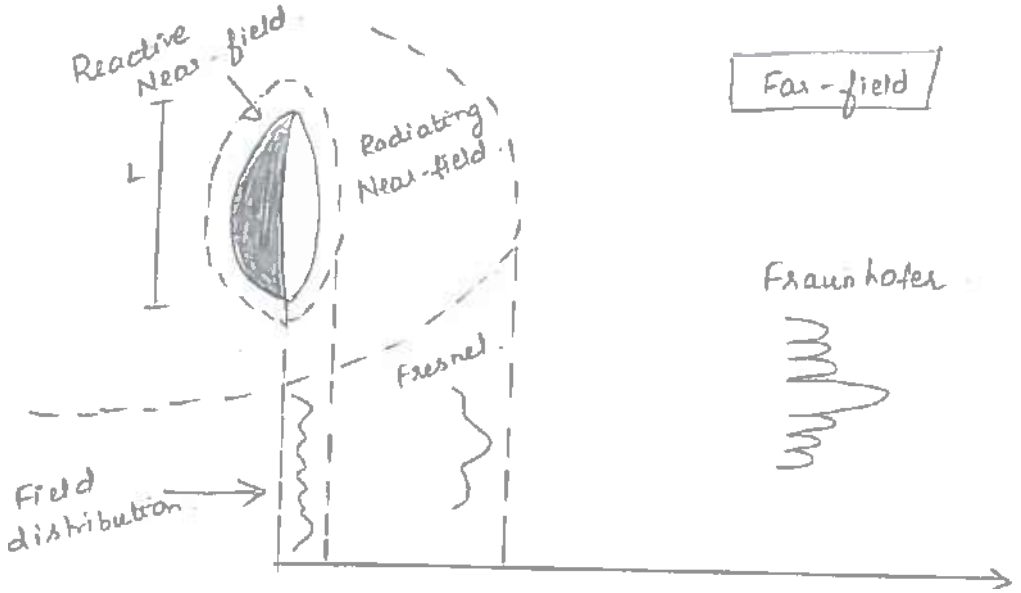
$$R_2 = \frac{2L^2}{\lambda} \text{ (m)}$$ from an antenna.

* The far-field patterns of certain antennas such as multibeam reflector antennas, are sensitive to the variations in phase over their apertures.



Radiation Pattern:

The radiation pattern of an antenna, as the observation distance is varied from the reactive near field to the far field, changes in ~~shang~~ shape because of variations of the fields in both magnitude and phase.



* A typical shape changes of radiation pattern from reactive near field toward the far field, with the largest dimension L as shown in figure.

* It is apparent that in the reactive near field region the pattern is more spread out and nearly uniform, with slight variations. In the radiating near-field region, the pattern begins to smooth and form lobes.

* In the far-field region, the pattern is well formed, usually consisting of few minor lobes and one, or more, major lobes.

FIELD AND POWER RADIATED BY AN ANTENNA:

Radiated Electric Field:

Consider an antenna located at the origin of a spherical coordinate system. At large distances where the localized near-zone fields are negligible, then the radiated electric field of an arbitrary antenna can then be expressed as,

$$\bar{E}(r,\theta,\phi) = [\hat{\theta} F_\theta(\theta,\phi) + \hat{\phi} F_\phi(\theta,\phi)] \frac{e^{-jkr}}{r} \text{ V/m} \rightarrow (1)$$

where $\bar{E}$ – Electric field vector.

$\hat{\theta}$ and $\hat{\phi}$ – Unit vectors in the spherical coordinate system.

r – Radial distance from an origin.

$k_0 = \frac{2\pi}{\lambda}$ – Free space propogation constant with wavelength $\lambda = c/f$ and

$F_\theta(\theta,\phi)$ and $F_\phi(\theta,\phi)$ – Antenna pattern function.

* Equation (1) represents that this electric field propogates in the radial direction with a phase variation of e^{-jk_0r} and an amplitude variation with distance of $1/r$. This is a TEM wave, so the electric field may be polarized in either the $\hat{\theta}$ and $\hat{\phi}$ direction but are not in the radial direction.

Radiated Magnetic Field :-

* The magnetic fields associated with an electric field of equation (1) can now be expressed as

$$H_\phi = \frac{E_\theta}{\eta_0} \longrightarrow (2a)$$

$$H_\theta = \frac{-E_\phi}{\eta_0} \longrightarrow (2b)$$

where wave impedance of free space, $\eta_0 = 377\Omega$.

* The magnetic field vector also polarized only in the transverse directions and the Poynting Vector for this wave is expressed as,

$$\bar{S} = \bar{E} \times \bar{H}^* \ W/m^2 \longrightarrow (3)$$

* And the time-averaged Poynting Vector is

$$\bar{S}_{avg} = \frac{1}{2} Re\{\bar{S}\} = \frac{1}{2} Re\{\bar{E} \times \bar{H}^*\} \ W/m^2 \longrightarrow (4)$$

Far-Field Distance

* The far-field distance is the distance where the spherical wave front radiated by an antenna becomes a close approximation to an ideal planar phase front of a plane wave.

* This approximation applies over the radiating aperture of an antenna which depends on the maximum dimension (L) of the antenna. Then the far-field distance is defined as

$$R_{ff} = \frac{2L^2}{\lambda} \text{ m} \longrightarrow (5)$$

* The above result is derived from the condition that the actual spherical wave front radiated by an antenna departs less than $\pi/8 = 22.5°$ from a true plane wave front over the maximum extent of an antenna.

* For electrically small antennas, such as short dipoles and small loops, this result (equation 5) may give a far-field distance that is too small; in the case, a minimum value of $R_{ff} = 2\lambda$ should be used.

Radiation Intensity :

* The radiation intensity gives the variation in radiated power versus position around the antenna. The radiation intensity of the radiated electromagnetic field is expressed as,

$$U(\theta,\phi) = r^2 |\overline{S_{avg}}| = \frac{r^2}{2} Re\{ E_\theta \hat{\theta} \times H_\phi^* \hat{\phi} + E_\phi \hat{\phi} \times H_\theta^* \hat{\theta} \} \longrightarrow (6)$$

By using equations (1), (2) and (4) then equation (6) becomes.

$$U(\theta,\phi) = \frac{r^2}{2\eta_0} [|E_\theta|^2 + |E_\phi|^2] = \frac{1}{2\eta_0} [|F_\theta|^2 + |F_\phi|^2] \; W \longrightarrow (7).$$

* The units of the radiation intensity are Watts or Watts per unit solid angle, since the radial dependence has been removed.

* We can find the total power radiated by an antenna by integrating the Poynting vector over the surface of a sphere of radius r that encloses the antenna which is equivalent to integrating the radiation intensity over a unit sphere:

$$P_{rad} = \int_{\phi=0}^{2\pi} \int_{\theta=0}^{\pi} S_{avg} \cdot \hat{r} \, r^2 \sin\theta \, d\theta \, d\phi$$

$$= \int_{\phi=0}^{2\pi} \int_{\theta=0}^{\pi} U(\theta,\phi) \sin\theta \, d\theta \, d\phi \longrightarrow (8).$$

ANTENNA PATTERN CHARACTERISTICS :-

* The radiation pattern (or) antenna pattern of an antenna is a plot of the magnitude of the far-zone field strength versus position around the antenna, at a fixed distance from an antenna.

(i) Lobes :-

* The pattern may exhibit several distinct lobes, with different maxima in different directions that is different part of the radiation pattern.

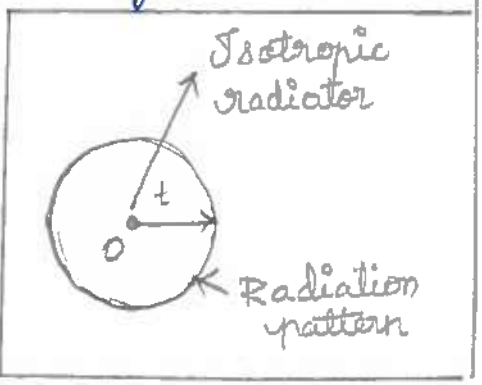
* The lobe having the maximum value in the desired direction is called main beam, which those lobes at lower levels other than main lobe are called the side lobes.

* A fundamental property of an antenna is its ability to focus power in a desired direction, not for the other directions. Thus an antenna with a broad main beam can transmit (or) receive power over a wide angular region, while it has a narrow main beam over a small angular region.

(ii) Isotropic radiator :-

* An isotropic radiator is a radiator which radiates uniformly in all the directions. It is also called as isotropic source or omni directional radiator or simply unipole.

* Basically it is a lossless ideal radiator or antenna. Generally, all the practical antennas are compared with the characteristics of isotropic radiator. So it is also called as reference antenna.



(b) Pencil Beam:-

Patterns that have relatively narrow main beams in both planes are known as pencil beam antennas and are useful in applications such as radar and point-to-point radio links.

(ii) Directivity (D):-

* Another measure of the focussing ability of an antenna is the directivity which is defined as, "the ratio of the maximum radiation intensity in the main beam to average radiation intensity over all space".

$$D = \frac{\text{Maximum Radiation Intensity in the main beam (or) test antenna}}{\text{Radiation Intensity of an Isotropic antenna.}}$$

$$D = \frac{U_{max}}{U_{avg}}.$$

* The average radiation intensity is equal to the total power radiated (P_{rad}) by an antenna divided by 4π.

$$D = \frac{4\pi U_{max}}{P_{rad}} = \frac{4\pi U_{max}}{\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi} U(\theta,\phi)\sin\theta\, d\theta\, d\phi.} \rightarrow (1)$$

* Directivity is a dimensionless ratio of power and it is usually expressed in dB as $D(dB) = 10\log(D)$ and an isotropic element $D=1$, or 0 dB.

* Typical directives for some common antennas are 2.2 dB for a wire dipole, 7.0 dB for a microstrip patch antenna, 23 dB for a waveguide horn antenna and 35 dB for a parabolic reflector antenna.

Relationship between Directivity and Beamwidth:

$$D \cong \frac{32400}{\theta_1 \theta_2} \longrightarrow (2)$$

* This approximation works well for antennas with antenna beam patterns.
* Here θ_1 and θ_2 are the beamwidth in two orthogonal planes of the main beam, in degrees.
* This approximation does not work well for the omnidirectional patterns because there is a well-defined main beam in only one plane for such patterns.

(iii) Antenna gain and Efficiency:

* The gain is an useful measure describes the performance of an antenna which acts as the figure of merit for an antenna. It is closely related to the directivity which is a measure that takes into account an antenna efficiency as well as its directional capabilities.
* The gain of the transmitting antenna is defined as, "the ability of an antenna to concentrate the radiated power in a given direction", where as for the receiving antenna, "it is an ability of absorbing incident power effectively from the particular radiation direction".

Radiation Efficiency (or) Antenna Efficiency:

The radiation efficiency of an antenna is defined as, "the ratio of the desired output power to the supplied input power".

$$\eta_{rad} = \frac{P_{rad}}{P_{in}} = \frac{P_{rad}}{P_{rad} + P_{loss}} \quad [\because P_{in} = P_{rad} + P_{loss}]$$

$$= \frac{P_{in} - P_{loss}}{P_{in}} = 1 - \frac{P_{loss}}{P_{in}} \longrightarrow (1)$$

where P_{rad} = Power radiated by the antenna

P_{in} = Power supplied to the input of the antenna.

P_{loss} = Power lost in the antenna.

* Other factors that can contribute to an effective loss of transmit power are the impedance mismatch at the input to the antenna and polarization mismatch with the receive antenna.

* These losses are external to an antenna and could be eliminated by the proper use of matching networks or the proper choice and positioning of the receive antenna.

* Antenna directivity is a function only of the shape of the radiation pattern which is not affected by losses in an antenna itself. An antenna having a radiation efficiency less than unity will not radiate all of its input power.

* The relation between antenna gain (G_1) and directivity (D) is expressed in terms of antenna efficiency (η) as,

$$G_1 = \eta_{rad} D \qquad 0 \leq \eta_{rad} \leq 1 \longrightarrow (2)$$

* In most well designed antennas, η_{rad} may be close to the unity (100%). In practice, G_1 is always less than D ($G < D$) due to Ohmic losses in the antenna.

* When an antenna efficiency is 100% ($\eta_{rad}=1$), the gain (G) and directivity (D) are used interchangeably. Gain of an antenna is expressed in decibal as,

$$G(dB) = 10\log_{10} G \longrightarrow (3)$$

(iv) Aperture Efficiency and effective area:

* Aperture antenna means that the antenna has a well-defined aperture area from which radiation occurs. Example: reflector antennas, horn antennas, lens antennas and array antennas.

* The maximum directivity that can be obtained from an electrically large aperture of area A and it is given as

$$D_{max} = \frac{4\pi A}{\lambda^2} \longrightarrow (1)$$

Aperture Efficiency:

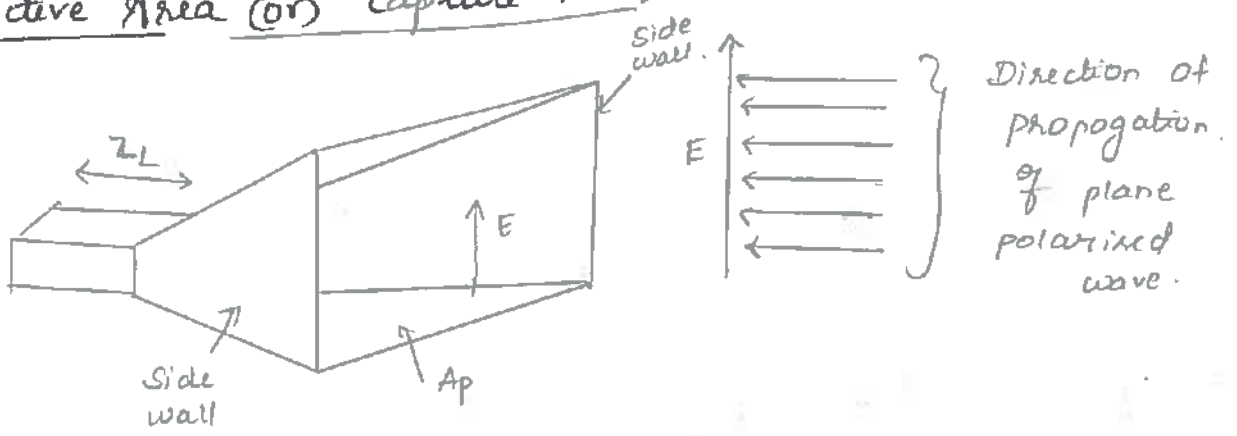
* An aperture efficiency is defined as, "the ratio of the actual directivity of an aperture antenna to the maximum directivity value possible to that of the antenna". Then the directivity of an aperture can be written as

$$D = \eta_{ap} \frac{4\pi A}{\lambda^2} \qquad 0 \le \eta_{ap} \le 1 \longrightarrow (2)$$

* An antenna efficiency is simply defined for an aperture antenna as, "the ratio of effective area (or) capture area to physical aperture of that antenna".

$$\eta_{ap} = \frac{A_e}{A_p} \quad \text{(dimensionless)} \longrightarrow (3)$$

Effective Area (or) Capture Area:



* Effective area (A_e) is an area over which an antenna extracts power from the incident radio waves. It may be defined as, "the ratio of power received at an antenna load terminal to the Poynting vector (power density) in W/m^2 of an incident wave.

$$A_e = \frac{\text{Power received by the antenna}}{\text{Poynting vector (or) power density of the incident wave.}}$$

$$A_e = \frac{P_r}{S_{avg}} \longrightarrow (4)$$

where P_r – Power received in watts.

S_{avg} – Power density [Power flow per sq. meter] or Poynting vector of an incident wave in W/m^2 and

A_e – Effective area in m^2

$$P_r = A_e S_{avg} \longrightarrow (5)$$

* The maximum effective aperture area of an antenna can be related to the directivity of an antenna as,

$$A_e = \frac{D\lambda^2}{4\pi} \longrightarrow (6)$$

where λ is the operating wavelength of the antenna and above expression does not include the effect of losses in the antenna. For electrically large aperture antennas, the effective aperature area is often close to the actual physical aperture area.

* For many other types of antennas, such as dipoles and loops, there is no simple relation between the physical cross-sectional area of an antenna and its effective aperture area.

(V) Antenna Noise Temperature:

The antenna temperature or antenna noise temperature for a lossless antenna is defined as, "the temperature of a far field region of space and near surroundings which are coupled to the antenna through radiation resistance".

Radiation Efficiency (η_{rad}):

It is the ratio of output power to input power of an antenna.

$$\eta_{rad} = \frac{P_o}{P_i} \longrightarrow (1)$$

* If a receiving antenna has dissipative loss and its radiation efficiency η_{rad} is less than unity ($\eta_{rad} < 1$). The power available at the the terminals of the antenna is reduced by the factor η_{rad} from that intercepted by the antenna.

* This reduction applies to received noise power as well as received signal power, so that the noise temperature (T_A) of an antenna will be reduced from the brightness temperature (T_b) by the factor η_{rad}.

* The thermal noise will be generated internally by resistive losses in the antenna which will increase the noise temperature of an antenna.

* In terms of noise power, a lossy antenna can be modeled as a lossless antenna and an attenuator having a power loss factor of $L = 1/\eta_{rad}$. Then, equivalent noise temperature of an attenuator, we can find the resulting noise temperature seen at the antenna terminals as,

$$T_A = \frac{T_b}{L} + \frac{(L-1)}{L} T_p = \eta_{rad} T_b + (1-\eta_{rad}) T_p \longrightarrow (1)$$

where T_A – Antenna noise temperature (K).
T_b – Brightness temperature (K) and
T_p – Antenna physical temperature (K).

* The antenna noise temperature (T_A) is a combination of the external brightness temperature seen by the antenna and the thermal noise generated by the antenna.

* For lossless antenna, $\eta_{rad} = 1$, then eqn (1) reduces to $T_A = T_b$. If the radiation efficiency is zero ($\eta_{rad} = 0$), it means that the antenna appears as a matched load and does not see any external background noise, then eqn (1) reduces to $T_A = T_p$, due to the thermal noise generated by the losses.

Thermal Noise:

Assuming no losses or other contributions between the antenna and the receiver then the noise power transferred to the receiver is given by

$$P_r = KT_A \Delta f \longrightarrow (2)$$

where P_r – Antenna noise Power (W)

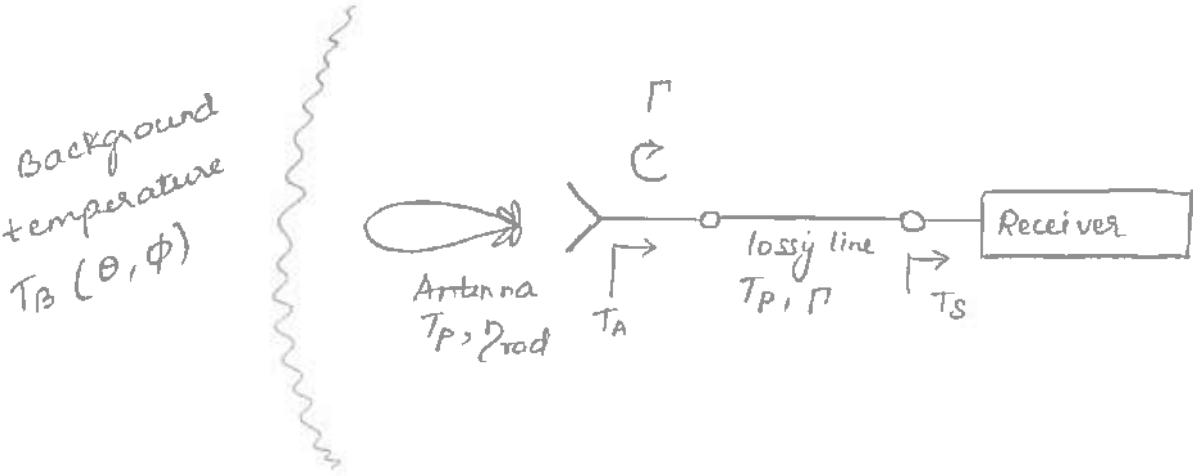
K – Boltzmann's Constant (1.38×10^{-23} J/k)

T_A – Antenna temperature (K)

Δf – Bandwidth (Hz).

Overall System Noise Temperature:

* The more general problem of a receiver is when it connected through a lossy transmission line to an antenna and viewing as a background noise temperature distribution T_B when an Impedance mismatch exists between the antenna and the line which can be represented by the system shown below.



* Here, the antenna is assumed to have a radiation efficiency η_{rad} and the connecting transmission line has a power loss factor of $L \geq 1$, with both at physical temperature T_p.

* The effect of an impedance mismatch between an antenna and the transmission line is represented by the reflection coefficient Γ.

* The equivalent noise temperature seen at the output terminals of the transmission line consists of three contributions:

(i) Noise power from an antenna due to internal noise and the background brightness temperature.

(ii) Noise power generated from the lossy line in the forward direction.

(iii) Noise power generated by the lossy line in the backward direction and reflected from an antenna mismatch toward the receiver.

* Due to above noise contributions, the noise due to the antenna is given by equation (1) is reduced by the loss factor of the line, $1/L$ and the reflection mismatch factor, $(1-|\Gamma|^2)$

* The forward noise power from the lossy line is reduced by the loss factor, $1/L$. The contribution from the lossy line reflected from the mismatched antenna is reduced by the power reflection coefficient, $|\Gamma|^2$ and the loss factor $1/L^2$.

* Therefore, the overall system noise temperature seen at an input to the receiver is given by

$$T_S = \frac{T_A}{L}(1-|\Gamma|^2) + (L-1)\frac{T_P}{L} + (L-1)\frac{T_P}{L^2}|\Gamma|^2 \rightarrow (3)$$

By substituting eqn (1) in eqn (3) we get

$$T_S = \frac{(1-|\Gamma|^2)}{L}[\eta_{rad} T_b + (1-\eta_{rad}) T_P] + \frac{(L-1)}{L}\left(1+\frac{|\Gamma|^2}{L}\right)T_P \rightarrow (4)$$

* For a lossless line ($L=1$), the effect of an antenna mismatch is to reduce the system noise temperature by the factor $(1-|\Gamma|^2)$ and the received signal power will be reduced by the same amount. For a matched antenna ($\Gamma=0$) equation (4) reduces to

$$T_S = \frac{1}{L}[\eta_{rad} T_b + (1-\eta_{rad}) T_P] + \frac{L-1}{L} T_P \rightarrow (5)$$

* Radiation efficiency accounts for the resistive losses, and thus involves the generation of thermal noise but aperture efficiency does not.

* Aperture efficiency applies to the loss of directivity in aperture antennas and by itself does not lead to any additional effect on noise temperature that would not be included through the pattern of an antenna.

(vi) Gain - Antenna Temperature Ratio: G/T:

* Another useful figure of merit for receive antennas is the G/T ratio and it is defined as,

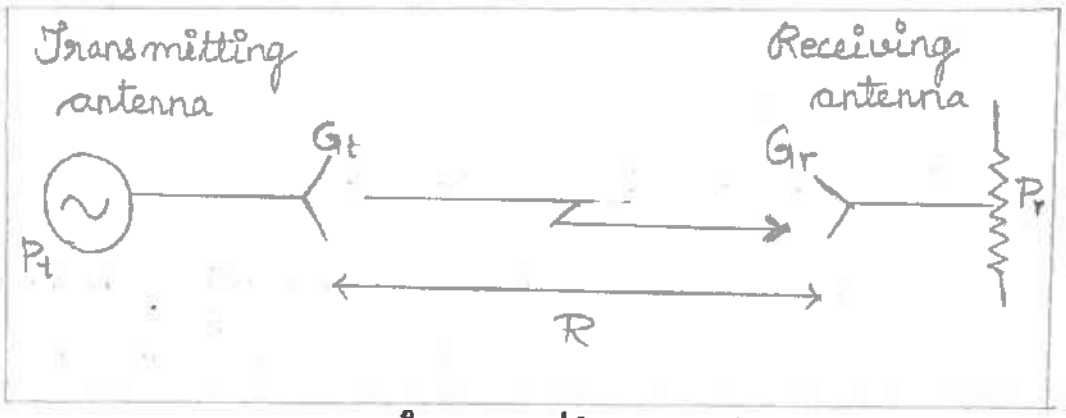
$$G/T\ (dB) = 10 \log \frac{G}{T_A}\ dB/k \longrightarrow (6)$$

* where G is the gain of the antenna and T_A is the antenna noise temperature.

* This quantity is important because, the Signal-to-Noise Ratio (SNR) at an input to a receiver is proportional to G/T_A. The ratio G/T can often be maximized by increasing the gain of an antenna and usually minimizes the reception of noise from hot sources at the low elevation angles.

* Higher gain requires a larger and more expensive antenna, and high gain may not be desirable for applications requiring Omnidirectional coverage (eg. cellular telephones or mobile data networks) so often a compromise must be made.

FRIIS TRANSMISSION EQUATION :



A Basic radio system

Where,

P_t - Transmit power

G_t - Transmit antenna gain

G_r - Receive antenna gain

P_r - Received Power which is delivered to a matched load

R - Distance between transmit and receive antennas.

* The power density radiated by an isotropic antenna at a distance R is given by

$$S_{avg} = \frac{P_t}{4\pi R^2} \ W/m^2 \longrightarrow (1)$$

* The power is distributed isotropically, and the area of Sphere is $4\pi R^2$. The general expression for the power density radiated by an arbitrary transmit antenna is,

$$S_{avg} = \frac{G_t P_t}{4\pi R^2} \ W/m^2 \longrightarrow (2)$$

* If this power density is incident on the receive antenna of effective aperture A_{er} is

$$P_r = A_{er} S_{avg} = \frac{G_t P_t A_{er}}{4\pi R^2} \ W \longrightarrow (3)$$

* The gain of the transmitting antenna can then be expressed as,

$$G_t = \frac{4\pi A_{et}}{\lambda^2} \longrightarrow (4)$$

Sub (4) in (3)

$$P_r = \frac{P_t A_{er}}{4\pi R^2} \times \frac{4\pi A_{et}}{\lambda^2}$$

$$= \frac{P_t A_{er} A_{et}}{R^2 \lambda^2}$$

$$\frac{P_r}{P_t} = \frac{A_{er} A_{et}}{R^2 \lambda^2} \text{ (Dimensionless)} \longrightarrow (5)$$

where, A_{et} = Effective aperture of transmitting antenna, m^2

A_{er} = Effective aperture of receiving antenna, m^2

From (4),

$$A_{er} = \frac{G_r \lambda^2}{4\pi} \longrightarrow (6)$$

Sub (6) in (3)

$$P_r = \frac{P_t G_t}{4\pi R^2} \times \frac{G_r \lambda^2}{4\pi}$$

$$\frac{P_r}{P_t} = \frac{G_r G_t \lambda^2}{4\pi R^2} \text{ W} \longrightarrow (7)$$

Eqn (5) and (7) are called as Friis transmission formula (or) Friis radio link formula.

* These equations include impedance mismatch at either antenna, polarization mismatch between the antennas, propagation effects leading to attenuation or depolarization and multipath effects that may cause partial cancellation of the received field.

* It is observed in eqn (4), that the received power decreases as $1/R^2$ as the separation between the transmitter and receiver increases. For long distance communications, radio links will perform better than the wired links.

* The received power is proportional to the product of $P_t G_t$, which characterizes the transmitter.

* In the main beam of an antenna, the product $P_t G_t$ can be interpreted equivalently as the power radiated by an isotropic antenna with input power $G_t P_t$. Thus this product is defined as the Effective Isotropic Radiated Power (EIRP).

$$EIRP = P_t G_t \ W \longrightarrow (8)$$

* For a given frequency, range and receiver antenna gain, the received power is proportional to the EIRP of the transmitter and can only be increased by increasing EIRP.

LINK BUDGET

* In a link budget, the various terms in the Friis transmission formula are often tabulated separately and each of the factors can be individually considered in terms of its net effect on the received power.

* In the link budget, the additional loss factors, such as line losses, or impedance mismatch at the antennas, atmospheric attenuation and polarization mismatch can also be added.

Path loss :-

* Path loss is one of the terms in a link budget which accounting for the free-space reduction in signal strength with distance between the transmitter and the receiver. It is defined as (in dB),

$$L_0(dB) = 20 \log\left(\frac{4\pi R}{\lambda}\right) > 0 \longrightarrow (1)$$

* Path loss depends on wavelength and it provides a normalization for the units of distance.

Other Terms:-

* The remaining terms of the Friis formula as shown in the following link budget:

Transmit power	P_t
Transmit antenna line loss	$(-) L_t$
Transmit antenna gain	G_{lt}
Path loss	$(-) L_o$
Atmospheric attenuation	$(-) L_A$
Receive antenna gain	G_{lr}
Receive antenna line loss	$(-) L_r$
Receive power	P_r

* Atmospheric attenuation and line attenuation loss terms also included in link budget, if all of the above quantities are expressed in dB (or) dBm, then the receive power is

$$P_r \text{ (dBm)} = P_t - L_t + G_{lt} - L_o - L_A + G_r - L_t \longrightarrow (2)$$

Impedance Mismatch Loss:

* If the transmit or receive antenna is not impedance matched to the transmitter or receiver or to their connecting lines, then an impedance mismatch will reduce the received power by the factor $(1 - |\Gamma|^2)$, where Γ is the appropriate reflection coefficient.

The resulting impedance mismatch loss can be included in the link budget to account for the reduction in received power will be expressed as,

$$L_{imp} \text{ (dB)} = -10 \log (1 - |\Gamma|^2) \geq 0 \longrightarrow (3)$$

Polarization Matching :-

* Maximum power transmission between transmitter and receiver requires both antennas to be polarized in the same manner. Therefore, the polarization matching of the transmit and receive antennas is an important entry in the link budget.

* For example, if a transmit antenna is vertically polarized, maximum power will only be delivered to a vertically polarized receiving antenna, while zero power would be delivered to a horizontally polarized receive antenna, and half the available power would be delivered to a circularly polarized antenna.

LINK MARGIN

* In practical communications systems, generally it is desired to have the received power level greater than the threshold level required for the minimum acceptable quality of service which is usually expressed as the minimum carrier to noise ratio (CNR) or minimum SNR.

* This design allowance for the received power is referred to as the Link margin and it can be expressed as the difference between the design value of received power and the minimum threshold value of the receive power.

$$\text{Link Margin (dB)} = LM = P_r - P_{r(min)} > 0 \longrightarrow (4)$$

* In equation (4) all the quantities are in dB. Link margin should be a positive number and its typical values may ranges from 3 to 20 dB.

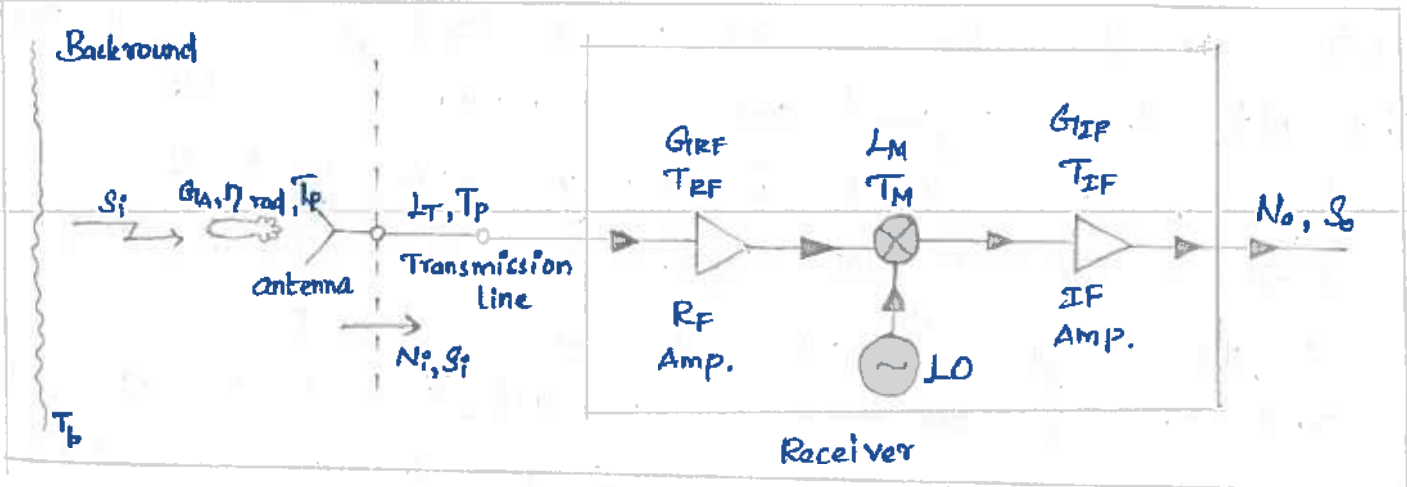
* Link Margin is used to account for fading effects and it is sometimes referred to as fade margin. For example, satellite links operating at frequencies above 10 GHz, often require fade margins of 20dB or more to account for attenuation during heavy rain.

* Link Margin for a given communication system can be improved by,

(i) Increasing the received power by increasing the transmit power or antenna gains.

(ii) Reducing the minimum threshold power by improving the design of the receiver or changing the modulation method.

NOISE CHARACTERIZATION OF A MICROWAVE RECEIVER.



Noise analysis of a microwave receiver.

* In this system, the total noise power at the output of the receiver is N_o, it will be due to the contributions from the antenna pattern, the loss in the antenna, the loss in the transmission line, and the receiver components.

* This noise power will determine the minimum detectable signal level for the receiver and for a given transmitter

Power, the maximum range of the communication link.

* The receiver components consists of an RF amplifier with gain G_{RF} and noise temperature T_{RF}, a mixer with an RF-to-IF conversion loss factor L_M and noise temperature T_M and an IF amplifier with gain G_{IF} and noise temperature T_{IF}.

* The noise effects of later signals in the microwave receiver can usually be ignored since the overall noise figure(F) is dominated by the characteristics of the first few stages.

* The component noise temperatures can be related to noise figures as $T = (F-1)T_0$. The noise temperature of the receiver can be expressed as,

$$T_{REC} = T_{RF} + \frac{T_M}{G_{RF}} + \frac{T_{IF} L_M}{G_{RF}} \longrightarrow (1)$$

* The transmission line connecting the antenna to the receiver has a loss L_T, and it is at a physical temperature T_P. Then its equivalent noise temperature is expressed as,

$$T_{TL} = (L_T - 1) T_P \longrightarrow (2)$$

* If the transmission line (TL) and receiver (REC) cascade, then the noise temperature at the antenna terminals that is the input to the transmission line is

$$T_{TL+REC} = T_{TL} + L_T T_{REC}$$

$$= (L_T - 1)T_P + L_T T_{REC} \longrightarrow (3)$$

* The entire antenna pattern can collect noise power. If antenna has a reasonably high gain with relatively low sidelobes, we can assume that all noise power comes via the main beam, so that the noise temperature of the antenna is given as,

$$T_A = \eta_{rad} T_b + (1 - \eta_{rad}) T_b \longrightarrow (4)$$

where η_{rad} - Efficiency of the antenna.

T_p - Physical temperature of the antenna

T_b - Equivalent brightness temperature of the back ground seen by the main beam.

* The noise power at the antenna terminals that is, the noise power delivered to the transmission line. is

$$N_i = kBT_A = kB[\eta_{rad} T_b + (1-\eta_{rad})T_p] \longrightarrow (5)$$

where B - system bandwidth.

* If S_r is the received power at the antenna terminals, then the input SNR at the antenna terminals is S_i/N_i. The O/P signal power is,

$$S_o = \frac{\Phi i \, G_{RF} \, G_{IF}}{L_T L_M} = S_i G_{sys} \longrightarrow (6)$$

* G_{sys} defines system power gain. The output noise power is

$$N_o = (N_i + kBT_{TL} + P_{REC}) G_{sys}$$

$$= kB(T_A + T_{TL} + P_{REC}) G_{sys} \longrightarrow (7)$$

By sub (3) and (4) in (7)

$$N_o = kB[\eta_{rad} T_b + (1-\eta_{rad})T_p + (L_T - 1)T_p + L_T T_{REC}] G_{sys}$$

$$= kBT_{sys} G_{sys} \longrightarrow (8)$$

* T_{sys} is the overall system noise temperature. Then the output SNR is written as,

$$\frac{S_o}{N_o} = \frac{S_i}{kBT_{sys}}$$

By using (8)

$$\frac{S_o}{N_o} = \frac{S_i}{kB[\eta_{rad} T_b + (1-\eta_{rad})T_p + (L_T - 1)T_p + L_T T_{REC}]} \longrightarrow (9)$$

* This SNR is improved by various signal processing techniques. It is very convenient to use an overall system noise figure to calculate the degradation in SNR from an I/P to O/P of the above system.

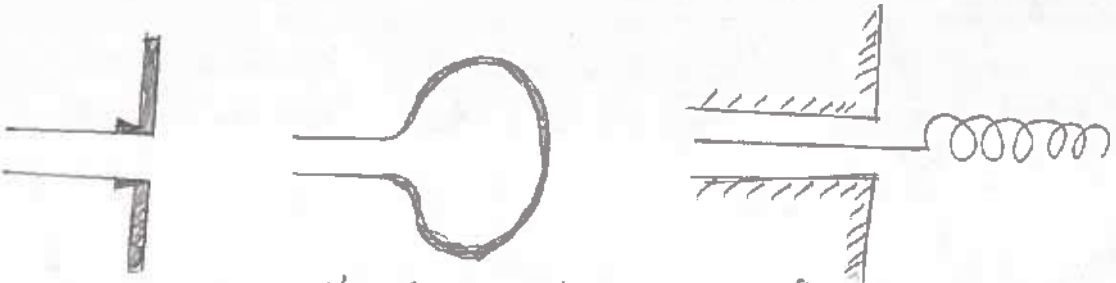
Unit - 2

Radiation Mechanisms and Design aspects.

Types of Antennas:-

(1) Wire Antennas:-

These antennas are used on automobile, buildings, ships, aircrafts, spacecraft and soon. There are various shapes of microantennas such as straight ~~line~~ wire (dipole), loop, helix as shown below.



(a) Dipole (b) circular loop (c) Helix

Loop antennas need not only be circular. They may take the form of rectangular, square, ellipse etc.,

(2) Aperture Antennas:-

These antennas are very useful for aircraft and space craft applications.

Types:-

(i) Rectangular Horn

(ii) Conical Horn

(iii) Rectangular waveguide.

(3) Microstrip Antennas:

* It is very popular in 1970's for spaceborne applications. Today they are used for government and commercial applications.

* These antennas are low profile, comfortable to planar and non-planar surfaces.

* Simple and inexpensive to fabricate.

* Can be mounted on the surface of high performance aircraft, spacecraft, stellites etc.

(4) Reflector Antennas:

Larger dimensions are needed to achieve the high gain required to transmit (or) receive signals after millions of miles of travel.

Linear wire Antennas:

* Antennas which is in the form of linear wire is called linear wire antenna.

Types of linear wire antenna.

1) Infinitesimal dipole $l < \lambda/50$
2) Short dipole $\lambda/50 < l < \lambda/10$.
3) Half wavelength dipole $l = \lambda/2$.

APERTURE ANTENNA :

The term aperture refers to an opening in a closed surface. The aperture antenna represent a class of antennas which are analysed by considering the antenna as an opening in a surface, in which radiation is considered to occur from an aperture. EM waves are transmitted or received through this opening.

Example for ~~Aperture~~ Aperture antennas are

1) Slot antenna.
2) Horn antenna.
3) Reflector antenna.
4) Lens antenna.

Horn antenna:

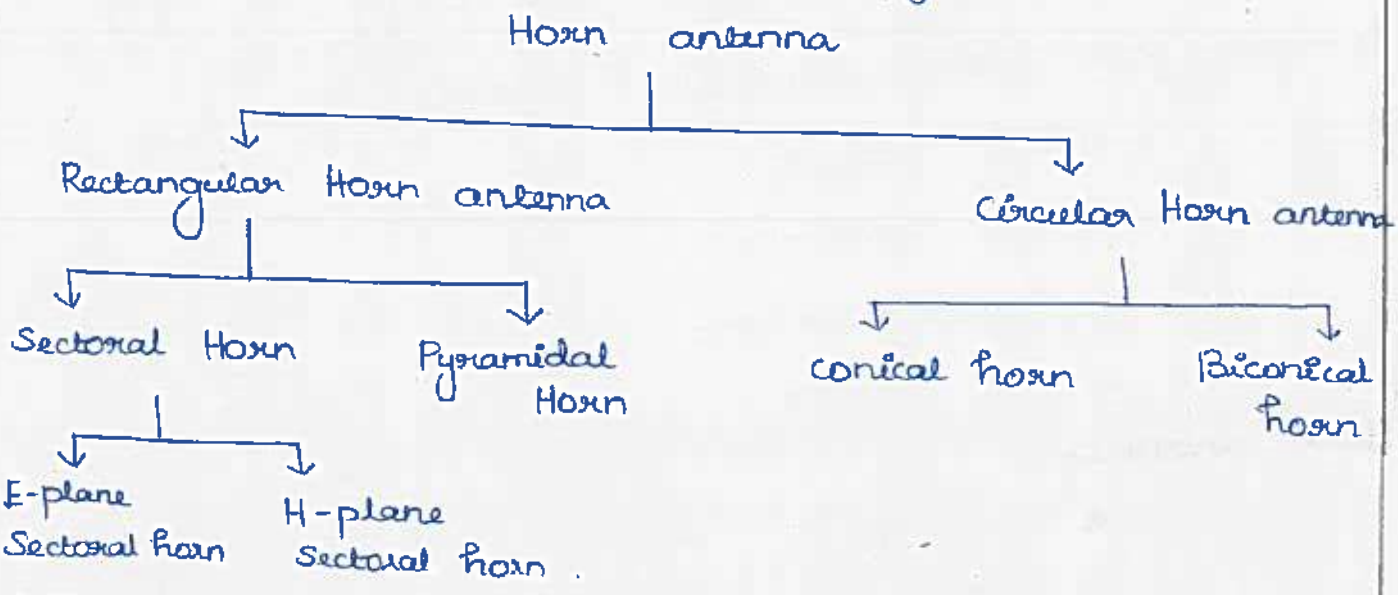
* One of the simplest and probably the most widely used microwave antenna is the horn antenna. It serves as a feed element for large radio astronomy, communication dishes and satellite tracking throughout the world.

* The function of horn is to produce a uniform phase front with a larger aperture than that of the magnitude and hence the greater directivity.

* A horn antenna (or) microwave horn is an antenna that consists of a flaring metal waveguide shaped like a horn to direct radio waves in a beam, Horns are widely used as antennas of UHF & microwave frequencies above 300MHz.

Types of horn antenna:

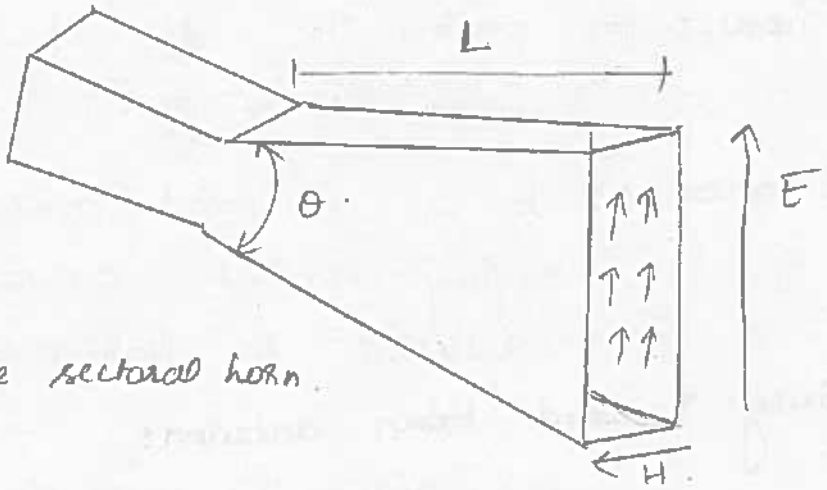
All the horn antennas are energized from either rectangular (or) circular waveguide. Basically horn antennas are classified as rectangular antennas and circular horn antennas. The rectangular horn antennas fed with rectangular waveguide, while circular horn antennas are fed with circular waveguide.



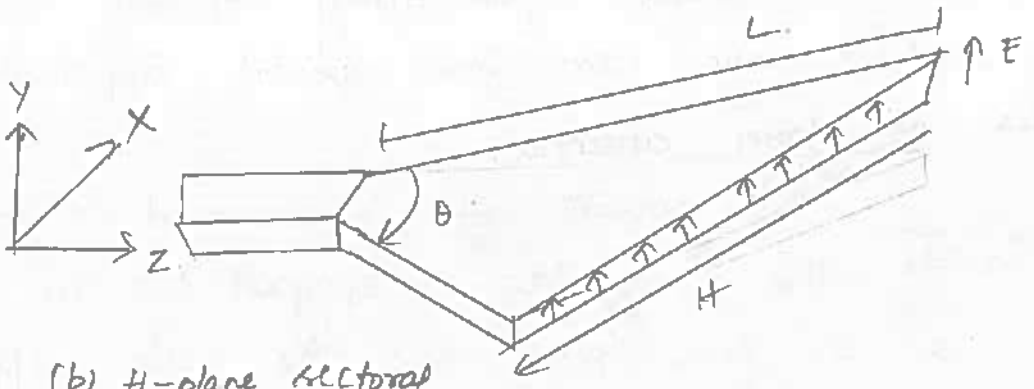
1) Rectangular Horn antenna:

* Depending upon the direction of flaring, the rectangular horns are classified as sectoral and pyramidal horn.

* The sectoral horn is a rectangular type waveguide with a flaring is done in only one direction. Te along the single wall of a rectangular waveguide.



(a) E-plane sectoral horn.



(b) H-plane sectoral Horn.

* It is further classified as E-plane sectoral horn and H-plane sectoral horn.

E-plane sectoral horn is obtained when the flaring is done in the direction of electric field vector.

H-plane sectoral horn is obtained if the flaring is done in the direction of magnetic field vector.

Pyramidal Horn: If the flaring is done along both the walls of the rectangular waveguide in the direction of both electric field and magnetic field vector, then horn obtained is called pyramidal horn and allow indicates E-field ...

2) Circular horn:

Conical horn: By flaring the walls of a circular waveguide, a conical horn is formed.

Biconical antenna: It is a broad bandwidth antenna made up of two roughly conical conductive objects, which is a non-directional in horizontal plane.

3) Exponentially Tapered Horn antenna:

To minimize the reflections of the guided wave, the transition region or horn between the waveguide at the throat and free space at the aperture could be given a gradual exponential taper. Such horns are called exponentially tapered horn antenna, which are used for special applications.

Principles of Horn antenna:

In general, it is observed that the fields inside the waveguide propagate in the same manner as in free space. But the main difference is that the propagation of wavefield is constrained by the walls of the waveguide from being spherically spreading.

Huygen's Principle:

* It says that each point on a primary wave front can be considered to be a new source of a secondary spherical wave and the secondary wavefront can be constructed as the envelope of these secondary spherical waves.

* At the end of the waveguide, the guided propagation changes to free space propagation, so this region is generally called as transition region.

(5)

* If the flare angle is small, results in small aperture area. Beam width is decreased and directivity is increased. D is inversely proportional to aperture.

* In a E plane of the Horn, δ is usually held to 0.25λ (or) less. However in the H-plane, δ can be larger (or) about 0.4λ.

For an optimum flare horn, the HPBW can be approximately.

$$\theta_H = \frac{67^\circ\lambda}{a_H} = \frac{67\lambda}{W} \text{ degree}$$

$$\theta_E = \frac{56^\circ\lambda}{a_E} = \frac{56\lambda}{a} \text{ degree}$$

Directivity $$D = \frac{4\pi A_e}{\lambda^2} = \frac{4\pi \varepsilon_{ap} A_p}{\lambda^2}$$

where,

A_e - effective aperture in m^2.

A_p - physical aperture in m^2 = Area of horn mouth opening.

$$\varepsilon_{ap} = \frac{A_e}{A_p} = \text{Aperture efficiency.}$$

For a pyramidal rectangular horn,

$$A_p = a_E a_H = a \times W$$

where,

a - height of the aperature.

W - height of the aperture.

a_E - E plane aperture.

a_H - H plane aperture.

similarly for a conical horn, $Ap = \pi r^2$.

For eg: if $a_E = a_H = \lambda = 1m$ and $\varepsilon_{ap} = 0.6$, thus the directivity of rectangular horn is given by

$$D = \frac{4\pi (0.6) Ap}{\lambda^2} = \frac{7.5\, Ap}{\lambda^2}$$

$$D(dB) = 10 \log \frac{7.5\, Ap}{\lambda^2}.$$

Advantages:

* The directivity of the pyramidal horn and conical horn is highest as they have more than one flare angle.

* It can be operated over a wide range of high frequency as there is no resonant element in the antenna.

Applications:

* Used in microwave applications.

* Used as feed element for large radio astronomy, satellite tracking, communication dishes in parabolic reflector.

* Used in short range radar systems.

* The waveguide impedance and free space impedance do not match with each other, the flaring (tapering) of the walls of the waveguide must be done so that the impedance matching is achieved along the concentrated radiation pattern with high directivity and narrow bandwidth.

Design of Horn antenna:

* Consider a pyramidal horn of length 'L' and aperture height 'h' with flaring along θ as shown in figure 2.6.2. The function of the horn is to produce a uniform phase front with a larger aperture in comparison to the waveguide. Because of this, the directivity increases.

* Consider an imaginary axis 'b' of horn as shown in figure 2.6.2. There is a path difference between a ray travelling along the side and along the axis of the horn.

Let 'δ' be the difference in the path of travel.

θ - flare angle (θ_E for E plane, θ_H for H plane) in degree.

h = a - Aperture (a_E for E plane, a_H for H plane) in m.

L - length of horn in (m).

$$\sin\frac{\theta}{2} = \frac{h/2}{L+\delta}$$

From the geometry ΔOBA, $\cos\frac{\theta}{2} = \frac{OB}{OA} = \frac{L}{L+\delta}$

$$\tan\frac{\theta}{2} = \frac{\sin\theta/2}{\cos\theta/2} = \frac{h/2}{L}.$$

$$\theta = 2\tan^{-1}\left(\frac{h}{2L}\right) = 2\cos^{-1}\left(\frac{L}{L+\delta}\right) \longrightarrow (1)$$

From figure,

$$(L+\delta)^2 = L^2 + (h/2)^2$$

$$L^2+\delta^2+2\delta L = L^2 + \frac{h^2}{4} \quad [\because \delta \text{ is small \& it can be neglected}]$$

$$2\delta L = \frac{h^2}{4}$$

$$\boxed{L = \frac{h^2}{8\delta}} \longrightarrow (2)$$

Equations (1) & (2) are the design equations of the horn antenna.

* If the flare angle 2θ is very large, the wavefront of the mouth of horn will be curved rather than plane. This will result in non-uniform phase distribution over the aperture, resulting increased beam width and decreased directivity.

Reflector Antenna

Reflector antennas (or) reflectors are widely used to modify the radiation pattern of a radiating element, for example backward radiation from an antenna may be eliminated by using a plane sheet reflector of large enough dimensions.

The antenna which is a radiating source in the reflector antenna is called primary antenna (or) feed, while the reflector antenna is called ~~Primary~~ the secondary antenna. the most common feeds are dipole, horn and slot etc.

Types of reflector Antennas

the reflector antennas are of several types and they are listed as:

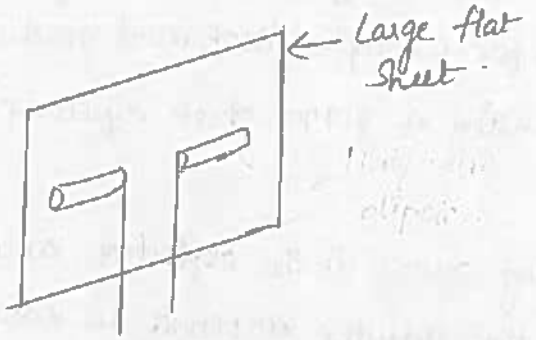
1) plane reflector (or) flat sheet reflector
2) corner reflector
3) parabolic reflector
4) Hyperbolic reflector
5) ~~Hy~~Elliptical reflector
6) Circular reflector.

a) Flat sheet reflector (or) plane reflector

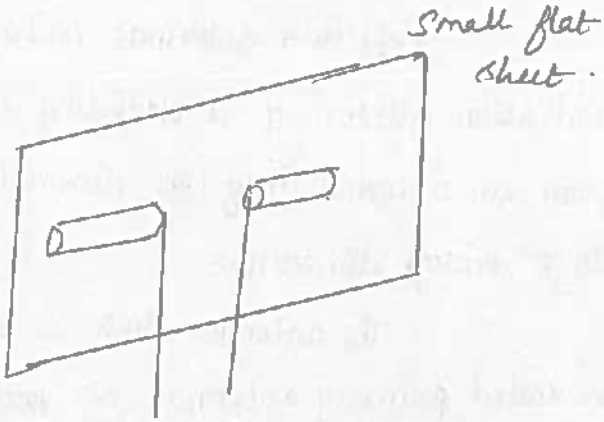
the plane reflector is the simplest form of the reflector antenna. the main advantage is that dipole (feed) backward directions are reduced and gain in the forward direction increases.

the gain can be increased further by reducing the spacing between the feed and sheet reflector, However bandwidth is narrow for small spacings.

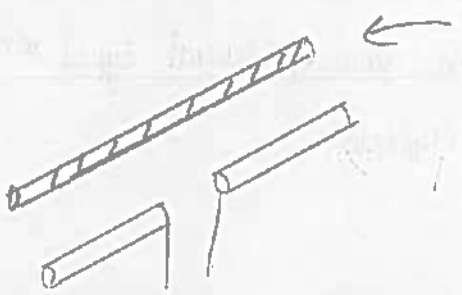
the polarization of the radiating source and its position relative to the reflecting surface can be used to control the radiating (pattern, impedence, directivity) of the overall system.



(a) Large sheet with half wave dipole fed.



(b) Small flat sheet with half wave dipole feed.



(C) Thin linear element with half wave dipole feed

Fig 3.6.3 various shapes of reflectors

(a) Large sheet with half wave dipole fed.

(b) small flat sheet with half wave dipole feed.

(c) thin linear reflector element with half wave dipole feed

~~Fig 2.6.4 various types of plane reflector.~~

A large flat sheet reflector can convert a bidirectional antenna array into a unidirectional system. small spacing between the antenna and the sheet will improve the gain in the forward direction.

The desirable properties of the sheet reflector may be largely preserved with the reflector reduced in size as shown in figure 2.6.4 (b)

Thin reflector element is used which is highly sensitive to the frequency changes was shown in figure 2.6.4 (c) which can be used to increase directivity

1) Analysis of plane reflector by method of images

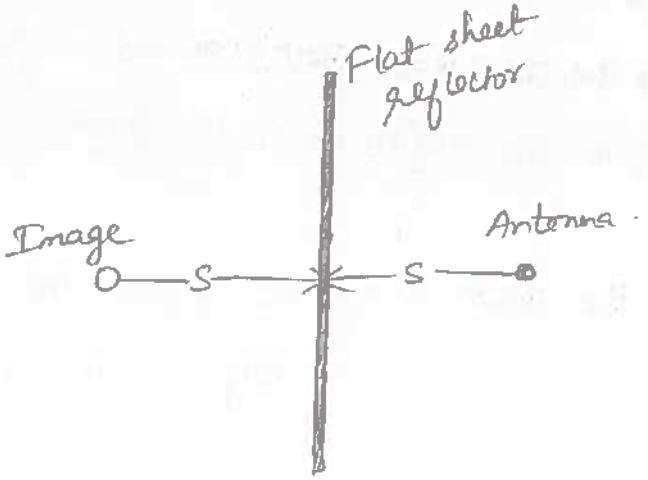


Fig 2.6.5 Antenna with flat sheet reflector & its image.

The problem of an antenna at a distance 's' from a perfectly conducting plane sheet reflector of infinite extent is readily handled by the method of images, here the reflector is replaced by an image of the antenna at a distance 2s from the antenna as shown in figure 2.6.5

Assuming zero reflector losses, the gain in terms of field intensity of a $\lambda/2$ dipole antenna at a distance 's' from an infinite plane reflector and it is expressed as

$$G_f(\phi) = 2\sqrt{\frac{R_{11} + R_L}{R_{11} + R_L - R_{12}}} \left|\sin(s_r \cos\theta)\right| \longrightarrow (1)$$

where $s_r = \frac{2\pi s}{\lambda}$

s - distance but reflector & feed

R_{11} - Impedance of antenna

R_{12} - Mutual impedance of the antenna & reflector

R_L - Antenna loss resistance.

When the reflecting sheet is reduced in size, the analysis can be done by considering three principal angular regions as shown in figure 2.6.6

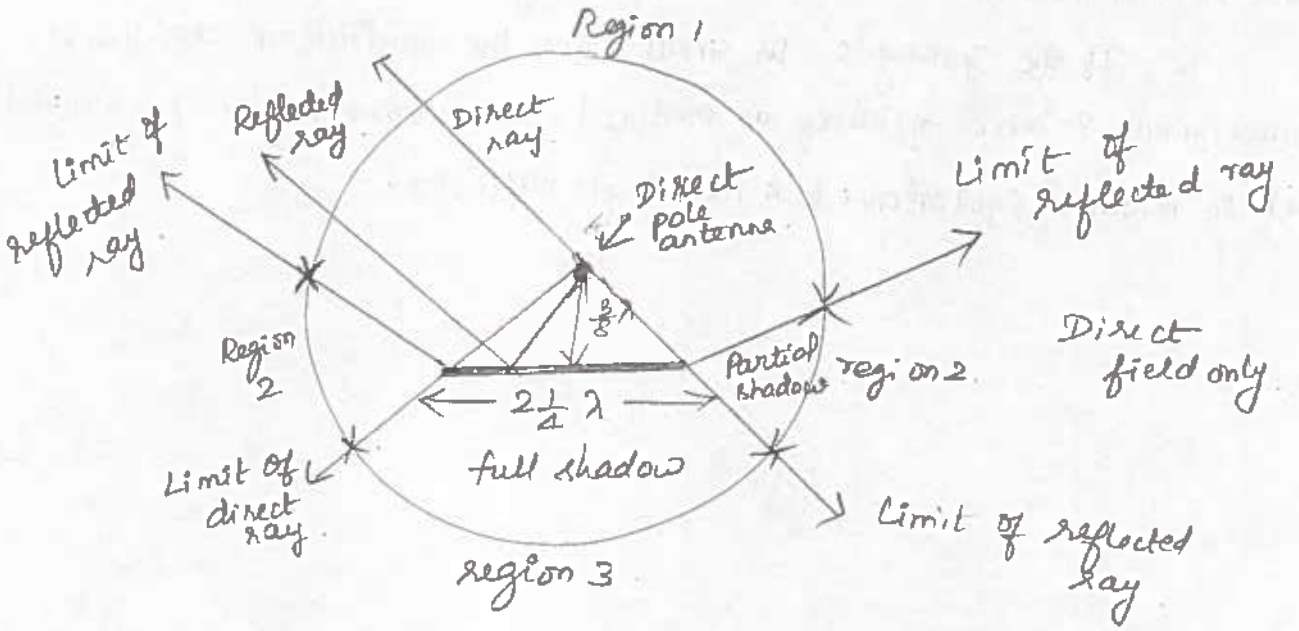
Region 1 (above (or) infront of the sheet): In this region, the radiated field is given by the addition of the resultant of the direct field of the dipole and reflected field from the sheet.

Region 2 (above (or) below at the sides of the sheet): In this region there is only the direct field from the dipole. this region is in the shadow of the reflected field.

Region 3 (below (or) behind the sheet): In this region, the sheet acts as a shield producing a full shadow, that is only the diffracted fields & there is no direct (or) reflected fields.

Fg: 2.6.6 Dipole antenna with 2.25 λ flat sheet reflector.

~~2) corner reflectors :-~~



2) Corner reflectors :-

The disadvantage of plane reflector is that there may be a radiation in back & side directions. In order to overcome this limitation, the geometrical shape of the plane reflector is modified as two plane reflectors were joined to form a corner with some angle, which reflects EM waves backtowards source. the reflector is known as corner reflector.

The angle between two plane reflectors were joined is called as included angle (or) corner angle (α). A corner reflector included with a driver element is called active corner reflector antenna at an angle of ($\alpha < 180°$) that produces a sharper radiation pattern than a flat sheet reflector ($\alpha = 180°$).

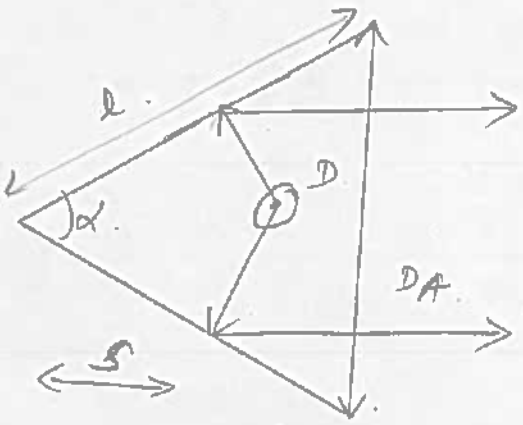
For most practical applications, $\alpha = 90°$ is used. A corner reflector without any driver element is called passive corner (or) retro (or) square corner reflector antenna.

For example, if the reflector is used as a passive target for radar (or) communication applications, it will return the signal exactly in the same direction as it is received when its included angle is 90°.

The system efficiency depends on the spacing between the vertex

of the corner reflector and the feed element is 's'. If 'α' decreases 's' must be increased to active desired efficiency.

If the spacing 's' is small, then the radiation resistance becomes small & hence efficiency is reduced. the corner angle of $\alpha = \pi$ radian (or) 180° which is equivalent to a flat sheet reflector.



D → Driven element
D_A → Aperture size
l → length
S → spacing between reflector and feed point location
α — corner angle.

Fig 2.6.7 Active corner reflector

Design equations of corner reflector :-

1) the aperture of the corner reflector $D\lambda$ is selected between one and two wavelength ($\lambda < D < 2\lambda$)

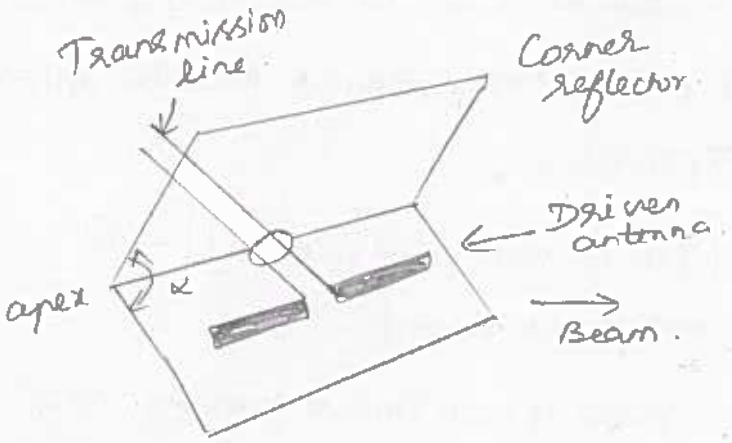
2) The spacing (s) between the vertex of the reflector & the feed element is selected as a fraction of wavelength. ($\lambda/4 < S < \frac{2\lambda}{3}$)

3) the length of the reflectors is approximately given as $l = 4s$ for $\alpha = 90°$ l $\alpha < 90°$ then $l > 2s$.

4) the height of the reflector (h) is generally selected as about 1.2 to 1.5 times greater than the total length of the feed element.

5) the radiation resistance is the function of 's'. If 's' is too large, the unwanted multiple lobes are produced & Hence the directivity of antenna is lost.

6) If 's' is very small, radiation resistance decreases.



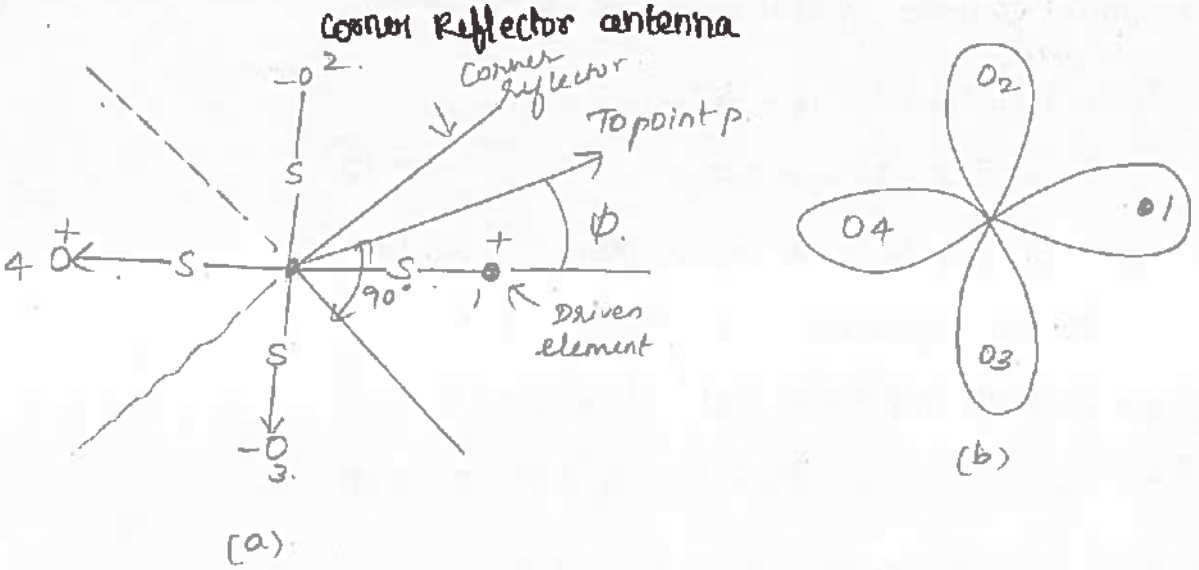


Fig. 2.6.8. Square corner Reflector with images used in analysis (a) 4-lobed pattern of driven element and images.

In general $D_A = \sqrt{l^2+l^2} = \sqrt{2l^2} = 1.414\,l$

but $\boxed{l = 2.5}$ ⟶ (1)

$\boxed{D_A = 2.828\,s}$ ⟶ (2)

The above equations are design equations of the corner reflector.

Consider a square corner reflector with corner angle $\alpha = 90°$ which consists of one driven element D. Based on one driven element, their corresponding images are represented by points (2) (3) & (4). the driven element along three images can carry current of same magnitude. the phase of current in 1 & 4 are

same as phase of currents in 3 & 2 but 180° out of phase.

when point 'p' is at large distance R from the antenna, then electric field intensity is expressed as,

$$E(\theta) = 2kI_1 \left| [\cos(s_r \cos\theta) - \cos(s_r \sin\theta)] \right| \longrightarrow (3)$$

where I_1 = current in each element

$s_r = \frac{2\pi s}{\lambda}$ → spacing of each element from the corner

k = propagation constant.

the terminal voltage V_1 at the centre of the driver element is given by

$$V_1 = I_1 Z_{11} + I_1 Z_{12} + I_1 Z_{14} - 2I_1 Z_{12}$$

$$= I_1 (Z_{11} + Z_{12} + Z_{14} - 2Z_{12}) \longrightarrow (4)$$

where Z_{11} – Self impedance of driver element (1/2 dipole)

Z_{12} – mutual impedance bet' element 1 & 2

Z_{14} – mutual impedance bet' element 1 & 4

Z_{12} – equation loss impedance of driver element.

interms of R, $R = \frac{V_1}{I_1} = R_{11} + R_{12} + R_{14} - 2R_{12}$

$$I_1 = \sqrt{\frac{P}{R_{11} + R_{12} + R_{14} - 2R_{12}}}$$

Sub I_1 value in eqn (3),

$$E(\theta) = 2k \sqrt{\frac{P}{R_{11} + R_{12} + R_{14} - 2R_{12}}} \left| [\cos(s_r \cos\theta) - \cos(s_r \sin\theta)] \right|$$

Field intensity at point p at a distance D from 1/2 dipole with reflector removed is $E_{HW}(\phi) = k \sqrt{\frac{P}{R_{11} + R_{12}}}$

$$\text{Gain } G_f(\phi) = \frac{E_\phi}{E_{HW}(\phi)} = 2 \sqrt{\frac{E_{11} + R_{12}}{E_{11} + R_{12} + R_{14} - R_{12}}} \left| \cos s_r(\cos\phi) - \cos(s_r \sin\phi) \right|$$

3) parabolic reflector

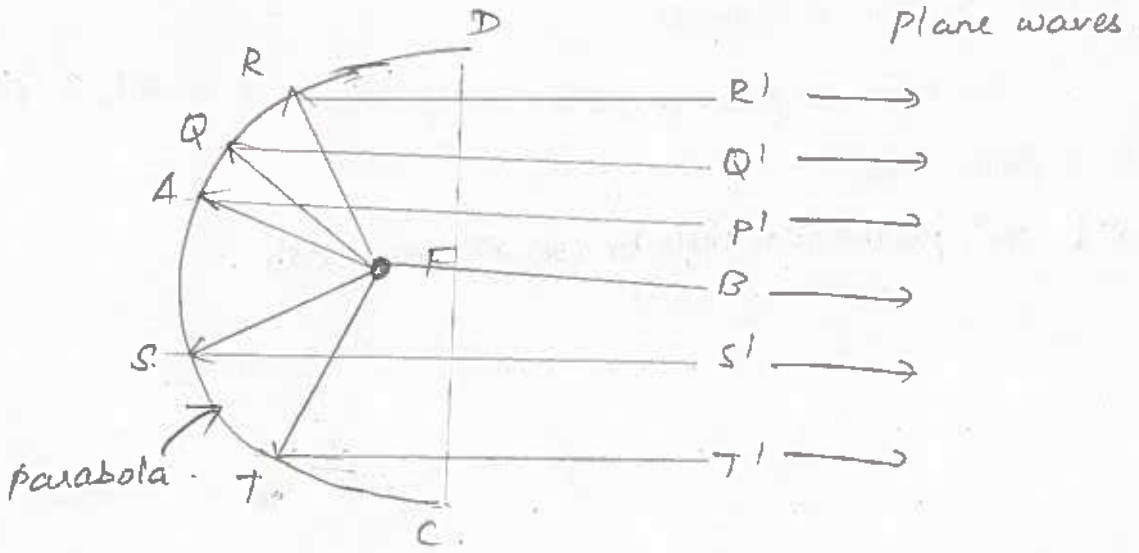


Fig 2.6.9 parabolic reflector.

The parabolic structure is used to improve the overall radiation characteristics such as antenna pattern, antenna efficiency, polarization etc, of the reflector antenna, the geometry of parabolic reflector in transmitting mode as shown in figure 2.6.9.

Here A_B ⟹ axis of parabole

f → focus

C_D → Direction

C_D – moth diameter D_A

A – Vertex

CAD – parabola

From the definition of a parabola, we have

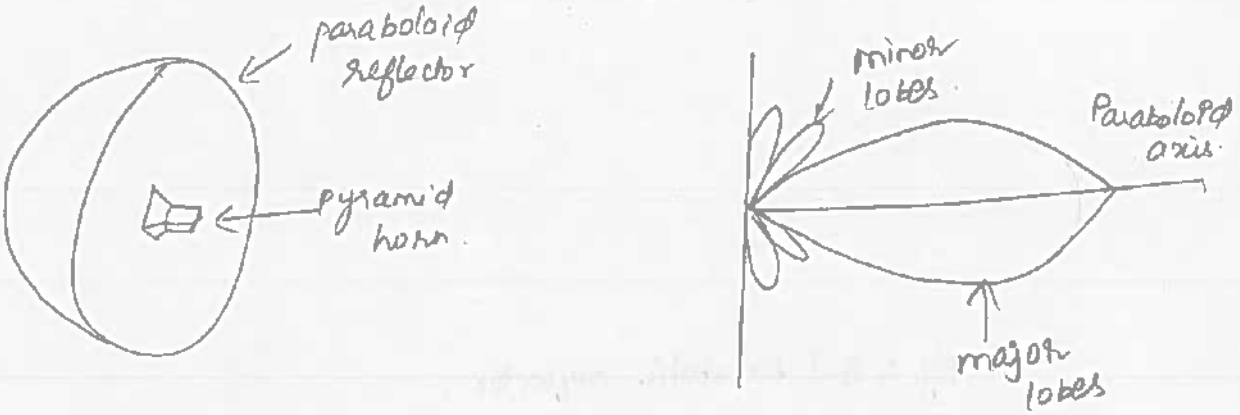
$$FP + PP' = FQ + QQ' = fS + SS' = \text{constant } (k)$$

Operation:- If a feed antenna is placed at the focus, all the waves are incident on the reflector and they are reflected track., forming a plane wavefront. By the time the reflected waves reach the directrix, all of them will be in phase, irrespective of the point on the parabola from which they are reflected. Hence the radiation is very high & is concentrated along the

axis of the parabola. same time, waves will be cancelled in other directions as result of path & phase differences.

The main purpose of parabolic reflector is to convert a spherical wave into a plane wave.

2) paraboloid (or) paraboloidal reflector (or) microwave dish.



(a) paraboloid (b) Radiation pattern.

Fig : 2.6.10 paraboloid with pyramided form as feed.

the paraboloid is also called as Microwave dish which produces sharp major lobe & smaller minor lobes.

If an isotropic source is placed at the focus point (F) of a parabolic reflector, then the radiation pattern B of source is intercepted & it is reflected as a plane wave of circular cross section.

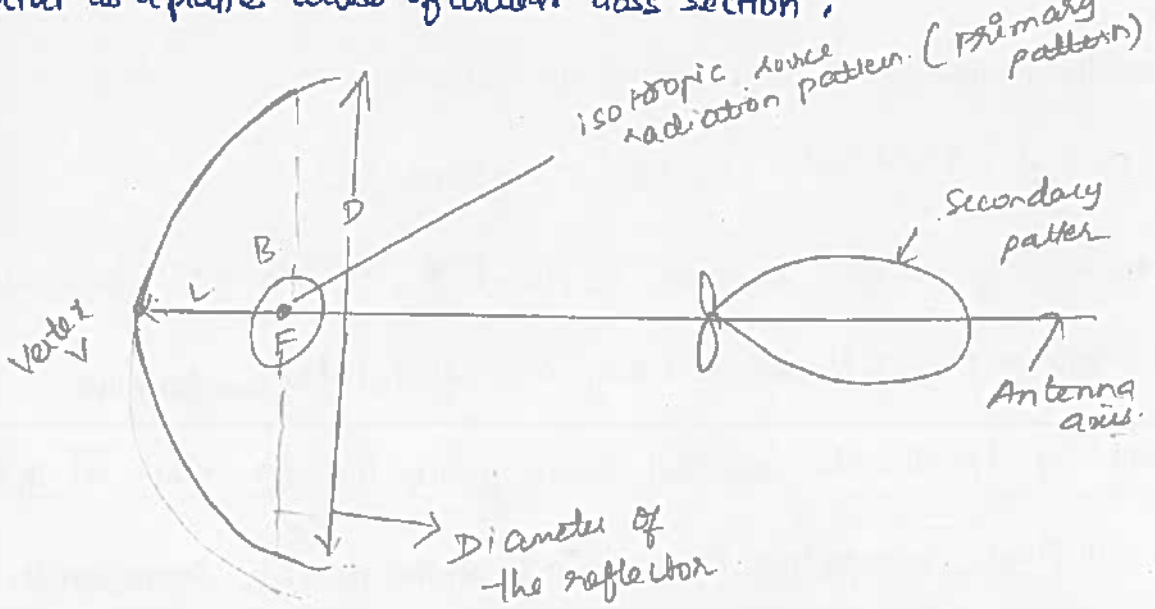


Fig 2.6.11 paraboloid & its radiation pattern.

The paraboloid and its radiation pattern is as shown in figure 26.11 : the radiation patter consists of very sharp major lobes & smaller minor lobes.

The distance bet' the focus (f) & vertex (v) is 'L' of the paraboloic is an even number of $\lambda/4$; then direct radiation in the axial direction from the source will be in opposite phase & will tend to cancel the central region of reflected wave.

$L = n\lambda/4$ where $n = 2, 4, 6$.

If n is odd ($n = 1, 3, 5$) . then the direct radiation in the axial direction from the source will be in the same phase & will tend to support the axial region of the reflected wave. When focal radio is small we will get the tapered field distribution (or) illumination . to get more uniform aperture field distribution ; it is necessary to make angle θ small by increasing the focal length 'L' which keeping the reflector diameter D constant & hence the focal ratio will be larger.

Patterns of large circular apertures with uniform illumination

According to Huggen's principle ; the normalized field pattern $E(\theta)$ as a function of θ & D is expressed as,

$$E(\theta) = \frac{2J_1}{\pi D} \frac{J_1\left[\left(\frac{\pi D}{\lambda}\right)\sin\theta\right]}{\sin\theta}$$

where , D → Diameter of aperture in m

θ → angle wrt the normal to the aperture

J_1 → first order Bessal function.

a) for circular aperture :-

$$BWFN = \alpha\theta_0 = \frac{140}{D\lambda} \text{ degrees.}$$

$$HPBW = \frac{58}{D\lambda}$$

The directivity (D) of the uniformly illuminated aperture is

given by $D = \frac{4\pi A_2}{\lambda^2}$

for circular Aperture $A_2 = \frac{\pi D^2}{4}$

$\therefore D = \frac{4\pi \frac{\pi D^2}{4}}{\lambda^2} = \pi^2 \left(\frac{D}{\lambda}\right)^2 \simeq 9.9 D\lambda^2$

Power gain $G_p = \frac{4\pi A_0}{\lambda^2}$

A_0 = capture area = k.A

K - constant dependent on feed antenna used.

$\therefore G_p = \frac{4\pi kA}{\lambda^2} = \frac{4\pi \times 0.65 \times A}{\lambda^2}$

for circular aperture $A = \frac{\pi D^2}{4}$ & $K = 0.65$ for dipole

$G_p = \frac{4\pi \times 0.65}{\lambda^2} \frac{\pi D^2}{4}$

$$\boxed{G_p = 6\pi \left(\frac{D}{\lambda}\right)^2 = 6 D_\lambda^2}$$

b) Rectangular aperture :-

$BWFN = \frac{115\lambda}{L} = \frac{115}{L_\lambda}$ degree , $L_\lambda = \frac{L}{\lambda}$

L - length of rectangular aperture in terms of λ, m

c) square aperture :-

$D = 4\pi \frac{L^2}{\lambda^2} = 12.6 L_\lambda^2$

and power gain over a $\lambda/2$ dipole is given as,

$G_p = 7.7 L_\lambda^2$

length of side in term of $\lambda = L_\lambda = \frac{L}{\lambda}$.

2.7 Feeding systems (or) structures :-

WKT parabolic reflector antenna consists of two basic parts namely (i) A source of radiation placed at the focus called primary radiator (or) feed. (ii) the reflector called secondary radiator.

The feed is said to be ideal feed, if it radiates entire energy towards the reflector. therefore, the entire surface of the reflector is illuminated & no energy is radiated in any unwanted direction.

practically there are number of possible feeds to the Parabolic reflector antenna. the secondary radiator used is a parabolic most of the times. the various feeds used as :-

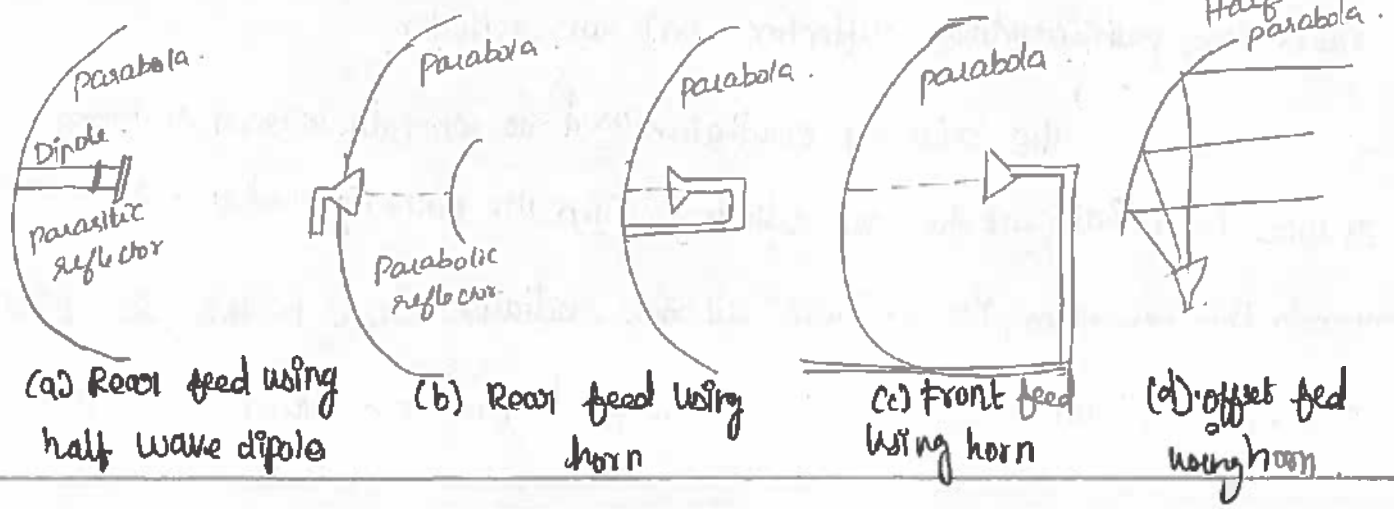
(i) Dipole antenna.
(ii) Horn antenna.
(iii) End fire antenna
(iv) casegrain feed
(v) offset feed.

(i) **Dipole feed :-** The simplest type of the feed.

It is not suitable feed for the parabolic reflector antenna

Instead of only dipole, a feed consisting dipole with parasitic reflectors (yagi-uda) can be used as a feed system.

The spacing bet' the dipole as a driver element & parabolic parastic reflector is 0.125 λ & for plane reflector 0.4 λ

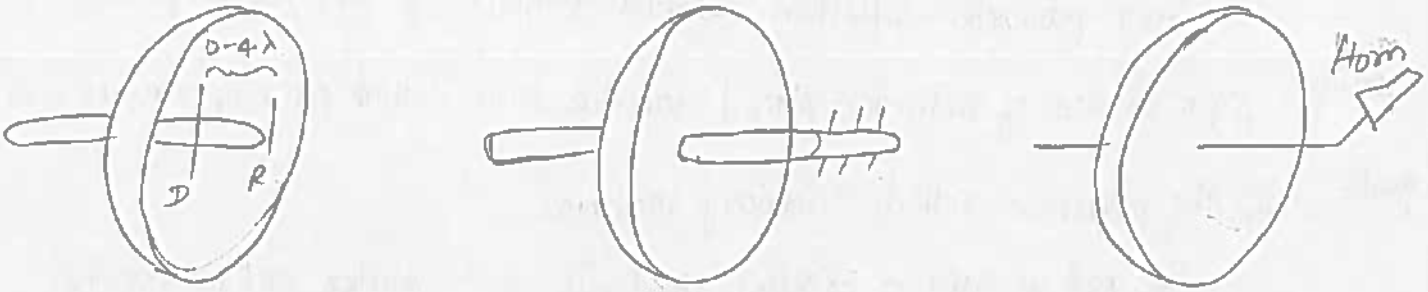


(a) Rear feed using half wave dipole

(b) Rear feed using horn

(c) Front feed using horn

(d) offset fed using horn



a) Dipole with plane reflector.

b) End fire array dipole

c) Horn with waveguide.

2.7.1 Different types of feed system.

2) End fire feed :- In some cases, an end fire array is used as feed radiator.

3) Horn Feed :- the most widely used feed system in the parabolic reflector antenna is waveguide turn antenna the horn antenna is fed with a waveguide. In case, if circular polarization is required, then the rectangular horn is replaced by a conical horn.

In all cases, the feed (or) primary radiator is placed at the focus to obtain maximum horn pattern.

4) cassegrain feed

the system of feeding paraboloid reflector is named after a mathematician prof. cassegrain. In this system, the feed is placed at the vertex of the parabolic reflector instead of placing it at the focus.

this system uses a hyperboloid reflector placed whose of the foci coincides with the focus of the parabolic reflector, this hyperboloid reflector is called cassegrain secondary reflector (or) sub-reflector.

the primary radiator used is generally a horn antenna, It aims to radiate at the sub-reflector, when the primary radiator radiates towards the cassegrain, it radiates all the radiations. Due to this, the parabolic reflector gets illuminated similar to the feed radiator placed at the focus.

then the parabolic reflector collimates all the radiations as previous feed system.

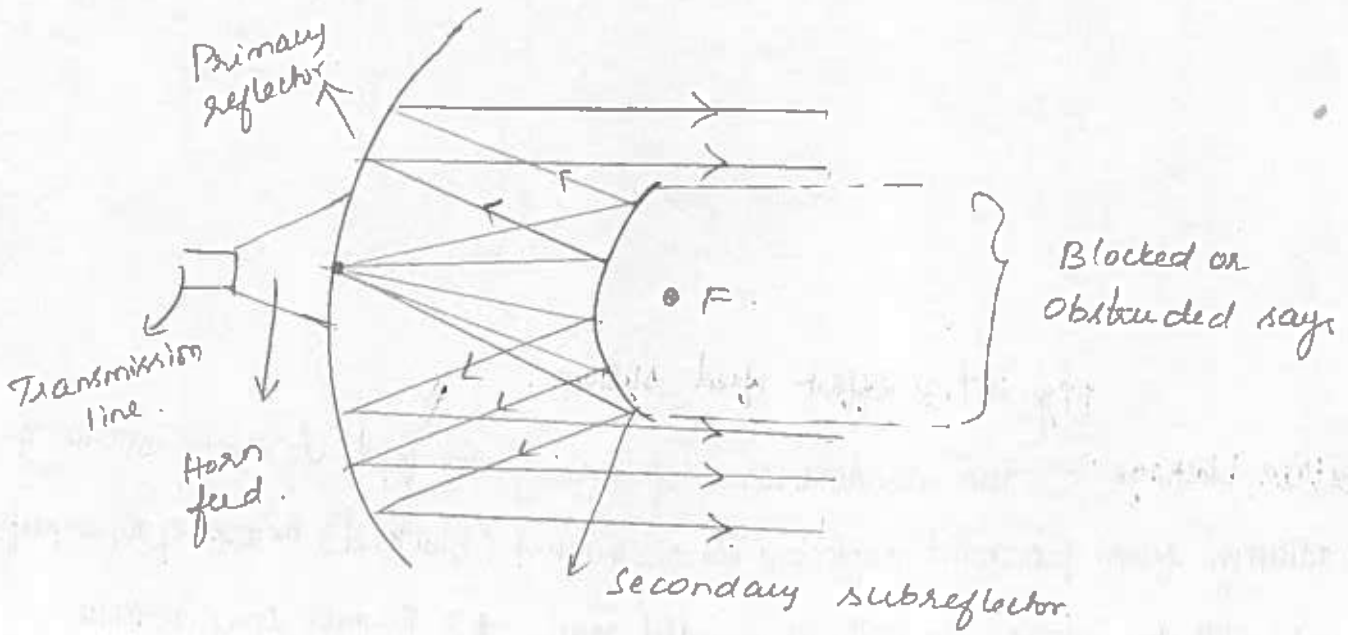


fig 2.7.2 Cassegrain feed system.

Advantages of cassegrain feed system:-

1) It reduces the spill over & minor lobe radiations.

2) the system has ability to place a feed at convenient place.

3) using this system, beam can be broadcast by adjusting one of the reflector surfaces.

offset feed system,

Due to aperture blockage effect, the mirror lobes are increases. Here the feed radiator is placed at the focus. With this system, all the rays are perfectly collimated without formation of the region of blocked rays.

5) Offset feed system:-

Due to aperture blockage effect, the minor lobes increase. Here the feed radiator is placed at the focus as shown in figure 2.7.3. With this system all the rays are perfectly collimated without formation of the region of blocked rays.

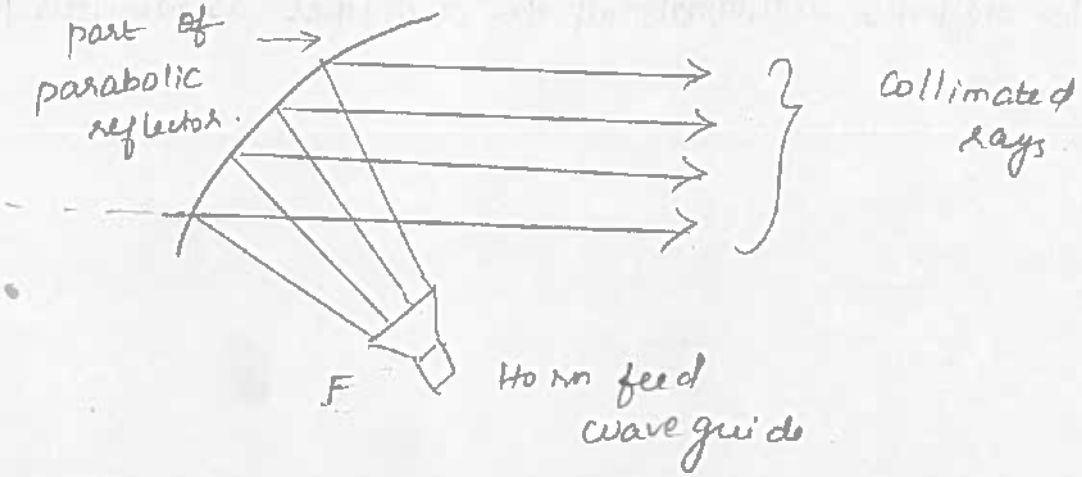


Fig: 2.7.3 offset feed system.

Aperture blockage:- the disadvantage of cassegrain feed is that some of the radiation from parabolid reflector is obstructed (blocked) because of the presence of sub reflector along the path of parallel rays. this is not very serious problem in case of a parabolic reflector of larger dimensions. But for smaller dimension parabolic reflector, it is the main drawback of the cassegrain feed system.

the aperture blockage effect can be avoided by using an offset reflector which is applicable to focal point feed.

Some of the radiations from the parabolic reflector are blocked by the hyperbolic reflector creating region of blocked rays.

Applications of parabolic reflector:-

It is used in 1) microwave communication

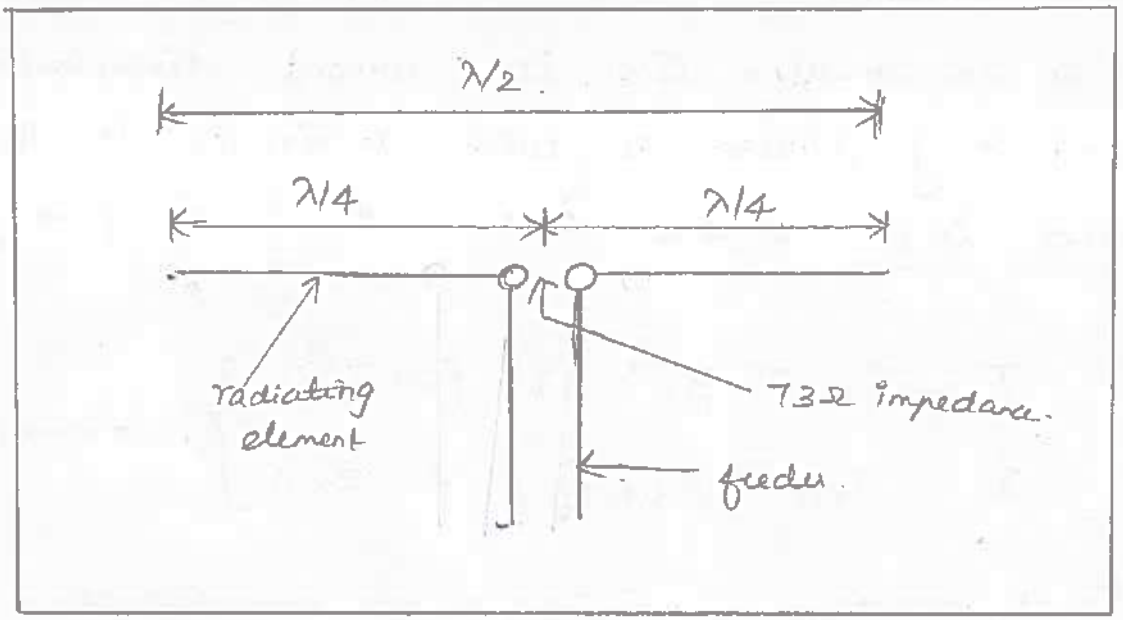
2) Radio astronomy

3) satellite transmission and reception.

HALF WAVE DIPOLE:

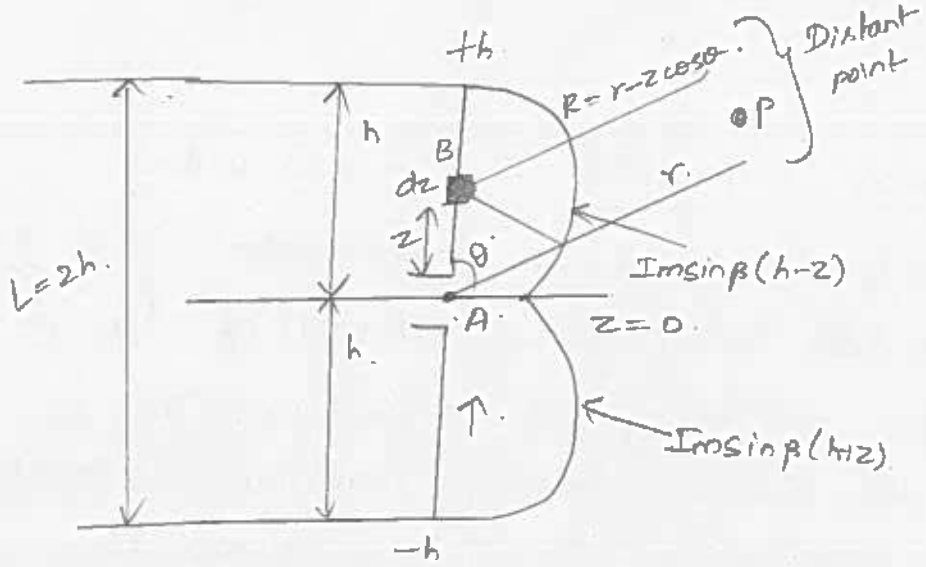
Definitions:-

Half wave dipole is the fundamental radio antenna which is made up of a metal rod or thin wire and it has a physical length of $\lambda/2$ in free space at the frequency of operation. This is usually fed at the centre having maximum current, that is, maximum radiation is in the plane normal to the axis.



Power radiation from Half Wave dipole:

It is a symmetrical antenna in which the two ends are at an equal potential relative to the midpoint. The overall specified length is $L=2h$ and Vertical antenna height is $h=\frac{L}{2}$.



In $\triangle ABC$, $AC = r - R$

$$\frac{r-R}{z} = \cos\theta$$

$$r - R = z\cos\theta$$

$$\boxed{R = r - z\cos\theta}$$

This antenna is fed at the centre with the help of a transmission line, its current distribution is approximately sinusoidal with maximum at the centre and zero at the ends and it is given by,

$$\left.\begin{aligned} I &= I_m \sin\beta(h-z) \quad \text{for } z>0 \\ I &= I_m \sin\beta(h+z) \quad \text{for } z<0 \end{aligned}\right\} \longrightarrow (1)$$

where I_m = current maximum at the current element $I\,dz$,

$\beta = \frac{2\pi}{\lambda}$ = phase constant and

$I\,dz$ - current element placed at a distance 'z' from $z=0$ plane.

Consider a point 'P' located at a far distance from the current element. Then the Vector potential at point P due to current element $I\,dz$ is given by,

$$dA_z = \frac{\mu}{4\pi R} I\,dz\, e^{-j\beta R} \longrightarrow (2).$$

Where R = Distance between Idz to distant point P.

By integrating 2nd equation, we get over the total of the antenna and it is given by,

$$\int dA_z = \int_{-h}^{0} \frac{\mu I dz e^{-j\beta R}}{4\pi R} + \int_{0}^{h} \frac{\mu I dz e^{-j\beta R}}{4\pi R}$$

$$A_z = \frac{\mu}{4\pi} \int_{-h}^{0} \frac{I_m \sin\beta(h+z) e^{-j\beta R}}{R} dz + \frac{\mu}{4\pi} \int_{0}^{+h} \frac{I_m \sin\beta(h-z) e^{-j\beta R}}{R} dz \quad \rightarrow (3)$$

Since the point 'P' is at a large distance, the lines to the point 'P' may be assumed to be parallel. We know that,

$$R = r - z\cos\theta.$$

R=r, When 'p' is at a large distance and replace R in the denominators of equation (3) only by r. But, in numerator, 'R' represents the phase factor and therefore the difference between R and r is very important.

$$A_z = \frac{\mu}{4\pi} \int_{-h}^{0} \frac{I_m \sin\beta(h+z)}{r} e^{-j\beta(r-z\cos\theta)} dz +$$

$$\frac{\mu}{4\pi} \int_{0}^{h} \frac{I_m \sin\beta(h-z)}{r} e^{-j\beta(r-z\cos\theta)} dz.$$

$$A_z = \frac{\mu I_m e^{-j\beta r}}{4\pi r} \left[\int_{-h}^{0} \sin\beta(h+z) e^{j\beta z\cos\theta} dz + \int_{0}^{h} \sin\beta(h-z) e^{j\beta z\cos\theta} dz \right]$$

For a $\lambda/2$ antenna, $L = 2h = \lambda/2$ & $z = h = \lambda/4 = \pi/2$ degrees

where

$$\sin\beta(h+z) = \sin\beta\left(\frac{\pi}{2}+z\right) = \cos\beta z$$

$$\sin\beta(h-z) = \sin\beta\left(\frac{\pi}{2}-z\right) = \cos\beta z$$

$$A_z = \frac{\mu I_m e^{-j\beta r}}{4\pi r}\left[\int_{-h}^{0} \cos\beta z\, e^{j\beta z\cos\theta}\, dz + \int_{0}^{h} \cos\beta z\, e^{+j\beta z\cos\theta}\, dz\right]$$

Now $\int_{-h}^{0} e^{j\theta}\, d\theta = \int_{0}^{h} e^{-j\theta}\, d\theta$. By using this property, change limits of integration of first term.

$$A_z = \frac{\mu I_m e^{-j\beta r}}{4\pi r}\left[\int_{0}^{h} \cos\beta z\, e^{-j\beta z\cos\theta}\, dz + \int_{0}^{h} \cos\beta z\, e^{j\beta z\cos\theta}\, dz\right]$$

$$A_z = \frac{\mu I_m e^{j\beta r}}{4\pi r}\left[\int_{0}^{h} \cos\beta z\left(e^{-j\beta z\cos\theta} + e^{j\beta z\cos\theta}\right) dz \times \frac{2}{2}\right]$$

WKT, $\dfrac{e^{-j\beta z\cos\theta} + e^{j\beta z\cos\theta}}{2} = \cos(\beta z\cos\theta)$

$$A_z = \frac{\mu I_m e^{-j\beta r}}{4\pi r}\int_{0}^{h} \cos\beta z\; 2\cos(\beta z\cos\theta)\, dz \longrightarrow (4).$$

$$2\cos\alpha\cos\beta = \cos(\alpha-\beta) + \cos(\alpha+\beta)$$

$$A_z = \frac{\mu I_m e^{-j\beta r}}{4\pi r}\int_{0}^{h} \cos\{\beta z(1+\cos\theta)\} + \cos\{\beta z(1-\cos\theta)\}\, dz$$

$\longrightarrow (5)$

Integrating eqn (5) & $h = z = \lambda/4$

$$= \frac{\mu I m e^{-j\beta r}}{4\pi r}\left[\frac{\sin(\beta z(1+\cos\theta))}{\beta(1+\cos\theta)} + \frac{\sin\beta z(1-\cos\theta)}{\beta(1-\cos\theta)}\right]_0^{\lambda/4}$$

$$= \frac{\mu I m e^{-j\beta r}}{4\pi\beta r}\left[\frac{(1-\cos\theta)\sin\beta z(1+\cos\theta) + (1+\cos\theta)\sin\beta z(1-\cos\theta)}{1-\cos^2\theta}\right]_0^{\lambda/4}$$

$$\left[\because \beta = \frac{2\pi}{\lambda}\cdot\frac{\lambda}{4} = \frac{\pi}{2}\right]$$

$$= \frac{\mu I m e^{-j\beta r}}{4\pi\beta r}\left[\frac{(1-\cos\theta)\sin(\pi/2+\pi/2\cos\theta) + (1+\cos\theta)\sin\beta z(1-\cos\theta)}{1-\cos^2\theta}\right]_0^{\pi/2}$$

$$= \frac{\mu I m e^{-j\beta r}}{4\pi\beta r}\left[\frac{(1-\cos\theta)\cos(\pi/2\cos\theta) + (1+\cos\theta)\cos(\pi/2\cos\theta)}{\sin^2\theta}\right]$$

$$A_z = \frac{\mu I m e^{-j\beta r}}{4\pi\beta r}\left[\frac{\cos(\pi/2\cos\theta)\left[1-\cos\theta+1+\cos\theta\right]}{\sin^2\theta}\right]$$

$$A_z = \frac{\mu I m e^{-j\beta r}}{2\pi\beta r}\left[\frac{\cos(\pi/2\cos\theta)}{\sin^2\theta}\right] \longrightarrow (6)$$

The next step is to find the magnetic field using Maxwell's equation for spherical co-ordinate system. The ϕ components of H is given by,

$$H_\phi = \frac{1}{\mu}(\nabla \times A)\phi = \frac{1}{\mu}\times\frac{1}{r}\left[\frac{\partial}{\partial r}(rA_\theta) - \frac{\partial}{\partial\theta}(Ar)\right] \longrightarrow (7)$$

But now the current element is placed along z-axis, then $A_r = -A_z\sin\theta$ and $A_\theta = 0$ and by substitute in eqn (7) we get

$$H_\phi = \frac{1}{\mu} \times \frac{1}{r}\left[\frac{\partial}{\partial r}(rA_z\cos\theta)\right] \longrightarrow (8)$$

By substituting, A_z of eqn (6) in eqn (8)

$$= \frac{1}{\mu} \times \frac{1}{r}\left\{\frac{\partial}{\partial r}\left[\frac{-r\mu I_m e^{-j\beta r}}{2\pi\beta r}\left\{\frac{\cos(\pi/2\cos\theta)}{\sin^2\theta}\right\}\sin\theta\right]\right\}$$

$$= \frac{-I_m}{2\pi\beta r}\left[\frac{\cos\pi/2\cos\theta}{\sin\theta}\right]\frac{\partial}{\partial r}\left[e^{-j\beta r}\right]$$

$$= \frac{-I_m}{2\pi\beta r}\left[\frac{\cos(\pi/2\cos\theta)}{\sin\theta}\right] e^{-j\beta r}(-j\beta)$$

$$H_\phi = \frac{j I_m e^{-j\beta r}}{2\pi\beta r}\left[\frac{\cos(\pi/2\cos\theta)}{\sin\theta}\right]$$

The magnitude of the magnetic field strength or magnetic field intensity for the radiation field of a half wave dipole is given by,

$$|H_\phi| = \frac{I_m}{2\pi r}\left\{\frac{\cos(\pi/2\cos\theta)}{\sin\theta}\right\} A/m \longrightarrow (9)$$

The electric field expression for the radiation field can be achieved from

$$\frac{E_\theta}{H_\phi} = \eta = 120\pi$$

$$|E_\theta| = 120\pi |H_\phi| \longrightarrow (10)$$

By substituting H_ϕ from eqn (9) in eqn (10), we get

$$|E_\theta| = 120\pi \times \frac{I_m}{2\pi r}\left\{\frac{\cos(\pi/2\cos\theta)}{\sin\theta}\right\}.$$

$$|E_\theta| = \frac{60 I_m}{r}\left\{\frac{\cos(\pi/2\cos\theta)}{\sin\theta}\right\} V/m \longrightarrow (11)$$

This is the amplitude of electric field intensity of radiation field of a $\lambda/2$ antenna (or) a $\lambda/4$ antenna.

Power radiated by a half wave dipole and its radiation resistance.

The product of magnitude values of E_θ and H_ϕ and it is given as

$$P_{max} = |E_\theta||H_\phi|$$

$$= \left[\frac{60 I_m}{r}\left\{\frac{\cos(\pi/2\cos\theta)}{\sin\theta}\right\}\right]\left[\frac{I_m}{2\pi r}\left[\frac{\cos\left(\frac{\pi}{2}\cos\theta\right)}{\sin\theta}\right]\right]$$

$$P_{max} = \frac{30 I_m^2}{\pi r^2}\left[\frac{\cos(\pi/2\cos\theta)}{\sin\theta}\right]^2 \longrightarrow (12)$$

Average power interms of effective or R.M.S current:

The average value of the power is half of the maximum power and it is expressed as,

$$P_{avg} = \frac{E_0}{\sqrt{2}} \cdot \frac{H_\phi}{\sqrt{2}} = \frac{1}{2} E_0 \cdot H_\phi = \frac{P_{max}}{2}$$

$$P_{avg} = \frac{15 I_m^2}{\pi r^2}\left[\frac{\cos\left(\frac{\pi}{2}\cos\theta\right)}{\sin\theta}\right]^2 W/m^2 \longrightarrow (13)$$

The effective or RMS current is related to the maximum current by the relation is given by,

$$I_{r.m.s} = \frac{I_m}{\sqrt{2}} \Rightarrow \frac{I_m}{1} = \sqrt{2}\, I_{r.m.s}$$

$$= 15\frac{(\sqrt{2}\, I_{rms})^2}{\pi r^2}\left\{\frac{\cos(\pi/2\cos\theta)}{\sin\theta}\right\}^2$$

$$\boxed{P_{avg} = \frac{30 I_{rms}^2}{\pi r^2}\left[\frac{\cos^2(\pi/2\cos\theta)}{\sin^2\theta}\right] W/m^2} \longrightarrow (14)$$

This is expression for average power interms of RMS current. This total radiated power is given by the surface integral of poynting vector over any surrounding surface,

ds = elemental area of spherical shell $= 2\pi r^2 \sin\theta\, d\theta$

Power radiated $(P_r) = \oint P_{avg} \cdot ds \longrightarrow (15)$

By substituting (14) in (15), we get

$$= \int_0^{\pi} \frac{30 I_{rms}^2}{\pi r^2}\left\{\frac{\cos^2(\pi/2\cos\theta)}{\sin^2\theta}\right\} 2\pi r^2 \sin\theta\, d\theta$$

$$= 60 I_{rms}^2 \int_0^{\pi} \frac{\cos^2(\pi/2\cos\theta)}{\sin\theta} d\theta$$

$$= 60 I_{rms}^2 \int_0^{\pi} \frac{1}{2}\left\{\frac{1+\cos(\pi\cos\theta)}{\sin\theta}\right\} d\theta$$

$$= 60 I_{rms}^2 \int_0^{\pi} \frac{1}{2}\left\{\frac{1+\cos(\pi\cos\theta)}{\sin\theta}\right\} d\theta$$

$$= 60 I_{rms}^2 \cdot I$$

Where $I = \frac{1}{2}\int_0^{\pi}\left\{\frac{1+\cos(\pi\cos\theta)}{\sin\theta}\right\} d\theta$. $\longrightarrow (16)$

The value of 'I' after integration is

$$I = 1.219$$

$$\text{Power radiated} = 60 I_{rms}^2 \times 1.219$$

$$P_r = 73.140 I_{rms}^2 \longrightarrow (17)$$

The equation (17) is an expression for the total power radiated by a half wave dipole in free space.

Radiation Resistance:

$$W_r = R_r \cdot I_{rms}^2 \longrightarrow (18)$$

By comparing equation (17) with equation (18), we get

$$R_r = 73.14 \simeq 73\Omega$$

The radiation resistance of a centre fed half wave dipole or simply dipole antenna is 73.14Ω (or) approximately 73Ω.

FREQUENCY INDEPENT ANTENNAS.

* A frequency - independent antenna is physically fixed size and operates on an instantaneous basis over a wide bandwidth (entire frequency band) with relatively constant impedance, pattern, polarization and gain.

* These antennas are broadband antennas which are using 10 to 10000 MHz region for practical applications such as TV, point to point communication feeds for reflectors and lenses.

Rumsey's Principle:

The condition of the frequency independent antenna was pointed out by V. H Rumsey. He stated that, "the performance that is, the impedance and pattern properties of a lossless antenna is independent of frequency if the dimensions of an antenna are specified in terms of angles such that they remain constant in terms of wavelength.

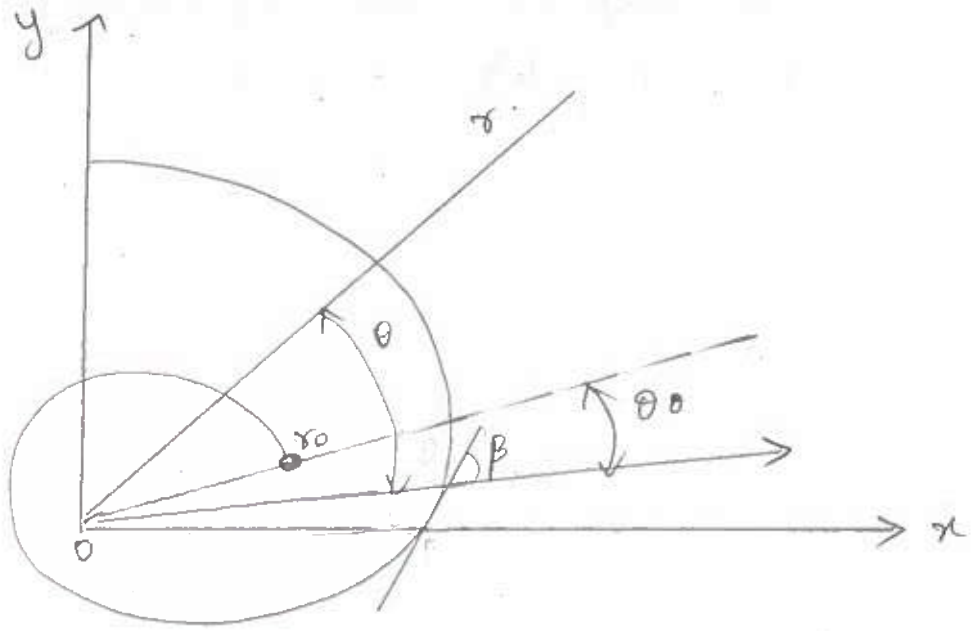
SPIRAL ANTENNA:-

* Spiral Antennas are a frequency independent antenna. It radiates a bidirectional main lobe perpendicular to the plane of the antenna. It produces circularly polarised waves within the band of operation and the radiation is elliptically polarized outside the band of operation.

* The surface of an equiangular spiral shape can be described completely by the angles. It fulfills all the necessary conditions that are employed to design the frequency independent antennas.

* When the total arm length is comparable with the wavelength, the frequency of an operation will be the lowest cut-off frequency and for all other frequencies above this, the pattern and impedance characteristics are frequency independent.

Planar Log-Spiral Antenna:-



* The equation of a logarithmic (or) log spiral is given by

$$r = a^{\theta} \longrightarrow (1a)$$

(or) $$\ln r = \theta \ln a \longrightarrow (1b)$$

where $r \rightarrow$ radial distance to point P on spiral
$\theta \rightarrow$ Angle with respect to x axis
$a \rightarrow$ constant.

* From equation (1a) the rate of change of radius with an angle is obtained as

$$\frac{dr}{d\theta} = a^{\theta} \ln a = r \ln a \longrightarrow (2)$$

* The constant 'a' in equation (2) is related to the angle β between the spiral and a radial line from the origin is given by,

$$\ln a = \frac{dr}{r d\theta} = \frac{1}{\tan\beta} \longrightarrow (3)$$

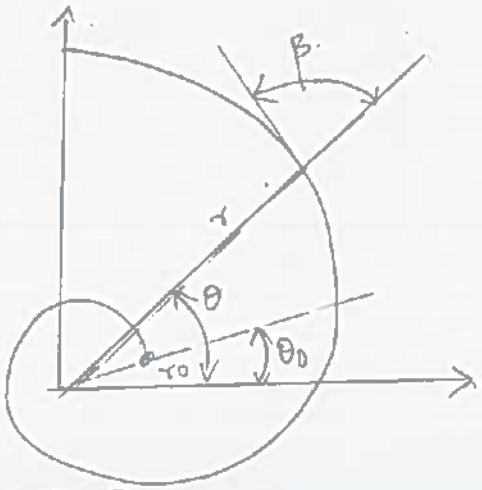
* From equation (1b),

$$\ln r = \theta \ln a$$

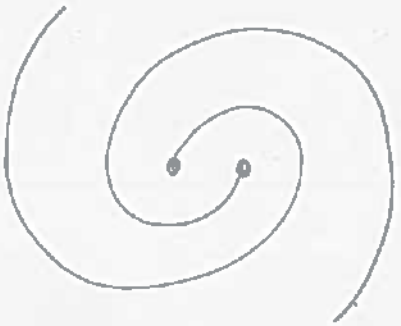
By substituting equation (3) in above expression we have

$$\ln r = \frac{\theta}{\tan\beta}.$$

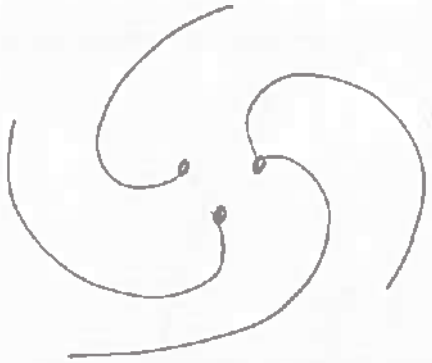
$$\theta = \tan\beta \ln r \longrightarrow (4)$$



(a) Single spiral.



(b) Two spiral ($\theta_0 = 0, \pi$)



(C) Multiple spiral
($\theta_0 = 0, \pi/2, \pi, 3\pi/2$)

(d) Multiple Spiral
($\theta_0 = 0, \pi/2, \pi, 3\pi/2$)

* The log spiral antenna in figure (a) is constructed so as to make $r = 1$ and $\theta = 0$ and $r = 2$ at $\theta = \pi$. These conditions determine the value of constant 'a' and 'β'. From equation (3) and (4) $\beta = 77.6°$ and $a = 1.247$.

* The shape of the spiral is determined by the angle β which is the same for all points on the spiral.

* For a second log spiral, which is identical in the form to the fig (a) and be generated by an angular rotation δ, so that equation (1) becomes,

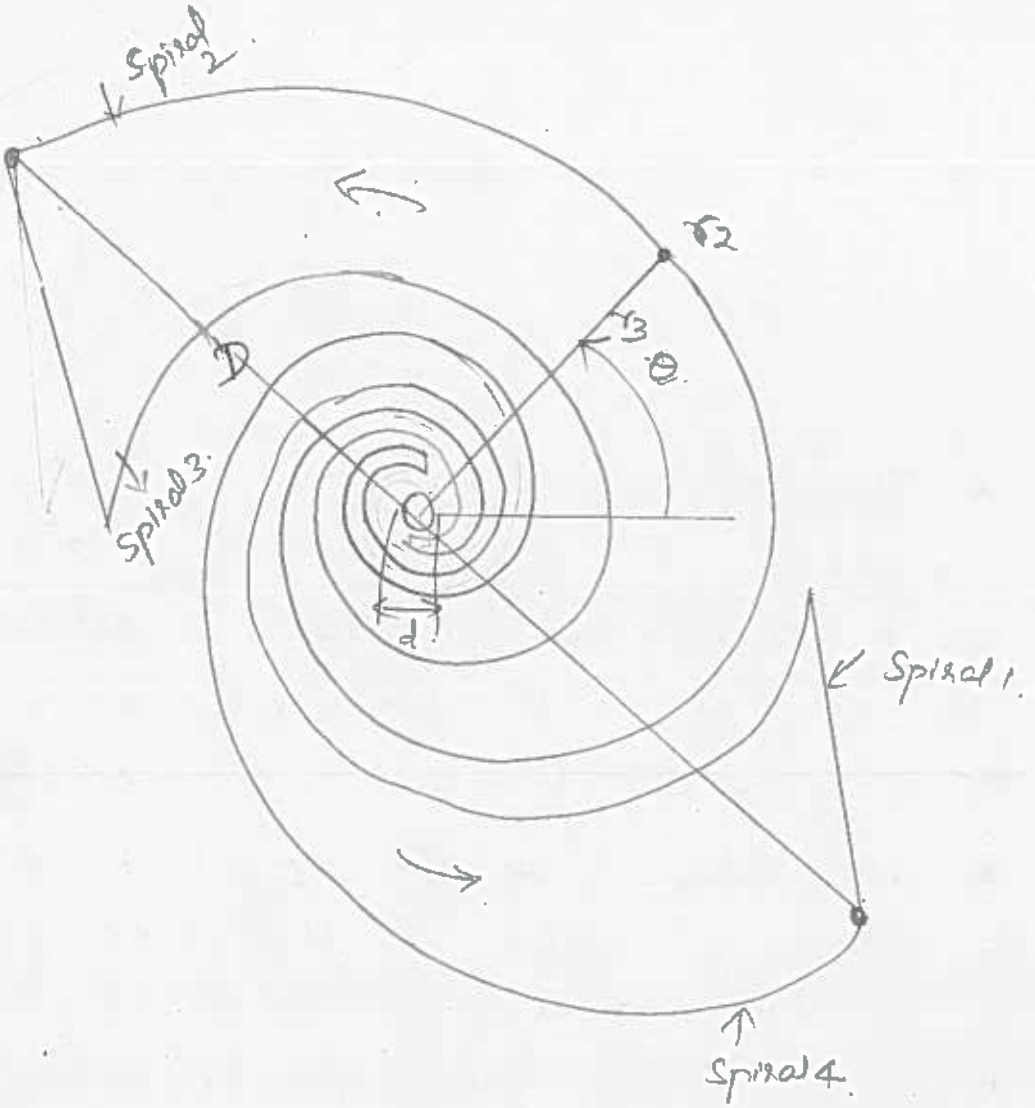
$$r_2 = a^{\theta - \delta} \longrightarrow (5a)$$

and a third and fourth spiral is given by,

$$r_3 = a^{\theta - \pi} \longrightarrow (5b)$$

and $$r_4 = a^{\theta - \pi - \delta} \longrightarrow (5c)$$

* Then for a rotation $\delta = \frac{\pi}{2}$ we have 4 spirals at 90° angles. Metalizing the areas between spirals 1 and 4 and 2 and 3, with the other areas open, self-complementary and the Congruence conditions are satisfied.

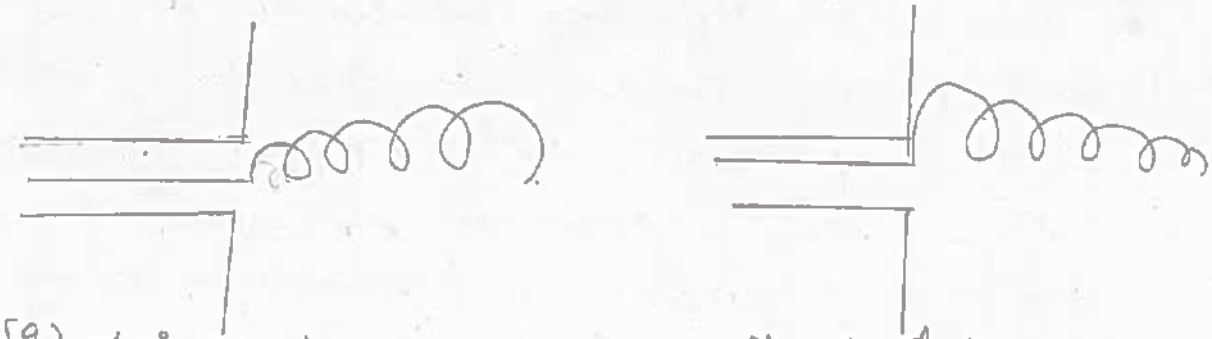


* From the above figure, the arrows indicate the direction of the outgoing waves travelling along the conductors resulting in right circularly polarized (RCP) radiation outward from the page and left-circularly polarized radiation in the pages.

* The high-frequency limit of operation is determined by the spacing 'd' of the input terminal and the low frequency limit by the overall diameter 'D'. The ratio D/d for the above figure is 25 to 1.

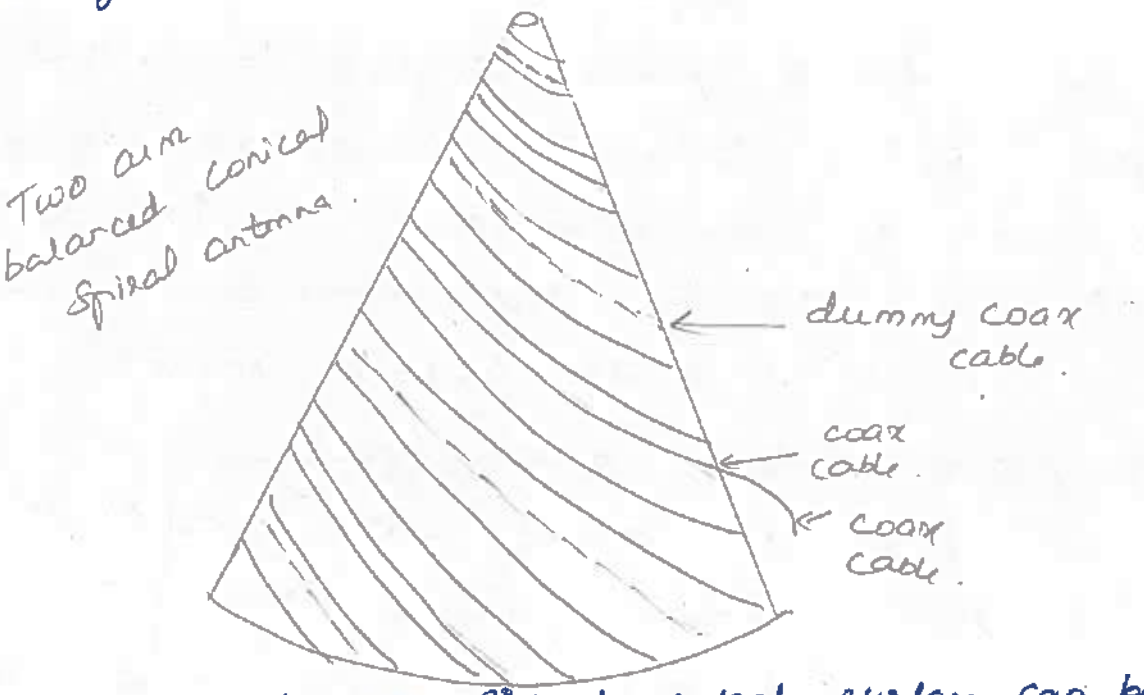
Conical - Spiral (CP) Antenna:

* A tapered helix is a conical-spiral antenna and these were described and investigated extensively in the years following 1947.



(a) α - increasing (b) α - decreasing

* The above figures shows tapered helical or conical spiral antennas in which the pitch angle is constant with diameter and turn spacing variable.



* The conducting signal spiral surface can be constructed conveniently using printed circuit technique, the conical ~~antenna~~ arms on dielectric cone which is also used as a support. The feed cable can be bounded to the metal arms which are wrapped around the cone as shown in the above figure.

* The conical equi-angular spiral antenna is fed at the apex by means of a balanced transmission line carried upinside the cone along the axis of the cone.

* The main difference between the conical spiral and planar antenna is that the conical spiral antenna provides unidirectional radiation in a single lobe towards the apex of the cone and with a maximum radiation along the axis.

* The ~~ant~~ conical antennas, the circular polarization and relatively constant impedance are preserved over large bandwidth.

* The input impedance is between 100 to 150 ohms for a pitch angle $\alpha = 17°$ and full angles 20° to 60°. The bandwidth depends on the ratio of base diameter to the truncated apex diameter and this ratio may be chosen arbitrarily larger such as 5:1 or more.

* The planar spiral antenna produces a bidirectional beam of about 50° to 60°. On the other hand, for a narrow angled cone, it produces an unidirectional beam of about 70° to 90° width.

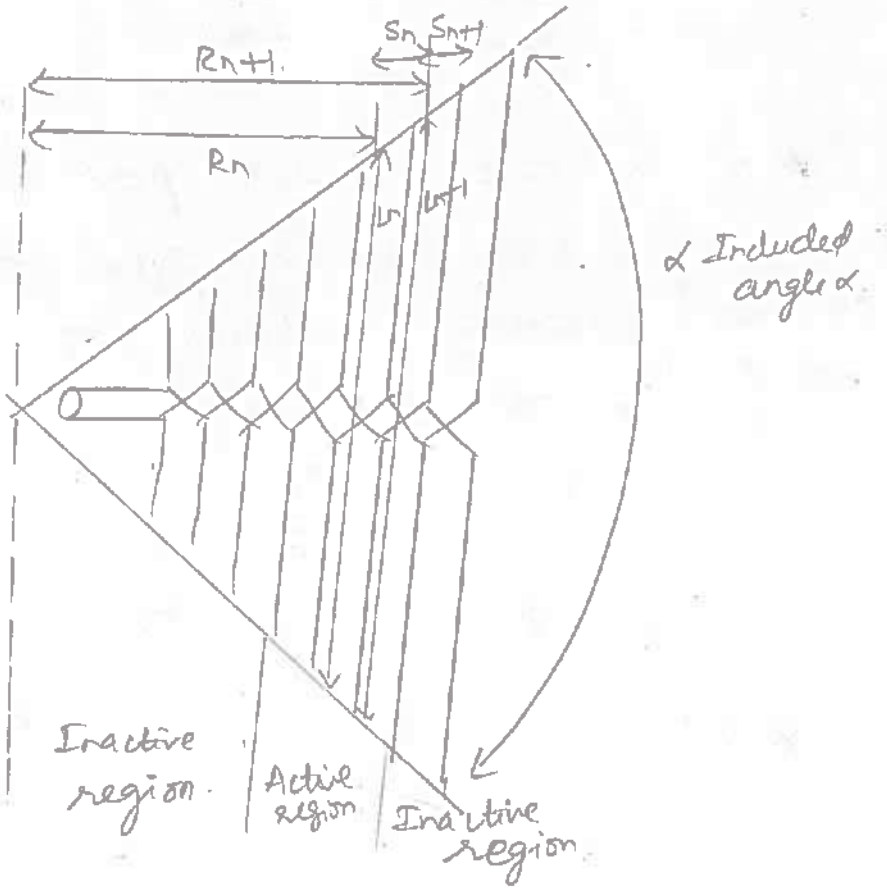
* Conical spirals can be used in conjunction with a ground plane, but with a reduction in the bandwidth when they are flush mounted on the plane.

LOG PERIODIC ANTENNA:-

* A log periodic antenna is a broadband narrow beam antenna. It is a frequency independent antenna.

* The geometry of an antenna structure is adjusted such that all the electrical properties of the antenna must repeat periodically with the logarithm of frequency. For every repetition, the structure size changes by a constant scale factor by which the structure can either be expanded or contracted. The log periodic principle can be understood with the help of the array of log periodic antenna known as log Periodic Dipole Array (LPDA).

Construction of LPDA:



* The LPDA consists of a number of dipoles of different lengths and spacings. The array is fed using a balanced transmission line which is connected at narrow end or apex of the array. Also the transmission line is transposed between each adjacent pairs of terminals of dipoles.

* The length of the dipoles increases from feed point towards other end such that the included angle α remains constant. The dipole lengths and the spacings between two adjacent dipoles are related through parameter called design ratio or scale factor denoted by τ. Thus the relationship between spacings 'S' and lengths 'L' of adjacent elements are scaled as

$$\frac{S_n}{S_{n+1}} = \frac{L_n}{L_{n+1}} = \tau \longrightarrow (1)$$

* τ is also called periodicity factor which is always less than 1. The above expression can be written in terms of constant K with the radii of the arm as

$$\frac{R_{n+1}}{R_n} = \frac{S_{n+1}}{S_n} = \frac{L_{n+1}}{L_n} = \frac{1}{\tau} = k; \quad k > 1 \longrightarrow (2)$$

where $n = 1, 2, 3, \ldots n$.

* The spacing factor (σ) is defined as

$$\sigma = \frac{S_n}{2L_n} \longrightarrow (3)$$

* The ends of the dipoles lie along the straight lines on both the sides. Thes two straight line meet at feed point or apex having an angle α which is angle included by two straight line. (Typical value of $\alpha = 30°$ and $\tau = 0.7$).

Working Principle of LPDA:-

* The Analysis of a log periodic dipole array can be done by considering three region of an antenna which is classified according to the length of the dipoles. They are

(i) Inactive transmission - line region ($L < \lambda/2$)

(ii) Active region $L \approx \lambda/2$.

(iii) Inactive reflective region ($L > \lambda/2$):-

(i) Inactive transmission line region ($L < \lambda/2$):-

* It is the region in which the length of the dipoles is less than the resonant length $\lambda/2$. Therefore, the elements present relatively high Capacitance impedance. The spacing between the elements are comparatively smaller.

* The current in the region will be very small and hence it is considered as inactive region. These currents leads to the voltage supplied by the transmission line. Transposition of transmission introduces 180° phase shift between adjacent dipoles.

* Hence currents in the elements of these regions are small and hence small radiation in backward direction (towards left).

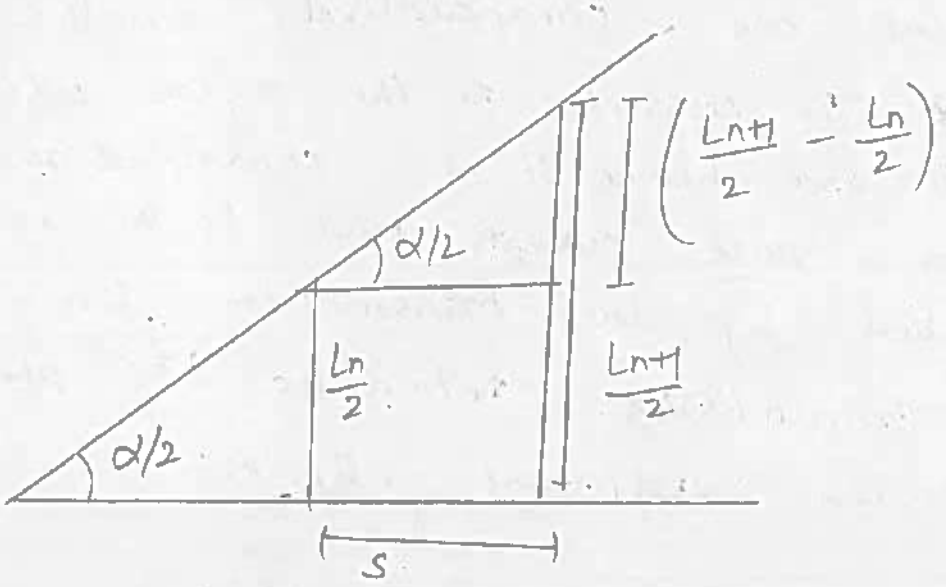
(ii) Active Region ($L \approx \lambda/2$):-

In this region, the dipole lengths are approximately equal to the resonant length ($\lambda/2$). Therefore, the dipoles in this region offer a resistive impedance. Thus the element currents are of large value and are in phase with the base voltage. Hence, there is a strong radiation towards left in backward direction and a little radiation towards right.

(iii) Inactive Reflective Region ($L > \approx \lambda/2$):-

The element (dipoles) lengths are longer than the resonant length (i.e) $L > \approx \lambda/2$). Hence the dipole offers an inductive impedance. The currents will be smaller in this region and also lags at the base voltage. Thus, any small amount of incident wave form an active region is reflected back towards the backward direction.

Design of LPDA:-



* Consider a part of a log periodic array as shown in figure. The performance of a log periodic dipole array depends on the following parameters.

(i) Apex angle (α)
(ii) Design ratio (τ) and
(iii) Spacing factor (σ)

* From the figure

$$\tan(\alpha/2) = \frac{\frac{L_{n+1} - L_n}{2}}{S} \longrightarrow (4)$$

$$\tan(\alpha/2) = \frac{L_{n+1} - L_n}{2S} = \frac{L_{n+1}\left[1 - \frac{L_n}{L_{n+1}}\right]}{2S} \longrightarrow (5)$$

But $\frac{L_{n+1}}{L_n} = K$ (i.e) $\frac{L_n}{L_{n+1}} = \frac{1}{K} \longrightarrow (6)$

* By substituting the equation (6) in equation (5)

$$\tan(\alpha/2) = \frac{(1 - 1/K) L_{n+1}}{2S} \longrightarrow (7)$$

For active region $L_{n+1} = \lambda/2 \longrightarrow (8)$.

* By substituting the equation (8) in (7)

$$\tan(\alpha/2) = \frac{(1 - \frac{1}{K})(\lambda/2)}{2S} = \frac{1 - 1/K}{4(S/\lambda)} = \frac{1 - 1/K}{4\sigma} \longrightarrow (9)$$

where $\sigma = S/\lambda$ = Spacing factor.

$\alpha \rightarrow$ Apex Angle

$K \rightarrow$ Scale factor

But $\tau = 1/K$

$$\tan(\alpha/2) = \frac{1 - \tau}{4\alpha} \longrightarrow (10)$$

* Out of the three parameters (σ, τ and α) two are specified and the third is determined. The number of elements in an array (n) can be obtained from an upper frequency (f_U) and lower frequency (f_L) and it is given as

$$\log(f_U) - \log(f_L) = (n-1)\log\left(\frac{1}{\tau}\right) \longrightarrow (ii)$$

Uses of Log Periodic Antenna:

(i) It is mainly used in the field of HF communicaton where the multiband steerable (rotatable) and fixd anternas are generally used. It has an advantage that no power is wasted in terminating resistance.

(ii) It is used for TV reception. Only one log periodic design will suffice for all the channels even upto an UHF band.

(iii) It is best suited for all round monitoring (i.e) a simple log periodic antenna will recover all the higher frequency bands, when there is no problem with the cost of installation.

Microstrip Antennas (or) patch antennas

The antenna which is made up of metal plates Placed on dielectric and fed by microstrip (or) coplaner transmission line is called microstrip antenna. It is also called as patch antenna (or) microstrip patch antenna.

As MSA are directly printed on to the circuit boards so it is called as printed antenna.

Applications

The MSA is generally preferred in high performance aircraft, spacecraft and missile applications where size, weight, cost, performance, ease of installation, and aerodynamic profile are the main constraints and low profile antennas are required.

Construction

MSAs are constructed on a dielectric substrate using a process similar to lithography in which patterns are printed on the substrate while fabricating on printed circuit board (or) integrating circuits.

The MSA consists of a very thin ($t << \lambda$) metallic strip(or) patch placed over a substrate. The substrate in between the patch and ground plane is a dielectric sheet or a dielectric constants are usually in the range of $2.2 \leq \varepsilon_r \leq 12$.

The height is very small ($h << \lambda$) as compared to the free space wavelength λ_0

$$0.003\lambda \leq h \leq 0.05\lambda$$

Length of the patch $\lambda/s < L < \lambda/L$

The size of MSA is inversely Proportional to its frequency. At frequencies lower than for an AM radio at 1 MHZ, the microstrip patch would be of the size of a football field. for MSA designed to receive an fm radio at 100MHz its length would be of the order of 1m. At X-band, MSA size will be of the order of 1 cm.

Rectangular Microstrip antenna

Rectangular shape is simplest and most widely used configuration for fabrication of microstrip antennas, Dimension of length L of Patch is always greater than the dimension of W.

* The shapes are useful for low cross polarization radiation and radiation pattern can be easily analyzed. square patch can generate Pencil beam while the rectangular patch can generate fan beam. Actually the circular patch is easy to fabricate but it is very difficult to calculate the current distribution in it.

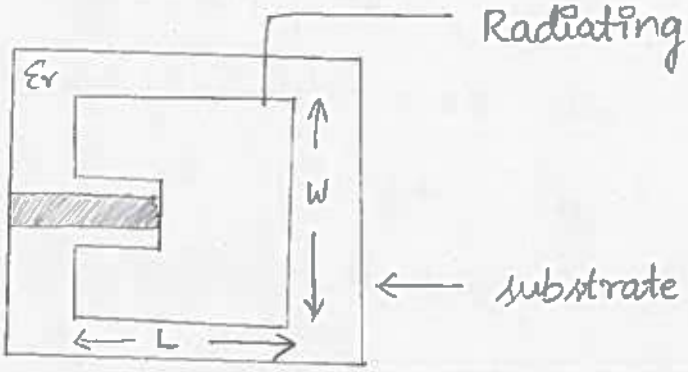
* To obtain linear and circular polarization either a single element (or) array of microstrip antenna can be used. To achieve greater directivity array of microstrip elements with single (or) multiple feed are used.

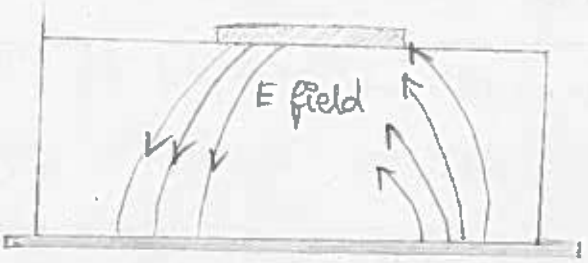
* Those E fields lines emerge out and Propagated in a direction which is normal to the substrate, they are now in the same direction. As the fields are in same phase both get added together.

* The frequency of operation of the patch antenna is generally determined by the length L. The critical frequency

$$f_c \simeq \frac{C}{2L\sqrt{\varepsilon_r}} = \frac{1}{2L\sqrt{\varepsilon_0 \varepsilon_r \mu_0}}$$

If patch length ($L = \lambda/2$) the electric field produced under the edges of opposite polarity

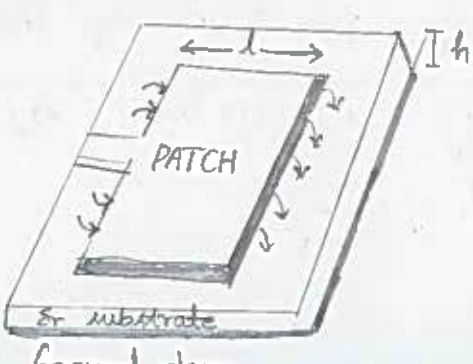


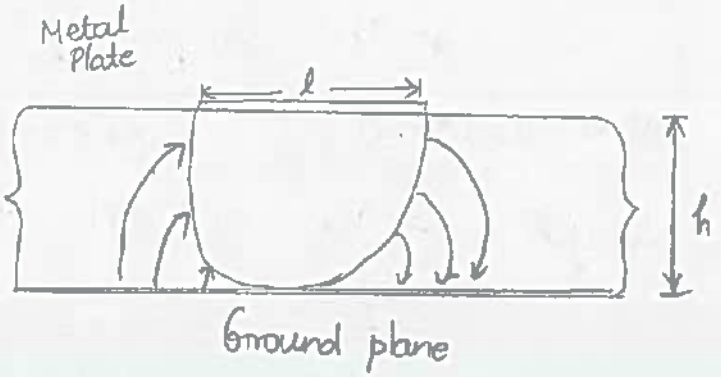


Electric field of patch antenna

Types of patches in MSA

The shapes of the radiating element (or) patch are out of the shapes, square, rectangular, triangular, dipole and circular are most commonly used shapes for the patch because of ease in fabrication.





Where c - velocity of light

ε_0 - permittivity of free space

ε_r - permittivity of dielectric substrate

μ_0 - permittivity of free space.

The expression for frequency of operation of patch antenna is considering L & W is given by

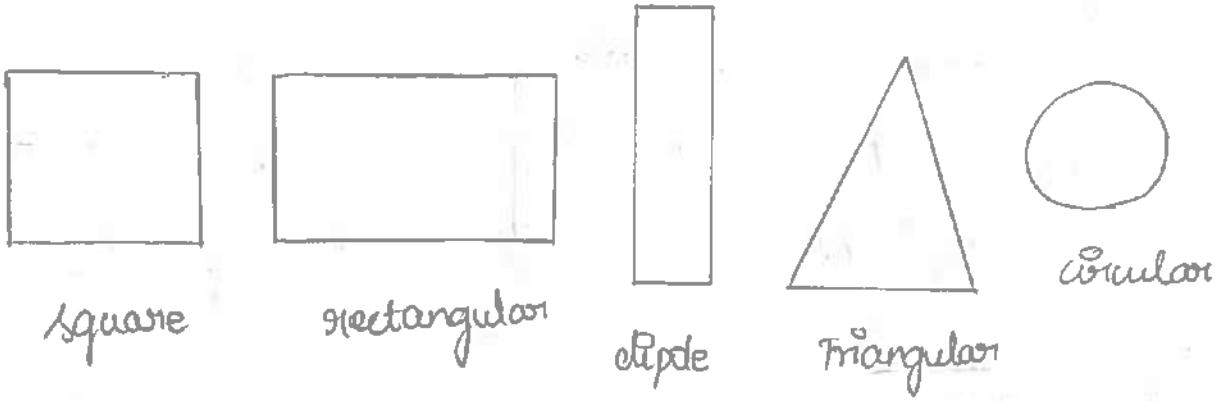
$$f_{r,nm} = \frac{c}{2\sqrt{\varepsilon_{r\,eff}}}\left[\left\{\frac{n}{L+2\Delta L}\right\}^2 + \left\{\frac{m}{W+2W}\right\}^2\right]^{1/2}$$

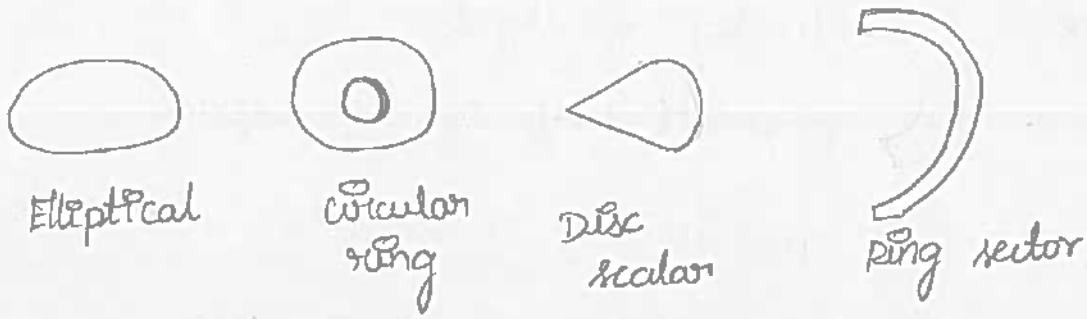
for dominating mode $n=1$, $m=0$

$$f_{r,nm} = \frac{c}{2(L+2\Delta L)\sqrt{\varepsilon_{r,eff}}}$$

W is the important parameter as it controls the i/p impedance of the antenna.

For square patch antenna ($L=W$), i/p impedance is typically same.

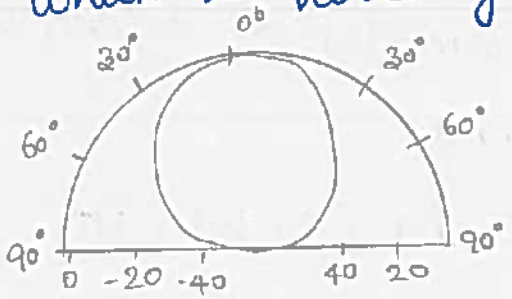




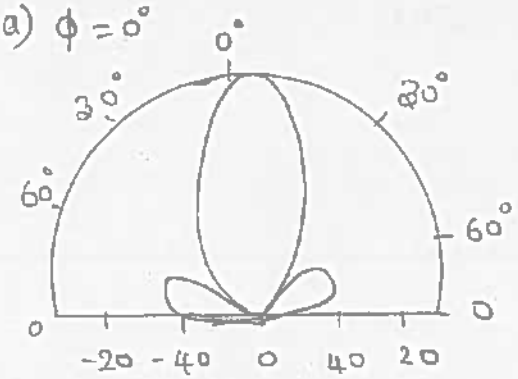
Various shapes of patchs

Radiation pattern of MSA

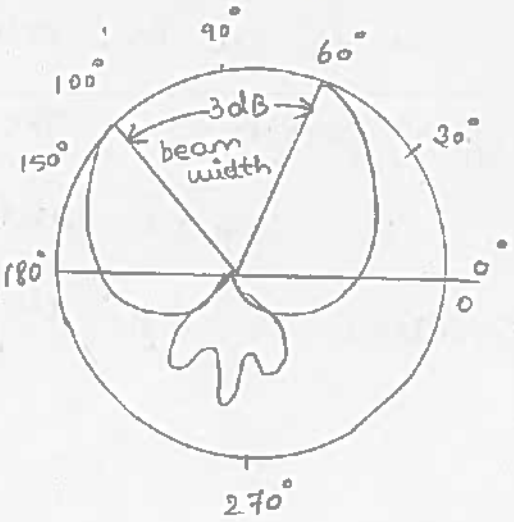
Consider a substrate of height $h << \lambda \simeq 0.05\lambda$ and length $L = 0.5\lambda$. The structure radiates from the fringing fields are exposed above the substrate at the edges of the patch. The patch acts as resonate the cavity with an electric field perpendicular to the patch. The magnetic field has tangential components which is vanishing at the four edges of the patch.



a) $\phi = 0°$



b) $\phi = 90°$



c) Rp for linearly polarized MSA

feed methods of MSA

1) contacting feed

In this method, the RF power is fed directly to the radiating patch which uses a connecting element such as microstrip (or) co-axial line. The commonly used feeds are microstrip feed and coaxial feed.

2) Non-contacting feed

Here electromagnetic coupling is done to transfer the power from feed line to the radiating patch. The most commonly used non contacting feed methods are aperture the fields of the lowest resonate mode $L > w$ are given by

$$E_Z = -E_0 \sin\left(\frac{\pi a}{L}\right) \quad -\frac{L}{2} \leq x \leq \frac{L}{2}$$

Most of the radiation from the MSA arises from sides 1 & 3. The sides 2 & 4 contribute little to the total radiation and they are normally negligible.

The normalized gain is given by

$$G_1(\theta, \phi) = \frac{|E(\theta, \phi)|^2}{|E(\theta, \phi)|^2_{max}}$$

The expression for E field components E_θ and E_ϕ are given by

$$E_\theta = \frac{\sin[(kW\sin\theta\sin\phi)/2]}{(kW\sin\theta\sin\phi)/2} \cos\left[kL\sin\theta\cos\frac{\phi}{2}\right]$$

$$E_\phi = \frac{\sin[(kW\sin\theta\sin\phi)/2]}{(kW\sin\theta\sin\phi)/2} \cos\frac{(kL\sin\theta\cos\phi)}{2} \sin\phi$$

where ϕ and θ are elevation and angle of radiation pattern.

$$k = \frac{2\pi}{\lambda} = \text{wave number}$$

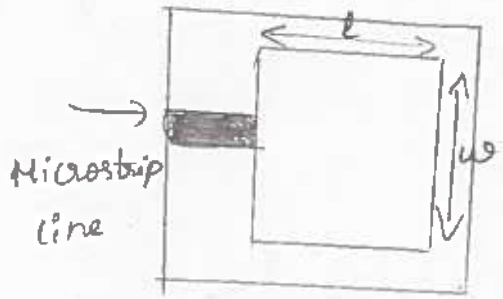
Thus the resonant field at any point is given by

$$E(\theta, \phi) = \sqrt{E_\theta^2 + E_\phi^2}$$

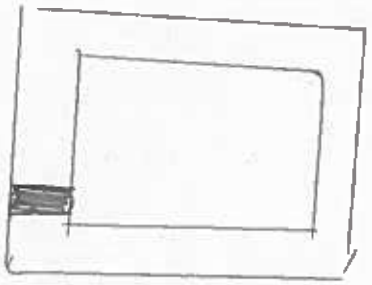
The normalised Radiation pattern is obtained by $L = W = \lambda/2$ $\sin\theta = 0^\circ$, $\phi = 90^\circ$ plane and linear polarized patch antenna

Coupling and proximity coupling

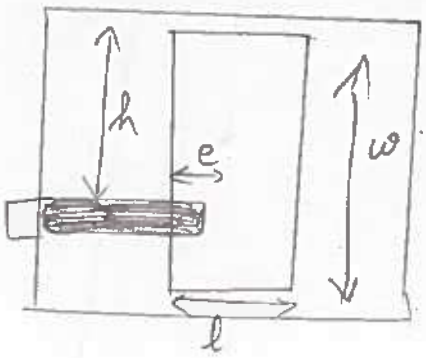
The microstrip feed methods are further subdivided into four main classes.



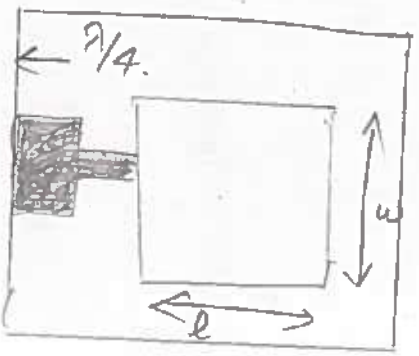
(a) center feed.



(b) offset feed



(C) Inset feed



(d) Quarter wave line feed

Advantages:-

* MSA are low profile antennas. They are smaller in size, light ~~and~~ weight antenna.
* Low fabrication cost.
* Simple and inexpensive.

Disadvantages:-

* MSA's are low gain and low efficiency antennas
* Low power handling capacity.
* Size is inversely proportional to frequency.

Applications:-

* Mobile and Satellite Communication
* RFID
* Radar application.
* Military and Space application.

UNIT-III

ANTENNA ARRAYS AND APPLICATIONS

Antenna Array:

An antenna array is simply defined as a system of similar antennas oriented similarly to get the greater directivity in a desired direction.

Linear Array:

The individual antenna of an antenna array system is termed as an 'element'. An antenna array is said to be linear, if the elements of an antenna array are equally spaced along a straight line.

Uniform Linear Array:

The linear antenna array is said to be an uniform linear array, if all the elements are fed with a current of equal magnitudes with the progressive uniform phase shift along the line.

Advantages:

- → High directivity is obtained
- → High SNR is obtained
- → Increase in overall gain
- → Power wastage is reduced
- → Better performance

Disadvantages:-

→ Mounting and maintenance is difficult
→ Large space required for placing antennas
→ High resistive losses.

* Practically, the various forms of the antenna array are used as radiating systems. Some of the practically used forms are as follows:

(i) Broadside array
(ii) End fire array
(iii) Collinear array
(iv) Parasitic array.

Two - Element Array: Arrays of Two Point Sources:

Array of two driven $\lambda/2$ elements.

* This is the simplest situation in the arrays of isotropic point sources in which it is assumed that the two point sources are separated by a distance (say 'd') and also have the same polarization.

* According to antenna theory, the superposition or addition of fields from the various sources at a great distance with the due regard to phases.

* Arrays of two isotropic point sources are different cases as follows:-

(i) Equal amplitude and phase
(ii) Equal amplitude and opposite phase
(iii) Unequal amplitude and opposite phase.

(1) Arrays of Two point sources with Equal Amplitude and Phase:

BROAD SIDE ARRAY:

* The Broadside array is defined as, "the array of antennas in which all the identical antennas are placed parallel to each other along axis of antenna array and each element is perpendicular to the axis of antenna array.

* The direction of maximum radiation is always perpendicular to the plane consisting elements (antenna array axis)."

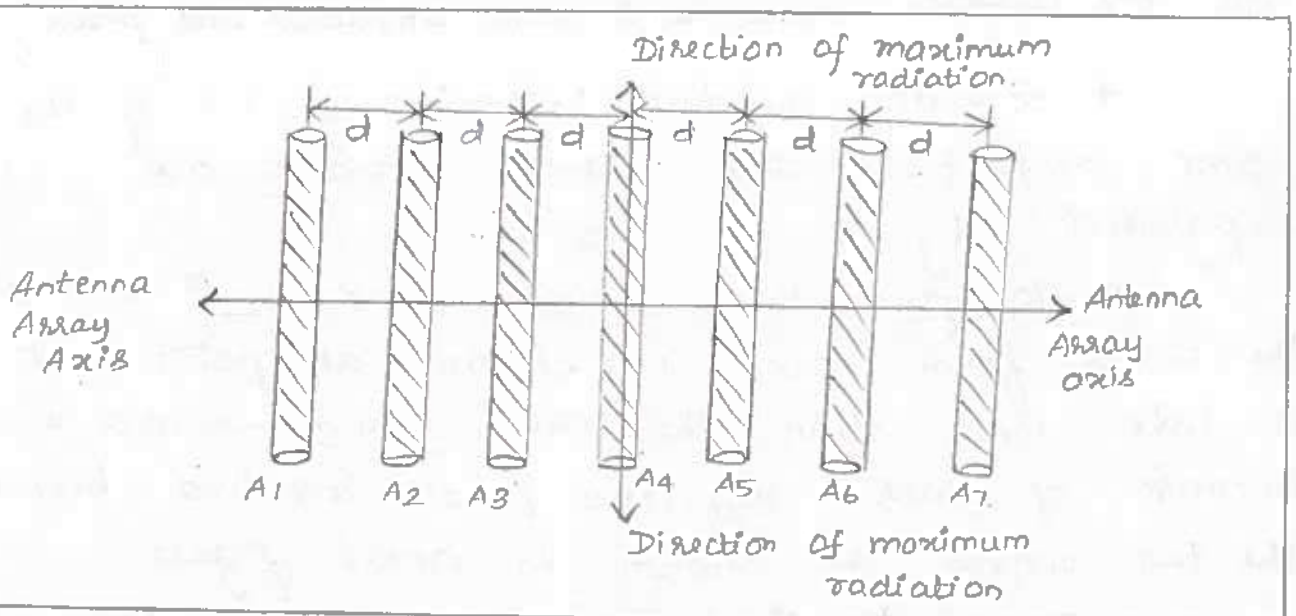


Fig. Broadside array of Antennas

* All individual antennas are spaced equally along the axis of antenna array. The spacing between any two elements is denoted by 'd' and they are fed with the currents of equal magnitude and same phase and the radiation pattern for the broadside array is bidirectional.

(1) Field Pattern :-

* Two isotropic point sources symmetrically situated with respect to an origin in the cartesian coordinate system as shown below.

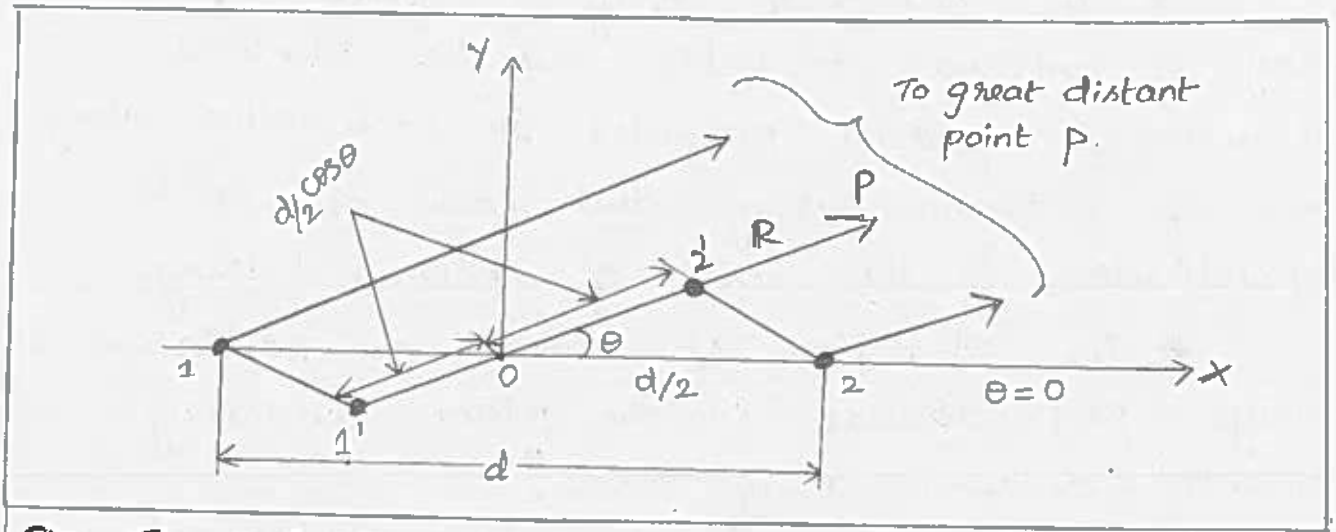


Fig. Two isotropic sources with same amplitude and phase

* Consider the two isotropic sources of the same amplitude and phase 1 and 2 are separated by a distance 'd'.

* To find the resultant field at P consider the waves from source 1 reaches at point P at a later time than the waves from source 2 because of path difference (1'2') involved between the two waves as shown in above figure.

* Thus the fields due to source 1 lags while that due to source 2 leads.

* Path difference between the two waves is expressed as, $= \frac{d}{2}\cos\theta + \frac{d}{2}\cos\theta \quad [\because (1'2')]$

$= d\cos\theta$ (meters)

Path difference $= \frac{d}{\lambda}\cos\theta$ (wavelengths) $\longrightarrow$ (1)

* Phase angle (ψ) = 2π × path difference

$$= 2\pi \times \frac{d}{\lambda} \cos\theta$$

$$= \frac{2\pi}{\lambda} d\cos\theta \text{ (radians)}$$

We know that $\beta = \frac{2\pi}{\lambda}$

$$\therefore \boxed{\psi = \beta d \cos\theta \text{ radians.}} \longrightarrow (2)$$

Now,

Let E_1 → Far Electric Field at distance P, due to source 1.

E_2 → Far Electric field at distance P, due to source 2.

E → Total Electric field at distance point.

* Then, total far field at distance point P, in the direction θ is given by

$$E = E_1 e^{-j\psi/2} + E_2 e^{+j\psi/2} \longrightarrow (3)$$

where $E_1 e^{-j\psi/2}$ → Field component due to source 1.

$E_2 e^{+j\psi/2}$ → Field component due to source 2.

* But in this case it is assumed that amplitudes are same (i.e)

$$E_1 = E_2 = E_0 \longrightarrow (4)$$

Then, equation (3) becomes

$$E = E_0 (e^{-j\psi/2} + e^{+j\psi/2}) \quad \left[\because \cos\theta = \frac{e^{j\theta} + e^{-j\theta}}{2}\right]$$

$$= 2E_0 \left(\frac{e^{-j\psi/2} + e^{+j\psi/2}}{2}\right)$$

$$E = 2E_0 \cos(\psi/2) \longrightarrow (5)$$

$$E = 2E_0 \cos\left(\frac{\beta d \cos\theta}{2}\right) \longrightarrow (6)$$

* This above equations are the equations of far-field pattern of two isotropic point sources of the same amplitude and phase.

* Here the total amplitude $2E_0$ whose maximum value may be 1. By substituting $2E_0 = 1$ or $E_0 = \frac{1}{2}$, the pattern is said to be normalised. Thus equation (6) becomes.

$$E = \cos\left(\frac{\beta d \cos\theta}{2}\right)$$

$$= \cos\left(\frac{2\pi}{\lambda} \cdot \frac{\lambda}{2} \cdot \frac{\cos\theta}{2}\right)$$

$$\boxed{E = \cos\left(\frac{\pi}{2}\cos\theta\right)} \longrightarrow (7)$$

* In order to draw the field pattern, the direction of maxima, minima and half power points must be known, which can be calculated with the help of equation (7) as follows:-

(ii) Maxima Direction for Major Lobe:-

E is maximum, when $\cos\left(\frac{\pi}{2}\cos\theta\right)$ is maximum (± 1).

$$\cos\left(\frac{\pi}{2}\cos\theta\right) = \pm 1$$

$$\frac{\pi}{2}\cos\theta = \cos^{-1}(\pm 1)$$

$$\frac{\pi}{2}\cos\theta_{max} = \pm N\pi \quad \text{where } N = 0, 1, 2, \ldots$$

If $N = 0$.

$$\frac{\pi}{2}\cos(\theta_{max})_{major} = \pm N\pi \quad [\text{Here } N=0]$$

$$\cos(\theta_{max})_{major} = 0.$$

$$(\theta_{max})_{major} = \cos^{-1}(0)$$

$$\boxed{(\theta_{max})_{major} = 90^\circ \text{ and } 270^\circ} \longrightarrow (8)$$

(iii) Minima Directions :-

E is minimum when $\cos\left(\frac{\pi}{2}\cos\theta\right)$ is minimum and its minimum value is 0.

$$\cos\left(\frac{\pi}{2}\cos\theta\right) = 0.$$

$$\frac{\pi}{2}\cos(\theta_{min}) = \pm(2N+1)\frac{\pi}{2} \quad \text{when } N = 0, 1, 2, \ldots$$

If $N = 0$,

$$\frac{\pi}{2}\cos(\theta_{min}) = \frac{\pi}{2}.$$

$$\cos(\theta_{min}) = \pm 1$$

$$\boxed{\theta_{min} = 0^\circ \text{ and } 180^\circ} \longrightarrow (9)$$

(iv) Half Power Point Directions :-

When the power is half, the voltage or current is $\frac{1}{\sqrt{2}}$ times the maximum value. Hence the condition for half power point is given by

$$\cos\left(\frac{\beta d\cos\theta}{2}\right) = \pm\frac{1}{\sqrt{2}}$$

Let $d = \lambda/2$ and $\beta = \frac{2\pi}{\lambda}$, then

$$\cos\left(\frac{\pi}{2}\cos\theta\right) = \pm\frac{1}{\sqrt{2}}$$

$$\frac{\pi}{2}\cos(\theta_{HPPD}) = \pm(2N+1)\frac{\pi}{4} \quad \text{where } N = 0, 1, 2, \ldots$$

If $N = 0 \Rightarrow \quad \frac{\pi}{2}\cos(\theta_{HPPD}) = \pm\frac{\pi}{4}.$

$$\cos(\theta_{HPPD}) = \pm\frac{1}{2}$$

$$\theta_{HPPD} = \cos^{-1}\left(\pm\frac{1}{2}\right)$$

$$\boxed{\theta_{HPPD} = 60^\circ \text{ or } 120^\circ} \longrightarrow (10)$$

* Now the field pattern between E and θ is drawn for the case $d = \lambda/2$ then the below figure is obtained which is bidirectional, 360° rotation of this figure around x-axis will generate the 3- dimensional space pattern - a doughnut shape.

* This is the simplest type of "Broadside array" and it is also known as "Broad side couplet" as two isotropic radiators in phase.

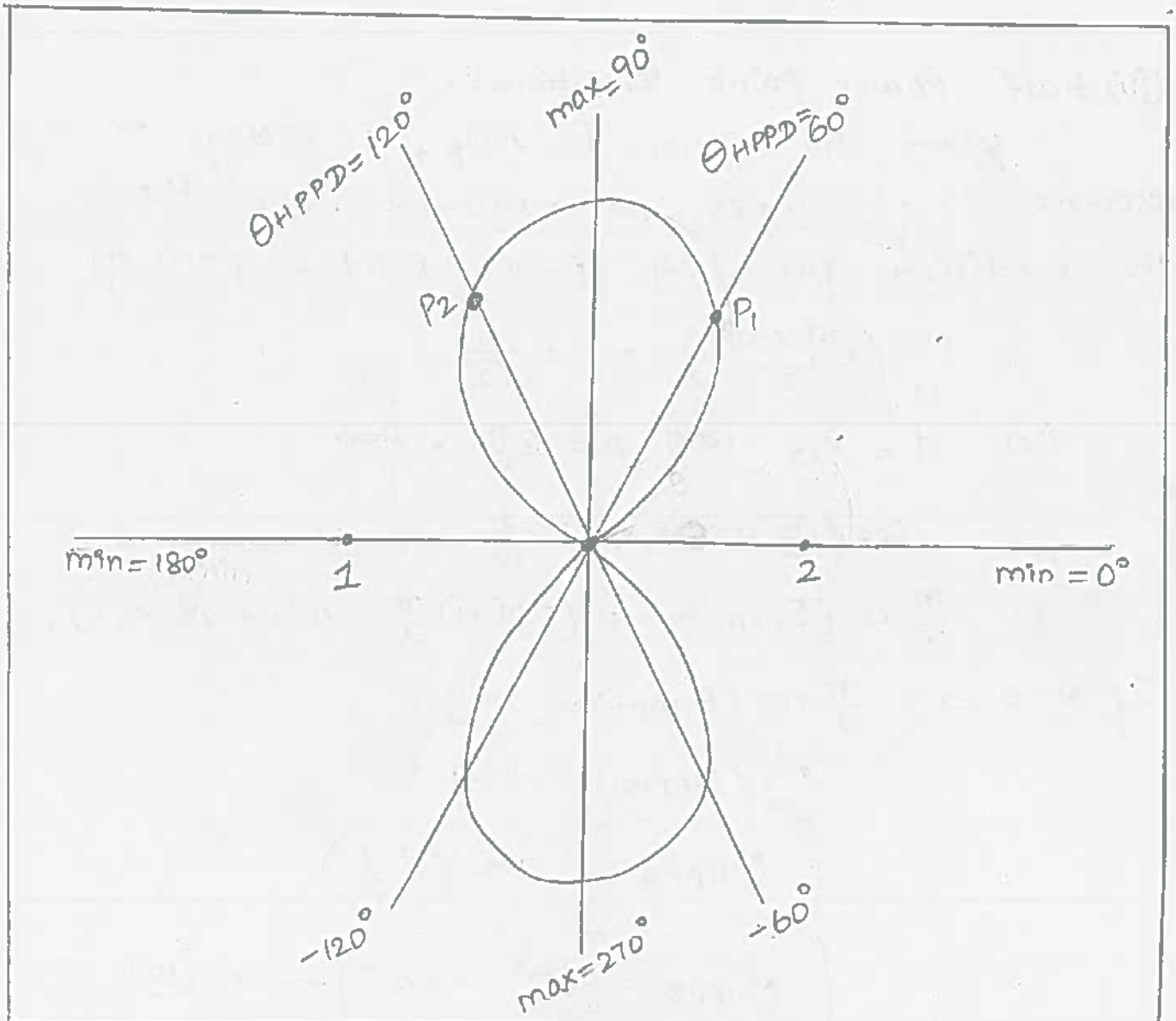


Fig. Field patterns of broadside array of two element with in-phase.

Arrays of two point sources with Equal Amplitude and Opposite Phase :

END FIRE ARRAY:

* An array is said to be end fire, if the direction of maximum radiation coincides with an axis to get unidirectional Radiation.

* In an end fire array, the number of identical antennas are spaced equally along a line All the antennas are fed individually with currents of equal magnitudes but their phases vary progressively along the line to make the entire arrangement to get unidirectional radiation along the axis of the array.

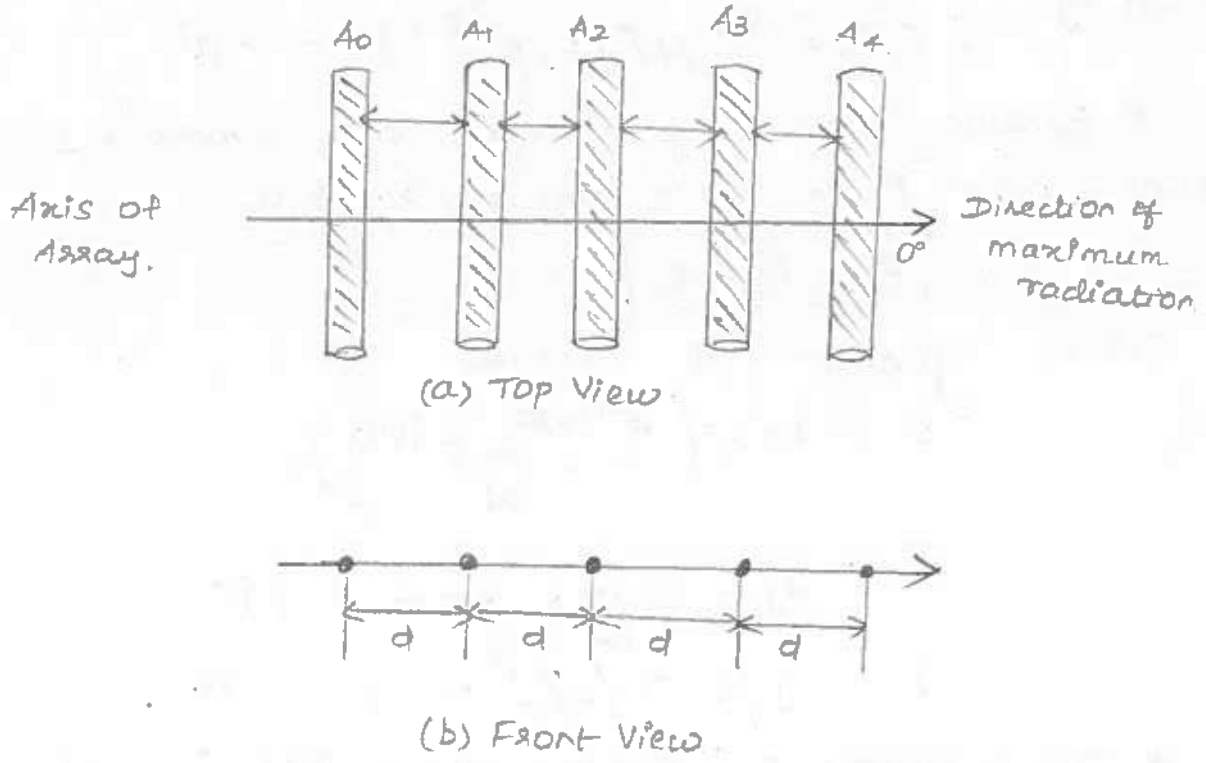


Fig End Fire Array.

(i) Field Pattern:

* Consider an array of two centre-fed vertical $\lambda/2$ elements (dipoles) in free-space arranged side by side with a spacing 'd' and equal currents in opposite phase (i.e) point source 1 is out of phase or opposite phase (180°) to as in below figure.

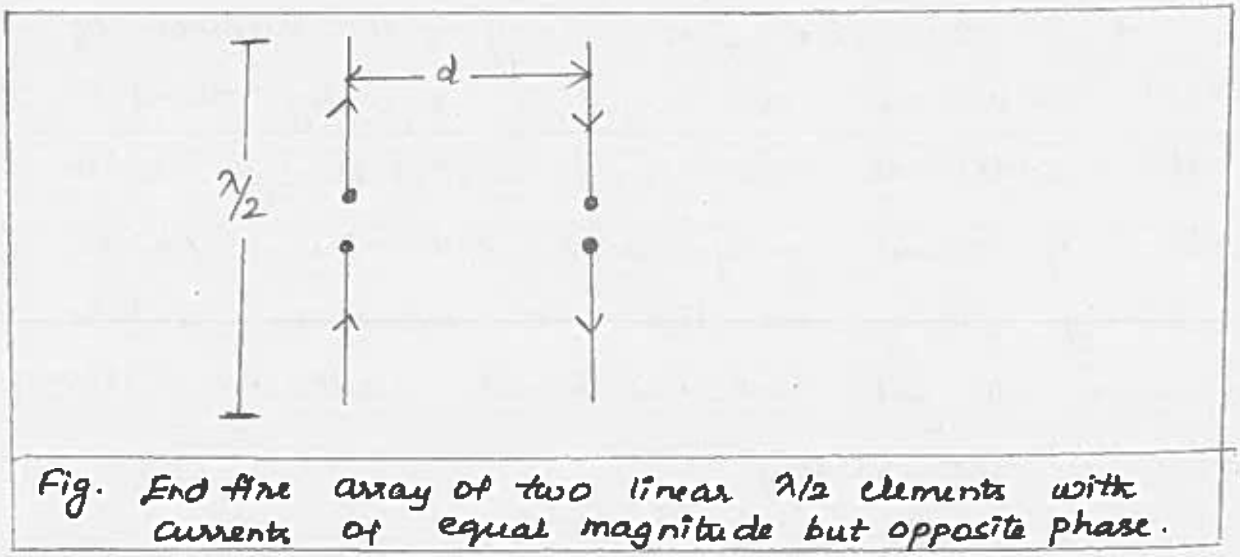


Fig. End fire array of two linear $\lambda/2$ elements with currents of equal magnitude but opposite phase.

* The total far field at a distant point 'P' is given by.

$$E = (-E_1 e^{-j\psi/2}) + (+E_2 e^{+j\psi/2}) \longrightarrow (11)$$

* Because, phase of source 1 and source 2 at distant point P is $-\psi/2$ and $+\psi/2$. But

$$\boxed{E_1 = E_2 = E_0}$$

Then equation (11) becomes.

$$E = E_0 2j\left(\frac{e^{+j\psi/2} - e^{-j\psi/2}}{2j}\right)$$

$$E = 2j\, E_0 \sin\left(\frac{\psi}{2}\right) \longrightarrow (12)$$

$$E = 2j\, E_0 \sin\left(\frac{\beta d}{2}\cos\theta\right) \longrightarrow (13)$$

* The operator 'j' simply means that an opposite phase brings a phase shift of 90° in the total field.

Let $d = \lambda/2$ and $2E_0 j = 1$.

$$E_{norm} = \sin\left(\frac{\beta d}{2}\cos\theta\right) \qquad \left[\because \beta = \frac{2\pi}{\lambda} \text{ and } d = \lambda/2\right]$$

$$= \sin\left(\frac{2\pi}{\lambda} \times \lambda/4 \cos\theta\right)$$

$$\boxed{E_{norm} = \sin\left(\frac{\pi}{2}\cos\theta\right)} \longrightarrow (14)$$

(ii) Maximum Directions :-

The maximum value of sine function is ±1.

$$\sin\left(\frac{\pi}{2}\cos\theta\right) = \pm 1.$$

$$\frac{\pi}{2}\cos(\theta_{max}) = \sin^{-1}(\pm 1)$$

$$\frac{\pi}{2}\cos(\theta_{max}) = \pm 1\,(2N+1)\frac{\pi}{2} \quad \text{where } N = 0, 1, 2, \ldots$$

If $N = 0 \Rightarrow$ $\cos(\theta_{max}) = \pm 1$

$$\boxed{\theta_{max} = 0° \text{ and } 180°} \longrightarrow (15)$$

(iii) Minima Directions:-

Minimum value of a sine function is 0.

$$\sin\left(\frac{\pi}{2}\cos\theta\right) = 0$$

$$\frac{\pi}{2}\cos(\theta_{min}) = \pm N\pi \quad \text{where } N = 0, 1, 2, \ldots$$

$$\cos(\theta_{min}) = 0$$

$$\boxed{\theta_{min} = 90° \text{ and } 270°} \longrightarrow (16)$$

(iv) Half Power Point Directions:-

$$\sin\left(\frac{\pi}{2}\cos\theta\right) = \pm \frac{1}{\sqrt{2}}.$$

$$\frac{\pi}{2}\cos(\theta_{HPPD}) = \sin^{-1}\left(\pm\frac{1}{\sqrt{2}}\right)$$

$$\frac{\pi}{2}\cos(\theta_{HPPD}) = \pm(2N+1)\frac{\pi}{4} \quad \text{where } N = 0, 1, 2, \ldots$$

If $N = 0 \Rightarrow$ $\cos(\theta_{HPPD}) = \pm\frac{1}{2}.$

$$\theta_{HPPD} = \cos^{-1}\left(\pm\frac{1}{2}\right)$$

$$\boxed{\theta_{HPPD} = 60°, \pm 120°} \longrightarrow (16)$$

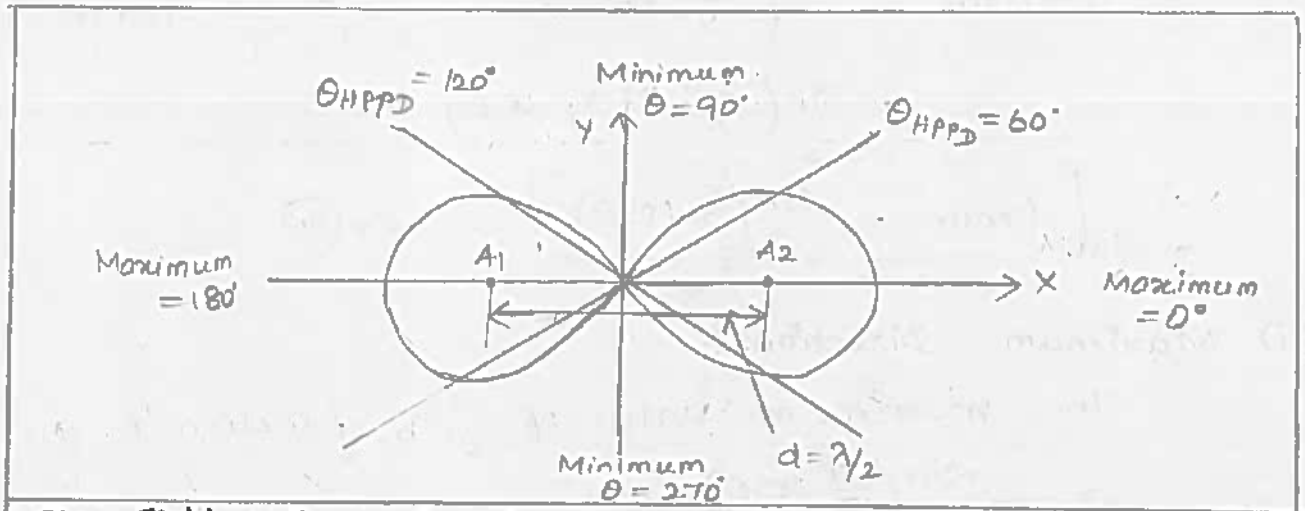


Fig. Field pattern for two point sources with spacing $d = \lambda/2$ and fed with current equal in magnitude but out of phase by 180°.

Arrays of Two Points Sources with Unequal Amplitude and Any Phase : Equal Currents of any phase Relation.

* Consider two points sources with an unequal amplitude and hence any phase difference say α. Assume the source 1 is taken as a reference for phase and amplitude.

* Fields due to source 1 and 2 at a distant point P and E_1 and E_2 where E_1 is greater than E_2 ($E_1 > E_2$). Then the total phase difference between the radiations of two sources at a point P is given by

$$\boxed{\psi = \frac{2\pi}{\lambda} d \cos\theta + \alpha} \longrightarrow (17)$$

where α is the phase angle by which the current (I_2) of source 2 leads to the current (I_1) of source 1.

If $\alpha = 0°$ or $180°$ and $E_1 = E_2 = E_0$, then the total fields at P is given by

$$E = E_1 e^{j0} + E_2 e^{j\psi}$$
$$= E_1 (1 + E_2/E_1 \, e^{j\psi}) \qquad \left[\because e^{j0} = e^0 = 1\right]$$

$$\boxed{E = E_1 (1 + k e^{j\psi})} \longrightarrow (18)$$

where $k = \frac{E_2}{E_1}$ and $E_1 > E_2$. Then $k < 1$ (i.e) $0 \leq k \leq 1$.

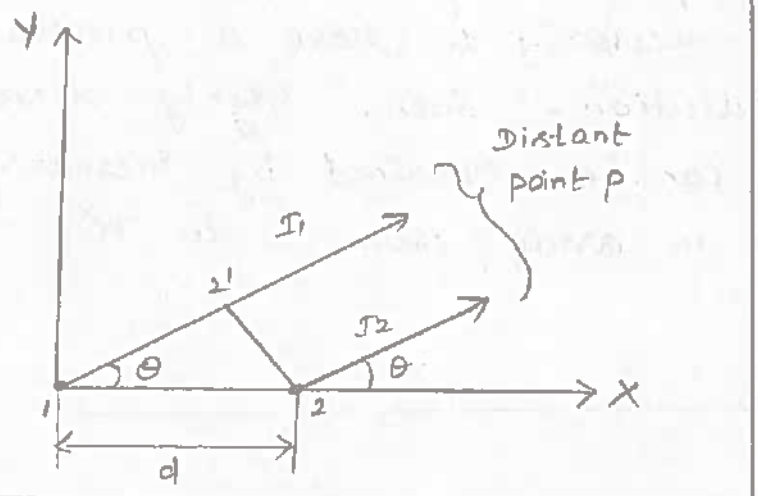


Fig (a). Two point sources with unequal amplitude and any phase difference α.

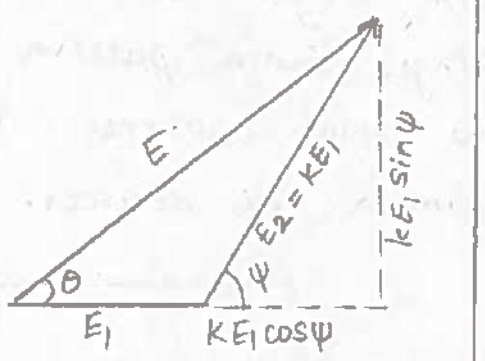


Fig (b) Vector diagram of fig (a).

From equation (18)

Magnitude and phase angle (θ) at point P is given by taking its modulus.

$$E = \left| E_1 \{ 1 + K(\cos\psi + j\sin\psi) \} \right|$$

(or)

$$E = E_1 \sqrt{(1 + k\cos\psi)^2 + (k\sin\psi)^2 / \theta} \longrightarrow (19)$$

where

Phase angle at P

$$\theta = \tan^{-1}\left(\frac{k\sin\psi}{1 + k\cos\psi}\right) \longrightarrow (20)$$

N-ELEMENT LINEAR ARRAY : UNIFORM AMPLITUDE AND SPACING.

* At higher frequencies, for a point to point communications, it is necessary to have a pattern with single beam radiation. Such, highly directive single beam pattern can be obtained by increasing the point sources in an array from 2 to 'n' number of sources.

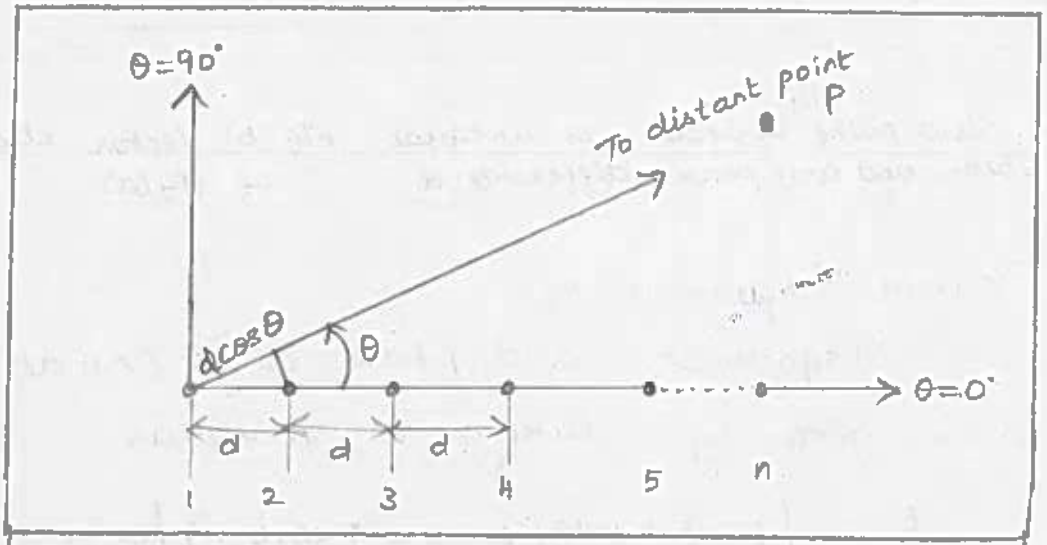


Fig. Uniform linear array of 'n' elements.

* An array of 'n' elements is said to be a linear array, "when all the individual elements are spaced equally along a line"

* An array is said to be an uniform array, "When the elements in the array are fed with the currents of equal magnitudes and uniform progressive phase shift along a line".

* Consider 'n' isotropic point sources of equal amplitude and spaced equally 'd' as a linear array. Here, sources are fed with in phase currents of equal Amplitudes (E_0).

* Total far-field at a distant point 'P' is obtained by adding the vectorically individual fields of 'n' sources as,

$$E_T = E_0 e^{0j\psi} + E_0 e^{j\psi} + E_0 e^{j2\psi} + E_0 e^{j3\psi} \ldots\ldots E_0 e^{j(n-1)\psi}$$

$$E_T = E_0 (1 + e^{j\psi} + e^{2j\psi} + e^{3j\psi} + \ldots + e^{j(n-1)\psi}) \longrightarrow (1)$$

* ψ is the total phase difference of the fields at a distant point 'P' from an adjacent sources and it is expressed as,

$$\psi = \beta d \cos\theta + \alpha \text{ (radian)} \longrightarrow (2)$$

where α = phase difference in adjacent point sources.

$\beta d \cos\theta$ = phase difference due to path difference and

$\beta = \frac{2\pi}{\lambda}$ = propogation constant.

* Consider, the source 1 is a phase reference. The field from source 2 is advanced in phase with respect to source 1 by ψ. The field from 3 is advanced in phase with respect to the source 1 by 2ψ etc. Multiplying equation (1) by $e^{j\psi}$ becomes,

$$E_T e^{j\psi} = E_0 (e^{j\omega} + e^{j2\psi} + e^{j3\psi} + e^{j4\psi} + \ldots e^{nj\psi}) \longrightarrow (3)$$

By subtracting equation (3) from an equation (1), we get.,

$$E_T - E_T e^{j\psi} = E_0 \left\{ \left[1 + e^{j\psi} + e^{j2\psi} + \ldots + e^{j(n-1)\psi} \right] - \left[e^{j\omega} + e^{j2\psi} + \ldots + e^{jn\psi} \right] \right\}.$$

$$E_T (1 - e^{j\psi}) = E_0 (1 - e^{jn\psi})$$

$$E_T = E_0 \left(\frac{1 - e^{jn\psi}}{1 - e^{j\psi}} \right) \longrightarrow (4)$$

Equation ④ may be written as,

$$E_T = E_0\left(\frac{1 - e^{jn\psi/2} \cdot e^{jn\psi/2}}{1 - e^{j\psi/2} \cdot e^{j\psi/2}}\right)$$

$$= \frac{E_0\left(e^{jn\psi/2} \cdot e^{-jn\psi/2} - e^{jn\psi/2} \cdot e^{jn\psi/2}\right)}{\left(e^{j\psi/2} \cdot e^{-j\psi/2} - e^{j\psi/2} \cdot e^{j\psi/2}\right)}$$

$$E_T = E_0\left[\frac{e^{\frac{j\psi n}{2}}\left(e^{-\frac{jn\psi}{2}} - e^{\frac{jn\psi}{2}}\right)}{e^{\frac{j\psi}{2}}\left(e^{-\frac{j\psi}{2}} - e^{\frac{j\psi}{2}}\right)}\right]$$

According to trigonometric identity,

$$e^{-j\theta} - e^{j\theta} = -2j\sin\theta \longrightarrow ⑤$$

* Using the equation ⑤, then the resultant field in equation ⑤ becomes,

$$E_T = E_0\left[\frac{\left(-2j\sin\frac{n\psi}{2}\right)e^{\frac{jn\psi}{2}}}{\left(-2j\sin\frac{n\psi}{2}\right)e^{\frac{j\psi}{2}}}\right]$$

$$E_T = E_0 \cdot e^{j\left(\frac{n-1}{2}\right)\psi}\frac{\sin\frac{n\psi}{2}}{\sin\frac{\psi}{2}} \longrightarrow ⑥$$

* Using equation ② the phase angle of the resultant field at point P is gives as,

$$\phi = \frac{(n-1)}{2}\psi = \left(\frac{n-1}{2}\right)(\beta d\cos\theta + \alpha) \longrightarrow ⑦$$

Then the equation ⑥ becomes.

$$E_T = E_0\left[\frac{\sin\frac{n\psi}{2}}{\sin\frac{\psi}{2}}\right]e^{j\phi} = E_0\left[\frac{\sin\frac{n\psi}{2}}{\sin\frac{\psi}{2}}\right](\cos\phi + j\sin\phi)$$

$$E_T = E_0\left[\frac{\sin\frac{n\psi}{2}}{\sin\frac{\psi}{2}}\right]\angle\phi \longrightarrow ⑧$$

* This equation (8) indicates the resultant field due to 'n' element linear array at a distant point P. The magnitude of the resultant field is gives as,

$$E_T = E_0 \left[\frac{\sin \frac{n\psi}{2}}{\sin \frac{\psi}{2}} \right] \longrightarrow (9)$$

* If the reference point (source 1) is shifted to the centre of an array, then the phase angle ϕ is automatically eliminated from an equation (8) and it is reduced to,

$$E_T = E_0 \left[\frac{\sin \frac{n\psi}{2}}{\sin \frac{\psi}{2}} \right] \longrightarrow (10)$$

ANTENNA ARRAY FACTOR (AF):

* AF is the ratio of the magnitude of the resultant field to the magnitude of the maximum field and it is given as

$$\text{Array factor} = \frac{|E_T|}{|E_{max}|} \longrightarrow (11)$$

* When, $\psi \to 0$, the field from an array is maximum in any direction. Thus the maximum value of E_T is 'n' times the field from a single source.

$$E_{T(max)} = E_0 n \longrightarrow (12)$$

* If E_0 is assumed to be unity for normalized ($E_0 = 1$) then an equation (12) becomes,

$$E_{T(max)} = n.$$

* The normalized field pattern (or) an antenna array factor may be obtained from the equations (10) and (12), we get

$$E_{NOR} = \frac{E_T}{E_{T(MAX)}}$$

$$= \frac{E_0 \frac{\sin\frac{n\psi}{2}}{\sin\frac{\psi}{2}}}{E_{0n}}$$

$$E_{NOR} = \frac{\sin\frac{n\psi}{2}}{n\sin\frac{\psi}{2}} = (\text{Array Factor})_n \longrightarrow (13)$$

PATTERN MULTIPLICATION:

* Multiplication of pattern or simply pattern multiplication in general can be states as follows:-

"The total field pattern of an array of non-isotropic but similar sources is the multiplication of the individual source patterns and the pattern of an array of isotropic point sources each located at the phase centre of individual source and having the relative amplitude and phase, whereas the total phase pattern is the addition of the phase pattern of the individual sources and that of the array of isotropic point sources."

* The total field pattern of an array of non-isotropic but similar sources may be expressed as

Total Field (E_T) = (Multiplication of field pattern) * (Addition of phase pattern).

$$E_T = \{E_i(\theta,\phi) \times E_a(\theta,\phi)\} \times \{E_{pi}(\theta,\phi) + E_{pa}(\theta,\phi)\} \rightarrow (1)$$

where $E_i(\theta,\phi)$ = Field pattern of individual source

$E_a(\theta,\phi)$ = Field pattern of array of isotropic point sources.

$E_{pi}(\theta,\phi)$ = phase pattern of individual source

$E_{pa}(\theta,\phi)$ = Phase pattern of array of isotropic point sources

θ — Polar angles

ϕ — Azimuth angles.

* This principle may be applied to any number of sources but they are similar. The word similar is used here to indicate the variation with absolute angle 'ϕ' of both the amplitude and phase of the field is the same.

* The maximum amplitudes of an individual source may be equal. If their maximum amplitudes are equal, then the sources are not only similar but are also identical.

Advantages:-

Advantages of pattern multiplication are,

(i) It is a speedy method for sketching the pattern of complicated arrays just by inspection.

(ii) It is a useful tool in the design of antenna arrays and.

(iii) The secondary lobes are determined from the number of nulls in the resultant pattern.

Radiation Pattern of 4- Isotropic elements fed in phase, spaced $\lambda/2$ Apart.

* Consider a four element of a isotropic (or) non-directive radiators are in a linear array antennas and are placed as shown in figure, in which the spacing between the elements is $\lambda/2$ and the currents are in phase (i.e) $\alpha = 0$.

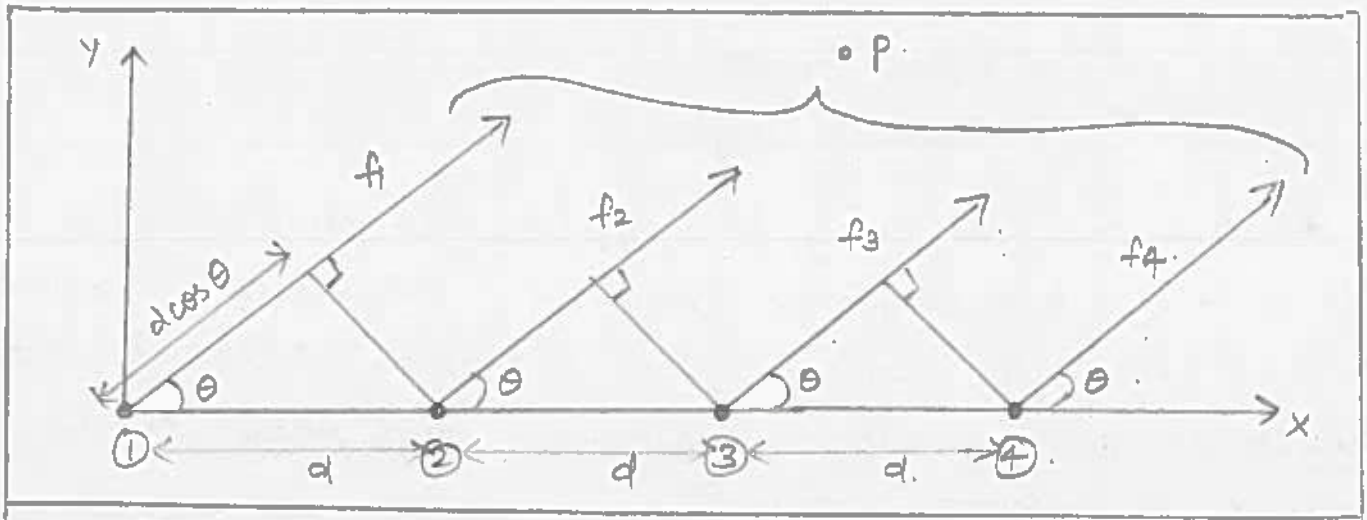


Fig. Linear array of 4 isotropic elements spaced $\lambda/2$ apart, fed in phase.

* One of the method to get the Radiation pattern of an array is adding the fields of an individual four elements at a distance point 'P' vectorially. However, the same radiation pattern can be obtained by pattern multiplication in the following manner.

* Now the elements 1 and 2 are considered as unit one (1) and this new unit is considered to be placed between the midway of elements 1 and 2. Similarly, the elements 3 and 4 are considered as unit two (2) and these are placed between the elements 3 and 4 and shown in figure.

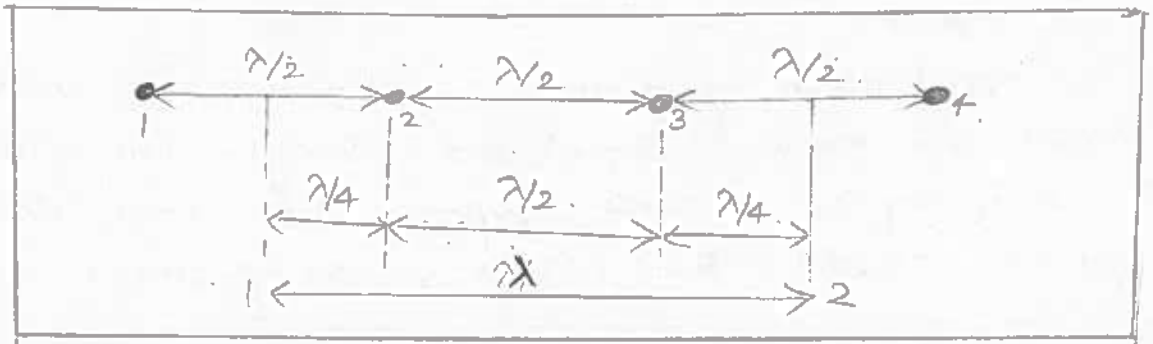


Fig. Two Units array spaced at λ.

* Now the four elements are spaced at λ/2 and have been replaced by at 2 units spaced λ and therefore the problem of determining the radiation of 4 elements has been reduced to find out the radiation pattern of 2 elements.

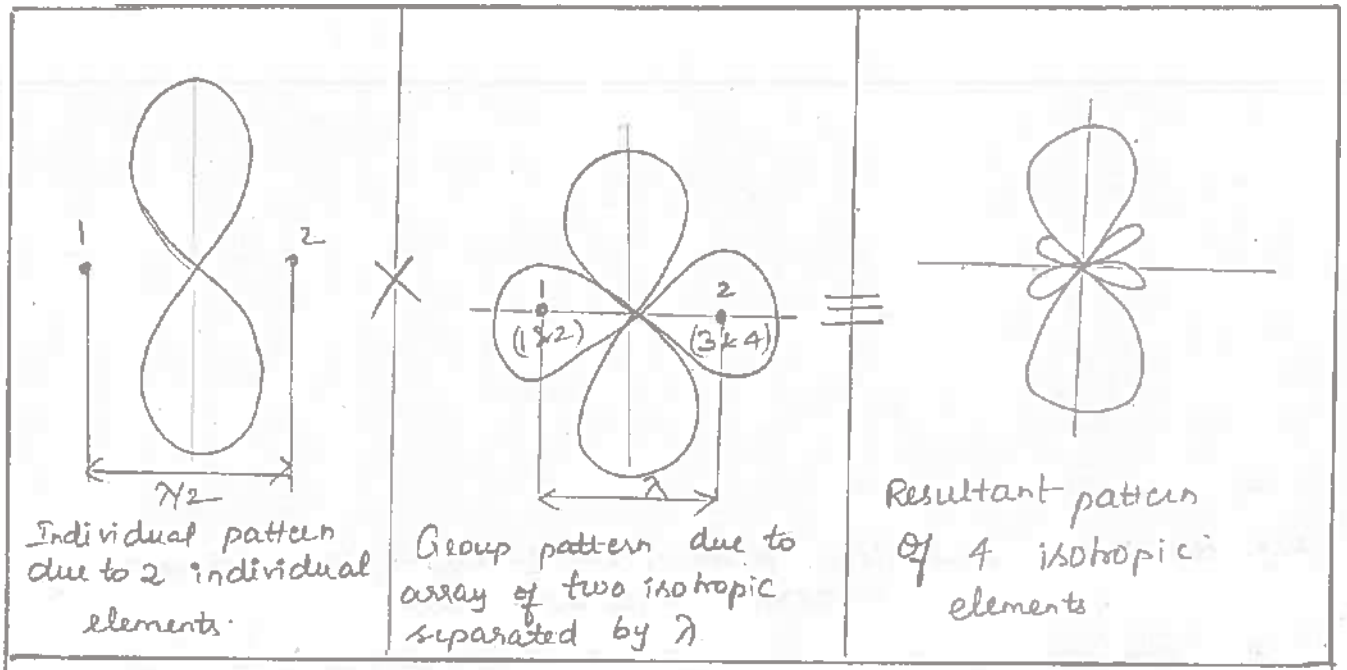


Fig. Resultant radiation pattern of 4 isotropic elements by pattern multiplication.

* Two isotropic point source spaced λ/2 apart fed in phase provides a bidirectional pattern. According to this pattern multiplication, the radiation pattern of 4 elements is obtained as,

$$\therefore \left(\begin{array}{c}\text{Resultant radiation}\\ \text{pattern of 4}\\ \text{elements}\end{array}\right) = \left\{\begin{array}{c}\text{Radiation pattern}\\ \text{of individual}\\ \text{elements}\end{array}\right\} \times \left\{\begin{array}{c}\text{Array of}\\ \text{two units}\\ \text{spaced '}\lambda\text{'}\end{array}\right\}$$

Radiation Pattern of 8-isotropic elements fed in Phase, Spaced $\lambda/2$ Apart.

* The application of pattern multiplication can now be extended to more complicated arrays. For example, consider 8 isotropic elements spaced $\lambda/2$ apart and fed in phase would be obtained as follows.

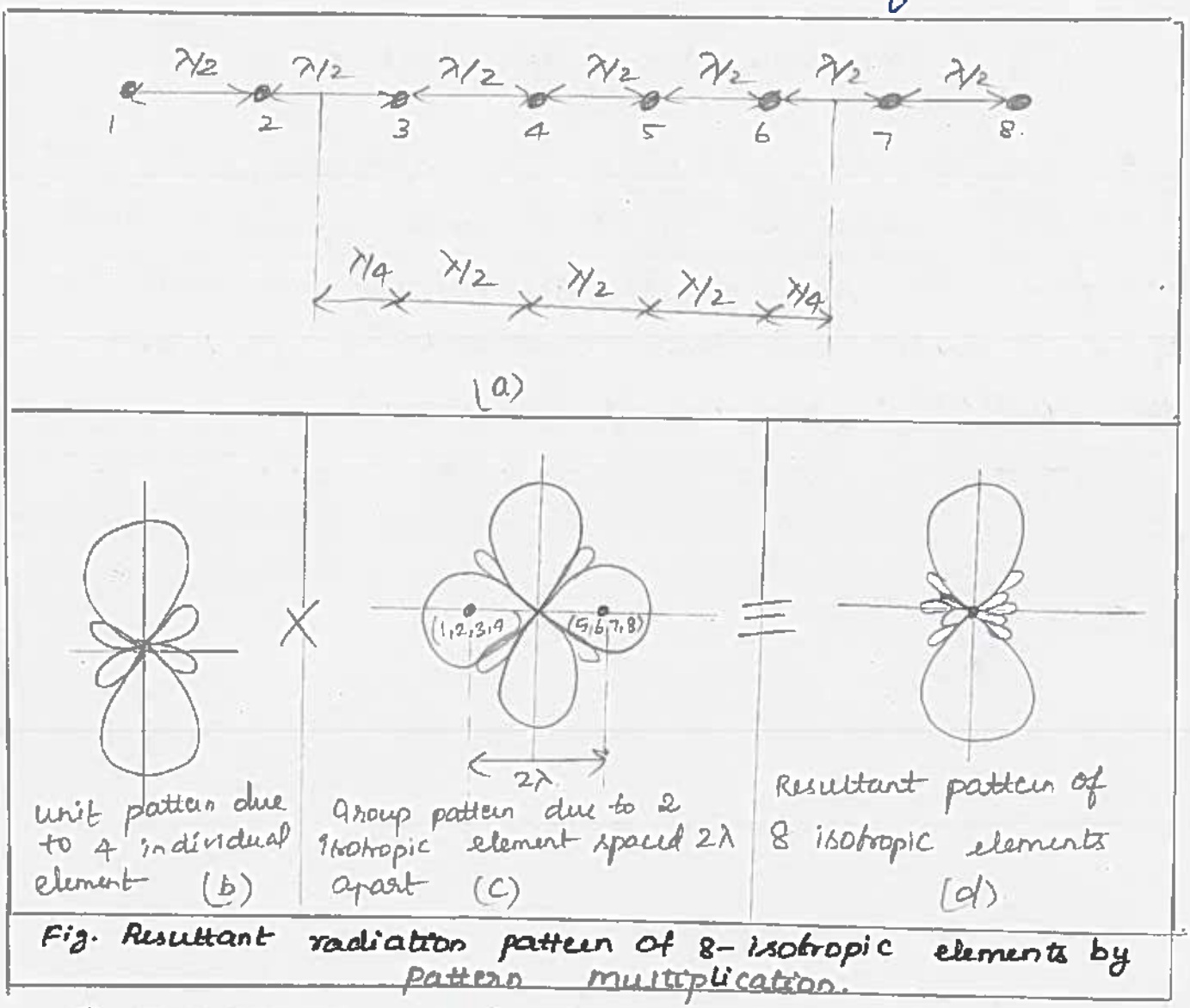


Fig. Resultant radiation pattern of 8-isotropic elements by pattern multiplication.

* Consider four elements (1, 2, 3 and 4) as one unit (1) and another four elements (5, 6, 7 and 8) as unit two (2). Now, the combined 8 elements array has been reduced to 2 elements (units) array spaced at a distance 2λ apart as shown in figure above. The resultant pattern for the 8 element array is obtained as shown in figure.

PHASE ARRAY:

* Phased array means an array of many elements with the phase of each element being a variable that provides the control of the beam direction, that is, maximum radiation in any desired direction and pattern shape including the side lobes.

* The applications for large phased arrays are mostly in the advanced radar systems and in radio astronomy.

* Smaller phased arrays and beam-forming arrays are used as feed systems to illuminate a reflector in satellite communication systems, when it is necessary to provide several spot beams, scanning beams, and/or wide-angle coverage beams from an one-antenna system.

* Some of the specialized phased arrays are
  (i) Frequency Scanning array,
  (ii) Retroarray and
  (iii) Adaptive array.

* In the frequency scanning array or scanning array, phase change is accomplished by varying the frequency. It is one of the simplest phased arrays since no phase control is required at an each element.

* A retro array or self-focussing array is an array that will receive a signal from any direction in space and automatically reflects an incoming signal back toward its source, usually after suitable modulation and amplitude.

Phased Array Designs:

Objectives:

The objectives of the phase array are:

(i) A phased array has to accomplish a beam steering without the mechanical and inertial problems of rotating an entire array and.

(ii) It has to provide a beam control at a fixed frequency (or) at any number of frequencies within a certain bandwidth in the frequency-independent manner.

* In the simplest form of a phased array, beam steering can be done by the mechanical switching. Consider a basic 3-elements array and each element be a $\lambda/2$ dipole antenna as in figure below.

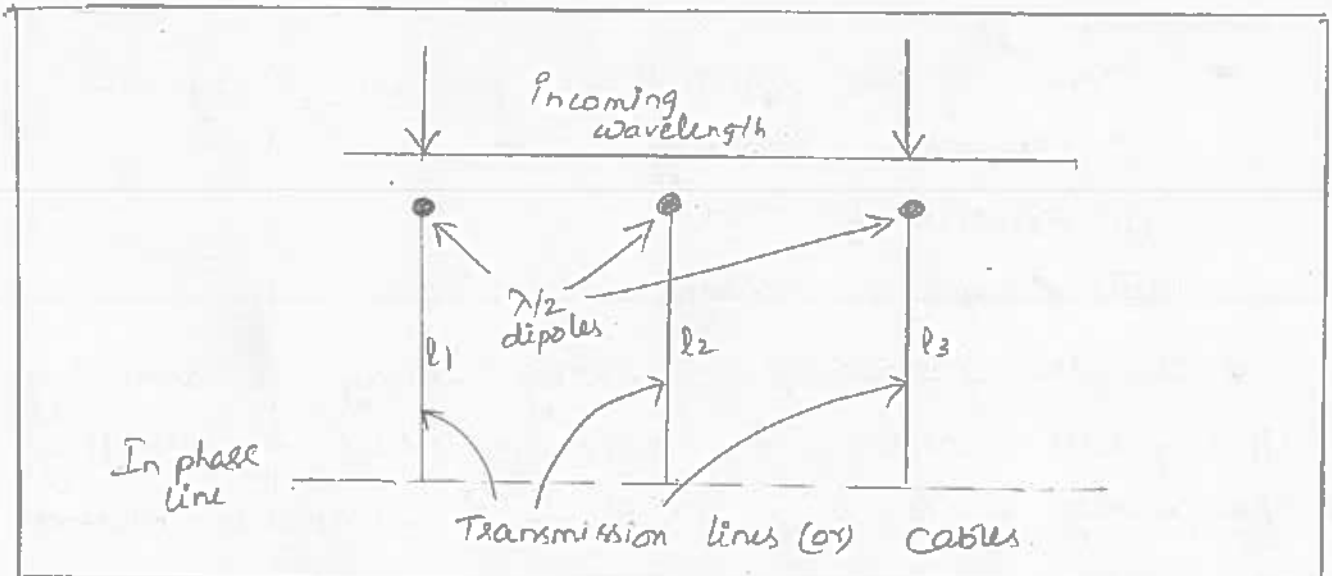


Fig. A simple based array of three $\lambda/2$ dipoles.

* An incoming wave will induce the voltages in all transmission lines in the same phase so that if all cables are of the same length ($l_1 = l_2 = l_3$). Then the voltages will be in phase at the inphase line.

* All three transmission cables are joined as a common point and this 3-element array will be operated as a broadside array.

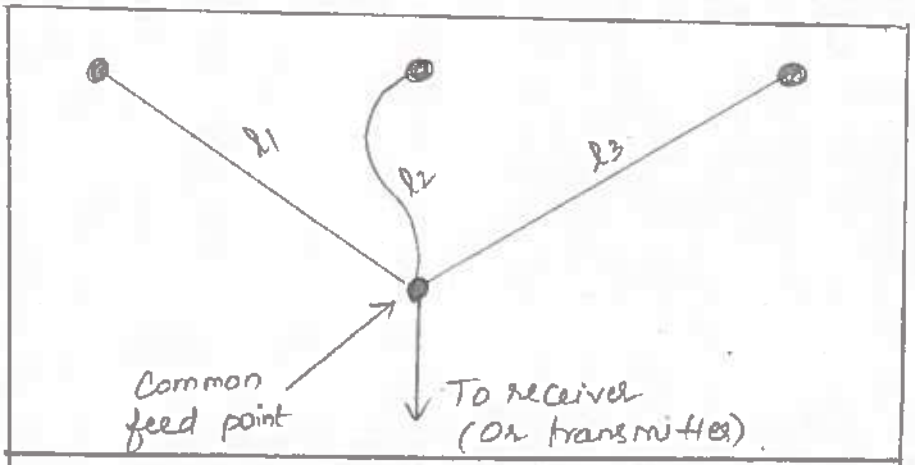
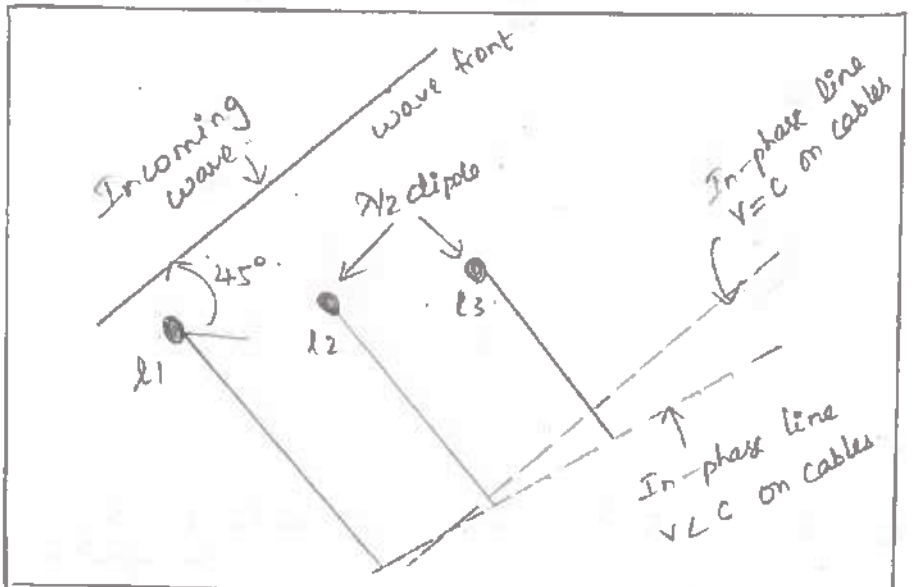


Fig. A Simple equal length cables joined to common point.

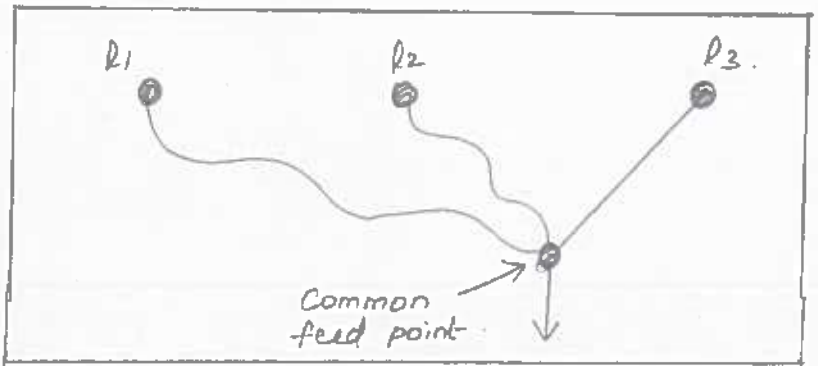
* For an impedance matching, the cable to the receiver or transmitter should be 1/3 of the impedance of these three cables, or a 3 to 1 impedance transformer can be inserted at the common junction point with all the cables of the small impedance.



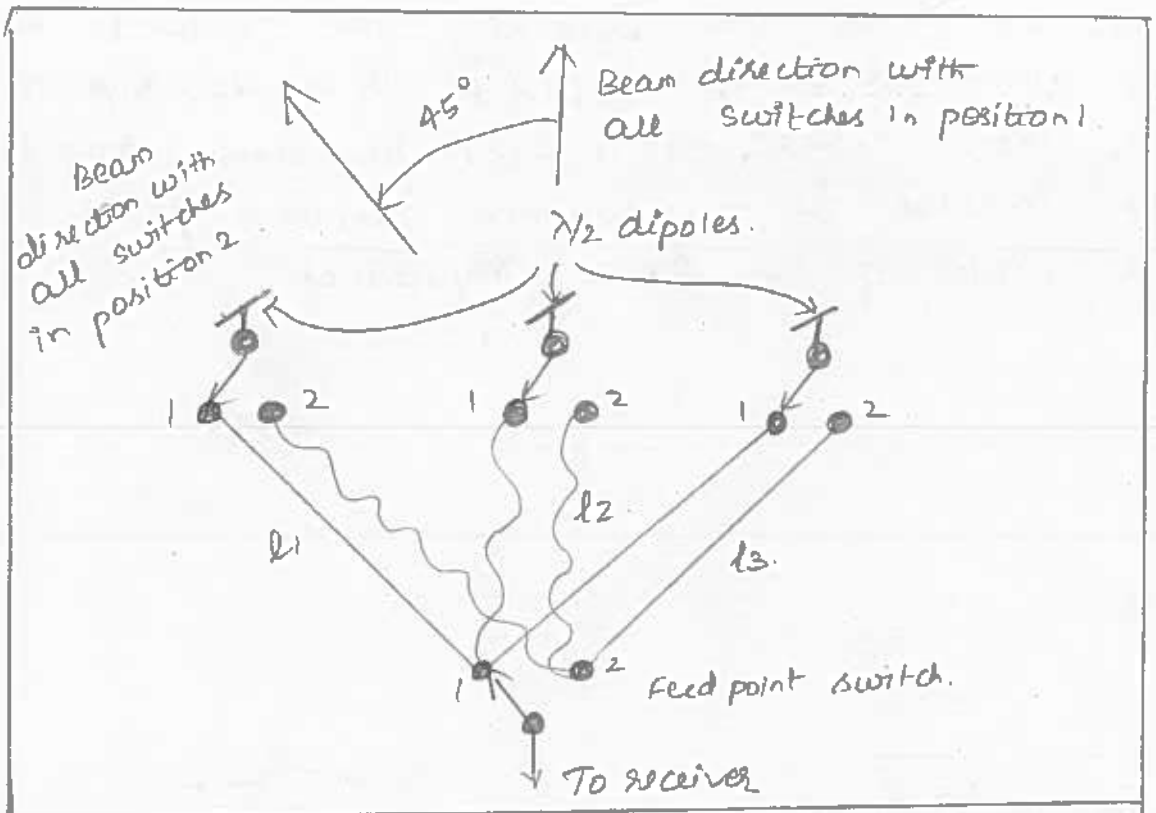
Incoming wave at 45° from broadside array.

* Now, consider a wave arriving at an angle of 45° from broadside as in figure above. If the wave velocity V = C (light velocity = 3×10^{-8} m) on the cables, then the inphase line is parallel to the wave front of the incoming wave.

* If v<c, the length l_2 and l_3 must be increased in order for all phases to be the same. Then the cables of these lengths are joined as shown below and the 3-element array will have its beam 45° from broadside.



Joining on all cables in a common point.



Switches for shifting from broadside to 45° reception

* By installing a switch at each antenna element and one at the common feed point as in above figure and mechanically ganging all the switches together, the beam can be shifted from broadside to 45° by operating the ganged switch.

* By adding more switch points and more cables of an appropriate length, the beam can be steered to an arbitrarily large number of directions. With more elements the narrower beams can be formed.

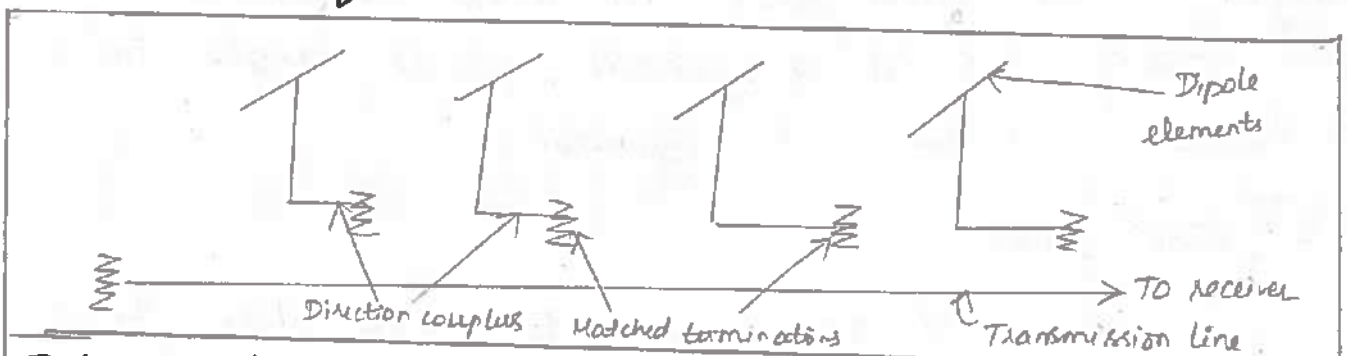
* With diodes (PIN type) in place of mechanical switches, control can be electronics. Computers can do the same thing by an appropriate programming of sampled signals.

* Instead of controlling the beam by switching the cables, a phase shifter can be installed at an each element. Thus, the phase shifting may be accomplished by a ferrite device.

* Insertion of cables of about $\lambda/4$, $\lambda/2$, $3\lambda/4$ by electronics switching provides the phase increments of 90°. For more precise phasing, cables with smaller incremental differences are used.

Different types of feed using phase array.

(i) Coupler feed.



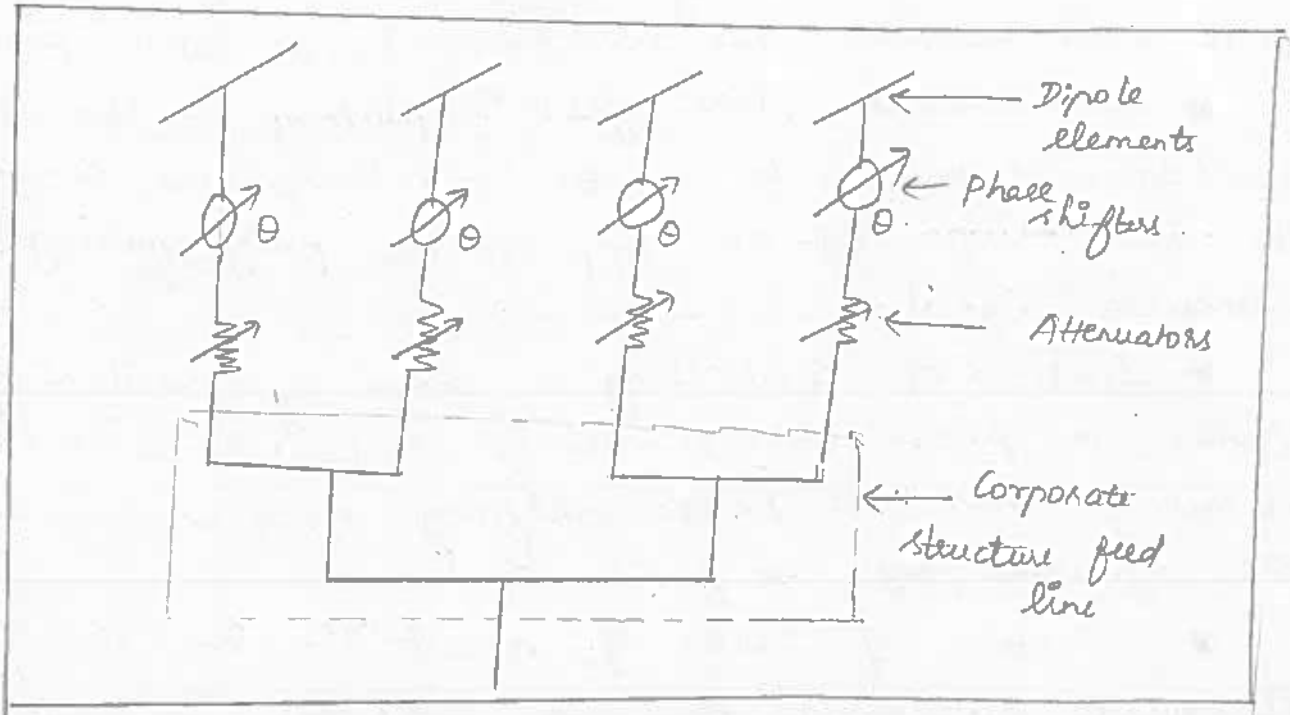
End-fed phase array with directional couplers feed.

* The transmission line has a matched termination for zero reflection so that they (ideally) passes a pure travelling wave on ~~time~~ line.

* Phasing is accomplished by physically sliding the directional couplers along the line. The element amplitude is controlled by changing the closeness of coupling.

~~Different Types of Feed using Phased Array~~.

(ii) Corporate Structure:-

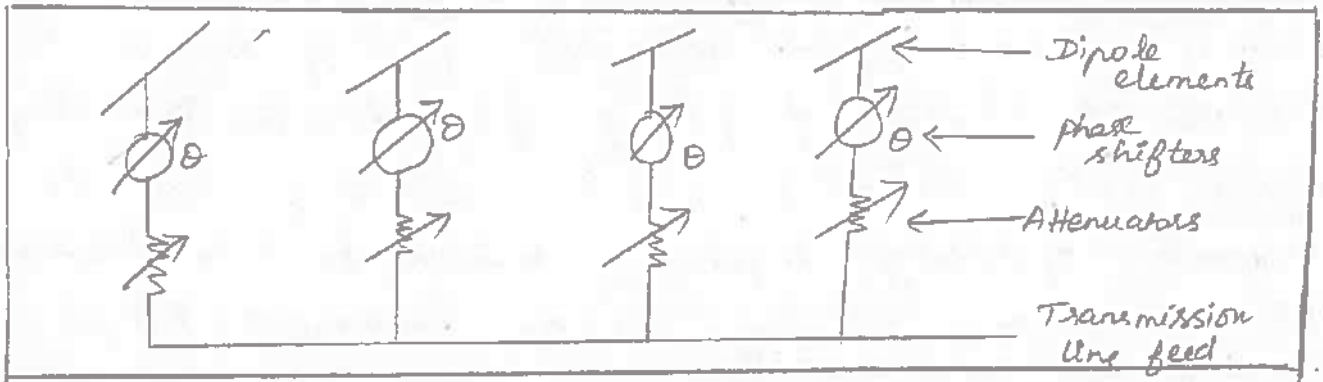


Schematic of phased array feed by corporate structure.

The above figure shows the schematic diagrams of a phased array with a phase shifter and attenuator at each element. The feed cables are all of about equal length in a corporate structure arrangement.

(iii) Line feed:-

All individual elements have phase the shifter and an attenuator. Since, a progressive phase shift is introduced between elements with a frequency change, the phase shifters must introduce an opposing change to compensate in addition to making the desired phase changes.



End-fed phase array with transmission line feed.

ADAPTIVE ARRAYS AND SMART ANTENNAS:

Adaptive arrays:

* Adaptive arrays are arrays that can automatically self-adapt to various incoming signals conditions so as to maximize the signal from a particular source or to null out interfering signals.

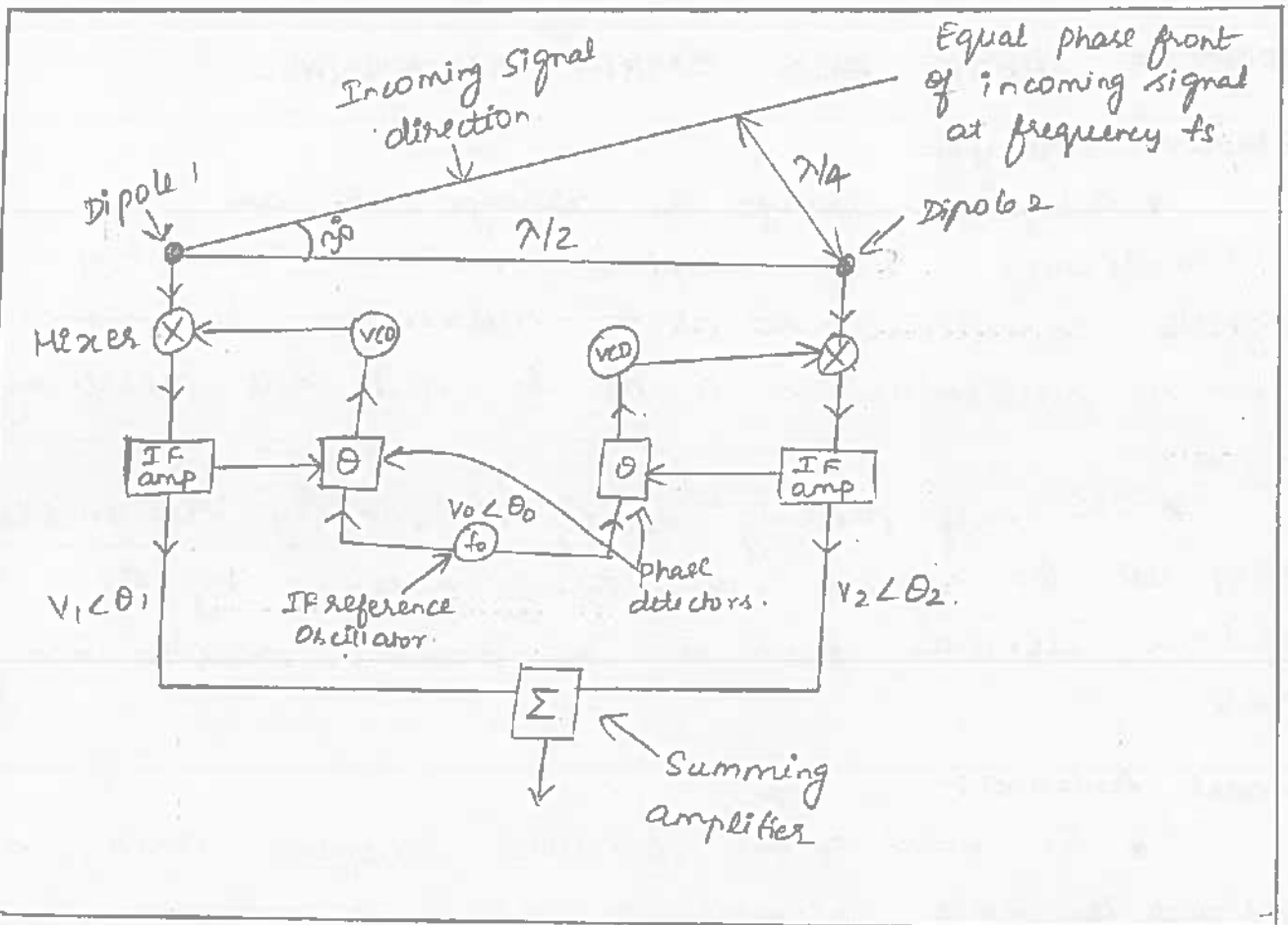
* The self phasing array is mainly designed to bring all the signals which are received by the various elements from a particular source into phase.

Smart Antenna:

* In most of the versatile adaptive array, an output of each element is sampled, digitized and processed by a computer which can be programmed to accomplish the tasks. Such an array may be called as "Smart antenna".

* Multiple beams may be simultaneously directed toward many signals arriving from different directions within the field of an antenna. These antennas are sometimes called as, "Digital Beam Forming (DBF) antennas".

* Considering a simple 2-element adaptive array as shown in below figure with $\lambda/2$ spacing between the elements at signal frequency f_s. Now the incoming signal arrive at 30° from broadside that is all the elements operating in phase and the beam in broadside and the wave arriving at an element 2 travels $\lambda/4$ further than to element 1, thus retarding the phase of the signal by 90° at element 2.



Two-element adaptive array with signal processing circuit

* The phase detector compares the phase of the downshifted signal with the phase of reference oscillator and produces an error voltage Vo with a magnitude proportional to the phase error between them that advances or retards the phase of VCO output so as to reduce the phase difference to zero. These objectives are achieved by using PPL principles.

* Now, all the signals at an intermediate frequency (IF) are in phase and may be added together in a summing amplifier.

* The voltage for the VCO of element 1 would ideally be equal in magnitude but of an opposite sign to the voltage for the VCO of element 2 so that the downshifted signals from both elements are locked in phase (i.e).

$$\theta_1 = \theta_2 = \theta_0.$$

where

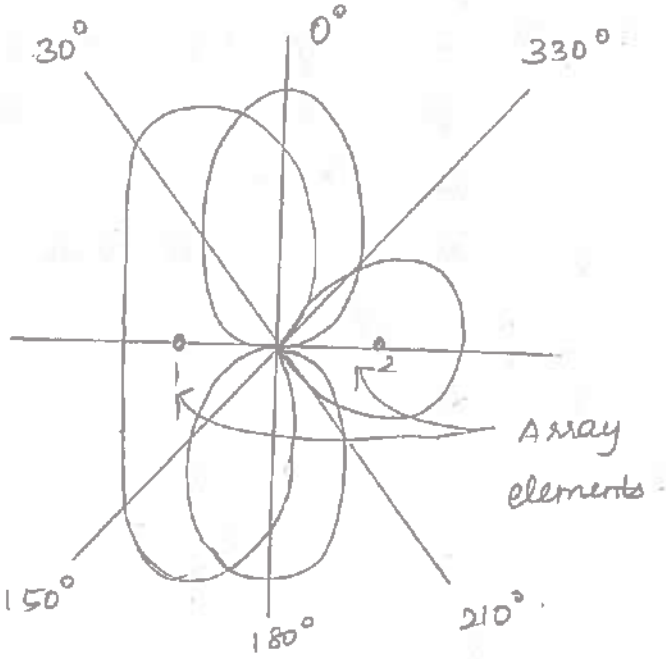
$\theta_1 \rightarrow$ phase of downshifted signal from elements 1.

$\theta_2 \rightarrow$ phase of downshifted signal from elements 2

$\theta_0 \rightarrow$ phase of reference oscillator.

* An equal gain from both IF amplifiers and the voltages V_1 and V_2 from both the elements should be equal so that,

$$V_1 \angle \theta_1 = V_2 \angle \theta_2.$$



Pattern of 2-element adaptive array for signals from 0° and 30°.

* The output from the summing amplifier is $2V_1$ $(=2V_2)$ and maximizing the response of an array to the incoming signal by steering the beam onto the incoming signal. For example, the beam will be in the 0° direction for a signal from the 0° direction and at 30° for a signal from that direction.

* For the 0° signal, nulls are at 90° and 270° while for the 30° signal, nulls are at 210° and 330°.

NON- UNIFORM EXCITATION AMPLITUDES; BINOMIAL ARRAYS

Need of Binomial Array:

The binomial array is needed for the following reasons:

(i) When the uniform linear array length is increased to increase the directivity, at that time secondary (or) minor lobes also appear along the desired radiation pattern.

(ii) In some of the special applications, it is desired to have single main lobe with no minor lobes. That means the minor lobes should be eliminated completely or reduced to minimum level as compared to main lobe because considerable amount of power is wasted in this directions.

* To reduce the sideband level, we use binomial array which deals with the non-uniform amplitude of elements. Here the amplitudes of the radiating sources are arranged to the coefficients of successive terms of binomial series.

The binomial series is

$$(a+b)^{(n-1)} = a^{(n-1)} + \frac{n-1}{1!} a^{(n-2)} \cdot b + \frac{(n-1)(n-2)}{2!} a^{(n-3)} \cdot b^2 + \frac{(n-1)(n-2)(n-3)}{3!} a^{(n-4)} b^3 + \ldots \rightarrow (1)$$

Where $n \rightarrow$ Number of radiating sources in the array.

Concepts of Binomial array:

* If an array is arranged in such a way that radiating sources are in the centre of the broadside array radiates more strongly than the radiating sources at the edges. The secondary lobes can be eliminated entirely, when the following two conditions are satisfied.

(i) The space between the two consecutive radiating sources does not exceed $\lambda/2$ and

(ii) The current amplitudes in radiating sources (from outer, towards centre source) are proportional to the coefficients of the successive terms of the binomial series.

* The above conditions are necessarily satisfied in the binomial arrays and the coefficients which corresponds to the ~~coefficients of the successive terms of the~~ Amplitude of the sources are obtained by putting $n = 1, 2, 3, \ldots$ in eqn (1).

* For example the relative amplitudes for the arrays of 1 to 10 radiating sources are given as follows:-

Number of sources	Relative Amplitude.
n=1	1
n=2	1 1
n=3	1 2 1
n=4	1 3 3 1
n=5	1 4 6 4 1
n=6	1 5 10 10 5 1
n=7	1 6 15 20 15 6 1
n=8	1 7 21 35 35 21 7 1
n=9	1 8 28 56 70 56 28 8 1
n=10.	1 9 36 84 126 126 84 36 9 1

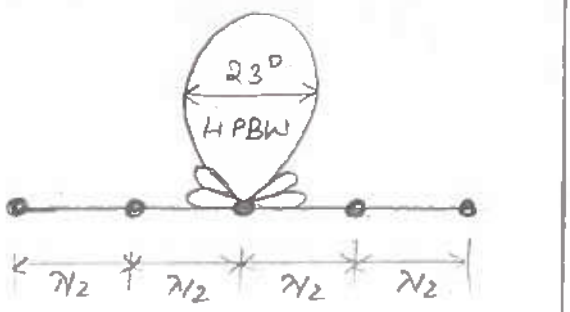
* These coefficients for any number of radiating sources can also be obtained from Pascal's triangle, where each internal Integer is the sum of above adjacent Integers.

PASCAL'S TRIANGLE.

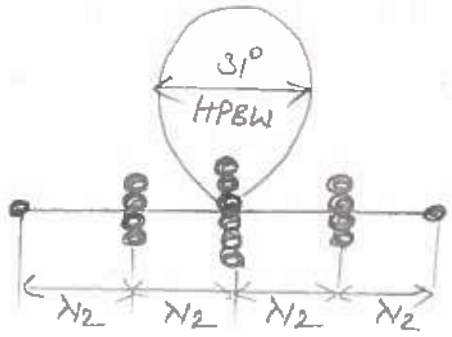
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
1 8 28 56 70 56 28 8 1
1 9 36 84 126 126 84 36 9 1.

* However the elimination of secondary lobes in a binomial array takes place in the cost of directivity. HPBW of binomial array is more that of uniform array for the same length of an array.

* For example, consider $n=5$, $d=\lambda/2$, HPBW of binomial array is 31° and HPBW of an uniform array is 23° as shown in figure.



(a) uniform array



(b) Binomial array with amplitude ratio 1:4:6:4:1

* If we reduce the spacing between the two elements to one half wavelength then only the primary lobes will obtain. The resultant pattern can be obtained by using the concept of the pattern multiplication. In general, the far field pattern for the binomial array of 'n' sources are

$$E_{nor} = \cos^{(n-1)}\left(\frac{\pi}{2}\cos\theta\right) \longrightarrow (2)$$

Disadvantages:

Disadvantages of binomial arrays are.

(i) HPBW increases and hence the directivity decreases.

(ii) For the design of a large array, the larger amplitude ratio of sources is required.

UNIT- IV

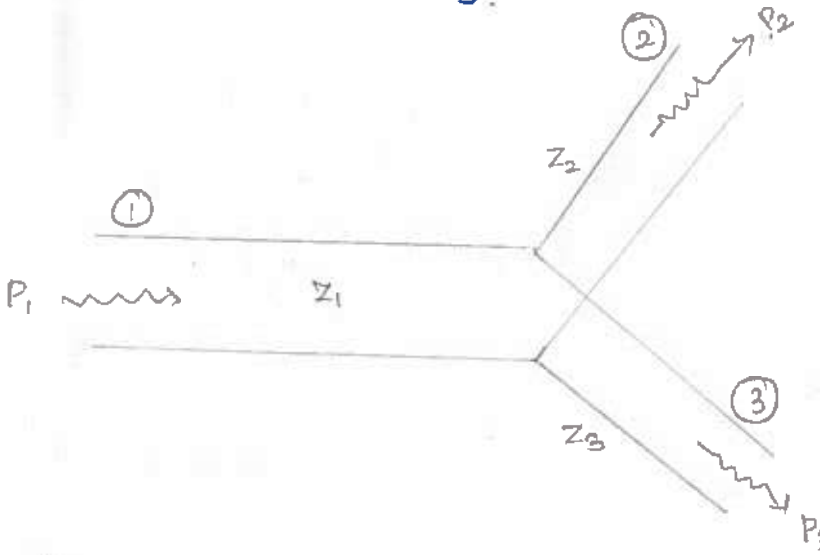
PASSIVE AND ACTIVE MICROWAVE DEVICES.

THREE PORTS JUNCTIONS (T - JUNCTIONS)

* A simple power divider is a T- junction network. When we are using as power divider, the port 1 acts as an input port and ports 2 and 3 are act as an output ports. The power is divided among the ports equally and it is expressed in terms of lossless as,

Input power = Output power

$$P_1 = P_2 + P_3.$$



* When used as power combiner, the port 1 is act as an output port and the ports 2 and 3 are act act as an input ports and it is expressed as

$$P_2 + P_3 = P_1$$

* The scattering matrix for an arbitrary three port network is given as

$$[S] = \begin{bmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{bmatrix} \longrightarrow (1)$$

For a lossless, the reciprocal three port junction is one where all three ports can be perfectly matched then $S_{11} = S_{22} = S_{33} = 0$ and its scattering matrix will be symmetric. Now the scattering matrix is given by,

$$[S] = \begin{bmatrix} 0 & S_{12} & S_{13} \\ S_{12} & 0 & S_{23} \\ S_{13} & S_{23} & 0 \end{bmatrix} \longrightarrow (2)$$

The scattering matrix of a reciprocal four port network that all the ports are perfectly matched and it is expressed as,

$$[S] = \begin{bmatrix} 0 & S_{12} & S_{13} & S_{14} \\ S_{12} & 0 & S_{23} & S_{24} \\ S_{13} & S_{23} & 0 & S_{34} \\ S_{14} & S_{24} & S_{34} & 0 \end{bmatrix} \longrightarrow (3)$$

Applications :-

* Used in the radiating elements of an array antenna.
* Used in the balanced amplifiers both as power dividers and power combiners.

FOUR PORTS NETWORKS : DIRECTIONAL COUPLERS

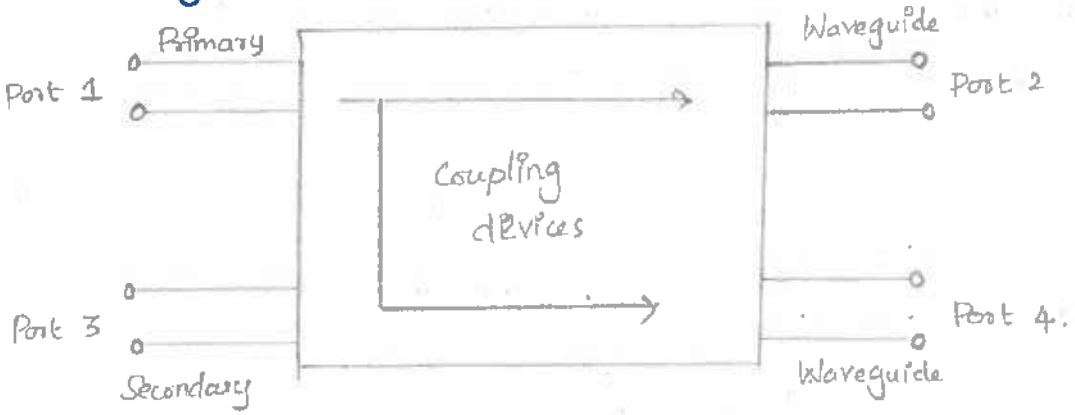
Definition :

A directional coupler is a four port passive device commonly used for coupling a known fraction of the micro wave power to a port in an auxiliary line while the power is flowing from an input port to an output port in the main line. The remaining port is ideally isolated port and matched terminated.

* Here, portions of the forward and reverse travelling waves on a line are separately coupled to two of the other ports.

* They can be designed to measure an incident and reflected power, SWR values, provide a signal path to a receiver or perform the other desirable operations.

* They can be unidirectional or bi-directional powers.

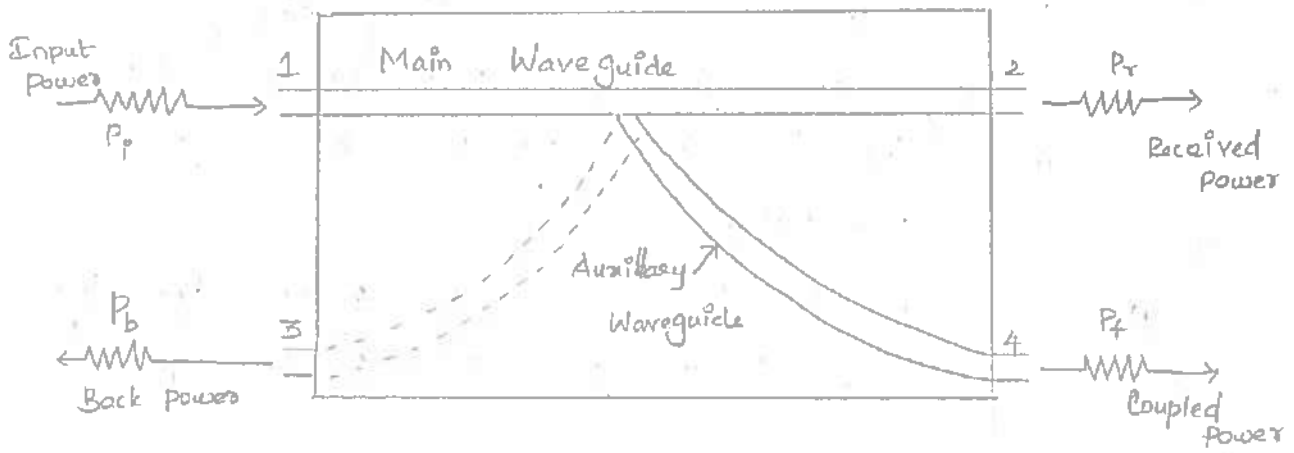


Properties :-

* A portion of power travelling from port 1 and port 2 is coupled to port 4, not to port 3.

* A portion of power travelling from port 2 and port 1 is coupled to port 3, not to port 4.

* A portion of power incident on port 3 is coupled to port 2 but not to port 1 and a portion of power incident on port 4 is coupled to port 1 but not to port 2. Also ports 1 and 3 are decoupled as ports 2 and 4.



Coupling factor (C) :-

* It is defined as the ratio of an incident power to the forward power P_f which is measured in dB.

$$C\ (dB) = 10 \log_{10} \left(\frac{P_i}{P_f}\right)$$

* The coupling factor is a measure of how much of an incident power is being sampled.

Directivity (D) :-

* It is defined as, the ratio of forward power to the backward power P_b as expressed in dB.

$$D\ (dB) = 10 \log_{10} \frac{P_f}{P_b}$$

* Directivity is a measure of how well the directional coupler distinguishes between the forward and reverse travelling powers.

Isolation (I) :-

* It is defined as the ratio of an incident power to the back power as expressed in dB.

$$I\ (dB) = 10 \log_{10} P_i/P_b$$

* The term Isolation is sometimes used to describe the directive properties of a coupler. Isolation (dB) equals coupling plus directivity.

Isolation (I) = Coupling factor (C) + Directivity (D).

Scattering Matrix of a Directional Coupler.

* Directional coupler is a four port network. Hence [S] is 4 x 4 matrix and it is expressed as,

$$[S] = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{41} & S_{42} & S_{43} & S_{44} \end{bmatrix} \longrightarrow (1)$$

* In a directional coupler all the four ports are perfectly matched to the junction. Hence, all the diagonal elements are zero.

$$S_{11} = S_{22} = S_{33} = S_{44} = 0 \longrightarrow (2)$$

* From symmetric property, $S_{ij} = S_{ji}$

$$S_{12} = S_{21},\ S_{23} = S_{32},\ S_{13} = S_{31},\ S_{24} = S_{42},\ S_{34} = S_{43},\ S_{41} = S_{14}. \longrightarrow (3)$$

* There is no coupling between port 1 and port 3.

$$S_{13} = S_{31} = 0 \longrightarrow (4)$$

* Also there is no coupling between port 2 and port 4.

$$S_{24} = S_{42} = 0 \longrightarrow (5)$$

* By substituting (2), (3), (4), (5) in (1),

$$[S] = \begin{bmatrix} 0 & S_{12} & 0 & S_{14} \\ S_{12} & 0 & S_{23} & 0 \\ 0 & S_{23} & 0 & S_{34} \\ S_{14} & 0 & S_{34} & 0 \end{bmatrix} \longrightarrow (6)$$

By applying an unity property of [S] matrix for (6), we then get,

$$[S][S^*] = I$$

$$\begin{bmatrix} 0 & S_{12} & 0 & S_{14} \\ S_{12} & 0 & S_{23} & 0 \\ 0 & S_{23} & 0 & S_{34} \\ S_{14} & 0 & S_{34} & 0 \end{bmatrix} \begin{bmatrix} 0 & S_{12}^* & 0 & S_{14}^* \\ S_{12}^* & 0 & S_{23}^* & 0 \\ 0 & S_{23}^* & 0 & S_{34}^* \\ S_{14}^* & 0 & S_{34}^* & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R_1 C_1 = |S_{12}|^2 + |S_{14}|^2 = 1 \longrightarrow (7)$$

$$R_2 C_2 = |S_{12}|^2 + |S_{23}|^2 = 1 \longrightarrow (8)$$

$$R_3 C_3 = |S_{23}|^2 + |S_{34}|^2 = 1 \longrightarrow (9)$$

By using zero property of [S] matrix

$$R_1 C_3 = S_{12} S_{23}^* + S_{14} S_{34}^* = 0 \longrightarrow (10)$$

By comparing (7) and (8)

$$|S_{12}|^2 + |S_{14}|^2 = |S_{12}|^2 + |S_{23}|^2$$

$$\boxed{S_{14} = S_{23}} \longrightarrow (11)$$

By (8) and (9)

$$|S_{12}|^2 + |S_{23}|^2 = |S_{23}|^2 + |S_{34}|^2$$

$$\boxed{S_{12} = S_{34}} \longrightarrow (12)$$

Let us assume that S_{12} is real and positive = 'P'

$$S_{12} = S_{34} = P = S_{34}^* \longrightarrow (13)$$

Sub (13) and (11) in (10)

$$S_{12} S_{23}^* + S_{14} S_{34}^* = 0 \qquad [\because S_{14} = S_{23}]$$

$$P(S_{23} + S_{23}^*) = 0$$

$$S_{23} + S_{23}^* = 0$$

$$\boxed{S_{23} = -S_{23}^*} \longrightarrow (14)$$

From (14), it is clear that S_{23} must be imaginary

$$S_{23} = jq$$

$$S_{23}^* = -jq$$

From (11) & (12)

$$S_{12} = S_{34} = P \longrightarrow (15a)$$

$$S_{13} = S_{14} = jq \longrightarrow (15b)$$

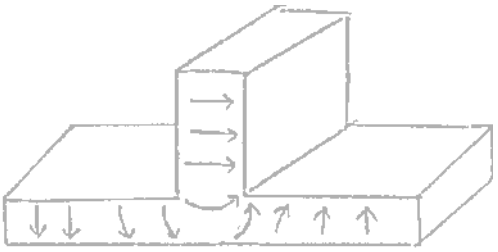
By substituting (15a) & (15b) in (7),

$$P^2 + q^2 = 1 \longrightarrow (16)$$

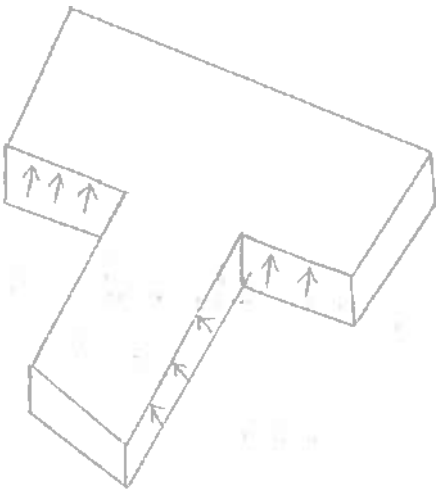
By substituting (15a) and (15b) in (6) then [S] matrix of a directional coupler is reduced to

$$[S] = \begin{bmatrix} 0 & P & 0 & jq \\ P & 0 & jq & 0 \\ 0 & jq & 0 & P \\ jq & 0 & P & 0 \end{bmatrix} \longrightarrow (17)$$

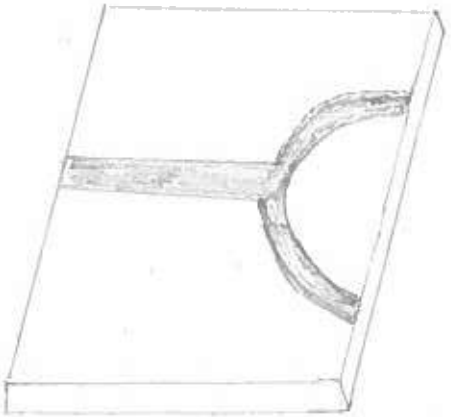
T - JUNCTION POWER DIVIDER.



(a) E-plane Waveguide T



(b) H - Plane waveguide T



(c) Microstrip line T-junction divider

* The T-junction power divider is a simple three port network that can for power division or power combining and it can be implemented in virtually any type of transmission line medium.

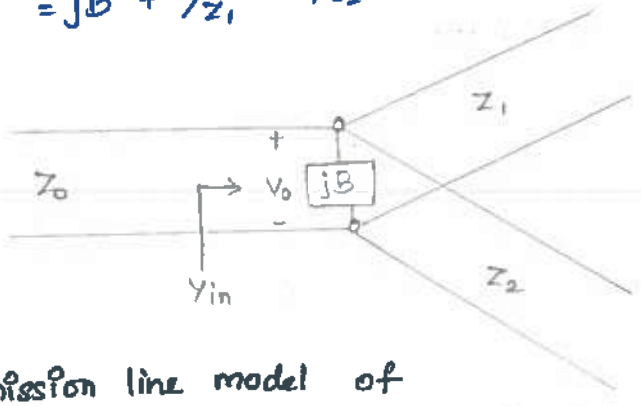
* Figure shows some commonly used T-junctions in waveguide and microstrip line or stripline from which are in the absence of transmission line loss, lossless junctions.

Lossless Divider

* In general, there may be fringing fields and higher order modes associated with the discontinuity at such a junction, leading to the stored energy that can be accounted for by a lumped susceptance, B.

* In order for the divider to be matched to an input line of characteristic impedance Z_0, we must have

$$Y_{in} = jB + 1/Z_1 + 1/Z_2 = 1/Z_0 \longrightarrow (1).$$



Transmission line model of a lossless T-junction divider.

* If the transmission lines are assumed to be lossless, then the characteristic impedance are real. If we also assume B=0, then (1) reduces to

$$\frac{1}{Z_1} + \frac{1}{Z_2} = \frac{1}{Z_0} \longrightarrow (2)$$

* The output line impedances Z_1 and Z_2 can be selected to provide various power division ratios. Thus for a 50Ω input line, a 3dB power divider can be made by using the two 100Ω output lines.

* If necessary, quarter - wave transformers can be used to bring an output line impedances back to the desired levels. If the output lines are matched, then an input line will be matched and there will be no isolation between the two output ports.

Resistive Divider.

* If a three port divider contains lossy components, it can be made to be matched at all ports although the two output ports may not be isolated.

* The resistive divider can easily be analyzed using circuit theory. Assuming that all ports are terminated in the characteristic impedance Z_0, then the impedance Z, seen looking into the $Z_0/3$ resistor followed by a terminated output line is

$$Z = \frac{Z_0}{3} + Z_0 = \frac{4Z_0}{3} \longrightarrow (3)$$

* Then input impedance of the divider is

$$Z_{in} = \frac{Z_0}{3} + \frac{2Z_0}{3} = Z_0 \longrightarrow (4)$$

* (4) shows that an input is matched to the feed line. Because the network is symmetric from all three ports, the output ports are also matched. Thus $S_{11} = S_{22} = S_{33} = 0$.

* If the voltage at port 1 is V_1, then by voltage division the voltage V at the center of the junction is given as

$$V = V_1 \frac{2Z_0/3}{\frac{Z_0}{3} + \frac{2Z_0}{3}} = \frac{2}{3} V_1$$

the output voltages by using voltage division are obtained as

$$V_2 = V_3 = V \frac{Z_0}{Z_0 + \frac{Z_0}{3}} = V \times \frac{Z_0}{\frac{3Z_0 + Z_0}{3}} = V \times \frac{3}{4} \longrightarrow (6)$$

Sub (5) in (6)

$$V_2 = \frac{3}{4} \times \frac{2}{3} V_1 = \frac{1}{2} V_1 \longrightarrow (7)$$

Thus $S_{21} = S_{31} = S_{23} = \frac{1}{2}$, so output powers are 6 dB below an input power level. The network is reciprocal, so the scattering matrix is symmetric but not a unitary matrix and it can be written as

$$[S] = \frac{1}{2}\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \longrightarrow (8)$$

* The power delivered to the input of the divider is

$$P_{in} = \frac{1}{2}\frac{V_1^2}{Z_0} \longrightarrow (9)$$

while the output powers are

$$P_2 = P_3 = \frac{1}{2}\frac{\left(\frac{1}{2}V_1\right)^2}{Z_0}$$

$$= \frac{1}{8}\frac{V_1^2}{Z_0} = \frac{1}{4}P_{in} \longrightarrow (10)$$

* Equations (10) represents that half of the supplied power is dissipated in the resistors.

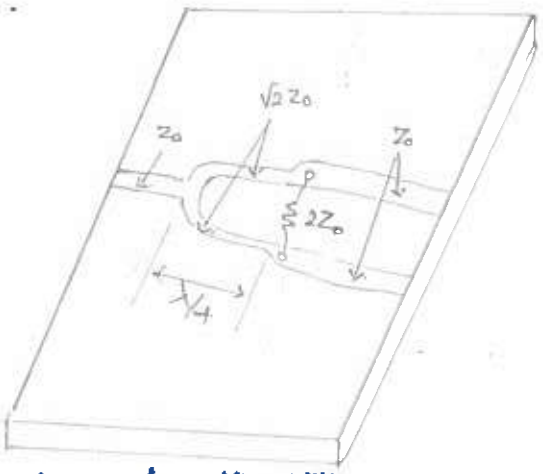
WILKINSON POWER DIVIDER.

(i) The lossless T-junction divider is not being matched at all the ports and it does not have any isolation between output ports.

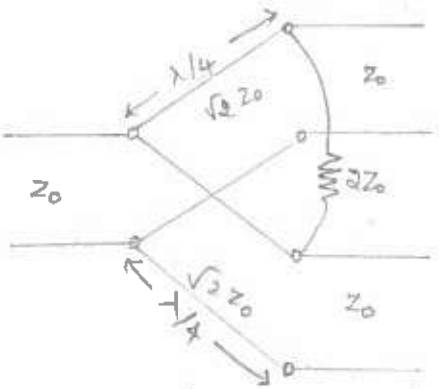
(ii) The resistive divider can be matched at all ports, but even though it is not lossless, isolation is still not achieved.

* The wilkinson power divider is a network with the useful property of an appearing lossless when the output ports are matched; that is only the reflected power from the output ports is dissipated.

* The wilkinson power divider can be made with the arbitrary power division which is often made in microstrip line or Stripline form, as depicted in figure.

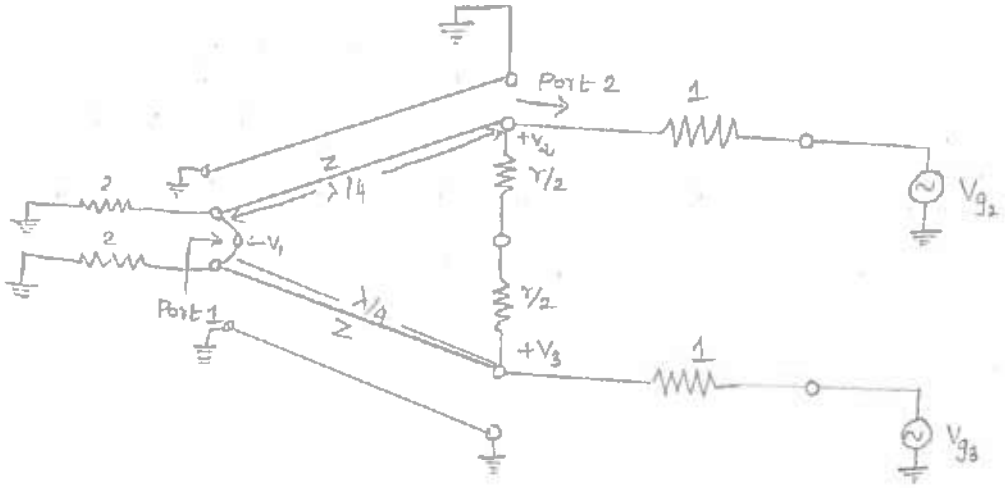


(a) An equal-split Wilkinson power divider in microstrip line form



(b) Equivalent transmission line circuit.

Even odd Mode Analysis.



* This network has been drawn in a form that is symmetric across the midplane; the two source resistors of a normalised value 2 which is combine in parallel to give a resistor of normalized value 1, representing the impedance of a matched source.

* The quarter wave lines have a normalized characteristic impedance Z, and the shunt resistor has a normalized value of r. For the equal-split power divider, ~~with~~. These values should be $Z = \sqrt{2}$ and $r = 2$.

* Now we define two separate modes of excitation for the circuit.

(i) Even mode : $V_{g_2} = V_{g_3} = 2V_0$

(ii) Odd mode : $V_{g_3} = -V_{g_3} = 2V_0$

* Superposition of these two modes effectively produces an excitation which can find by using scattering parameters of the network is $V_{g_2} = 4V_0$ and $V_{g_3} = 0$.

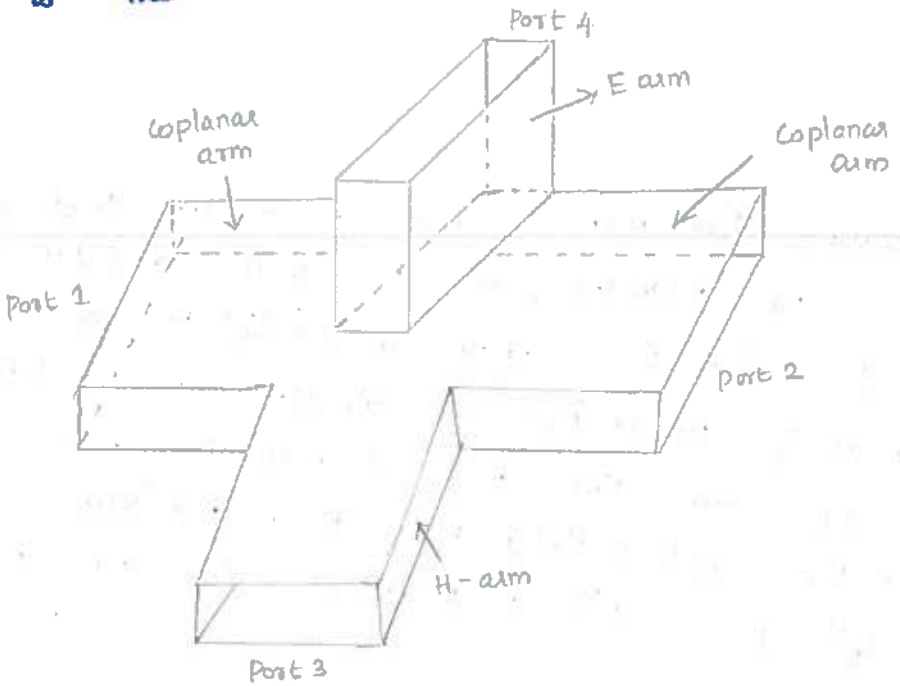
Drawback :-

This divider requires crossovers for their resistors with $N \geq 3$, which makes the fabrication difficult in planar form.

WAVE GUIDE MAGIC - T

* A hybrid junction is a four - port network in which a signal incident on any one of the port divides between two output ports with the remaining port being isolated.

* A magic tee is a combination of the E-plane tee and H - plane tee. Ports 1 and 2 are collinear arms, port 3 is the H-arm and port 4 is the E - arm.



Characteristics of Magic Tee.

* The magic - T has the following characteristics when all the ports are terminated with matched load.

(i) If two in phase waves of equal magnitude are fed into ports 1 and 2, the output at port 4 is subtractive and

hence zero and the total output will appear additively at port 3. Hence, port 4 is called the difference (or) E-arm and port 3 is the sum (or) H-arm.

* A wave incident at port 4 divides equally between the ports 1 and 2 but opposite in phase with no coupling to port 3.

* A wave incident at port 3 divides equally between the ports 1 and 2 are in phase with no coupling to port 4.

$$S_{43} = S_{34} = 0 \longrightarrow (1)$$

* A wave fed into one collinear port 1 or 2 will not appear in the other collinear port 2 or 1. Hence two collinear ports 1 and 2 are isolated from each other

$$S_{12} = S_{21} = 0 \longrightarrow (2)$$

* A magic T can be matched by putting screws suitably in the E and H arms without destroying the symmetry of the junction. For an ideal, lossless magic-T matched at ports 3 and 4.

$$S_{33} = S_{44} = 0 \longrightarrow (3)$$

S-Matrix for magic Tee.

* [S] is a 4×4 matrix since there are 4 ports,

$$[S] = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{41} & S_{42} & S_{43} & S_{44} \end{bmatrix} \longrightarrow (4)$$

* From symmetric property $S_{ij} = S_{ji}$

$$S_{12} = S_{21},\ S_{13} = S_{31},\ S_{14} = S_{41},\ S_{23} = S_{32},\ S_{24} = S_{42},\ S_{34} = S_{43} \longrightarrow (5)$$

* Port 3 has H-plane tee section,

$$S_{23} = S_{13} \longrightarrow (6)$$

Ily, Port 4 has E-plane tee section

$$S_{24} = -S_{14} \longrightarrow (7)$$

* By sub ①, ③, ⑤, ⑥ and ⑦ in ④,

$$[S] = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{12} & S_{22} & S_{13} & -S_{14} \\ S_{13} & S_{13} & 0 & 0 \\ S_{14} & -S_{14} & 0 & 0 \end{bmatrix} \longrightarrow (8)$$

Using unitary property on eqn (8),

$$[S][S^*] = I$$

$$\begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{12} & S_{22} & S_{13} & -S_{14} \\ S_{13} & S_{13} & 0 & 0 \\ S_{14} & -S_{14} & 0 & 0 \end{bmatrix} \begin{bmatrix} S_{11}^* & S_{12}^* & S_{13}^* & S_{14}^* \\ S_{12}^* & S_{22}^* & S_{13}^* & -S_{14}^* \\ S_{13}^* & S_{13}^* & 0 & 0 \\ S_{14}^* & -S_{14}^* & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R_1C_1 = |S_{11}|^2 + |S_{12}|^2 + |S_{13}|^2 + |S_{14}|^2 = 1 \longrightarrow (9)$$

$$R_2C_2 = |S_{12}|^2 + |S_{22}|^2 + |S_{13}|^2 + |S_{14}|^2 = 1 \longrightarrow (10)$$

$$R_3C_3 = |S_{13}|^2 + |S_{13}|^2 = 1 \longrightarrow (11)$$

$$R_4C_4 = |S_{14}|^2 + |S_{14}|^2 = 1 \longrightarrow (12)$$

By equating (9) and (10)

$$|S_{11}|^2 + |S_{12}|^2 + |S_{13}|^2 + |S_{14}|^2 = |S_{12}|^2 + |S_{22}|^2 + |S_{13}|^2 + |S_{14}|^2$$

$$|S_{11}| = |S_{22}| \longrightarrow (13)$$

From (11),

$$|S_{13}|^2 + |S_{13}|^2 = 1$$

$$2|S_{13}|^2 = 1$$

$$\boxed{|S_{13}| = 1/\sqrt{2}} \longrightarrow (14)$$

From (12)

$$|S_{14}|^2 + |S_{14}|^2 = 1$$

$$2|S_{14}|^2 = 1$$

$$\boxed{|S_{14}| = 1/\sqrt{2}} \longrightarrow (15)$$

* By sub (14) and (15) in (9)

$$|S_{11}|^2 + |S_{12}|^2 + \frac{1}{2} + \frac{1}{2} = 1$$

$$|S_{11}|^2 + |S_{12}|^2 = 0$$

which is valid if, $S_{11} = S_{12} = 0 \longrightarrow (16)$

* From (13) and (16),

$$\boxed{S_{22} = 0} \longrightarrow (17)$$

* The [S] of magic tee is obtained by substituting the Scattering parameters from (16) and (17) in (8)

$$[S] = \begin{bmatrix} 0 & 0 & S_{13} & S_{14} \\ 0 & 0 & S_{13} & -S_{14} \\ S_{13} & S_{13} & 0 & 0 \\ S_{14} & -S_{14} & 0 & 0 \end{bmatrix} \longrightarrow (18)$$

* Using (14), (15)

$$[S] = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{bmatrix} \longrightarrow (19)$$

where

$$|S_{13}| = \frac{1}{\sqrt{2}} = |S_{14}|$$

* Hence in any four ports junction, if any two ports are perfectly matched to the junction, then the remaining two ports are automatically matched to the junction. Such a junction where in all the four ports are perfectly matched to the junction is called a magic tee.

Applications:-

* Measurement of impedance
* As duplexer
* As mixer
* As an isolator.

ATTENUATORS :-

* An attenuator is basically a passive device which controls the amount of microwave power transferred from one point to another without causing a big distortion to its waveform on a microwave transmission system. It result in decreasing the power level of a microwave signal.

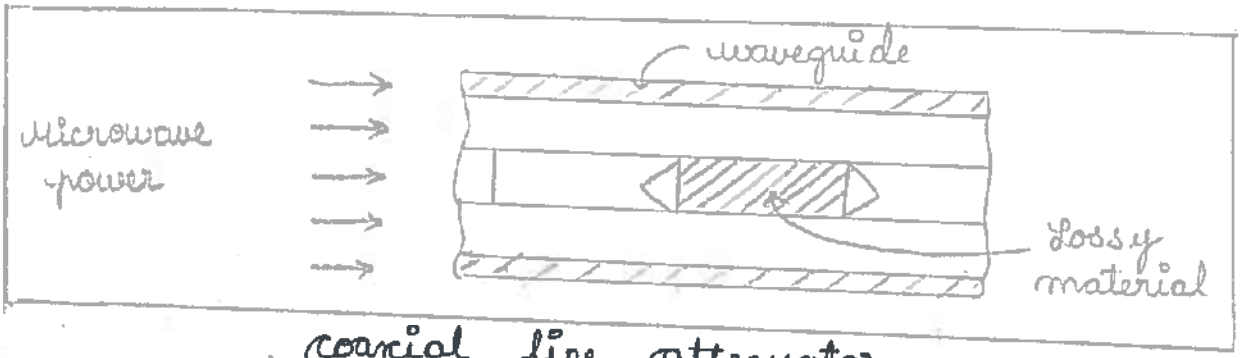
* Microwave attenuators control the flow of microwave power either reflecting it or absorbing it and it is expressed in decibels of the relative power.

* Attenuator which attenuates the RF signal in a waveguide system is referred as a waveguide attenuator.

There are two main types namely,

* fixed attenuator
* variable attenuator

Fixed attenuator :-



coaxial line attenuator

* Fixed attenuators are used where a fixed amount of attenuation is to be provided. If such a fixed attenuators absorbs all the energy entering into it and it is called as waveguide terminator.

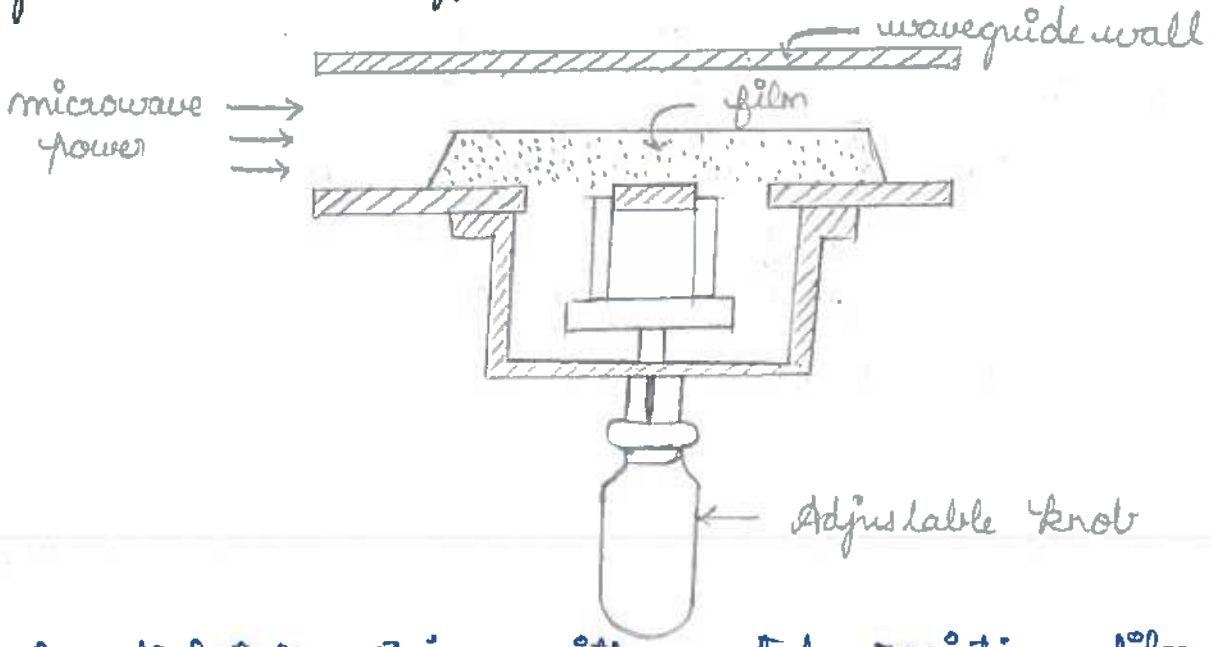
* The coaxial line based fixed type of attenuator, Here, resistive film is fixed at the centre of a conductor which absorbs the power and as a result of power loss, that is microwave signal passes through it also gets attenuated.

Variable attenuator :-

* Variable attenuator provide continuous or step wise variable attenuation. For rectangular waveguides, these attenuators can be flap type or vane type. For circular waveguides, the rotary type is normally used.

* The most commonly used variable attenuators are,
  * waveguide variable type
  * Rotary - vane attenuators

waveguide variable type :-



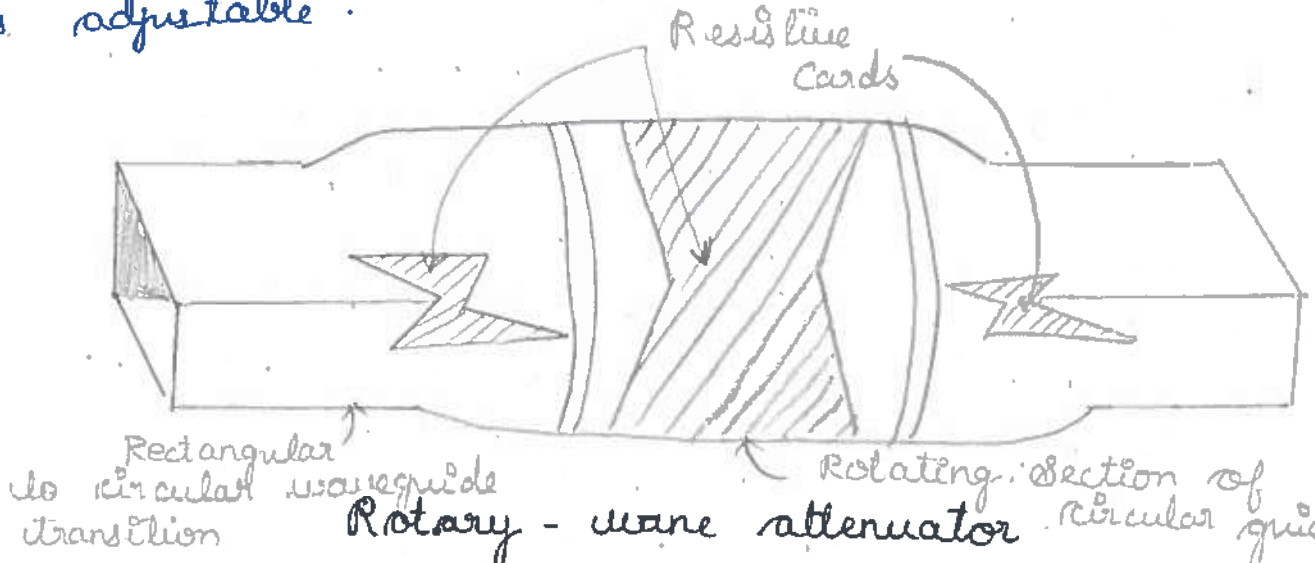
* Thin dielectric strip with coated resistive film is placed at the centre of the waveguide. Thus film is placed in the waveguide which is parallel to the maximum E - field.

* Resistive vane is moved from one side of the wall to the centre by using a screw where 'E' field is considered to be maximum. This resistive film is shaped to give a linear attenuation variation.

Rotary - vane attenuator :-

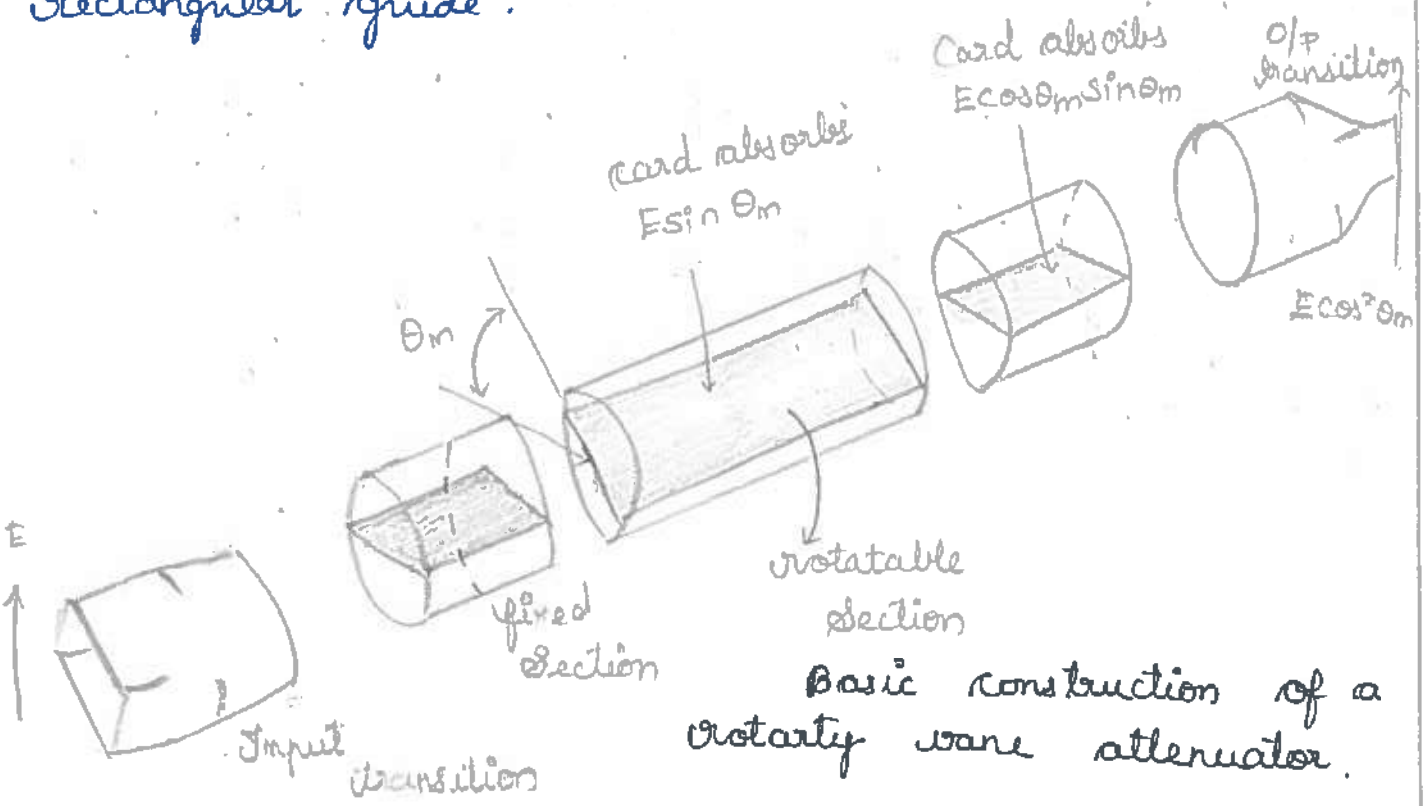
* Rotary - vane attenuator is so accurate that it is used as a calibration standard in most of the microwave laboratories

* Rotary - vane attenuator is a simple form of attenuator consists of a thin tapered resistive card, whose depth of penetration waveguide is adjustable.



Rotary - vane attenuator

* Rotary vane attenuator has three circular waveguide section, two fixed and one rotatable It also includes input and output transitions that provides low SWR connections to standard rectangular guide.



Basic construction of a rotarty vane attenuator.

* The attenuation is controlled by rotation of the center section. Here an attenuation is a function of rotation angle θ_m only. The minimum loss occuring with $\theta_m = 0$ and the maximum loss occurs when $\theta_m = 90^\circ$. The principle of operation is based upon an interaction between the plane-polarized waves and the thin resistive cards.

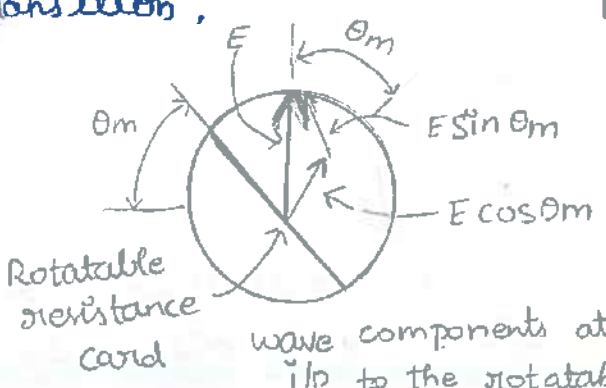
* The input transition is which converts the TE_{10} wave into a vertically polarized TE_{11} wave in a circular guide. The electric field associated with this wave is denoted by E.

* In the first fixed section, the resistive card perpendicular to an electric field and the wave propagates without any loss.

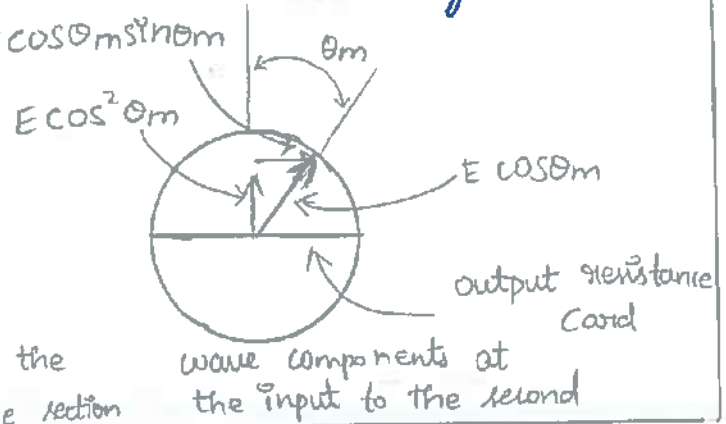
* When the card in the rotable section is horizontal ($\theta_m = 0$), then the wave passes through it and also to an output fixed section without any loss. Thus for $\theta_m = 0$, the total loss is 0dB.

* For any other angle, the component which is parallel to the rotable card ($E\sin\theta_m$) is absorbed and the perpendicular component ($E\cos\theta_m$) arrives at the second fixed section with its polarization at an angle of θ_m with respect the vertical plane.

* The portion of the wave that is parallel to the output card ($E\cos\theta_m\sin\theta_m$) is absorbed, while the perpendicular components ($E\cos^2\theta_m$) proceeds to an output port via the circular to rectangular transition.



wave components at the i/p to the rotatable section



wave components at the input to the second fixed section

MICROWAVE RESONATOR :-

* Microwave resonators are tunable circuits which are used in microwave oscillators, amplifiers, wavemeters and filters.

* At the tuned frequency the circuit resonates where the average energies stored in an electric field, W_e and magnetic field, W_m are equal and the circuit impedance becomes purely real.

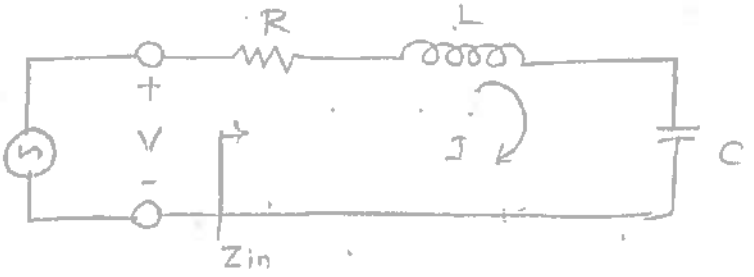
Quality factor Q which is a measure of the frequency selectivity of a cavity and it is defined as.

$$Q = \frac{2\pi \times \text{Maximum Energy Stored}}{\text{Energy dissipated per cycle.}}$$

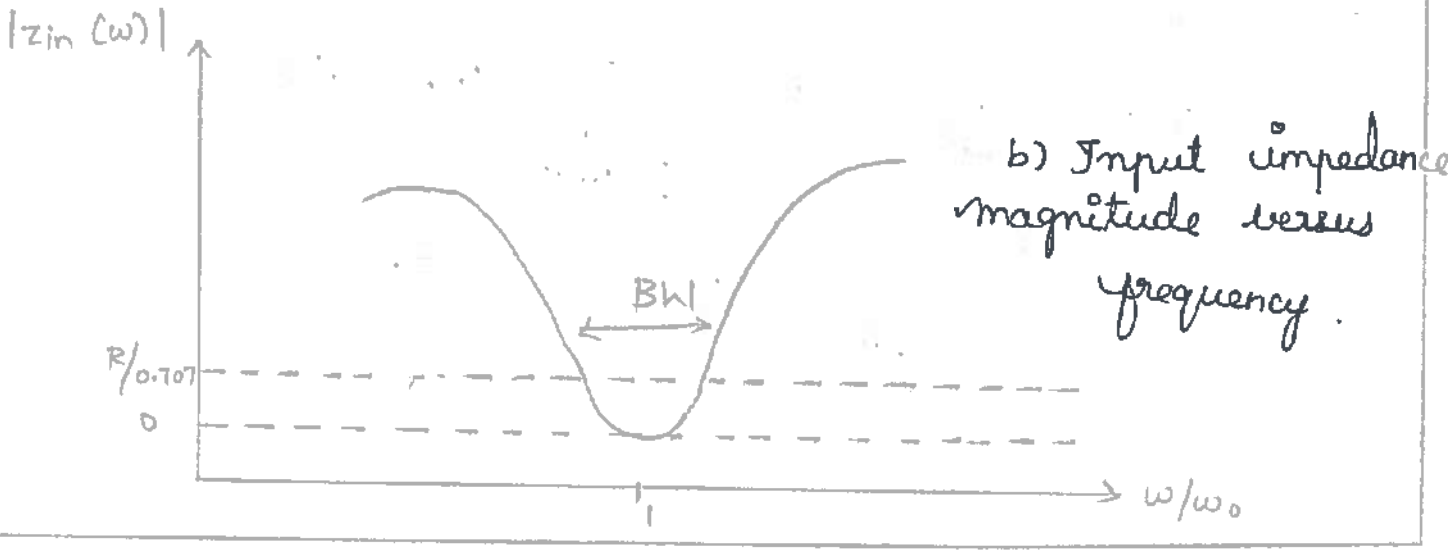
i) Series resonant circuits :-

At frequencies near resonance, a microwave resonator can usually be modeled by either a series or parallel RLC lumped-element equivalent circuit.

$$Z_{in} = R + j\omega L - j\frac{1}{\omega C} \longrightarrow (1)$$



a) A series RLC resonator circuit



b) Input impedance magnitude versus frequency.

* The complex power delivered to the resonator is given as.

$$P_{in} = \frac{1}{2} VI^* = \frac{1}{2} Z_{in} |I|^2 = \frac{1}{2} Z_{in} \left|\frac{V}{Z_{in}}\right|^2 \longrightarrow (2)$$

* By substituting eqn ① in eqn. ② we get

$$P_{in} = \frac{1}{2} |I|^2 \left(R + j\omega L - j\frac{1}{\omega C} \right) \longrightarrow (3)$$

* The power dissipated by the resistor R is,

$$P_{loss} = \frac{1}{2} |I|^2 R \longrightarrow (4a)$$

* Average magnetic energy stored in the inductor L is,

$$W_{in} = \frac{1}{4} |I|^2 L \longrightarrow (4b)$$

* Average electric energy stored in the capacitor C is,

$$W_c = \frac{1}{4} |V_c|^2 C = \frac{1}{4} |I|^2 \frac{1}{\omega^2 C} \longrightarrow (4c)$$

where $V_c \rightarrow$ voltage across the capacitor

Then the complex power of (3) can be rewritten as,

$$P_w = P_{loss} + 2j\omega (W_m - W_e) \longrightarrow (5)$$

and the input impedance of (1) can be rewritten as,

$$Z_{in} = \frac{2P_{in}}{|I|^2} = \frac{P_{loss} + 2j\omega (W_m - W_e)}{\frac{1}{2}|I|^2} \longrightarrow (6)$$

$$Z_{in} = \frac{P_{loss}}{\frac{1}{2}|I|^2} = R \longrightarrow (7)$$

* From equation 4b & 4c, $W_m = W_e$ implies that the resonant frequency, ω_0 can be defined as.

$$\omega_0 = 1/\sqrt{LC} \longrightarrow (8)$$

* Quality factor of a resonant circuit is defined as,

$$Q = \omega \frac{\text{Average energy stored}}{\text{Energy loss / second}} = \omega \frac{W_m + W_e}{P_{loss}} \longrightarrow (9)$$

* Thus Q is a measure of the loss of a resonant circuit that is, lower loss implies a higher Q.

* The Q of the resonator itself, disregarding external loading effects, is called as unloaded Q and it is denoted as Q_0.

$$Q_0 = \omega_0 \frac{2W_m}{P_{loss}} = \omega_0 \frac{2 \times \frac{1}{4} |I|^2 L}{\frac{1}{2} |I|^2 R}$$

$$= \frac{\omega_0 L}{R} = \frac{1}{\omega_0 RC} \longrightarrow (10)$$

[Average energy stored = $W_m + W_e$
At resonance, $W_m = W_e$
So, $2W_m$ or $2W_e$]

* Eqn (10) show that Q increases as R decreases, At resonance, bandwidth is

$$B.W = \frac{1}{Q_0} \longrightarrow (11)$$

ii) Parallel Resonant circuits :-

The parallel RLC resonant circuit which is the dual of the series RLC circuit and an input impedance is expressed as

$$Z_{in} = \left(\frac{1}{R} + \frac{1}{j\omega L} + j\omega C\right)^{-1} \longrightarrow (12)$$

The complex power delivered to the resonator is given as

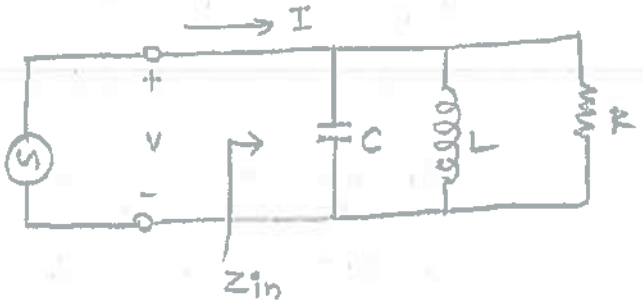
$$P_{in} = \frac{1}{2} V I^* = \frac{1}{2} Z_{in} |I|^2 = \frac{1}{2} |V^2| \frac{1}{Z_{in}^*} \longrightarrow (13)$$

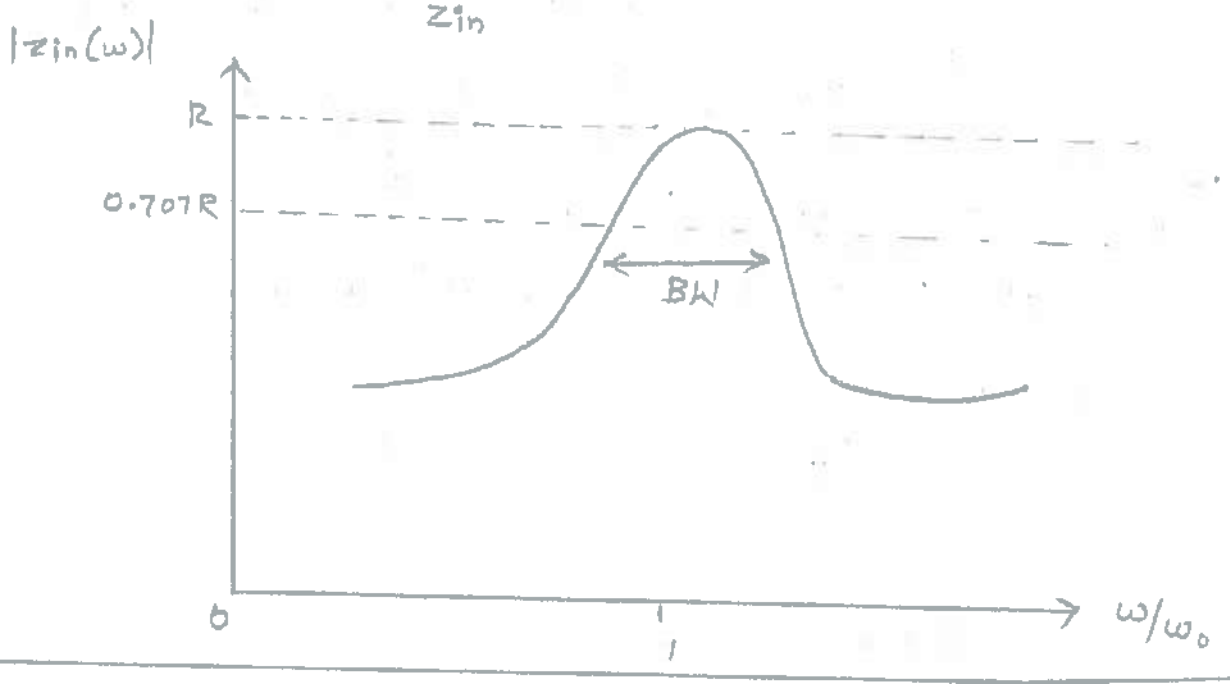
By substituting eqn. (12) in eqn (13) we get

$$P_{in} = \frac{1}{2} |V|^2 \left(\frac{1}{R} + \frac{j}{\omega L} - j\omega C\right) \longrightarrow (14)$$

The power dissipated by the resistor R is,

$$P_{loss} = \frac{1}{2} \frac{|V|^2}{R} \longrightarrow (15a)$$





* The average electric energy stored in the capacitor C is

$$W_e = \frac{1}{4} |V|^2 C \longrightarrow (15b)$$

* The average magnetic energy stored in the inductor L is,

$$W_m = \frac{1}{4} |I_L|^2 L = \frac{1}{4} |V|^2 \frac{1}{\omega^2 L} \longrightarrow (15c)$$

* where I_L is the current through the inductor Then, the complex power from eqn. (14) can be rewritten as.

$$P_{in} = P_{loss} + 2j\omega (W_m - W_e) \longrightarrow (16)$$

* Similarly, the input impedance can be expressed as,

$$Z_{in} = \frac{2 P_{in}}{|I|^2} = \frac{P_{loss} + 2j\omega (W_m - W_e)}{\frac{1}{2} |I|^2} \longrightarrow (17)$$

Resonance occurs when $W_m = W_e$. Then from eqn (17) and (15a), the input impedance at a resonance is expressed as,

$$Z_{in} = \frac{P_{loss}}{\frac{1}{2} |I|^2} \qquad [\because |V|^2 = |I|^2 R^2]$$

$$= \frac{\frac{1}{2} \frac{|V|^2}{R}}{\frac{1}{2} |I|^2} = R \longrightarrow (18)$$

This is a purely real impedance. From eqn. (15b) and eqn. (15c), we get $W_m = W_e$

implies that the resonant frequency ω_0 can be defined as,

$$\omega_0 = \frac{1}{\sqrt{LC}} \longrightarrow (19)$$

$$Q_0 = \omega_0 \frac{2W_m}{P_{loss}}$$

By using (15a) and (15c) in the above eqn. we obtained,

$$= \omega_0 \frac{2 \times \frac{1}{4} |V|^2 \frac{1}{\omega_0^2 L}}{\frac{1}{2} \frac{|V|^2}{R}}$$

$$= \frac{R}{\omega_0 L}$$

$$= \omega_0 RC \longrightarrow (20)$$

Eqn. 20 show that the Q of the parallel resonant circuit increases as R increases.

The behaviour of the magnitude of the input impedance versus frequency. The half-power bandwidth edges occur at frequencies $\left(\frac{\Delta\omega}{\omega_0} = \frac{B.W}{2}\right)$ such that

$$|Z_{in}|^2 = \frac{R^2}{2}$$

which implies that

$$B.W = \frac{1}{Q_0} \longrightarrow (21)$$

Quantity	Series Resonator	Parallel Resonator
Input impedance/ admittance	$Z_{in} = R + j\omega L - j\frac{1}{\omega C}$ $\approx R + j\frac{2RQ_0\Delta\omega}{\omega_0}$	$Y_{in} = \frac{1}{R} + j\omega C - j\frac{1}{\omega L}$ $\approx \frac{1}{R} + j\frac{2Q_0\Delta\omega}{R\omega_0}$
Power Loss	$P_{loss} = \frac{1}{2}\lvert I\rvert^2 R$	$P_{loss} = \frac{1}{2}\frac{\lvert V\rvert^2}{R}$
Stored magnetic energy	$W_m = \frac{1}{4}\lvert I\rvert^2 L$	$W_m = \frac{1}{4}\lvert V\rvert^2\frac{1}{\omega^2 L}$
Stored electric energy	$W_e = \frac{1}{4}\lvert I\rvert^2\frac{1}{\omega^2 C}$	$W_e = \frac{1}{4}\lvert V\rvert^2 C$
Resonant frequency	$\omega_0 = \frac{1}{\sqrt{LC}}$	$\omega_0 = \frac{1}{\sqrt{LC}}$
unloaded Q	$Q_0 = \frac{\omega_0 L}{R} = \frac{1}{\omega_0 RC}$	$Q_0 = \omega_0 RC = \frac{R}{\omega_0 L}$
External Q	$Q_e = \frac{\omega_0 L}{R_L}$	$Q_e = \frac{R_L}{\omega_0 L}$

iii) Loaded and unloaded Q :-

* The unloaded Q that is Q_o is defined as a characteristic of the resonator itself in the absence of any loading effects caused by an external activity.

* A resonator is invariably coupled to an other circuitry, which will have the effect of lowering the overall or loaded Q, that is Q_L of the circuit.

* A resonator coupled to an external load resistor R_L. If the resonator is a series RLC circuit, the load resistor R_L, adds in the series with R so that an effective resistance $R + R_L$.

If the resonator is a parallel RLC circuit the load resistor R_L combines in parallel with R R, so an effective resistance is $RR_L/(R+R_L)$.
A external Q is defined as

$$Q_e = \begin{cases} \frac{\omega_o L}{R_L} & \text{for series circuits} \\ \frac{R_L}{\omega_o L} & \text{for parallel circuits} \end{cases}$$

Then the loaded Q can be expressed as,

$$\frac{1}{Q_L} = \frac{1}{Q_e} + \frac{1}{Q_o}$$

Transmission Line Resonators :-

* Ideal lumped circuit elements are often unattainable the at microwave frequencies, so distributed elements are frequently used
* The different sections of transmission line are used with various lengths and terminations to form resonators.

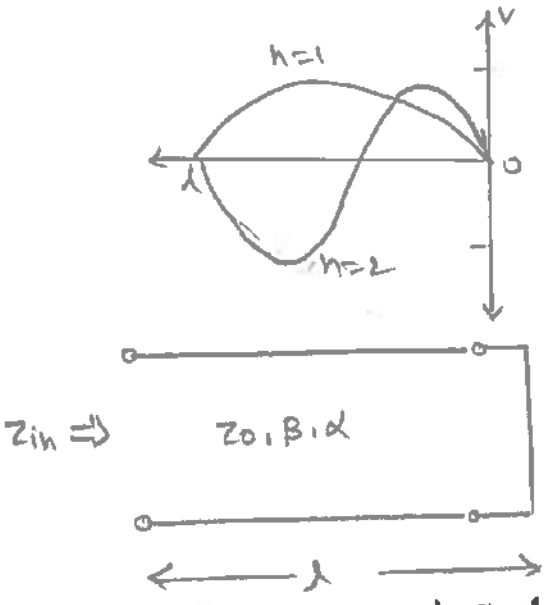
1) Short-Circuited $\lambda/2$ line :-

* A length of lossy transmission line, short circuit at one end is shown in Fig 6-31. The line has a characteristic impedance, Z_0, propagation constant, β, and attenuation constant, α.
* At the resonant frequency $\omega = \omega_0$, the length of the line is $l = \lambda/2$ and an input impedance is expressed as

$$Z_{in} = Z_0 \tanh(\alpha + j\beta)l \longrightarrow (1)$$

using an identity for the hyperbolic tangent and it gives

$$Z_{in} = Z_0 \frac{\tanh\alpha l + j\tan\beta l}{1 + j\tan\beta l \tanh\alpha l} \longrightarrow (1a)$$



Short circuited length of lossy transmission line

* For a lossless line $\alpha = 0$ and $Z_{in} = jZ_0 \tan\beta l$. In practice it usually desirable to use a low-loss transmission line, so we assume that $\alpha l << 1$ and then $\tanh\alpha l \approx \alpha l$. For $\alpha_{coal} l / \omega_0 << 1$, the input impedance becomes

$$Z_{in} = R + 2jL\Delta\omega \longrightarrow (2)$$

where

$$R = Z_0 \alpha l \longrightarrow (3)$$

and inductance of the equivalent circuit as

$$L = \frac{Z_0 \pi}{2\omega_0} \longrightarrow (4)$$

* The capacitance of the equivalent circuit is given as

$$C = \frac{1}{\omega_0^2 L} \longrightarrow (5)$$

* The resonator of thus resonates for $\Delta\omega = 0$ ($l = \lambda/2$) and its input impedance at resonance is $Z_{in} = R = Z_0 \alpha l$

* Resonance also occurs for $l = n\lambda/2$, $n = 1, 2, 3, \ldots$

$$Q_0 = \frac{\omega_0 L}{R} = \frac{\pi}{2\alpha L} = \frac{\beta}{2\alpha} \longrightarrow (6)$$

2) Short-circuited $\lambda/4$ line :-

* The input impedance of a shorted line of length l is expressed as

$$Z_{in} = Z_0 \tanh(\alpha + j\beta) l = Z_0 \frac{\tanh\alpha l + j\tan\beta l}{1 + j\tan\beta l \tanh\alpha l}$$

$$Z_{in} = Z_0 \frac{1 - j\tanh\alpha l \cot\beta l}{\tanh\alpha l - j\cot\beta l} \longrightarrow (7)$$

* For $l = \lambda/4$ at $\omega = \omega_0$, for small loss $\tanh\alpha l \approx \alpha l$ and for $\alpha l \pi \Delta\omega / 2\omega_0 \ll 1$, then an input impedance is expressed as

$$Z_{in} = \frac{1}{(1/R) + 2j\Delta\omega C} \longrightarrow (8)$$

* The resistance of the equivalent circuit is expressed as

$$R = \frac{Z_0}{\alpha l} \longrightarrow (9)$$

capacitance

$$C = \frac{\pi}{4\omega_0 Z_0} \longrightarrow (10)$$

Inductance

$$L = \frac{1}{\omega_0^2 C} \longrightarrow (11)$$

* The Resonator has a parallel-type resonance for $l = \lambda/4$, with an Input Impedance at resonance of $Z_{in} = R = \frac{Z_0}{\alpha l}$.

By using equation (9) and (10) we get

$$= \omega_0 \times \frac{Z_0}{\alpha L} \times \frac{\pi}{4\omega_0 Z_0} = \frac{\pi}{4\alpha l} \longrightarrow (12)$$

At resonance,

$$l = \frac{\pi}{2\beta} = \frac{\pi}{4\alpha \times \frac{\pi}{2}\beta} = \frac{\beta}{2\alpha} \longrightarrow (13)$$

3) Open circuit $\lambda/2$ line :-

* The Input Impedance of an open-circuited lossy transmission line of length is

$$Z_{in} = Z_0 \coth(\alpha + j\beta) l = Z_0 \frac{1 + j \tan\beta l \tanh \alpha l}{\tanh \alpha l + j \tan \beta l} \longrightarrow (14)$$

when $l = \lambda/2$ at $\omega = \omega_0$, at resonance and $\tanh \alpha l \approx \alpha l$

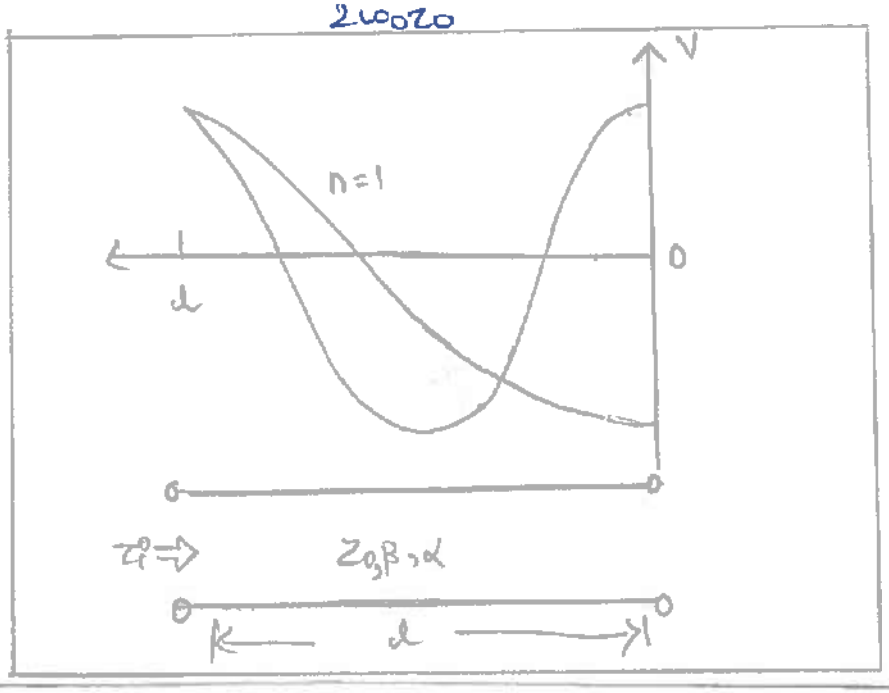
$$Z_{in} = \frac{Z_0}{\alpha l + j(\Delta\omega\pi/\omega_0)} \longrightarrow (15)$$

* The resistance of the equivalent RLC circuit is

$$R = \frac{Z_0}{\alpha l} \longrightarrow (16)$$

and capacitance

$$C = \frac{\pi}{2\omega_0 Z_0} \longrightarrow (17)$$



Inductance of the equivalent

$$L = \frac{1}{\omega_0^2 C} \longrightarrow (18)$$

* From equation (16) and (17), the unloaded Q is obtained as

$$Q_0 = \omega_0 RC = \omega_0 \times \frac{Z_0}{\alpha l} \times \frac{\pi}{2\omega_0 Z_0} = \frac{\pi}{2\alpha l} \longrightarrow (19)$$

At resonance $l = \pi/\beta$ then equation (16) becomes

$$Q_0 = \frac{\beta}{2\alpha} \qquad [\beta = \pi/l] \longrightarrow (20)$$

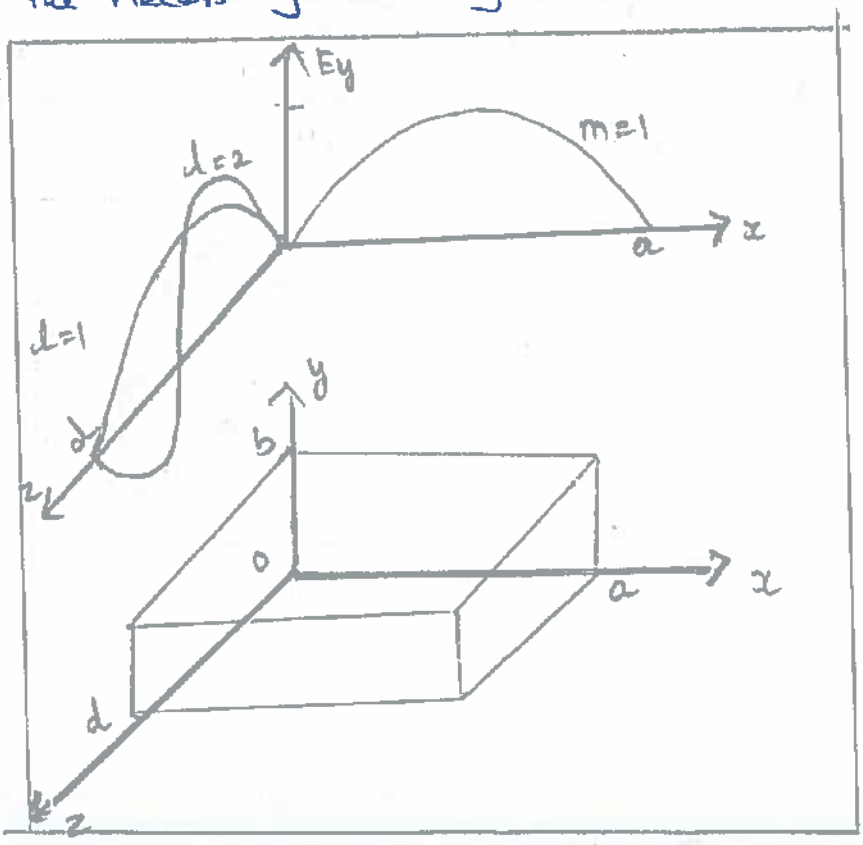
Rectangular waveguide - cavity Resonator:-

* microwave resonators can also be constructed from the closed sections of waveguide. It is because the radiation loss from an open-ended waveguide can be significant.

* Coupling to a cavity resonator may be by a small aperture, or a small probe or loop. The mode having the lowest resonant frequency is known as the dominant mode.

1) Resonant Frequencies:-

* To find the fields of the TE or TM waveguide mode that ~~best~~ satisfy the necessary boundary conditions on the walls of the cavity.



* The transverse electric fields (E_x, E_y) of the TE_{mn} or TM_{mn}, rectangular waveguide mode can be written as

$$\bar{E}_t(x,y,z) = \bar{e}(x,y)\left(A^+ e^{-j\beta_{mn}z} + A^- e^{j\beta_{mn}z}\right) \longrightarrow (1)$$

* where $\bar{e}(x,y)$ is the transverse variation of the mode, and A^+, A^- are arbitrary amplitude of the forward and backward traveling waves

* The propagation constant of the m,n^{th} TE or TM mode is obtained as

$$\beta_{mn} = \sqrt{k^2 - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2} \longrightarrow (2)$$

where $k = \omega\sqrt{\mu\varepsilon}$, and μ and ε are the permeability and permitivity of the material filling the cavity.

* A resonance wave number for the rectangular cavity can be defined as

$$k_{mnl} = \sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 + \left(\frac{l\pi}{d}\right)^2} \longrightarrow (3)$$

In the TM_{mnl} or TE_{mnl} resonant mode of the cavity m,n,l indicate the number of variations in the standing wave pattern in the x,y,z directions respectively

* The resonant frequency frequency of the TE_{mnl} and TM_{mnl} mode is given by

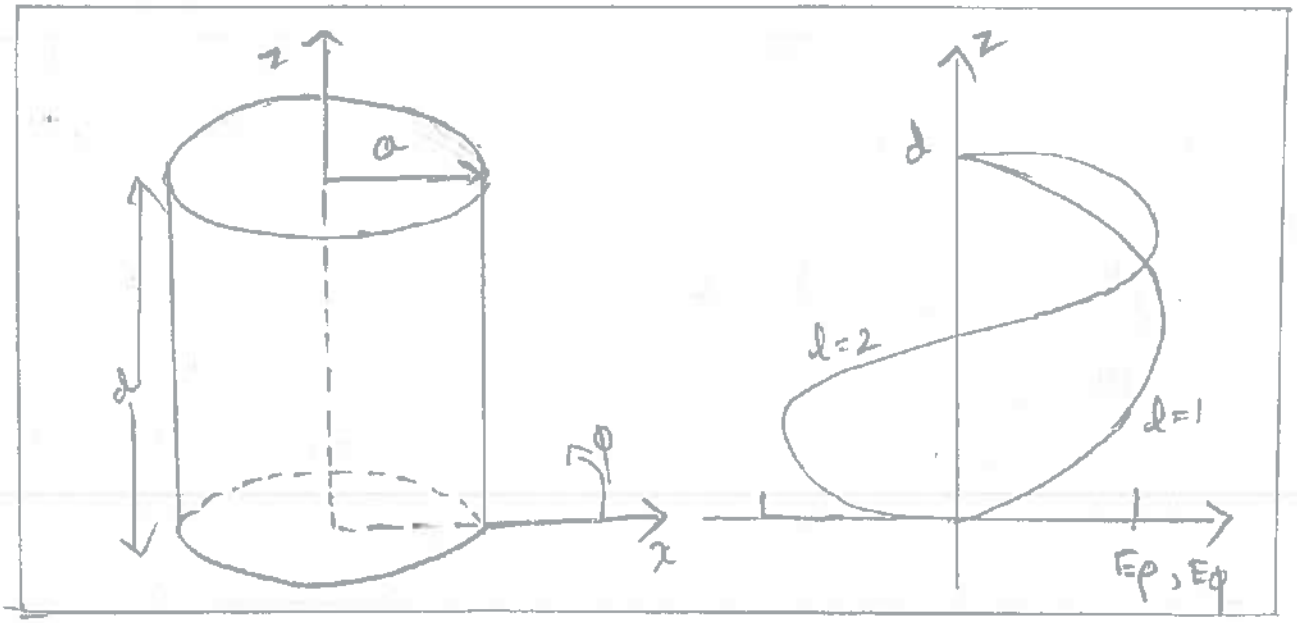
$$f_{mnl} = \frac{ck_{mnl}}{2\pi\sqrt{\mu_r\varepsilon_r}} = \frac{c}{2\pi\sqrt{\mu_r\varepsilon_r}}\sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 + \left(\frac{l\pi}{d}\right)^2} \longrightarrow (4)$$

If $b < a < d$, the dominant resonant mode will be the TE_{101} mode, corresponding to the TE_{10} dominant waveguide mode in a shorted guide of the length $\lambda_g/2$, and it is similar to the short-circuited $\lambda/2$ transmission line resonator. The dominant TM resonant mode is the TM_{110} mode

Circular waveguide _ cavity Resonator:-

* A cylindrical cavity resonator can be constructed from a section of a circular waveguide which is shorted at both ends, similar to rectangular cavities. Because the dominant circular waveguide mode is the TE_{11} mode, the dominant cylindrical cavity mode is the TE_{111} mode

* In operation, power will be absorbed by the cavity as it is tuned to an operating frequency of the system and this absorption can be monitored with a power meter elsewhere in the system



1) Resonant frequencies

* The geometry of a cylindrical cavity is shown. The transverse electric fields (E_ρ, E_ϕ) of the TE_{nm} or TM_{nm} circular waveguide mode can be written as

$$\bar{E}_t(\rho,\phi,z) = \bar{e}(\rho,\phi)\left(A^+ e^{-j\beta_{nm}z} + A^- e^{+j\beta_{nm}z}\right) \longrightarrow (5)$$

where $\bar{e}(\rho,\phi)$ represents the transverse variation of the mode, and A^+ and A^- are arbitrary amplitudes of the forward and backward traveling waves

* The propagation constant of the TE_{nm} mode is

$$\beta_{nm} = \sqrt{k^2 - \left(\frac{p'_{nm}}{a}\right)^2} \longrightarrow (6)$$

* The propagation constant of the TM_{nm} mode is

$$\beta_{nm} = \sqrt{k^2 - \left(\frac{p_{nm}}{a}\right)^2} \longrightarrow (7)$$

where $k = \omega\sqrt{\mu\varepsilon}$

* The resonant frequency of the TE_{nml} mode is

$$f_{nml} = \frac{c}{2\pi\sqrt{\mu_r\varepsilon_r}}\sqrt{\left(\frac{p'_{nm}}{a}\right)^2 + \left(\frac{l\pi}{d}\right)^2} \longrightarrow (8)$$

for $l = 0, 1, 2, 3, \cdots$

and the resonant frequency of the TM_{nml} mode is

$$f_{nml} = \frac{c}{2\pi\sqrt{\mu_r\varepsilon_r}}\sqrt{\left(\frac{p_{nm}}{a}\right)^2 + \left(\frac{l\pi}{d}\right)^2} \longrightarrow (9)$$

PRINCIPLES OF MICROWAVE SEMICONDUCTOR DEVICES:

GUNN DIODE OSCILLATOR:

* Transferred electron oscillator or Gunn diode oscillator makes use of two terminal devices based on the phenomenon known as "transferred electron effect."

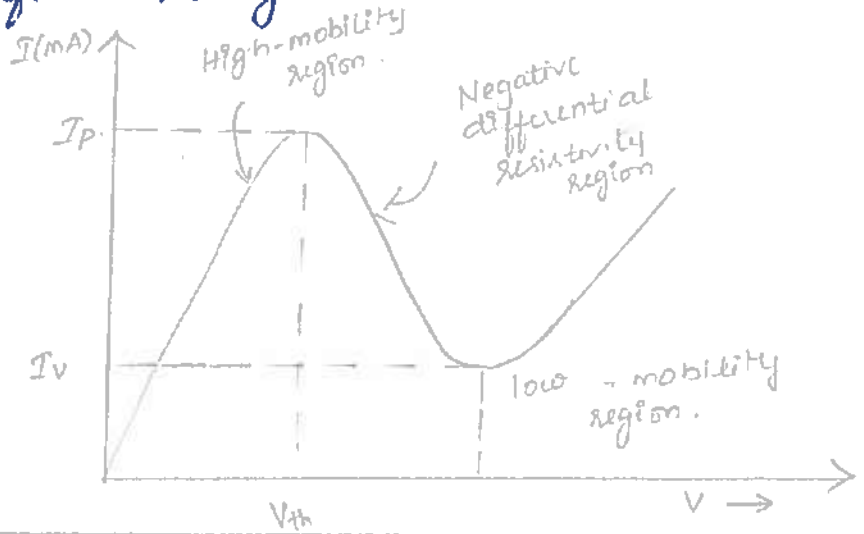
* Gunn diodes are negative resistance devices which are normally used as a low power oscillator at microwave transmitter and also as local oscillator in receivers.

Transferred Electron Effect:

→ Some materials like gallium arsenide (GaAs) exhibits a negative differential mobility (ie a decrease in the carrier velocity with an increase in the electric field) when biased above a threshold value of an electric field.

→ The electrons in the low-energy band will be transferred into the higher energy band. The behaviour is called Transferred electron effect (or) Gunn effect and the device is called Transferred electron device (TED) or transferred electron oscillator or Gunn diode or Gunn oscillator.

* The conductivity is directly proportional to the mobility and hence the current decreases with an increase in electric field strength.



Applications of Gunn Diode:

(i) Gunn diodes are negative resistance devices which are normally used as a low power oscillator at the microwave frequencies in transmitter and also as local oscillator in receiver front ends.

(ii) Used in parametric amplifiers as pump source

(iii) Used in radar transmitters (police radar, CW Doppler radar.

(iv) In broadband microwave amplifiers.

(v) Pulsed Gunn diode oscillators used in transponders for air traffic control (ATC) and in industry telemetry systems.

(vi) Fast combinational and sequential logic circuits

(vii) Low and medium power oscillator in microwave receivers.

* TED's are fabricated from compound semiconductors such as Gallium arsenide (GaAS), Indium phosphide (InP) or Cadmium telluride (CdTe).

* The positive resistances absorb power (passive devices) whereas negative resistances generate power (active devices).

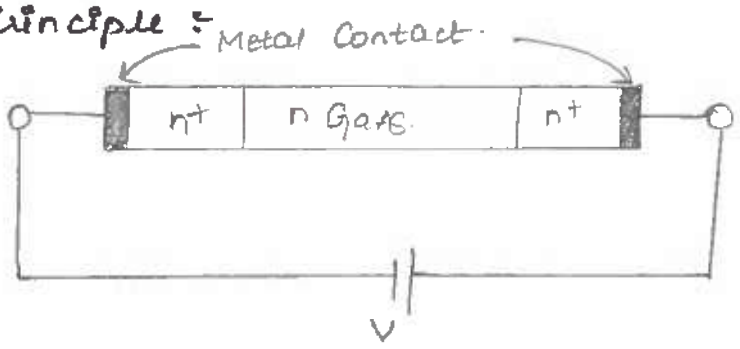
Features of TED's:

(i) TED's are bulk devices without junctions.

(ii) TED's are operate with hot electrons having more thermal energy and

(iii) TED's are tunable over a wide frequency range with low a noise characteristics.

Working Principle :

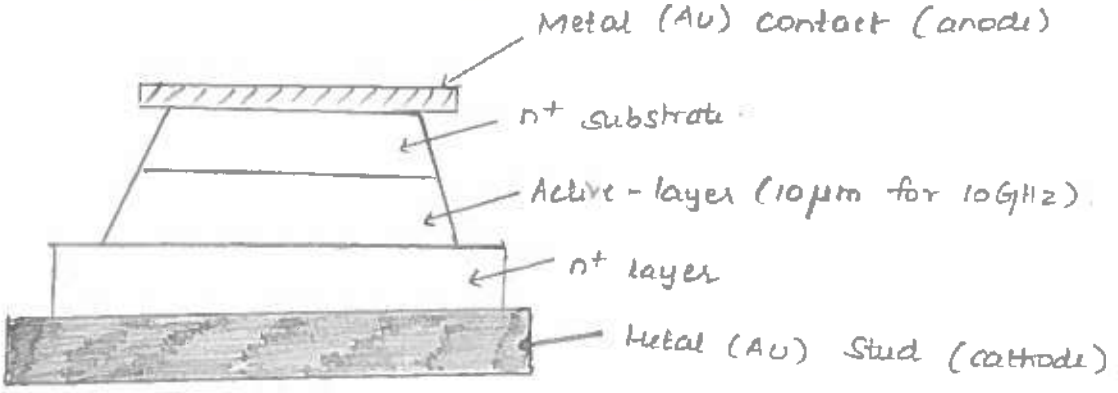


A Simple Gunn Oscillator

* The basic structure of a Gunn diode as shown in figure which consists of n-type GaAs Semiconductor with regions of high doping (n+).

* Even though there is no junction this is called a diode with reference to the positive end (anode) and negative end (cathode) of the dc voltage (V) which is applied across the device.

* If a dc (or) diode voltage (or) an electric field at low level is applied to the GaAs an electric field is established across it. Initially, the current will increase with a rise in the voltage.



Basic Construction of Gunn diode

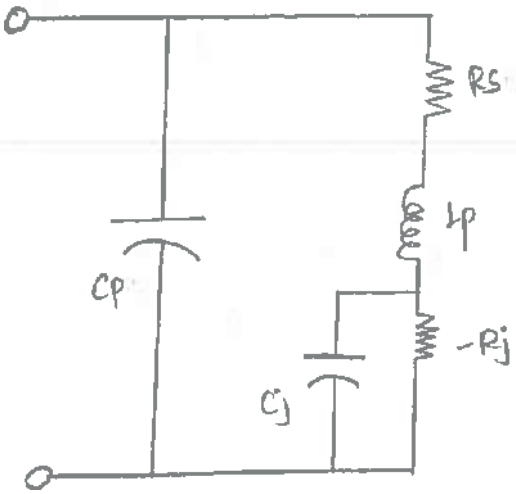
* At low E-field in the material, most of the electrons will be located in the lower energy band.

* When the diode voltage exceeds a certain threshold value (V_{th}) a high electric field (3.2 kV/m for GaAs) is produced across the active regions and thus electrons are excited from their initial lower valley to the higher valley where they become virtually immobile.

* If the rate at which the electrons transferred is very high, the current will decrease with an increase in voltage, resulting in an equivalent negative resistance effect.

Negative Resistance:

The carrier drift velocity is linearly increased from zero to a maximum when an electric field is varied from zero to a threshold value. When the electric field is beyond the threshold value of 3000V/cm, the drift velocity is decreased and the diode exhibits negative resistance.



C_j – Diode capacitance
$-R_j$ – Diode resistance
R_s – Total resistance of leads, Ohmic contact, bulk resistance of diode.
L_p – Package inductance and
C_p – Package capacitance.

* GaAs is a poor conductor, considerable heat is generated in the diode. The diode should be well bonded into a heat sink. The negative resistance has a value that typically lies in the range −5 to −20 ohm.

AVALANCHE TRANSIT – TIME DEVICES:

* Avalanche transit – time devices (W.T. Read, 1958) are p-n junction diode with the highly doped p and n regions. They could produce a negative resistance at microwave frequencies by using a carrier Impact Ionization avalanche breakdown and carriers drift in the high field intensity region under the reverse biased condition.

Modes of Avalanche Device:

There are three distinct modes of avalanche Oscillators.

(i) IMPATT: Impact Ionisation Avalanche Transit Time Device

(ii) TRAPATT: Trapped Plasma Avalanche Triggered Transit Device

(iii) BARITT: Barrier Injected Transit – Time Device

* It has long drift regions similar to those of IMPATT diodes. The carriers traversing the drift regions of BARITT diodes. However they are generated by minority carrier injection from forward-biased junctions rather than being extracted from the plasma of an avalanche region.

* BARITT diodes have low noise figures of 15dB, but their bandwidth is relatively narrow with low output power.

IMPATT DIODE OSCILLATOR AND AMPLIFIER:

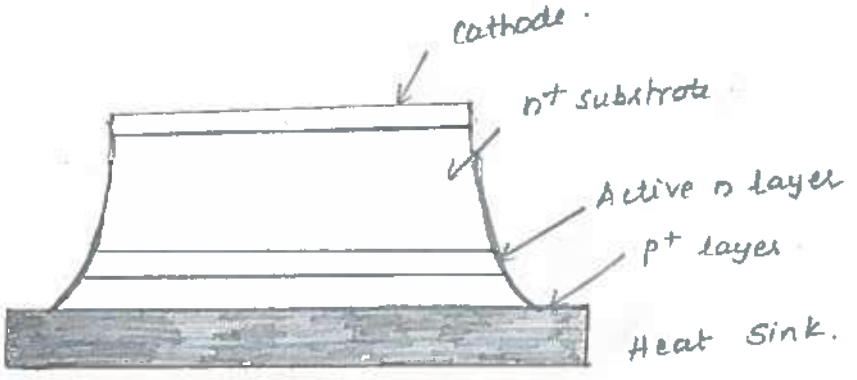
* The word IMPATT is an acronym for Impact Ionization Avalanche Transit Time. These diodes employ impact ionization and transit time properties of semiconductor structure to produce negative resistance at microwave frequencies.

* IMPATT diodes have many forms, n^+pip^+ or p^+nin^+ read device, p^+nn^+ abrupt junction and p^+in^+ diode.

* Many IMPATT diodes consist of a high doping avalanche region followed by a drift region where the field is low enough that the carriers can transverse through it without avalanching.

* IMPATT diodes can be manufactured from Ge, Si, GaAs or InP. Among these, GaAs provides the highest efficiency the highest operating frequency and least noise figure. But the fabrication process is more difficult and it is more expensive than Si.

* A typical construction and package of a simple IMPATT diode is shown in figure. An n-type expitaxial layer is formed over the n^+ substrate. On top of this is the diffused p^+ layer. A metallised cathode and plated heat sink as anode are also included.



Construction and package of p^+nn^+ IMPATT diode.

Principles of Operation:

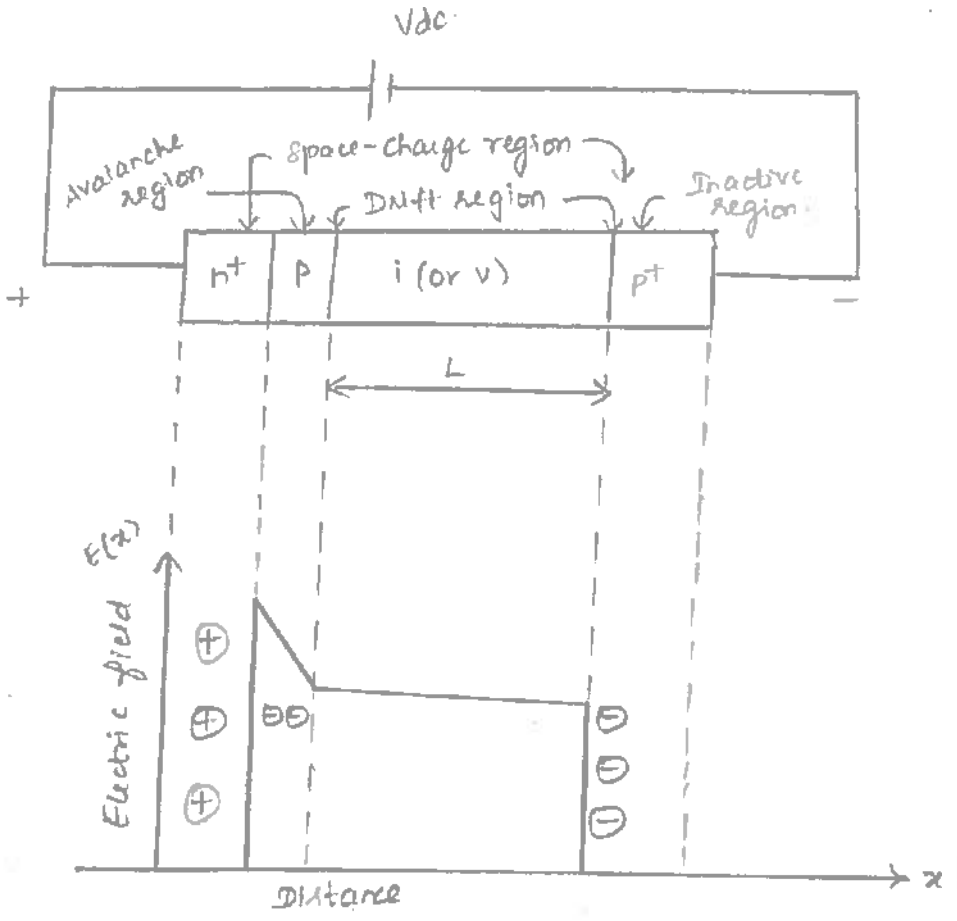
* From the figure, IMPATT diode is an $n^+-p-i-p^+$ structure where the superscript plus sign denotes very high doping and 'i' or 'v' refers to intrinsic material.

* The devices consists essentially two regions.

(i) The thin 'p' region at which avalanche multiplication occurs. This region is also called the high-field region or the avalanche region.

(ii) The 'i' or 'v' region through which the generated holes must drift in moving to the p^+ contact. This region is also called as the instrinsic region or the drift region.

* The space between the n^+-p junction and $i-p^+$ junction is called the Space charge region.



(a) Silicon structure

(b) Field distribution.

IMPATT diode operation

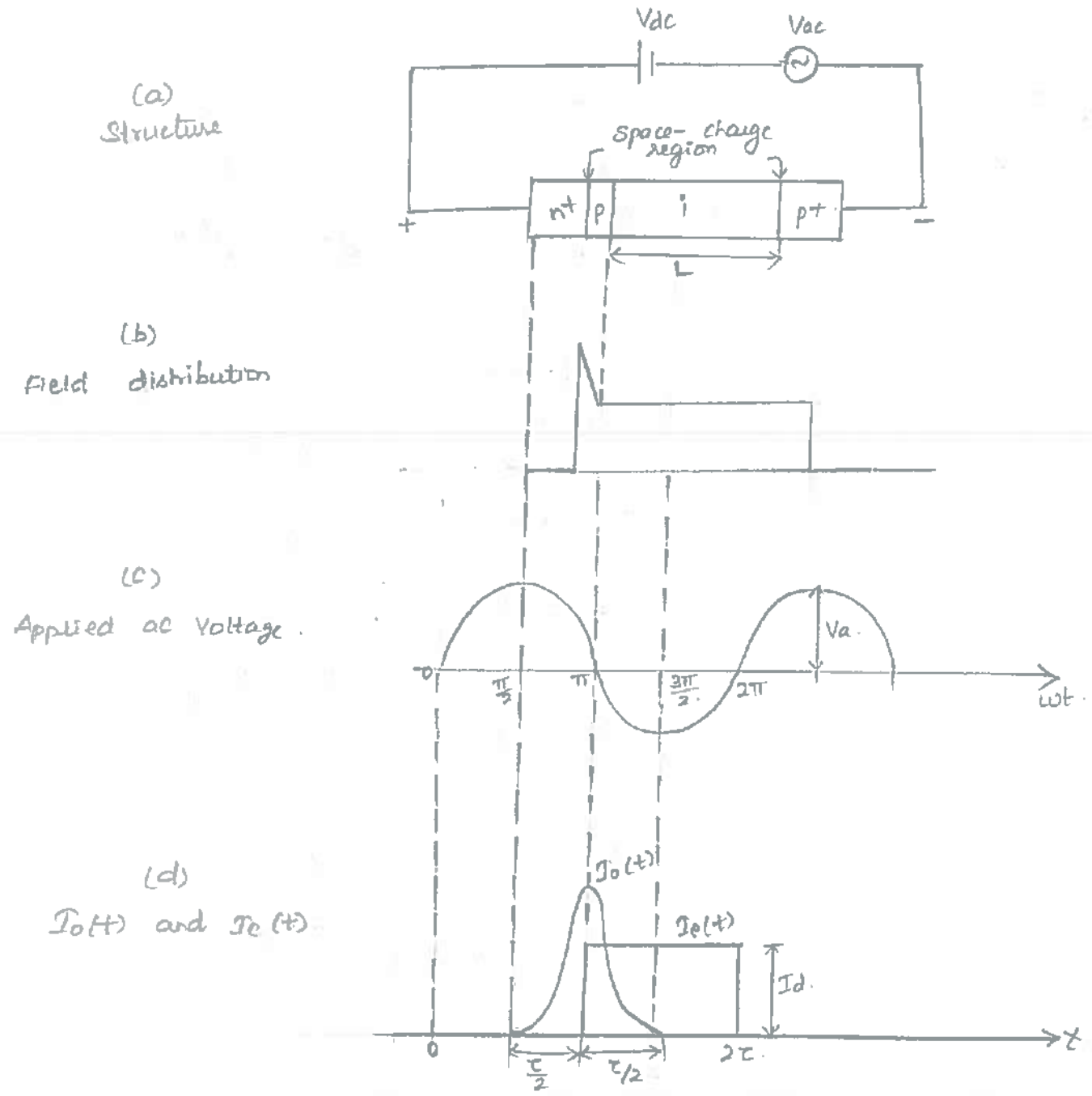
Avalanche multiplication

The avalanche multiplication factor M is expressed as

$$M = \frac{1}{1-\left(\frac{V_{dc}}{V_b}\right)^n} \longrightarrow (1)$$

Mechanism of Oscillations:-

When the IMPATT diode is mounted in a microwave resonant circuit an ac voltage can be maintained at a given frequency in the circuit which is shown below.



Field, voltage and currents in IMPATT diode.

Carrier Current $I_o(t)$ and External Current $I_e(t)$:-

* The total field across the diode is the sum of the dc and ac fields. This total field causes break down at the $n^+ - p$ junction during the positive half of the ac voltage cycle.

* If the field is above the breakdown voltage and the carrier current $I_o(t)$ generated at the $n^+ - p$ junction by an avalanche multiplication grows exponentially with time while the field is above the critical value.

* During the negative half cycle when the field is below the break down voltage, the carrier current $I_o(t)$ decays exponentially to a small steady - state value.

* The carrier current $I_o(t)$ reaches its maximum on the middle of the ac voltage cycle. Under the influence of an electric field the generated holes are injected into the space - charge region towards the negative terminal.

* The injected holes traverse the drift space and they can include a current $I_e(t)$ in the external circuit.

$$I_e(t) = \frac{Q}{\tau} = \frac{V_d Q}{L} \longrightarrow (2)$$

where Q – Total charge of the moving holes

V_d – Hole drift velocity

L – Length of the drift i region

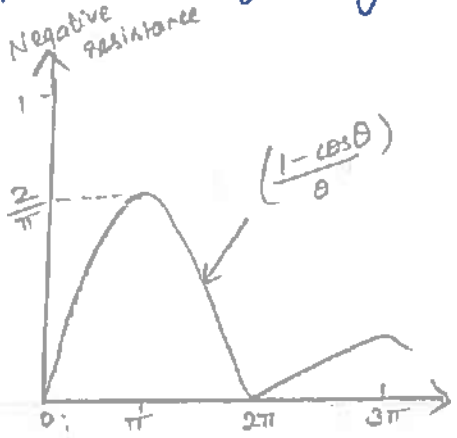
* When the pulse of hole current $I_o(t)$ is suddenly generated in the n^+-p junction, a constant current $I_e(t)$ starts flowing in an external circuit and continues to flow during the time 'τ' in which the holes are moving across the space-charge region.

Negative Resistance:

These diodes exhibit a diffuential negative resistance mainly by two effects.

(i) The impact ionization avalanche effect, which causes the carrier current $I_o(t)$ and the ac voltage to be out of phase by 90°.

(ii) The transit-time effect, which further delays the external current $I_e(t)$ relative to the ac voltage by 90°.



Negative resistance versus transit angle.

* θ is the transit angle and it is given as

$$\theta = \omega\tau = \omega\frac{L}{V_d} \longrightarrow (5)$$

* The peak value of the negative resistance occurs nears $\theta = \pi$. For transit angles larger than π, the negative resistance of the diode decreases rapidly.

Resonant Frequency:

* The resonant frequency of the cavity is given by

$$f = \frac{1}{2\tau} = \frac{V_d}{2L} \rightarrow (4)$$

* If the resonator is tuned to this frequency, IMPATT diodes provide a high power CW and pulsed microwave source.

Efficiency:

The efficiency of the IMPATT diode is given by

$$\eta = \frac{P_{ac}}{P_{dc}} = \frac{\text{RF power output}}{\text{dc input power}} = \left(\frac{V_a}{V_d}\right)\left(\frac{I_a}{I_d}\right) \rightarrow (5)$$

where V_a & I_a — ac voltage and current

V_d & I_d — dc voltage and current

Output power = $P_{out} = P_{ac} = \eta P_{dc}$

Performance Characteristics:

Theoretically, $\eta = 30\%$ ($< 30\%$ in practical) and

15% for Si;

23% for GaAs.

* GaAs IMPATTs have higher power and efficiency in the 40-to-60 GHz region whereas Si IMPATT's are produced with higher reliability and yield in the same frequency region.

IMPATT Diode Power Amplifier:

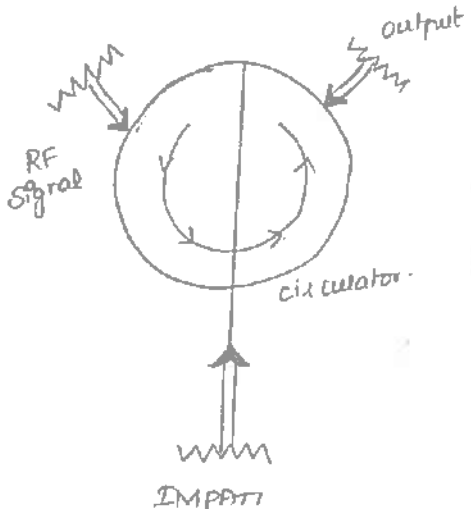
* The IMPATT diode can be used as an amplifier with the same basic circuit arrangement as oscillator, provided $R_L > |R_d|$ where R_L is the load resistance ~~results~~ and R_d is the diode negative resistance.

* The circulator is incorporated with IMPATT diode as shown in figure. The negative resistance is used to terminate one port of the circulator and an actual load is connected to the other port.

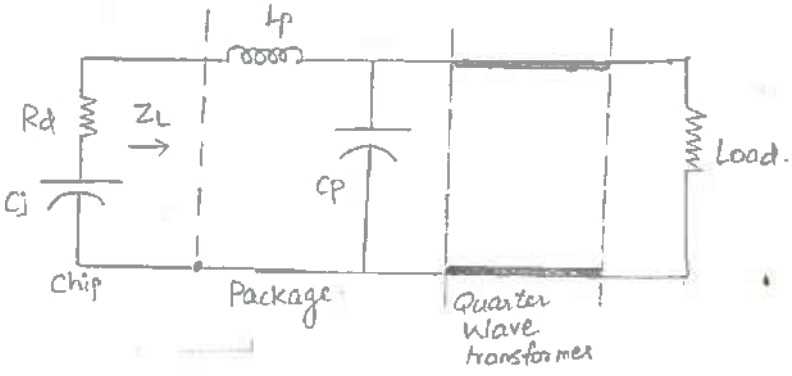
* The input RF power is fed from the remaining port. The negative resistance results in voltage reflection coefficient at the port which is greater than unity.

* Thus the average power from the source Pav Circulates to the negative resistance port and the reflected power is greater than an incident power. The reflected power is delivered to the load.

IMPATT Circulator type amplifier



Equivalent Circuit of IMPATT diode.



* Here, R_d is the diode negative resistance and L_p, C_p are the Package lead Inductance and capacitance, respectively.

Advantages:

IMPATT diodes provide potentially reliable, compact, inexpensive and moderatively efficiency microwave power sources.

Disadvantages:

(i) IMPATT diodes have low efficiency.

(ii) It tend to be noisy due to an avalanche process and requires the high level of operating current.

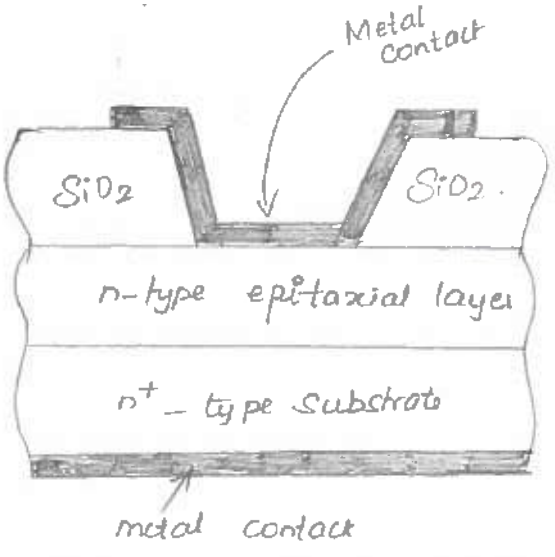
(iii) A typical noise figure is 30 dB which is worse than that of the Gunn diodes.

Applications:

(i) Used in microwave generators
(ii) Used in modulated output oscillators.
(iii) Used in receiver local oscillators.
(iv) Used in paramentric amplifier as pumps.
(v) IMPATT diodes are suitable for negative resistance amplification.

SCHOTTKY BARRIER DIODE (SBD)

SBD is a simple metal semiconductor barrier diode that exhibiting a non-linear impedance and it is basically an extension of the point contact diode.



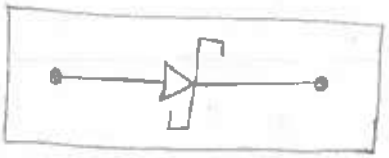
Schottky diode.

Construction :-

* The diode is constructed on a thin silicon (n^+- type) substrate by growing epitaxially on n-type active layer of about 2-micron thickness. A thin SiO_2 layer is grown thermally over this active layer.

* Metal- Semiconductor junction is formed by depositing metal over SiO_2 - Schottky diodes also exhibit a square-law characteristic and have a higher burn out rating low 1/f noise and better reliability than point contact diodes.

Operation :-

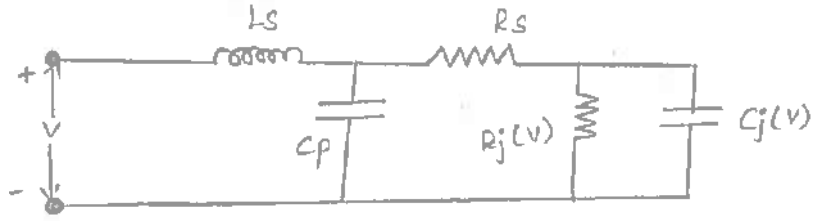


Symbol of SBD

* When the device is forward biased, the barrier height gets reduced. The major carriers (electrons) can be easily injected from the highly doped n-semiconductor material into the metal with an approximately V-I characteristic.

* When it is reverse-biased, the barrier height becomes too high for the electrons to cross and thus no conduction takes place.

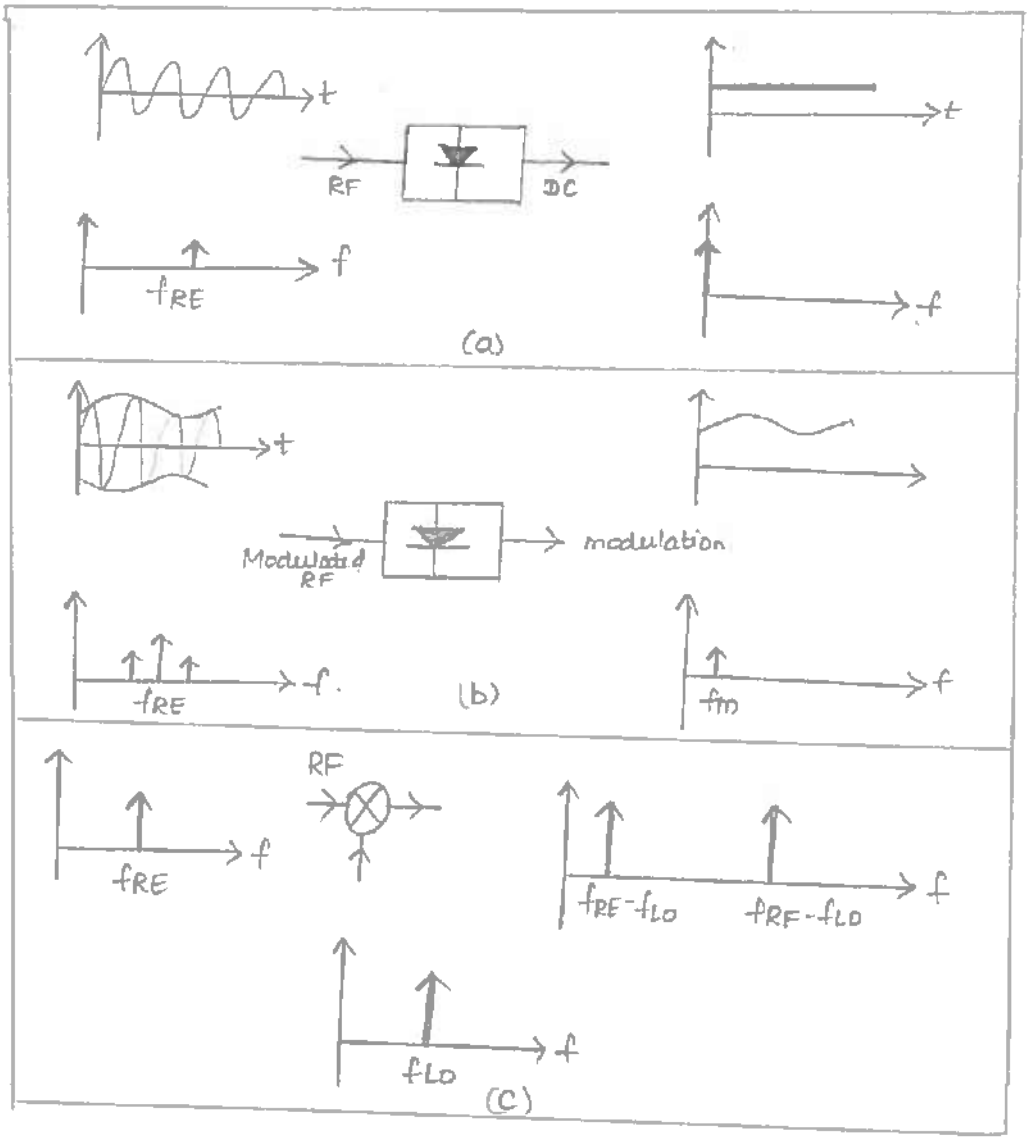
Equivalent Circuit :-



where Rj – Junction resistance of the diode.
Cj – Junction capacitance
Rs – Series resistor.
Ls – Series inductance.
Cp – Shunt Capacitance.

Applications:

The primary applications of Schottky diodes is in frequency conversion of an input signal. The below figure illustrates the three basic frequency conversion operations of rectification (conversion to DC), detection (demodulation of an amplitude-modulated signal) and mixing (frequency shifting).



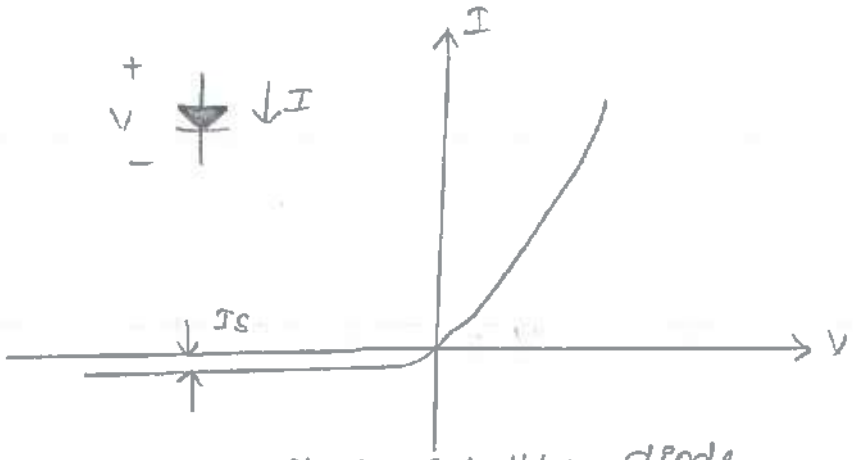
Basic frequency conversion operations of rectification, detection and mixing.

(a) Diode rectifier (b) Diode detector (c) Mixer

* A junction diode can be modeled as a nonlinear resistor with a small-signal V-I relationship expressed as

$$I(V) = I_S(e^{\alpha V} - 1)$$

where $\alpha = q/nkT$ and q is the charge of an electron, k is Boltzmann's constant, T is temperature, n is the ideality factor and I_S is the saturation current.



V-I characteristics of a Schottky diode.

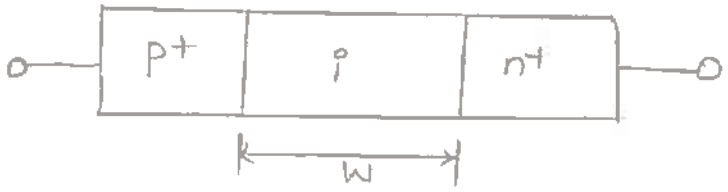
PIN DIODES:-

* PIN diodes can be used to construct an electronic switching element easily integrated with planar circuitry and it is capable of high-speed operation.

* Switching speeds typically ranges from 1 to 10μs, although speeds as fast as 20ns are possible with careful designed of the diode driving circuit. PIN diodes can also be used as power limiters, modulators and variable attenuators.

PIN diode characteristics:-

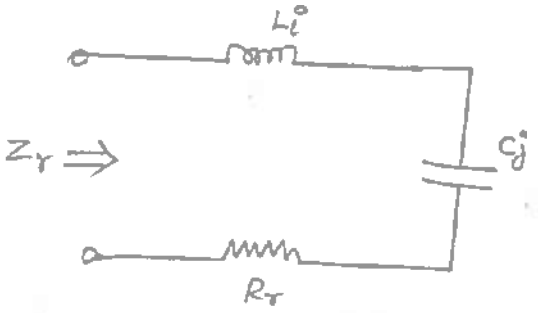
A PIN diode consist of a high-resistivity intrinsic (i) semiconductor layer between two highly doped p^+ and n^+ Si layers as shown in figure. The device acts as electrically variable resistor which is related to the 'i' layer thickness.



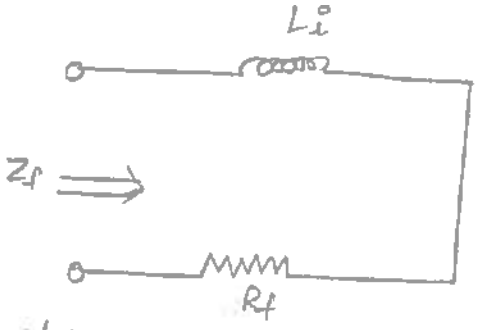
* The intrinsic layer has a very large resistance in reverse bias and it decreases in forward bias. When mobile carriers from p and n regions are injected into I layer, carriers take time such that the diode ceases to act as a rectifier at the microwave frequency and appears as a linear resistance.

* The property makes it suitable for being a variable attenuator at the microwave frequencies.

Equivalent Circuit:-



(a) Reverse bias state



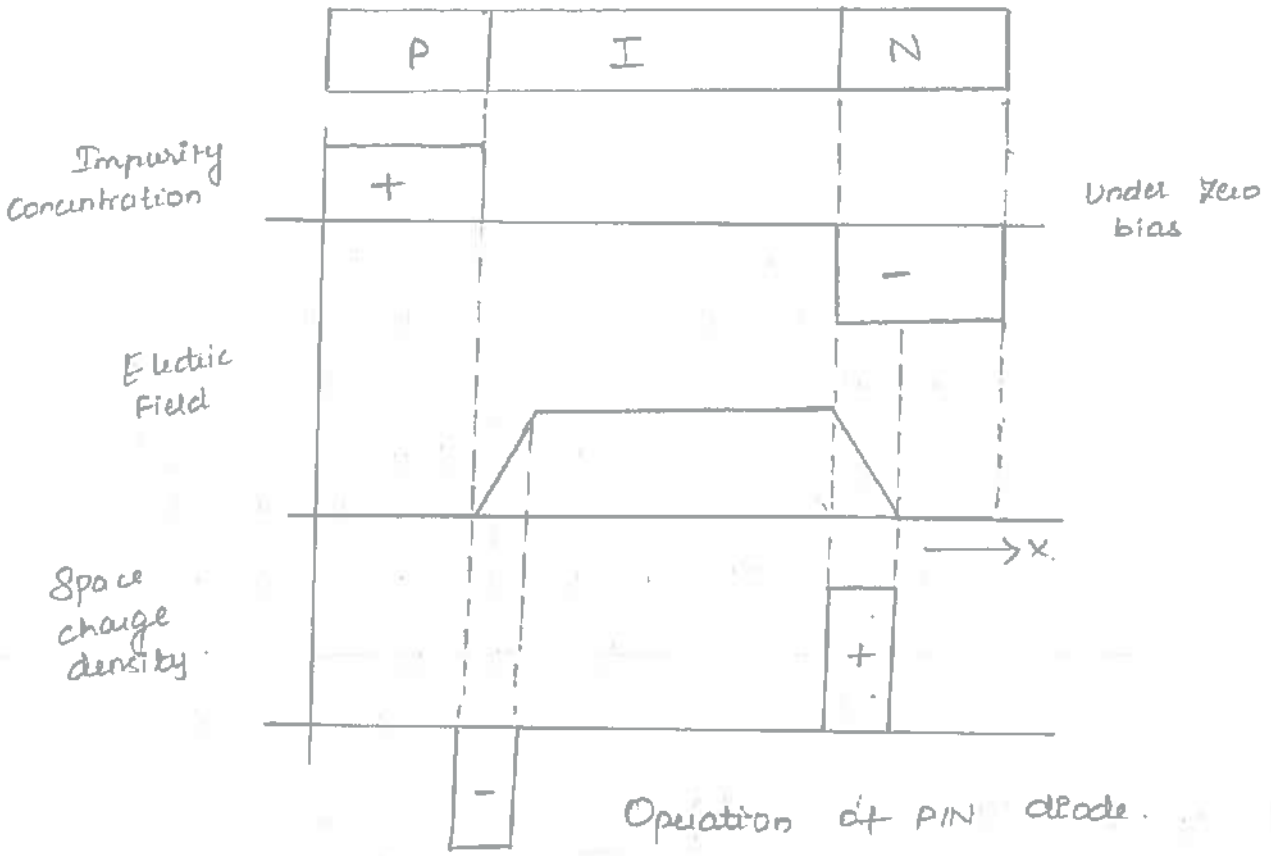
(b) Forward bias state

* The parasitic inductance, L_i, is typically less than 1 nH. The reverse resistance, R_r, is usually small relative to the series reactance due to the junction capacitance and is often ignored.

Operation of PIN diode:-

The operation can be explained by considering

→ Zero bias
→ reverse bias
→ forward bias



Operation of PIN diode.

(i) Zero bias:

* At the zero bias, the diffusion of the holes and electrons across the junction causes space charge (density) region of thickness which is inversely proportional to the impurity concentration.

* An ideal i layer has no depletion region (i.e) p layer has a fixed negative charge and n layer has a fixed positive charge under zero bias.

(ii) Reverse bias:

* As reverse bias is applied, the space charge regions in the p and n layers will become thicker. The reverse resistance will be very high and almost constant.

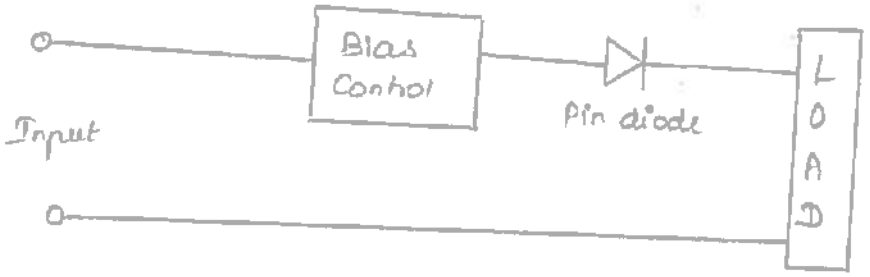
(ii) forward bias:

* With the forward bias, carriers will be into the i layer and the p and n space charge regions will become thinner (i.e) electrons and holes are injected into the P layer from p and n layers respectively.

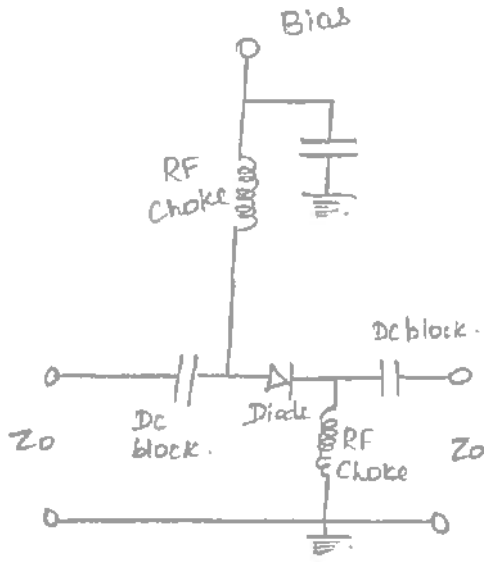
* This results in the carrier concentration in the i layer becoming raised above an equilibrium levels and the resistivity drops as the forward bias is increased. Thus, the low resistance is offered in the forward direction.

Applications of PIN Diodes:

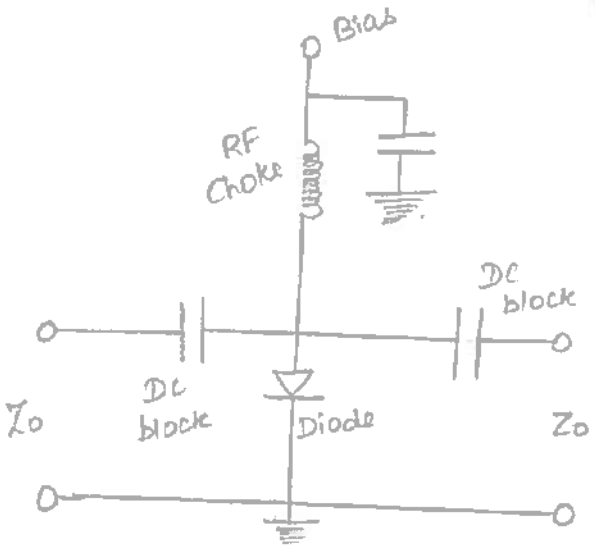
(i) Single - Pole PIN diode Switches:



* A PIN diode can be used in either a series or a shunt configuration to form a single-pole, single-throw RF Switch as shown below

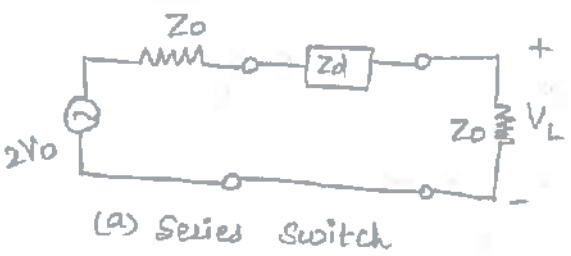


(a) Series Configuration

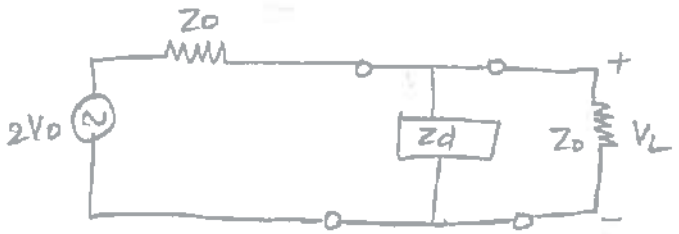


(b) Shunt Configuration

* In the series configuration of (a) the switch is in 'ON' when the diode is forward biased, while in the shunt configuration of (b) the switch is in 'ON' when the diode is reverse biased.



(a) Series switch



(b) Shunt switch

* In both the above cases, input power is reflected when the switch is in the OFF state. The DC blocking capacitors should have a relatively low impedance at the RF operating frequency, while the RF choke inductors should have a relatively high RF impedance.

* In some designs, high-impedance quarter-wavelength lines can be used in the place of the chokes to provide RF blocking.

* The insertion loss in terms of the actual load voltage, V_L and V_0 which is the load voltage that would appear if the switch (Z_d) were absent:

$$I_L = -20\log\left|\frac{V_L}{V_0}\right| \longrightarrow (1)$$

* Simple circuit analysis applied to the two cases from the above figure gives the following results

$$I_L = -20\log\left|\frac{2Z_0}{2Z_0 + Z_d}\right| \text{ (series switch)} \longrightarrow (2a)$$

$$I_L = -20\log\left|\frac{2Z_d}{2Z_d + Z_0}\right| \text{ (Shunt switch)} \longrightarrow (2b)$$

* In both cases, Z_d is the diode impedance for either the reverse or forward bias state. Thus,

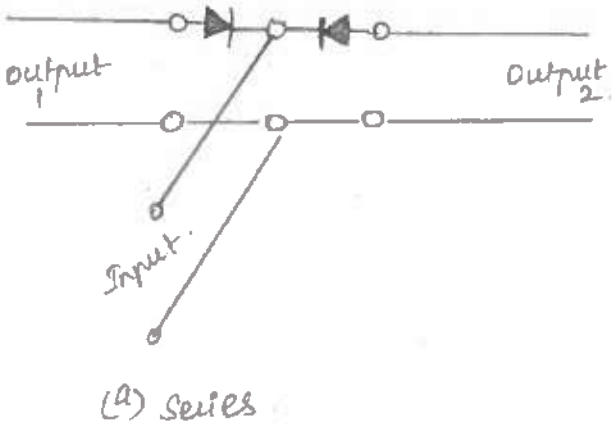
$$Z_d = \begin{cases} Z_r = R_r + j(\omega L_i - \frac{1}{\omega C_j}) & \text{for reverse bias} \\ Z_f = R_f + j\omega L_i & \text{for forward bias} \end{cases} \rightarrow (3)$$

* The ON-state or OFF state insertion loss of a switch can usually be improved by adding an external reactance in series or in parallel with the diode, to compensate for the diode reactance. This technique is usually used reduces the bandwidth.

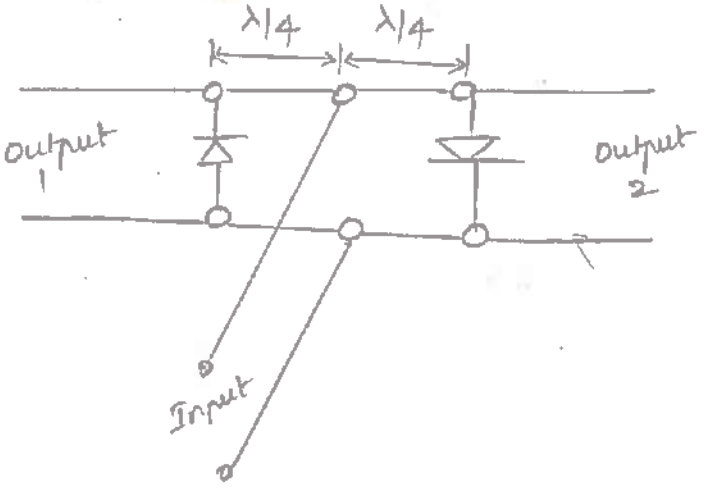
(2) Double Switches:

* Several single-throw switches can be combined to form a variety of multiple pole and/or multiple-throw configurations.

* The PIN diode double switch circuit uses two diodes called as Single-pole Double-Throw (SPDT). The below figure shows a series and shunt circuits for a single-pole, double-throw switch; such a switch requires atleast two switching elements.



(a) Series



(b) Shunt

* In an operation, one diode is forward biased in the low-impedance state, with the other diode reverse biased in the high-impedance state. The input signal is switched from one output to the other by reversing the diode bias states.

* The Quarter-wave lines of the shunt circuit limits the bandwidth of this configuration.

(3) PIN Diode Phase Shifters:

Several types of microwave phase shifters can be constructed with the PIN diode switching elements.

Advantages:

Compared with ferrite phase shifters, diode phase shifters have the advantages of small size, integrability with planar circuitary and high speed.

Drawbacks:

* The power requirements for diode phase shifters are generally greater than those for a latching ferrite phase shifter because diodes require continuous bias current, while a latching ferrite device requires only a pulsed current to change its magnetic state.

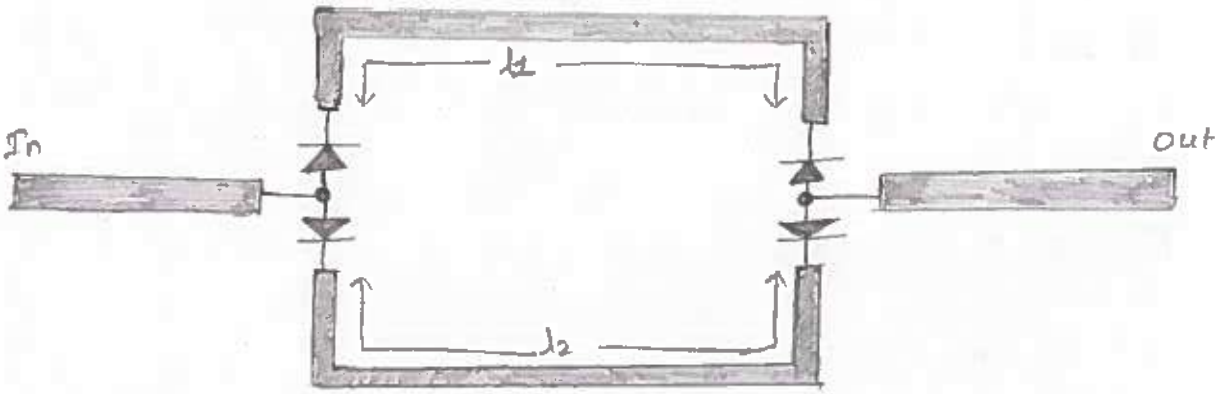
* There are basically three types of PIN diode phase shifters:

(i) Switched line

(ii) Loaded line

(iii) Reflection.

(i) Switched - line Phase shifter:



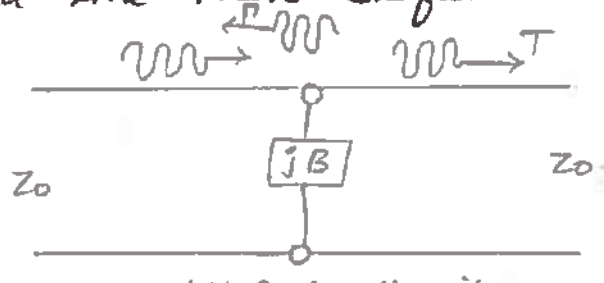
* The switched-line phase shifter is the most straightforward type using two single-pole, double-throw switches to route the signal flow between one of two transmission lines of the different length.

* The different phase shift between the two paths is expressed as.

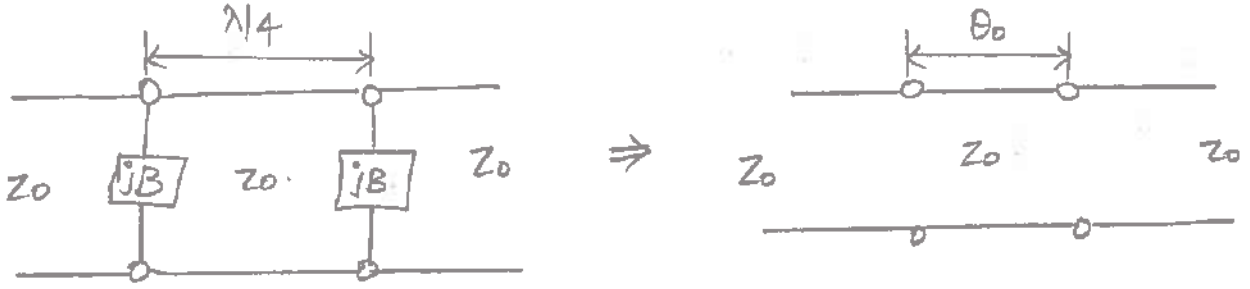
$$\Delta\phi = \beta(l_2 - l_1) \longrightarrow (4)$$

where β is the propogation constant of the line.

(ii) Loaded-Line Phase Shifter:



(a) Basic circuit



(b) Practical loaded-line phase shifter and its equivalent circuit

* A design that is useful for small amounts of phase shift (generally 45° or less) is the loaded-line phase shifter which is illustrated above figure.

* The reflection and transmission coefficients can be written as

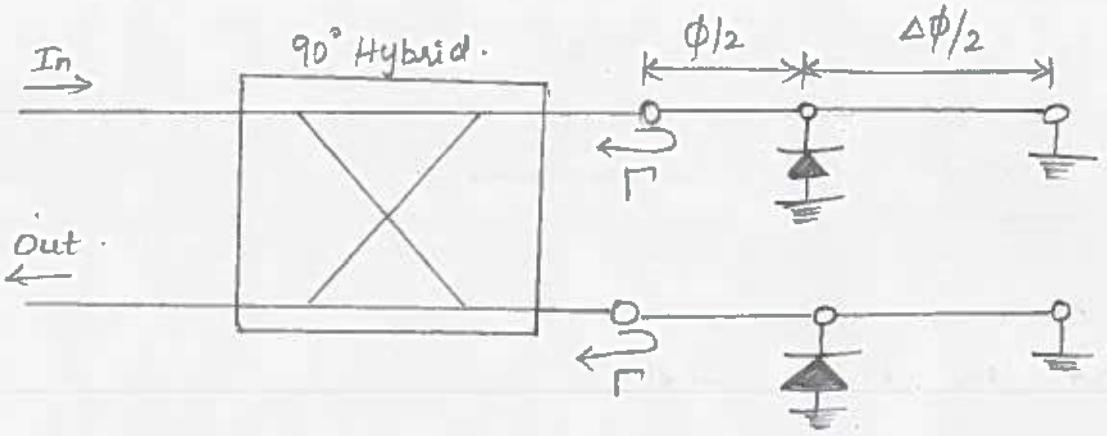
$$\Gamma = \frac{1-(1+jb)}{1+(1+jb)} = \frac{-jb}{2+jb} \longrightarrow (5a)$$

$$T = 1+\Gamma = \frac{2}{2+jb} \longrightarrow (5b).$$

The phase shift in the transmitted wave introduced by the load is

$$\Delta\phi = \tan^{-1}\left(\frac{b}{2}\right) \longrightarrow (6)$$

(iii) Reflection phase shifter:-



* The reflection phase shifter which uses an SPST switch to control the path length of a reflected signal.

* Ideally the diodes would look like short circuits in their ON state and open circuits in their OFF state, so that the reflection coefficients at the right side of hybrid written as

$$\Gamma = \begin{cases} e^{-j(\phi+\pi)} & - \text{Diodes 'ON' state} \\ e^{-j(\phi+\Delta\phi)} & - \text{Diodes 'OFF' state} \end{cases} \longrightarrow (7)$$

MICROWAVE TUBES

Microwave tubes are connected to overcome the limitations of conventional electronic vaccum tubes such as triodes, tetrodes and pentodes.

These conventional electronic Vaccum tubes fails to operate above the 1GHz.

In microwave tubes the electron transit time is utilized for microwave oscillation or amplification.

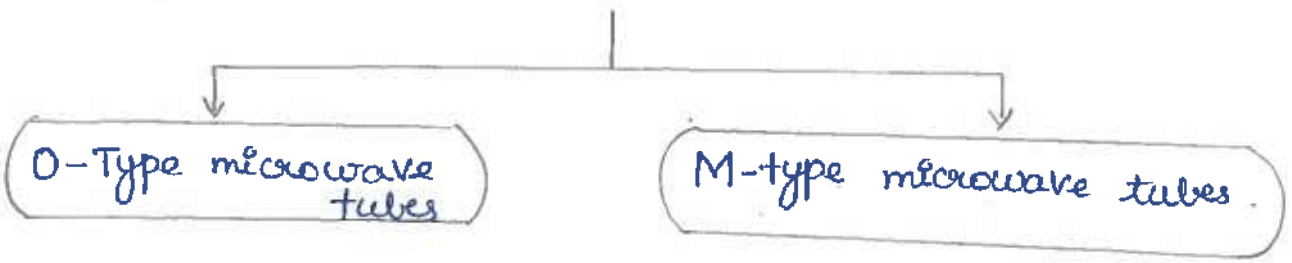
Transit time:-

The transit time is the time taken for an electron to travel from cathode to anode.

The principle used in the microwave tubes are an electron beam on which Space-charge waves interact with an electromagnetic fields in the microwave cavities to transfer energy to an output circuit of the cavity (Klystrons and Magnetrons) (or)

interact with an electromagnetic fields in a Slow-wave structure to give an amplification through the transfer of energy (traveling wave tubes)

Microwave tubes.



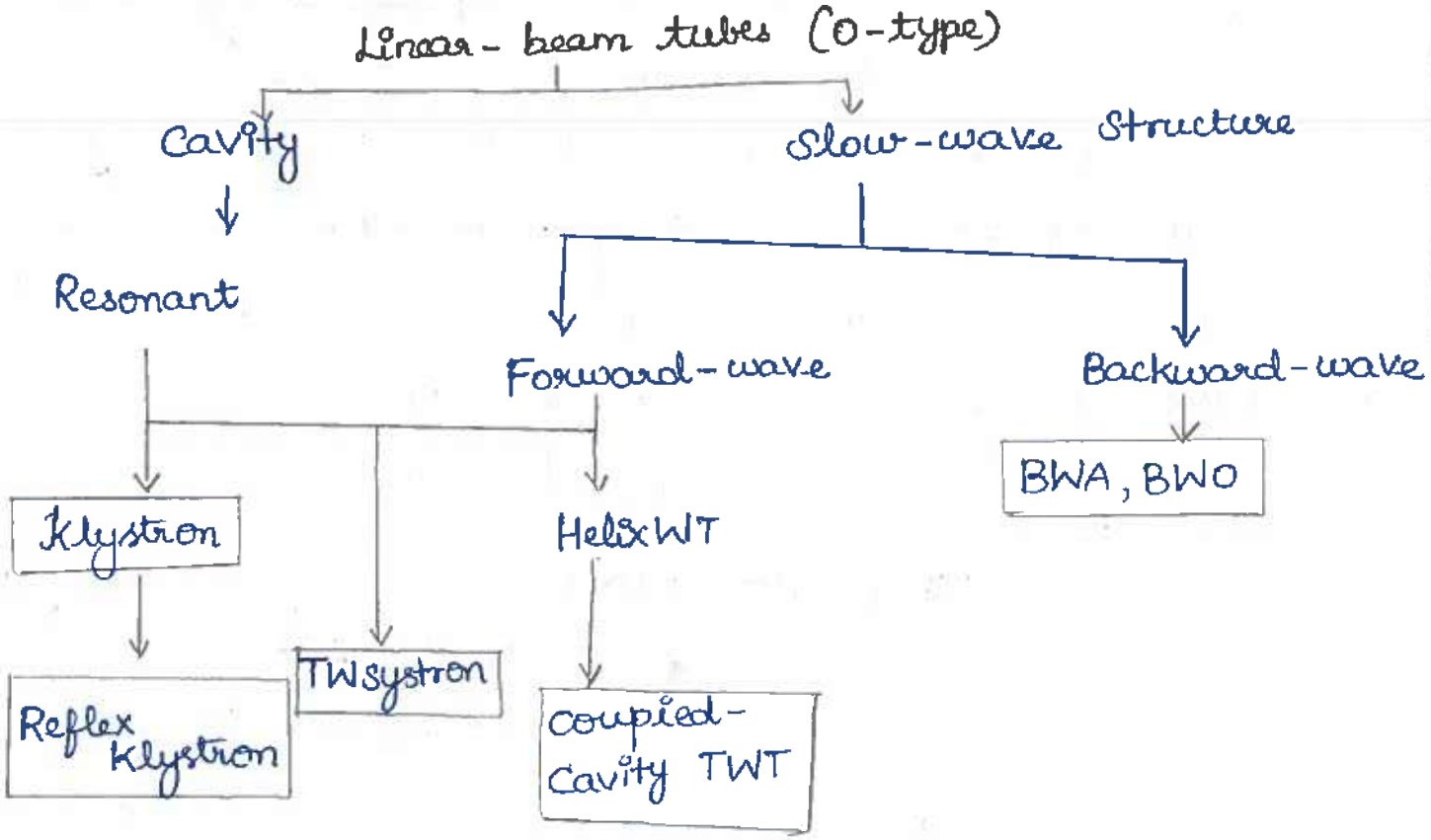
M-type Microwave tube:-

Magnetrons are crossed field devices (M-type) where the static magnetic field is perpendicular to an electric field. In this tube, the electrons can travel in a curved path.

O-Type microwave Tube:-

The most important microwave tubes are linear beam or 'O'-type tubes in recognition of the straight path taken by an electron beam.

Klystrons and TWTs are linear beam tubes in which an accelerating electric field in the same direction as the static magnetic field used to focus an electron beam.



KLYSTRONS

Definition:

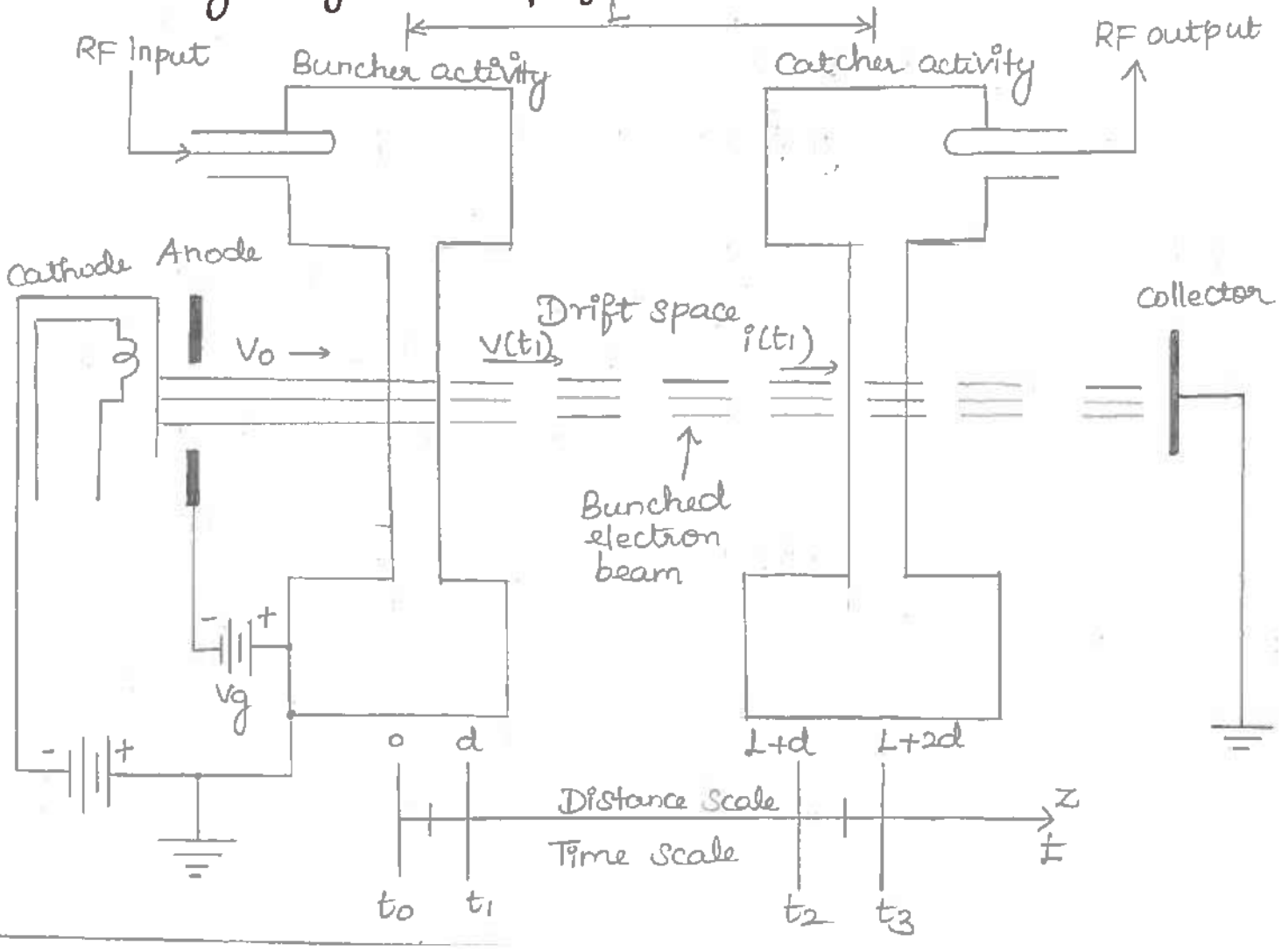
A Klystron is a Vaccum tube that can be used either as a generator or as an amplifier of power at the microwave frequencies operated by the principles of Velocity and current modulation.

There are 2 basic configuration of klystron tubes,

(i) Reflex Klystron - It is used as power microwave oscillator and

(ii) Two cavity (or) Multicavity klystron - It is used as low power microwave amplifier.

Two cavity klystron amplifier:-



Introduction:

A two-cavity klystron amplifier is a Velocity modulated tube in which the velocity modulation process produces a density modulated stream of electrons. It consists of two cavities namely, buncher (input) cavity and catcher (output) cavity.

Drift Space:

The separation between the buncher and catcher grids is called as drift space.

Operation:-

→ Cathode emits an electrons beam. This electron beam first reaches the anode. The accelerating anode produces a high velocity electrons beam.

→ The input RF signal to be amplified excites the branch activity with a coupling loop.

Bunching:-

The electrons beam passing the buncher activity gap at zeros of the gap voltage Vg (voltage between buncher grids) passes through an unchanged velocity.

The electrons beam passing through the positive half cycles of the gap voltage undergoes an increase in velocity, those passing through the negative swings of the gap voltage undergoes a decrease in velocity.

As a result of these actions, the electrons gradually bunch together as they travel down the drift space. This is called bunching.

The first cavity act as the buncher and Velocity - modulates the beam. Thus the electron beam is velocity modulated to form bunches or undergoes density modulation in accordance with the input RF signal cycle.

Velocity modulation:

→ The variation in electron velocity in the drift space is known as the velocity modulation.

When this density modulated electron beam passing through the catcher cavity grid, it induces RF current (ac current) and thereby excite the RF field in an output activity at an input signal cycle.

→ The ac current on the beam is such that the level of excitation of the second cavity is much greater than that in the buncher activity and hence the amplification takes place.

→ If desired, a portion of an amplified output can be fed back to the buncher cavity in a regenerative manner to obtain a Self-Sustained oscillations.

The maximum bunching should occur approximately at a midway between the second activity grids during its retarding phase, thus the kinetic energy is transferred from the electrons to field of the second cavity.

The electrons then emerge from the second cavity with reduced velocity and terminates at the collector.

Catcher activity:

The output cavity catches energy from the bunched electron beam. Therefore, it is also called as catcher activity.

Characteristics and applications:

Characteristics:

(i). Efficiency = 40%.

(ii) power output:

(a) continuous wave average power = 500kW

(b). pulsed power 30MW at 10GHz.

(iii). power gain ≃ 30dB

Applications:

Used in Troposphere scatter transmitters

Satellite communication ground stations.

used in UHF TV transmitters

Radar transmitters.

Reflex Klystron Oscillator :- Single cavity Klystron.

Introduction:

* The reflex klystron is an oscillator with a built in feedback mechanism.

* It uses the same activity for both the bunching and the output.

* The repeller electrode is a negative potential and sends the bunched electron beam back to the reasonator cavity.

* This provides a positive feedback mechanism which supports oscillations.

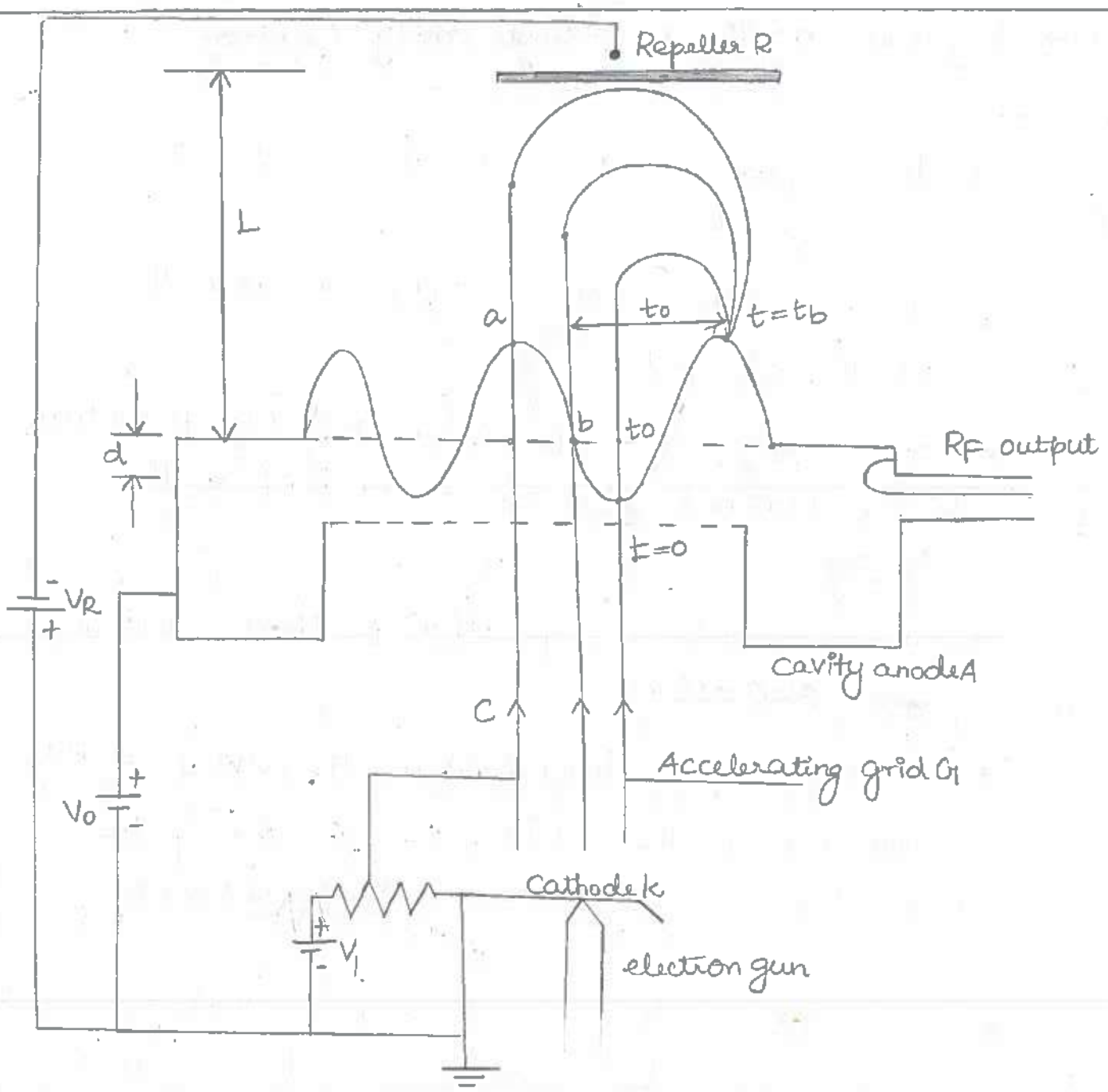
* Due to dc voltage (Vo) in the cavity circuit, RF noise is generated in the cavity, The electromagnetic noise field in the cavity act as a cavity resonant frequency.

* When the oscillation frequency is varied, the resonant frequency of activity and the feedback path phase shift must be readjusted for a positive feedback.

Mechanism of oscillation:

* The electron beam injected from the cathode is first velocity - modulated by cavity - gap voltage.

* The electron which encountered the positive half cycle of the RF field, where in the cavity gap velocity will be accelerated.



The electrons which encountered zero RF field will pass with unchanged original velocity, and the electrons which encountered the negative half cycle velocity will be decelerated.

All these velocity modulated electrons will be repelled back to the cavity by the repeller due to its negative potential.

The repeller distance 'L' and the voltages (beam & repeller voltage) can be adjusted to receive all the velocity modulated electrons at a same time on the positive peak of the cavity RF field.

The velocity modulated electrons are bunched together and lose their kinetic energy when they encounter the positive cycle of the RF field. This loss of energy is transferred to the cavity in order to conserve the total power.

If the power delivered by the bunched electrons to the cavity is greater than the power loss in the cavity, an electromagnetic field amplitude at the resonant frequency of the cavity will increases to produce the microwave oscillations

The electrons passing through the buncher grids are accelerated / passed through which an unchanged initial dc velocity depending upon whether they encounter the RF signal field at the buncher cavity gap at positive / negative / zero crossing phase of the cycle respectively as shown by distance-time plot. This is called the applegate diagram.

Explanation:

When the gap voltage is at a positive peak, an electron passing at this moment is called early electron. This electron is accelerated towards

the repeller and travels at a distance, which is longer comparatively.

The electron at a neutral zero of gap Voltage is called the reference electron. When the gap Voltage is at a negative peak the corresponding electron is called the late electron. This electron decelerated and travels at a less distance.

These electrons have different Velocities cover a different distances and forms a bunch at the cavity gap.

Modes of oscillation:

The condition for oscillation of reflex klystron is

$$t_0 = \left(n + \frac{3}{4}\right) T = NT$$

where $N = n + \frac{3}{4}$

Mode of oscillation = $n = 0, 1, 2, 3, \ldots$

T is the time period at the resonant frequency

t_0 is the time taken by the reference electron to travel in the repeller space.

Characteristics:

Frequency range : 1 to 25 GHz

Power output : It is a low-power generator of 10 to 500mW.

Efficiency : About 20 to 30%

Applications:

The main applications of reflex klystrons are, This type is widely used in the laboratory for microwave measurements.

In microwave receivers, as local oscillators in commercial and military applications.

Also plays a role in airborne Doppler radars as well as missiles.

Drawbacks of klystrons:

The klystrons are having the following drawbacks:

(i) Klystrons are essentially narrow band devices as they utilize cavity resonators to velocity modulate the electron beam over a narrow gap.

(ii) In klystrons and magnetrons, the microwave circuit consists of a resonant structure which limits the bandwidth (or the operating frequency range) of the tube.

Helix - Traveling Wave tube (or) Travelling wave tube amplifier:

Definition:

A Traveling wave Tube amplifier (TWTA) circuit uses a helix slow wave non resonant microwave guiding structure and thus a broad band microwave amplifier.

Two main constituents of TWT are,

(i) An electron beam and

(ii) A structure supporting a slow electromagnetic wave (slow-wave structure).

In the case of TWT, the microwave circuit is non-resonant and the wave propagate with the same speed as the electron in the beam.

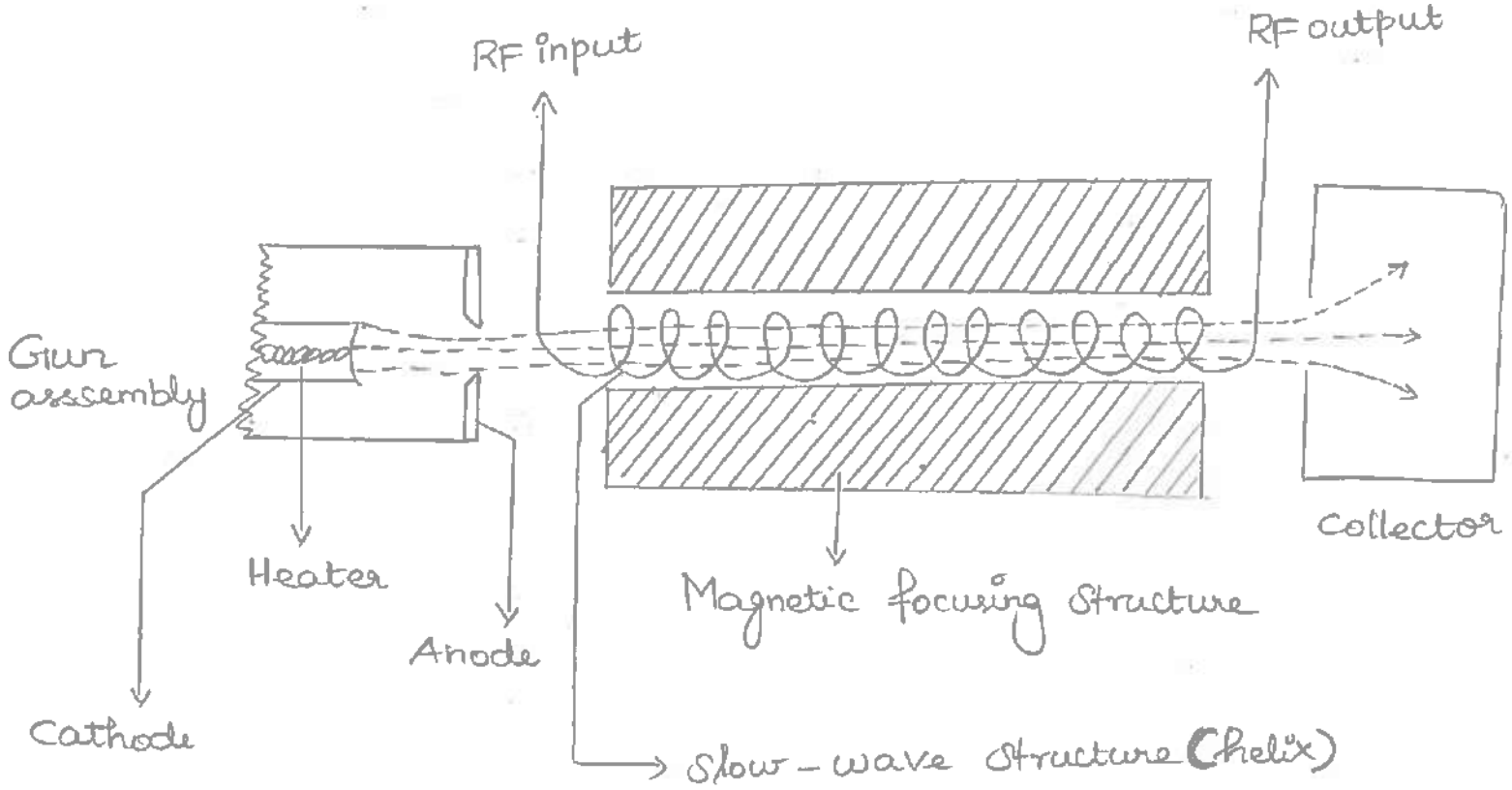
The initial effect of on the beam is a small amount of velocity modulation caused by the weak electric fields associated with the the traveling wave.

This velocity modulation later translates to current modulation, which then induces an RF current in circuit, causes an amplification.

Operation:-

The electron beam is focused axially by a static magnetic field and collected in a collector circuit.

The microwave input signal is injected on the helix slow-wave circuit surrounding an electron beam, which produces an axial electric field of the signal at the centre of helix and it can interacts with the electron beam.



Simplified TWTA circuit.

The dc beam voltage is adjusted so that the beam velocity is slightly greater than the axial component of a field on the slow-wave structure.

During transit along the axis, an electron beam transfers great amount of energy to the traveling signal wave and thus signal field amplitude increases.

Attenuator!-

An attenuator is placed over a part of the helix near an output end to attenuate any reflected waves due to impedance mismatch that can be fedback to an input to cause the oscillations.

Magnet :-

The magnet produces produces an axial magnetic field to prevent spreading of an electron beam as it travels down the tube.

Need of slow-wave structures (Helix Tube):

Slow-wave structures are special circuits that are used in microwave tubes to reduce the wave velocity in a certain direction so that an electron beam and the signal wave can interact.

Characteristics of TWTA:-

Frequency range : 3GHz and higher

Bandwidth : about 0.8GHz

Efficiency : 20 to 40%.

Power output : upto 10kW average

Power gain : upto 60dB.

Applications of TWTA:-

The main applications of TWTA are,

→ Used in medium power satellite
→ Used in high power satellite transponder output,
→ Used in radar transmitters, and
→ Used in broadband microwave amplifier.

Microwave crossed-Field Tubes (M-Type)

Introduction

A crossed field microwave tube is a device that converts dc into microwave energy using an electronic energy-conversion process.

M-type devices or crossed field tubes in which the dc magnetic field and dc electric field are perpendicular to each other. The principal tube in this type is called Magnetron.

In all crossed-field tubes, the dc magnetic field plays a direct role in the RF interaction process. A magnetron oscillator is used to generate high microwave power.

Power output and efficiency

A magnetron can deliver a peak power output of upto 40 MW with the dc voltages of 50kV at 10GHz. The average power output is 800kW.

The magnetron possesses a very high efficiency ranging from 40 to 70%. Magnetrons are commercially available for peak power output from 3kW and higher.

Applications

Radar transmitter with high output power
Satellite and missiles of telemetry
Industrial heating and
Microwave Ovens.

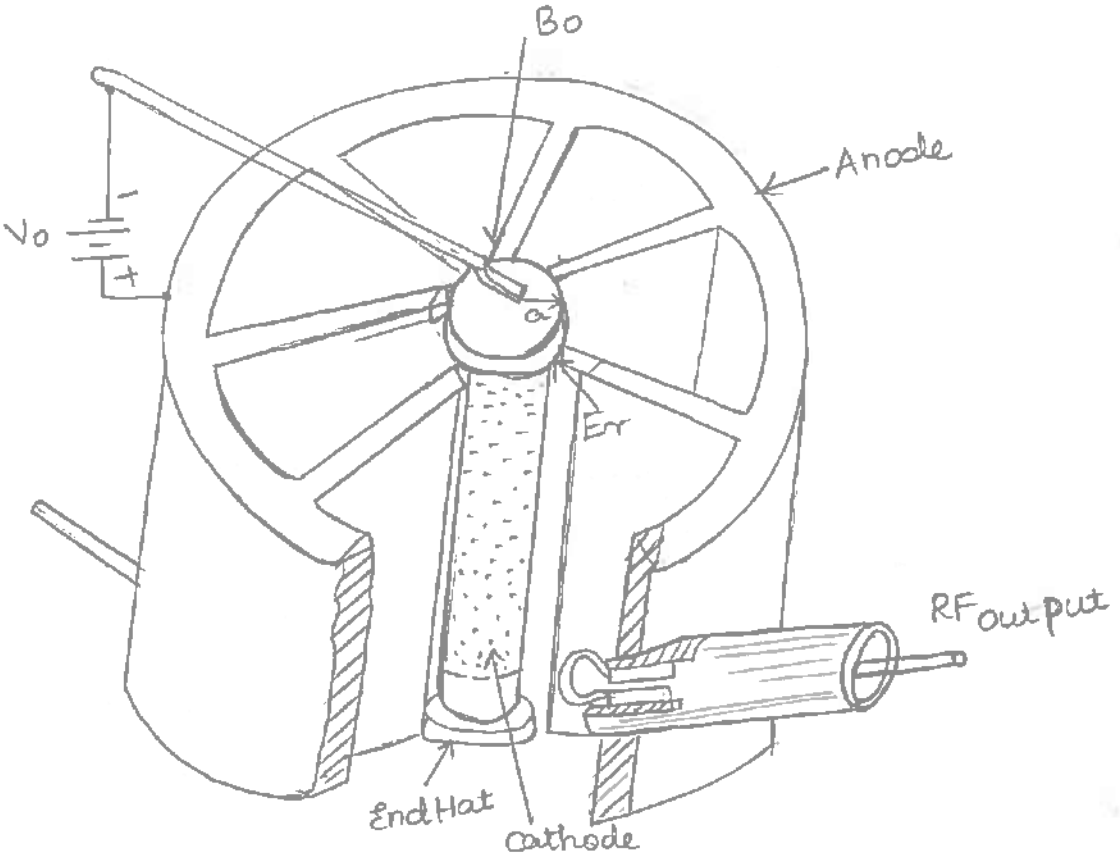
Cylindrical Magnetron: Magnetron:

This type of magnetron is also called as a conventional magnetron. It consists of an cylindrical cathode of finite length and radius 'a' at the centre surrounded by a cylindrical anode of radius b.

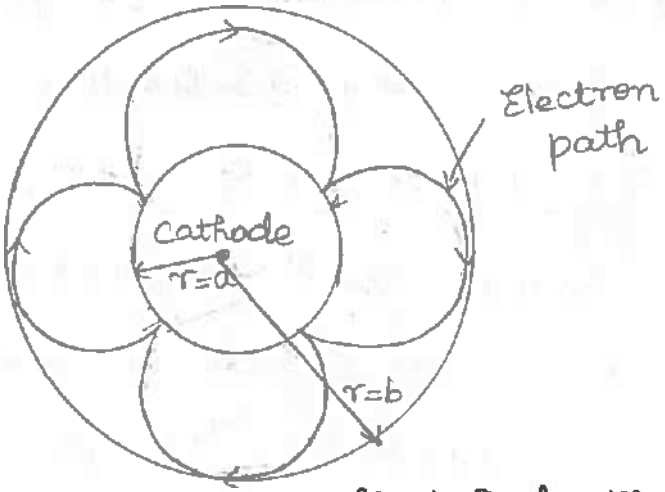
The anode is a slow wave structure consisting of several reentrant cavities equi-spaced around the circumstance and coupled together through the anode cathode space by means of slots.

The dc voltage Vo is applied between the cathode and the anode and a dc magnetic flux density Bo is maintained in the positive z direction by means of a permanent magnet or an electromagnet.

The accelerated electrons in the curved trajectory, when retarded by the RF field, the transfer energy from an electron to the cavities to grow RF oscillations till the system RF losses balances the RF oscillations for stability.



Schematic diagram of a cylindrical magnetron



Electron path in a cylindrical magnetron

Impedance Matching

- The basic idea of impedance matching is illustrated in Figure 1, which shows an impedance matching network placed between a load impedance and a transmission line.

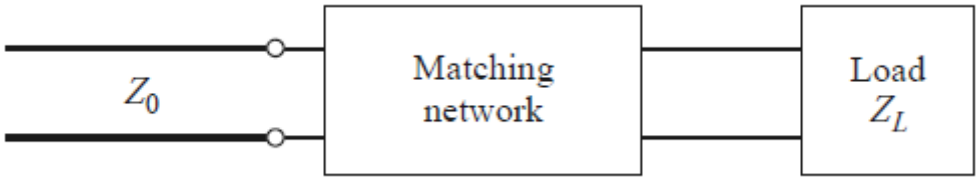


Figure 1

- The matching network is ideally lossless, to avoid unnecessary loss of power, and is usually designed so that the impedance seen looking into the matching network is Z_0.
- Then reflections will be eliminated on the transmission line to the left of the matching network, although there will usually be multiple reflections between the matching network and the load. This procedure is sometimes referred to as tuning.
- Impedance matching or tuning is important for the following reasons:
  - Maximum power is delivered when the load is matched to the line (assuming the generator is matched), and power loss in the feed line is minimized.
  - Impedance matching sensitive receiver components (antenna, low-noise amplifier, etc.) may improve the signal-to-noise ratio of the system.
  - Impedance matching in a power distribution network (such as an antenna array feed network) may reduce amplitude and phase errors.
- Factors that may be important in the selection of a particular matching network includes complexity, bandwidth, implementation and adjustability.

Matching with Lumped Elements (L Networks)

- In either of the configurations of Figure 2, the reactive elements may be either inductors or capacitors, depending on the load impedance.

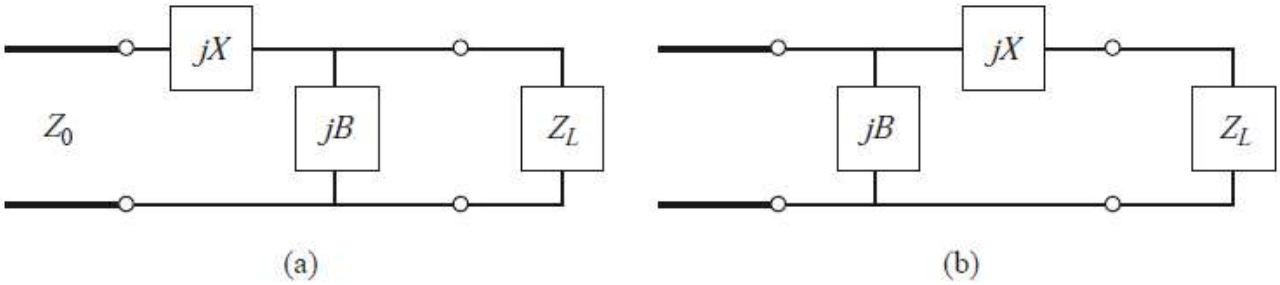


Figure 2

- Thus, there are eight distinct possibilities for the matching circuit for various load impedances.
- If the frequency is low enough and/or the circuit size is small enough, actual lumped-element capacitors and inductors can be used.

- This may be feasible for frequencies up to about 1 GHz or so, although modern microwave integrated circuits may be small enough such that lumped elements can be used at higher frequencies as well.
- There is, however, a large range of frequencies and circuit sizes where lumped elements may not be realizable.
- This is a limitation of the L-section matching technique.

Single-Stub Tuning

- Another popular matching technique uses a single open-circuited or short-circuited length of transmission line (a stub) connected either in parallel or in series with the transmission feed line at a certain distance from the load, as shown in Figure 3.

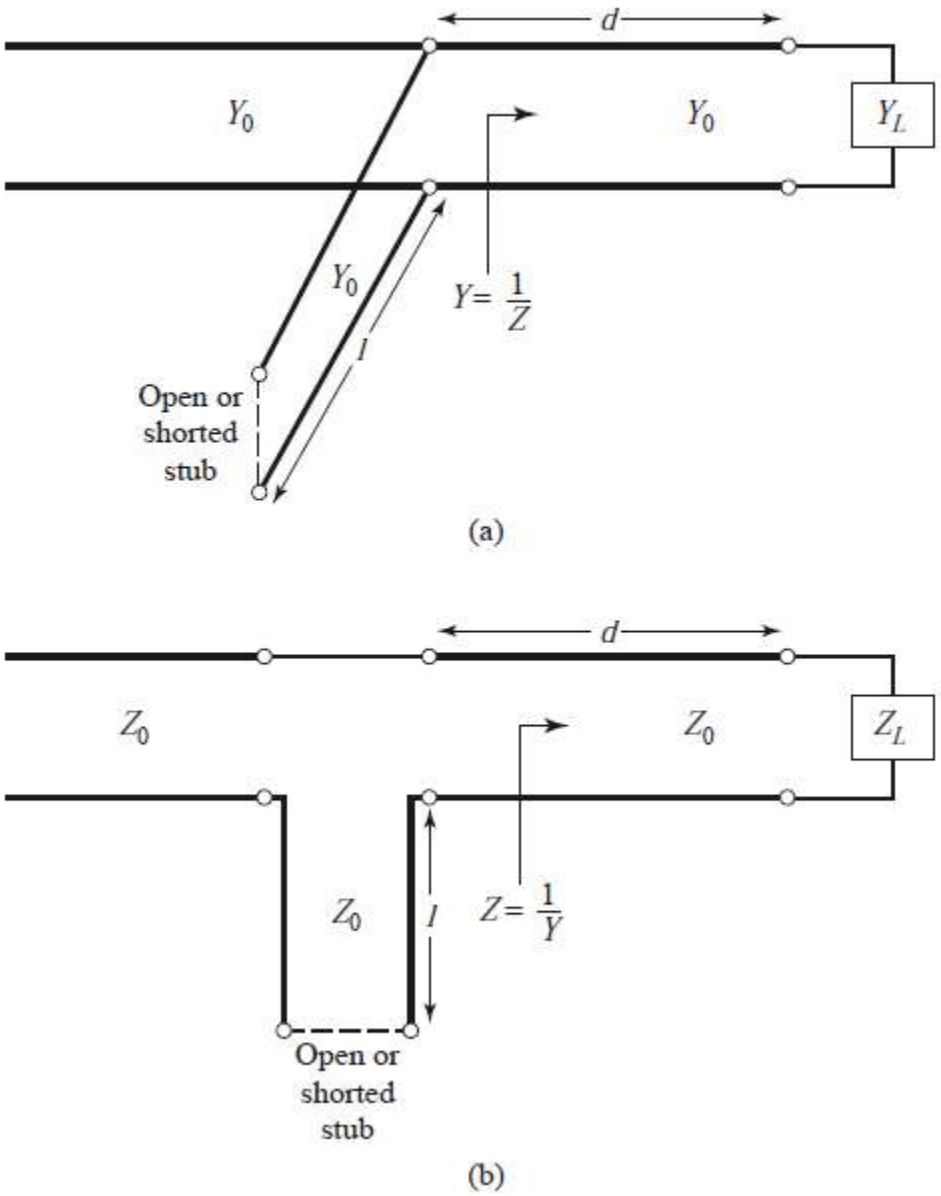


Figure 3

- Such a single-stub tuning circuit is often very convenient because the stub can be fabricated as part of the transmission line media of the circuit, and lumped elements are avoided.

- Shunt stubs are preferred for microstrip line or stripline, while series stubs are preferred for slotline or coplanar waveguide.
- In single-stub tuning the two adjustable parameters are the distance, d, from the load to the stub position, and the value of susceptance or reactance provided by the stub.
- For the shunt-stub case, the basic idea is to select d so that the admittance, Y, seen looking into the line at distance d from the load is of the form $Y_0 + jB$.
- Then the stub susceptance is chosen as $-jB$, resulting in a matched condition.
- For the series-stub case, the distance d is selected so that the impedance, Z, seen looking into the line at a distance d from the load is of the form $Z_0 + jX$.
- Then the stub reactance is chosen as $-jX$, resulting in a matched condition.
- The proper length of an open or shorted transmission line section can provide any desired value of reactance or susceptance.
- For a given susceptance or reactance, the difference in lengths of an open- or short-circuited stub is $\lambda/4$.
- For transmission line media such as microstrip or stripline, open-circuited stubs are easier to fabricate since a via hole through the substrate to the ground plane is not needed.
- For lines like coax or waveguide, however, short-circuited stubs are usually preferred because the cross-sectional area of such an open-circuited line may be large enough (electrically) to radiate, in which case the stub is no longer purely reactive.

Double-Stub Tuning

- The single-stub tuner of the previous section is able to match any load impedance (having a positive real part) to a transmission line, but suffers from the disadvantage of requiring a variable length of line between the load and the stub.

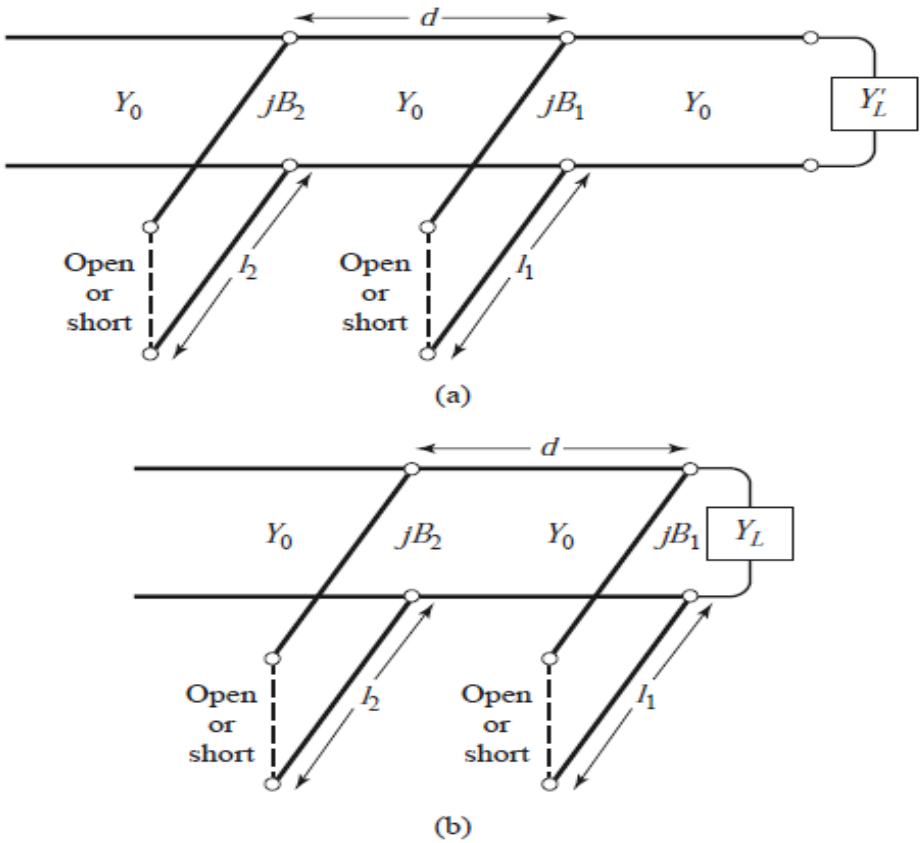


Figure 4

- This may not be a problem for a fixed matching circuit, but would probably pose some difficulty if an adjustable tuner was desired.
- In this case, the double-stub tuner, which uses two tuning stubs in fixed positions, can be used.
- Such tuners are often fabricated in coaxial line with adjustable stubs connected in shunt to the main coaxial line. However, a double-stub tuner cannot match all load impedances.
- The double-stub tuner circuit is shown in Figure 4a, where the load may be an arbitrary distance from the first stub.
- Although this is more representative of a practical situation, the circuit of Figure 4b, where the load Y_L' has been transformed back to the position of the first stub, is easier to deal with and does not lose any generality.
- The shunt stubs shown in Figure 4 can be conveniently implemented for some types of transmission lines, while series stubs are more appropriate for other types of lines.
- In either case, the stubs can be open-circuited or short-circuited.

Microwave Filters

- A filter is a two-port network used to control the frequency response at a certain point in an RF or microwave system by providing transmission at frequencies within the passband of the filter and attenuation in the stopband of the filter.
- Typical frequency responses include low-pass, high-pass, bandpass, and band-reject characteristics.

FILTER DESIGN BY THE IMAGE PARAMETER METHOD

- The image parameter method of filter design involves the specification of passband and stopband characteristics for a cascade of simple two-port networks.
- The method is relatively simple but has the disadvantage that an arbitrary frequency response cannot be incorporated into the design.
- The image parameter method also finds application in solid-state traveling-wave amplifier design.

Image Impedances and Transfer Functions for Two-Port Networks

- We begin with definitions of the image impedances and voltage transfer function for an arbitrary reciprocal two-port network; these results are required for the analysis and design of filters by the image parameter method.
- Consider the arbitrary two-port network shown in Figure 1, where the network is specified by its ABCD parameters.
- Note that the reference direction for the current at port 2 has been chosen according to the convention for ABCD parameters.
- The image impedances, Z_{i1} and Z_{i2}, are defined for this network as follows:
    - Z_{i1} = input impedance at port 1 when port 2 is terminated with Z_{i2}
    - Z_{i2} = input impedance at port 2 when port 1 is terminated with Z_{i1}.
- Thus both ports are matched when terminated in their image impedances.

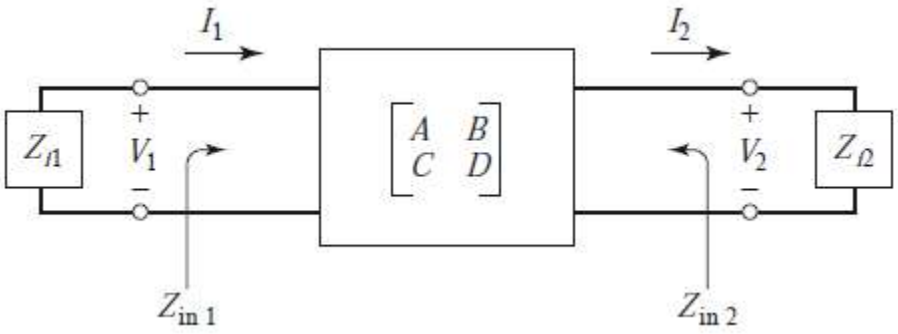
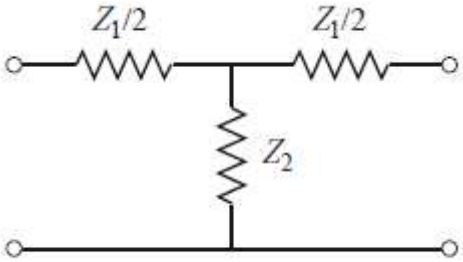
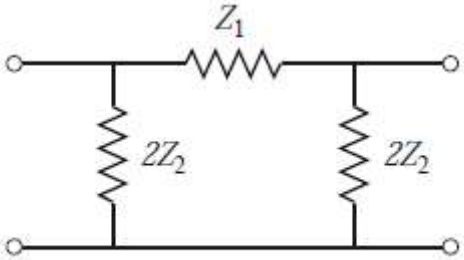
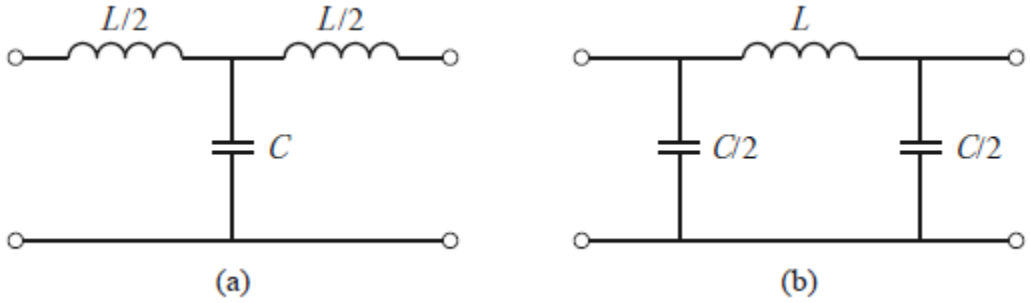


Figure 1

Image Parameters for T and π networks

 T-Network	 π-Network
ABCD Parameters: $A = 1 + \frac{Z_1}{2Z_2}$ $B = Z_1 + \frac{Z_1^2}{4Z_2}$ $C = \frac{1}{Z_2}$ $D = 1 + \frac{Z_1}{2Z_2}$	**ABCD Parameters:** $A = 1 + \frac{Z_1}{2Z_2}$ $B = Z_1$ $C = \frac{1}{Z_2} + \frac{Z_1}{4Z_2^2}$ $D = 1 + \frac{Z_1}{2Z_2}$
Z Parameter: $Z_{11} = Z_{22} = Z_2 + \frac{Z_1}{2}$ $Z_{12} = Z_{21} = Z_2$	**Y Parameter:** $Y_{11} = Y_{22} = \frac{1}{Z_1} + \frac{1}{2Z_2}$ $Y_{12} = Y_{21} = \frac{1}{Z_1}$
Image Impedance: $Z_{iT} = \sqrt{Z_1 Z_2}\sqrt{1 + \frac{Z_1}{4Z_2}}$	**Image Impedance:** $Z_{i\pi} = \frac{\sqrt{Z_1 Z_2}}{\sqrt{1 + \frac{Z_1}{4Z_2}}} = \frac{Z_1 Z_2}{Z_{iT}}$
Propagation Constant: $e^{\gamma} = 1 + \frac{Z_1}{2Z_2} + \sqrt{\frac{Z_1}{Z_2} + \frac{Z_1^2}{4Z_2^2}}$	**Propagation Constant:** $e^{\gamma} = 1 + \frac{Z_1}{2Z_2} + \sqrt{\frac{Z_1}{Z_2} + \frac{Z_1^2}{4Z_2^2}}$

Constant k Low pass Filter



Cut off Frequency

$$\omega_c = \frac{2}{\sqrt{LC}}$$

Characteristics Impedance

$$R_0 = \sqrt{\frac{L}{C}} = k$$

Image Impedance

$$Z_{iT} = R_0 \sqrt{1 - \frac{\omega^2}{\omega_c^2}}$$

Propagation Factor

$$e^{\gamma} = 1 - \frac{2\omega^2}{\omega_c^2} + \frac{2\omega}{\omega_c} \sqrt{\frac{\omega^2}{\omega_c^2} - 1}$$

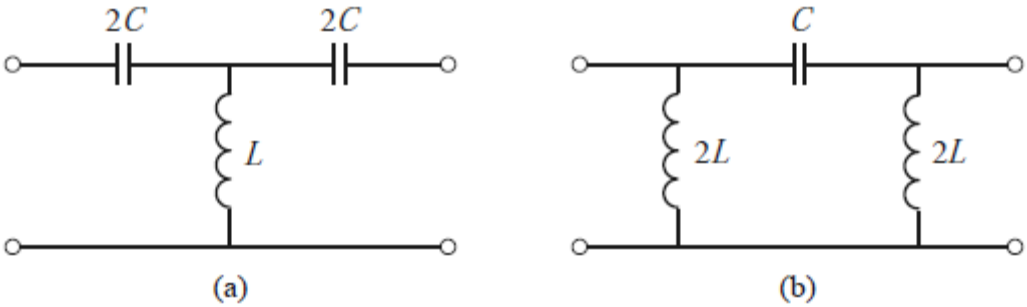
Constant k High Pass Filter

Cut off Frequency

$$\omega_c = \frac{1}{2\sqrt{LC}}$$

Characteristics Impedance

$$R_0 = \sqrt{\frac{L}{C}} = k$$



FILTER DESIGN BY THE INSERTION LOSS METHOD

- A perfect filter would have zero insertion loss in the passband, infinite attenuation in the stopband, and a linear phase response (to avoid signal distortion) in the passband.
- Of course, such filters do not exist in practice, so compromises must be made; herein lies the art of filter design.
- The image parameter method of the previous section may yield a usable filter response for some applications, but there is no methodical way of improving the design.
- The insertion loss method, however, allows a high degree of control over the passband and stopband amplitude and phase characteristics, with a systematic way to synthesize a desired response.
- The necessary design trade-offs can be evaluated to best meet the application requirements.
- If, for example, a minimum insertion loss is most important, a binomial response could be used; a Chebyshev response would satisfy a requirement for the sharpest cutoff.
- If it is possible to sacrifice the attenuation rate, a better phase response can be obtained by using a linear phase filter design.
- In addition, in all cases, the insertion loss method allows filter performance to be improved in a straightforward manner, at the expense of a higher order filter.
- For the filter prototypes to be discussed below, the order of the filter is equal to the number of reactive elements.

Characterization by Power Loss Ratio

- In the insertion loss method a filter response is defined by its insertion loss, or power loss ratio, P_{LR}:

$$P_{LR} = \frac{Power\ available\ from\ source}{Power\ delivered\ to\ load} = \frac{P_{inc}}{P_{load}} = \frac{1}{1-|\Gamma(\omega)|^2} \qquad \longrightarrow (1)$$

- Observe that this quantity is the reciprocal of $|S_{12}|^2$ if both load and source are matched.
- The insertion loss (IL) in dB is

$$IL = 10 \log P_{LR} \qquad \longrightarrow (2)$$

- We know that $|\Gamma(\omega)|^2$ is an even function of ω; therefore it can be expressed as a polynomial in ω^2.
- Thus we can write

$$|\Gamma(\omega)|^2 = \frac{M(\omega^2)}{M(\omega^2)+N(\omega^2)} \qquad \longrightarrow (3)$$

where M and N are real polynomials in ω^2. Substituting this form in (1) gives the following:

$$P_{LR} = \frac{1}{1-\dfrac{M(\omega^2)}{M(\omega^2)+N(\omega^2)}} = \frac{M(\omega^2)+N(\omega^2)}{M(\omega^2)+N(\omega^2)-M(\omega^2)} = \frac{M(\omega^2)+N(\omega^2)}{N(\omega^2)}$$

$$= 1 + \frac{M(\omega^2)}{N(\omega^2)} \qquad \longrightarrow (4)$$

- For a filter to be physically realizable its power loss ratio must be of the form in (4).
- Notice that specifying the power loss ratio simultaneously constrains the magnitude of the reflection coefficient, $|\Gamma(\omega)|$.
- We now discuss some practical filter responses.

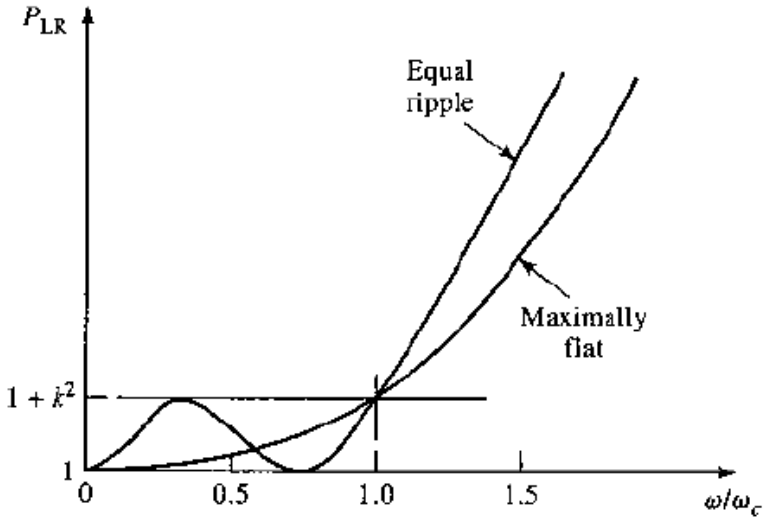
Maximally flat:

- This characteristic is also called the binomial or Butterworth response, and is optimum in the sense that it provides the flattest possible passband response for a given filter complexity, or order.
- For a low-pass filter, it is specified by

$$P_{LR} = 1 + k^2 \left(\frac{\omega}{\omega_c}\right)^{2N} \qquad \longrightarrow (5)$$

where N is the order of the filter and ω_c is the cutoff frequency.

- The passband extends from $\omega = 0$ to $\omega = \omega_c$; at the band edge the power loss ratio is $1 + k^2$.
- If we choose this as the −3 dB point, as is common, we have $k = 1$, which we will assume from now on.
- For $\omega > \omega_c$, the attenuation increases monotonically with frequency, as shown in Figure 1.



- For $\omega \gg \omega_c$, $P_{LR} \cong k^2(\omega/\omega_c)^{2N}$, which shows that the insertion loss increases at the rate of 20N dB/decade.
- Like the binomial response for multisection quarter-wave matching transformers, the first (2N−1) derivatives of (5) are zero at ω = 0.

Equal ripple:

- If a Chebyshev polynomial is used to specify the insertion loss of an N^{th} order low-pass filter as

$$P_{LR} = 1 + k^2 T_N^2 \left(\frac{\omega}{\omega_c}\right)$$

then a sharper cutoff will result, although the passband response will have ripples of amplitude $1 + k^2$, as shown in Figure 1, since $T_N(x)$ oscillates between ± 1 for $|x|^2$.

- Thus, k^2 determines the passband ripple level.
- For large x, $T_N(x) \cong \frac{1}{2}(2x)^N$, so for $\omega \gg \omega_c$ the insertion loss becomes

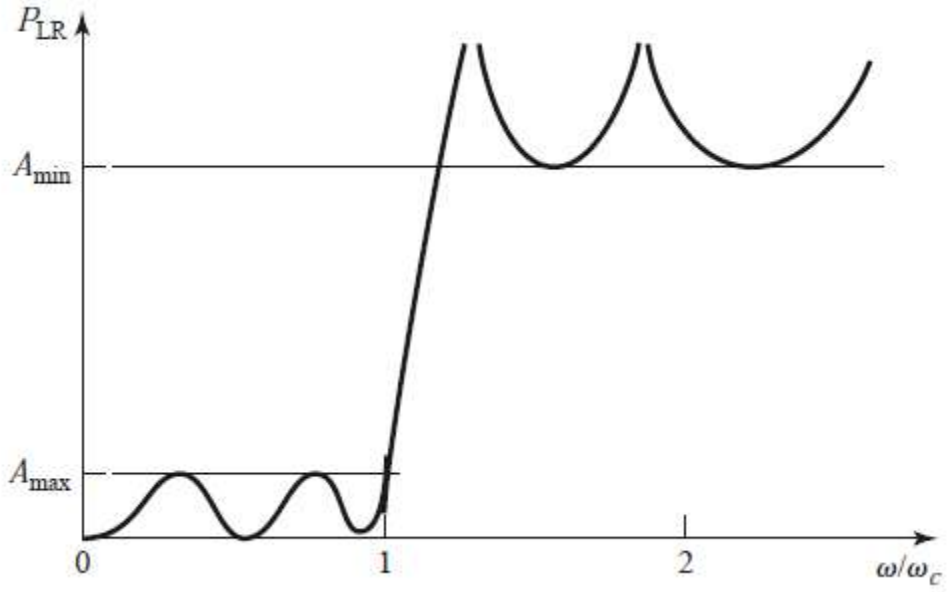
$$P_{LR} \cong \frac{k^2}{4}\left(\frac{2\omega}{\omega_c}\right)^{2N}$$

which also increases at the rate of 20N dB/decade.

- However, the insertion loss for the Chebyshev case is $(2^{2N})/4$ greater than the binomial response at any given frequency where $\omega \gg \omega_c$.

Elliptic function:

- The maximally flat and equal-ripple responses both have monotonically increasing attenuation in the stopband.
- In many applications it is adequate to specify a minimum stopband attenuation, in which case a better cutoff rate can be obtained.
- Such filters are called elliptic function filters, and they have equal-ripple responses in the passband as well as in the stopband, as shown in Figure 2.
- The maximum attenuation in the passband, A_{max}, can be specified, as well as the minimum attenuation in the stopband, A_{min}.
- Elliptic function filters are difficult to synthesize, so we will not consider them further; the interested reader is referred to reference



Linear phase:

- The above filters specify the amplitude response, but in some applications (such as multiplexing filters for communication systems) it is important to have a linear phase response in the passband to avoid signal distortion.
- Since a sharp-cutoff response is generally incompatible with a good phase response, the phase response of a filter must be deliberately synthesized, usually resulting in an inferior attenuation characteristic.
- A linear phase characteristic can be achieved with the following phase response:

$$\emptyset(\omega) = A\omega\left[1 + p\left(\frac{\omega}{\omega_c}\right)^{2N}\right]$$

where $\varphi(\omega)$ is the phase of the voltage transfer function of the filter, and p is a constant.

- A related quantity is the group delay, defined as

$$\tau_d = \frac{d\emptyset}{d\omega} = A\left[1 + p(2N+1)\left(\frac{\omega}{\omega_c}\right)^{2N}\right]$$

which shows that the group delay for a linear phase filter is a maximally flat function.

- More general filter specifications can be obtained, but the above cases are the most common.

Microwave Amplifier Design:

- Signal amplification is one of the most basic and prevalent circuit functions in modern RF and microwave systems.
- Early microwave amplifiers relied on tubes, such as klystrons and traveling-wave tubes, or solid-state reflection amplifiers based on the negative resistance characteristics of tunnel or varactor diodes.

Two Port Power Gains
Power gain:

$$G = \frac{P_L}{P_{in}}$$

- It is the ratio of power dissipated in the load Z_L to the power delivered to the input of the two-port network.
- This gain is independent of Z_S, although the characteristics of some active devices may be dependent on Z_S.

Available power gain:

$$G_A = \frac{P_{avn}}{P_{avs}}$$

- It is the ratio of the power available from the two-port network to the power available from the source.
- This assumes conjugate matching of both the source and the load, and depends on Z_S, but not Z_L.

Transducer power gain:

$$G_T = \frac{P_L}{P_{avs}}$$

- It is the ratio of the power delivered to the load to the power available from the source.
- This depends on both Z_S and Z_L.

These definitions differ primarily in the way the source and load are matched to the two port device; if the input and output are both conjugately matched to the two-port device, then the gain is maximized and

$$G = G_A = G_T$$

Stability

Unconditional stability: The network is unconditionally stable if $|\Gamma_{in}| < 1$ and |_out| < 1 for all passive source and load impedances (i.e., $|\Gamma_S| < 1$ and $|\Gamma_L| < 1$).

Conditional stability: The network is conditionally stable if $|\Gamma_{in}| < 1$ and $|\Gamma_{out}| < 1$ only for a certain range of passive source and load impedances. This case is also referred to as potentially unstable.

Microwave Power Amplifier Design

- Power amplifiers are used in the final stages of radar and radio transmitters to increase the radiated power level.
- Typical output powers may be on the order of 100–500 mW for mobile voice or data communications systems, or in the range of 1–100 W for radar or fixed point radio systems.
- Important considerations for RF and microwave power amplifiers are efficiency, gain, intermodulation distortion, and thermal effects.
- Single transistors can provide output powers of 10–100 W at UHF frequencies, while devices at higher frequencies are generally limited to output powers less than 10 W.
- Various power-combining techniques can be used in conjunction with multiple transistors if higher output powers are required.
- So far we have considered only small-signal amplifiers, where the input signal power is low enough that the transistor can be assumed to operate as a linear device.
- The scattering parameters of linear devices are well defined and do not depend on the input power level or output load impedance, a fact that greatly simplifies the design of fixed-gain and low noise amplifiers.
- For high input powers (e.g., in the range of the 1 dB compression point or third-order intercept point), transistors do not behave linearly.
- In this case the impedances seen at the input and output of the transistor will depend on the input power level, and this greatly complicates the design of power amplifiers.

Characteristics of Power Amplifiers and Amplifier Classes

- The power amplifier is usually the primary consumer of DC power in most hand-held wireless devices, so amplifier efficiency is an important consideration.
- One measure of amplifier efficiency is the ratio of RF output power to DC input power:

$$\eta = \frac{P_{out}}{P_{dc}} \qquad \longrightarrow (1)$$

- This quantity is sometimes referred to as drain efficiency (or collector efficiency).
- One drawback of this definition is that it does not account for the RF power delivered at the input to the amplifier.
- Since most power amplifiers have relatively low gains, the efficiency of (1) tends to overrate the actual efficiency.
- A better measure that includes the effect of input power is the power added efficiency, defined as

$$\eta_{PAE} = PAE = \frac{P_{out} - P_{in}}{P_{DC}} = \left(1 - \frac{1}{G}\right)\frac{P_{out}}{P_{DC}} = \left(1 - \frac{1}{G}\right)\eta \qquad \longrightarrow (2)$$

  where G is the power gain of the amplifier.
- Silicon bipolar junction transistor amplifiers in the cellular telephone band of 800–900 MHz band have power added efficiencies on the order of 80%, but efficiency drops quickly with increasing frequency.
- Power amplifiers are often designed to provide the best efficiency, even if this means that the resulting gain is less than the maximum possible.
- Another useful parameter for power amplifiers is the compressed gain, G_1, defined as the gain of the amplifier at the 1 dB compression point. Thus, if G_0 is the small-signal (linear) power gain, we have

$$G_1(dB) = G_0(dB) - 1 \quad \longrightarrow (3)$$

- Nonlinearities can lead to the generation of spurious frequencies and intermodulation distortion.
- This can be a serious issue in wireless transmitters, especially in a multicarrier system, where spurious signals may appear in adjacent channels.
- Linearity is also critical for nonconstant envelope modulations, such as amplitude shift keying and higher order quadrature amplitude modulation methods.
- Class A amplifiers are inherently linear circuits, where the transistor is biased to conduct over the entire range of the input signal cycle.
- Because of this, class A amplifiers have a theoretical maximum efficiency of 50%.
- Most small-signal and low-noise amplifiers operate as class A circuits.
- In contrast, the transistor in a class B amplifier is biased to conduct only during one-half of the input signal cycle.
- Usually two complementary transistors are operated in a class B push-pull amplifier to provide amplification over the entire cycle.
- The theoretical efficiency of a class B amplifier is 78%.
- Class C amplifiers are operated with the transistor near cutoff for more than half of the input signal cycle, and generally use a resonant circuit in the output stage to recover the fundamental.
- Class C amplifiers can achieve efficiencies near 100% but can only be used with constant envelope modulations.
- Higher classes, such as class D, E, F, and S, use the transistor as a switch to pump a highly resonant tank circuit, and may achieve very high efficiencies.
- The majority of communications transmitters operating at UHF frequencies or above rely on class A, AB, or B power amplifiers because of the need for low distortion products.

Low Noise Amplifier Design

- Besides stability and gain, another important design consideration for a microwave amplifier is its noise figure.
- In receiver applications especially it is often required to have a preamplifier with as low a noise figure as possible since the first stage of a receiver front end has the dominant effect on the noise performance of the overall system.
- Generally it is not possible to obtain both minimum noise figure and maximum gain for an amplifier, so some sort of compromise must be made.
- This can be done by using constant-gain circles and circles of constant noise figure to select a usable trade-off between noise figure and gain.
- Here we will derive the equations for constant–noise figure circles and show how they are used in transistor amplifier design.
- The noise figure of a two-port amplifier can be expressed as

$$F = F_{min} + \frac{R_N}{G_S}\left|Y_S - Y_{opt}\right|^2 \quad \longrightarrow (1)$$

where the following definitions apply:
$Y_S = G_S + jB_S$ = source admittance presented to transistor.
Y_{opt} = optimum source admittance that results in minimum noise figure.
F_{min} = minimum noise figure of transistor, attained when $Y_S = Y_{opt}$.

R_N = equivalent noise resistance of transistor.
G_S = real part of source admittance.

$$Y_S = \frac{1}{Z_0}\frac{1-\Gamma_S}{1+\Gamma_S}$$

$$Y_{opt} = \frac{1}{Z_0}\frac{1-\Gamma_{opt}}{1+\Gamma_{opt}}$$

$$\left|Y_S - Y_{opt}\right|^2 = \frac{4}{Z_0^2}\frac{\left|\Gamma_S - \Gamma_{opt}\right|^2}{|1+\Gamma_S|^2\left|1+\Gamma_{opt}\right|^2}$$

$$G_S = \frac{1}{Z_0}\frac{1-|\Gamma_S|^2}{|1+\Gamma_S|^2}$$

Hence

$$F = F_{min} + \frac{R_N}{\frac{1}{Z_0}\frac{1-|\Gamma_S|^2}{|1+\Gamma_S|^2}}\frac{4}{Z_0^2}\frac{\left|\Gamma_S - \Gamma_{opt}\right|^2}{|1+\Gamma_S|^2\left|1+\Gamma_{opt}\right|^2}$$

$$F = F_{min} + \frac{4R_N}{Z_0}\frac{\left|\Gamma_S - \Gamma_{opt}\right|^2}{(1-|\Gamma_S|^2)\left|1+\Gamma_{opt}\right|^2} \quad \longrightarrow (2)$$

- For a fixed noise figure F we can show that this result defines a circle in the Γ_S plane.
- First define the noise figure parameter, N, as

$$N = \frac{\left|\Gamma_S - \Gamma_{opt}\right|^2}{1-|\Gamma_S|^2} = \frac{(F-F_{min})\left|1+\Gamma_{opt}\right|^2}{4R_N/Z_0} \quad \longrightarrow (3)$$

- which is a constant for a given noise figure and set of noise parameters.
- Then rewrite (3) as

$$\left|\Gamma_S - \Gamma_{opt}\right|^2 = N(1-|\Gamma_S|^2)$$

$$\left(\Gamma_S - \Gamma_{opt}\right)\left(\Gamma_S^* - \Gamma_{opt}^*\right) = N(1-|\Gamma_S|^2)$$

$$\Gamma_S\Gamma_S^* - \left(\Gamma_S\Gamma_{opt}^* + \Gamma_S^*\Gamma_{opt}\right) + \Gamma_{opt}\Gamma_{opt}^* = N - N|\Gamma_S|^2$$

$$\Gamma_S\Gamma_S^* - \frac{\left(\Gamma_S\Gamma_{opt}^* + \Gamma_S^*\Gamma_{opt}\right)}{N+1} = N - \frac{\left|\Gamma_{opt}\right|^2}{N+1}$$

- Add $\frac{\left|\Gamma_{opt}\right|^2}{(N+1)^2}$ to both sides to complete the square to obtain

$$\left|\Gamma_S - \frac{\Gamma_{opt}}{N+1}\right| = \frac{\sqrt{N\left(N+!-\left|\Gamma_{opt}\right|^2\right)}}{N+1}$$

- This result defines circles of constant noise figure with centers at

$$C_F = \frac{\Gamma_{opt}}{N+1}$$

and radii of

$$R_F = \frac{\sqrt{N\left(N+!-\left|\Gamma_{opt}\right|^2\right)}}{N+1}$$

Microwave Mixer Design

- A mixer is a three-port device that uses a nonlinear or time-varying element to achieve frequency conversion.
- An ideal mixer produces an output consisting of the sum and difference frequencies of its two input signals.
- Operation of practical RF and microwave mixers is usually based on the nonlinearity provided by either a diode or a transistor.
- As we have seen, a nonlinear component can generate a wide variety of harmonics and other products of input frequencies, so filtering must be used to select the desired frequency components.
- Modern microwave systems typically use several mixers and filters to perform the functions of frequency up-conversion and down-conversion between baseband signal frequencies and RF carrier frequencies.

Mixer Characteristics

- The symbol and functional diagram for a mixer are shown in Figure 1.

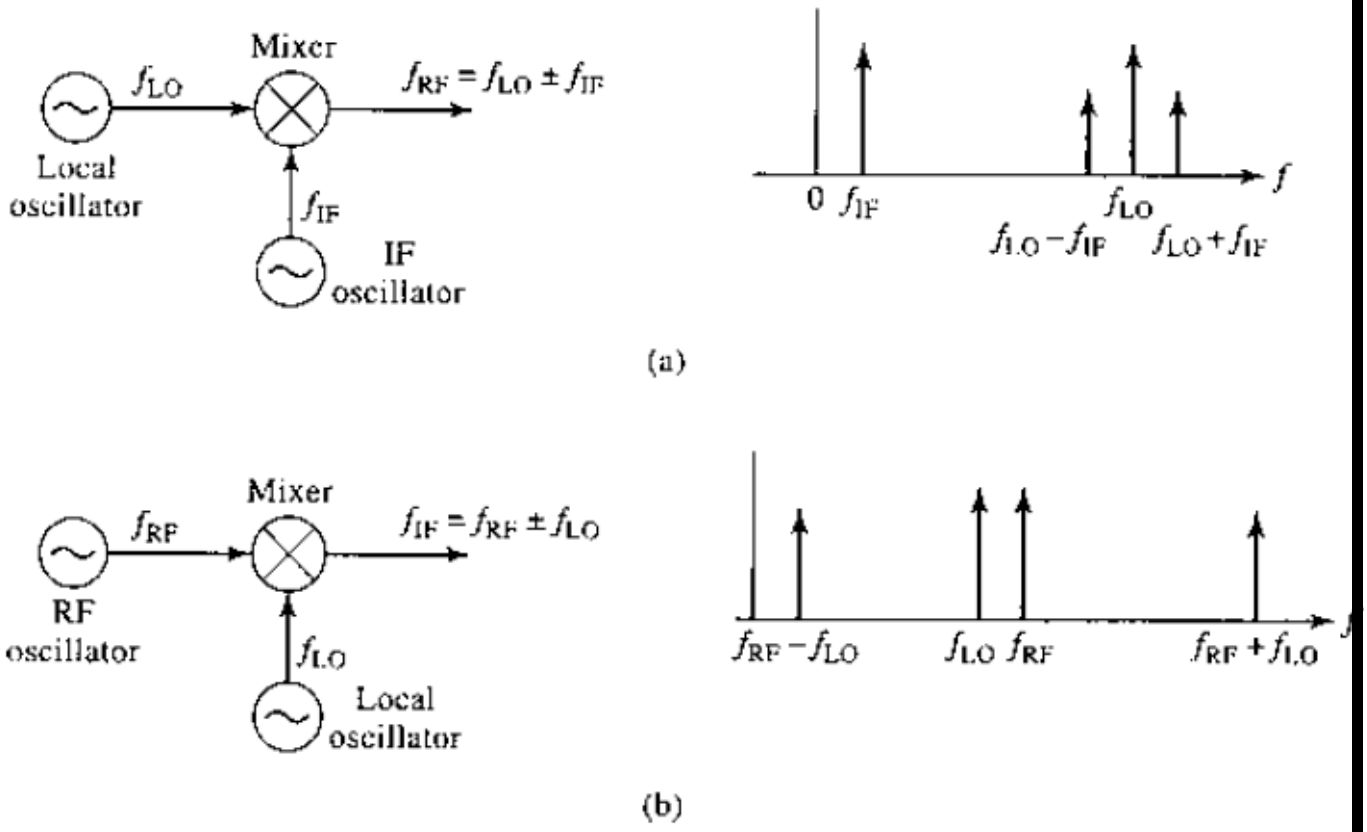


Figure 1

- The mixer symbol is intended to imply that the output is proportional to the product of the two input signals.
- We will see that this is an idealized view of mixer operation, which in actuality produces a large variety of harmonics and other undesired products of the input signals.

Up Conversion

- Figure 1a illustrates the operation of frequency up-conversion, as occurs in a transmitter.
- A local oscillator (LO) signal at the relatively high frequency f_{LO} is connected to one of the input ports of the mixer.
- The LO signal can be represented as

$$v_{LO}(t) = \cos 2\pi f_{LO} t$$

A lower frequency baseband or intermediate frequency (IF) signal is applied to the other mixer input.

This signal typically contains the information or data to be transmitted, and can be expressed for our purposes as

$$v_{IF}(t) = \cos 2\pi f_{IF} t$$

The output of the idealized mixer is given by the product of the LO and IF signals:

$$v_{RF}(t) = K v_{LO}(t) v_{IF}(t) = K \cos 2\pi f_{LO} t \cos 2\pi f_{IF} t$$

$$v_{RF}(t) = \frac{K}{2}[\cos 2\pi (f_{LO} - f_{IF})t + \cos 2\pi (f_{LO} + f_{IF})t]$$

where K is a constant accounting for the voltage conversion loss of the mixer.

The RF output is seen to consist of the sum and differences of the input signal frequencies:

$$f_{RF} = f_{LO} \pm f_{IF}$$

Down Conversion

Conversely, Figure 1b shows the process of frequency down-conversion, as used in a receiver.

In this case an RF input signal of the form

$$v_{RF}(t) = \cos 2\pi f_{RF} t$$

is applied to the input of the mixer, along with the LO signal.

The output of the mixer is

$$v_{IF}(t) = K v_{RF}(t) v_{LO}(t) = K \cos 2\pi f_{RF} t \cos 2\pi f_{LO} t$$

$$v_{IF}(t) = \frac{K}{2}[\cos 2\pi (f_{RF} - f_{LO})t + \cos 2\pi (f_{RF} + f_{LO})t]$$

Thus the mixer output consists of the sum and difference of the input signal frequencies.

The spectrum for these signals is shown in Figure 1b.

In practice, the RF and LO frequencies are relatively close together, so the sum frequency is approximately twice the RF frequency, while the difference is much smaller than f_{RF}.

The desired IF output in a receiver is the difference frequency, $f_{RF} - f_{LO}$, which is easily selected by low-pass filtering:

$$f_{IF} = f_{RF} - f_{LO}$$

Single-Ended Diode Mixer

A basic diode mixer circuit is shown in Figure 2.

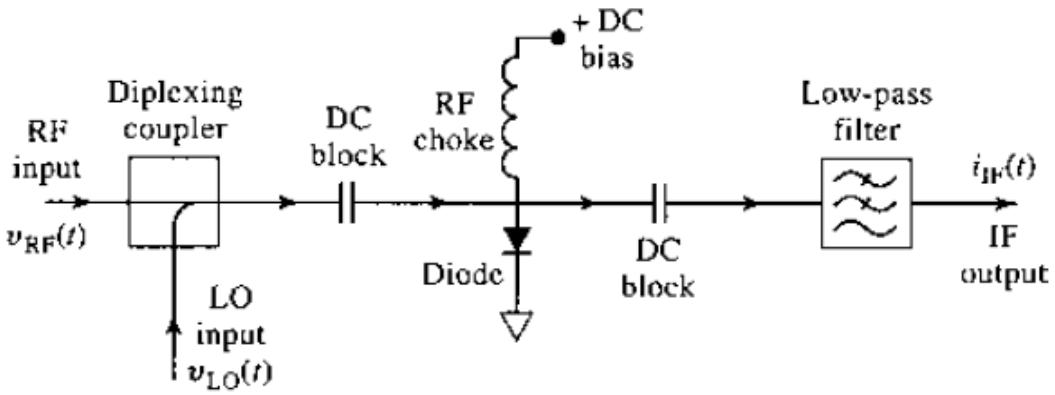


Figure 2

This type of mixer is called a single ended mixer because it uses a single diode element.
The RF and LO inputs are combined in a diplexer, which superimposes the two input voltages to drive the diode.
The duplexing function can be implemented using a directional coupler or hybrid junction to provide signal combining as well as isolation between the two inputs.
The diode may be biased with a DC bias voltage, which must be decoupled from the RF signal paths.
This is done by using DC blocking capacitors on either side of the diode, and an RF choke between the diode and the bias voltage source.
The AC output of the diode is passed through a low pass filter to provide the desired IF output voltage.
This description is for application as a down-converter, but the same mixer can be used for up-conversion since each port may be used interchangeably as an input or output port.

Single Ended FET Mixer

The circuit for a single-ended FET mixer is shown in Figure 3.
A duplexing coupler is again used to combine the RF and LO signals at the gate of the FET.
An impedance matching network is also usually required between the inputs and the FET, which typically presents a very low input impedance.
RF chokes are used to bias the gate at a negative voltage near pinch-off, and to provide a positive bias for the drain of the FET.
A bypass capacitor at the drain provides a return path for the LO signal, and a low-pass filter provides the final IF output signal.

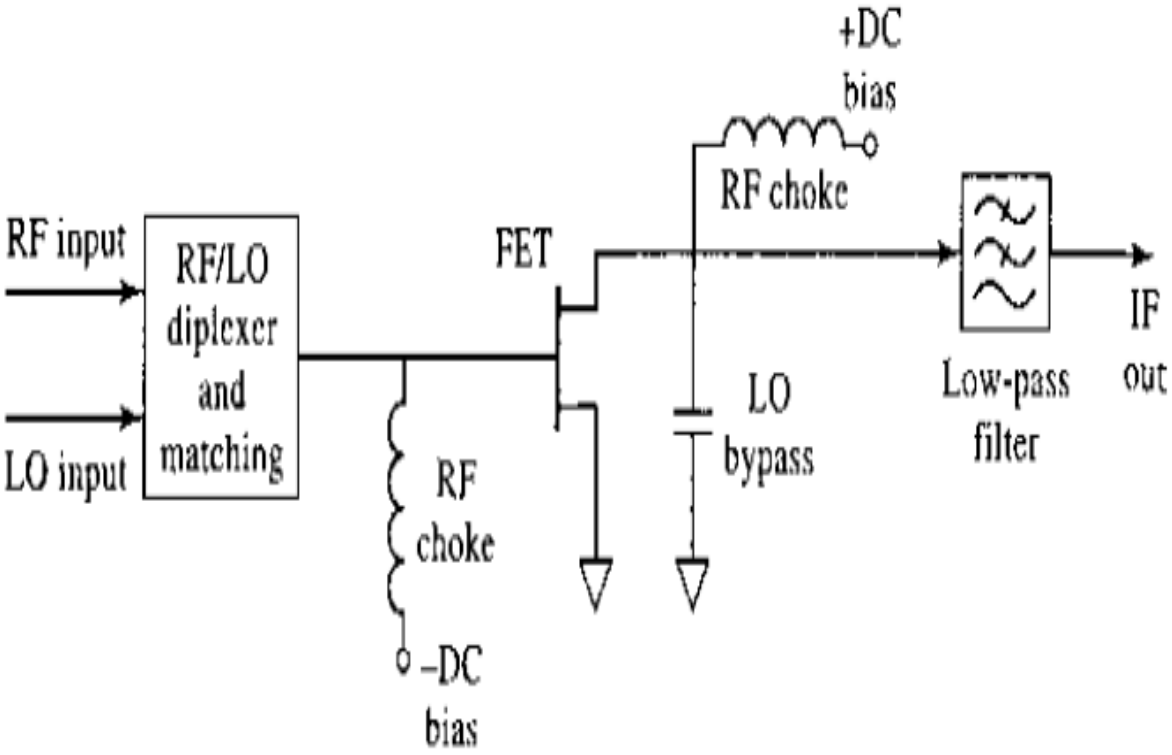


Figure 3

Balanced Mixer

RF input matching and RF-LO isolation can be improved through the use of a balanced mixer, which consists of two single-ended mixers combined with a hybrid junction.

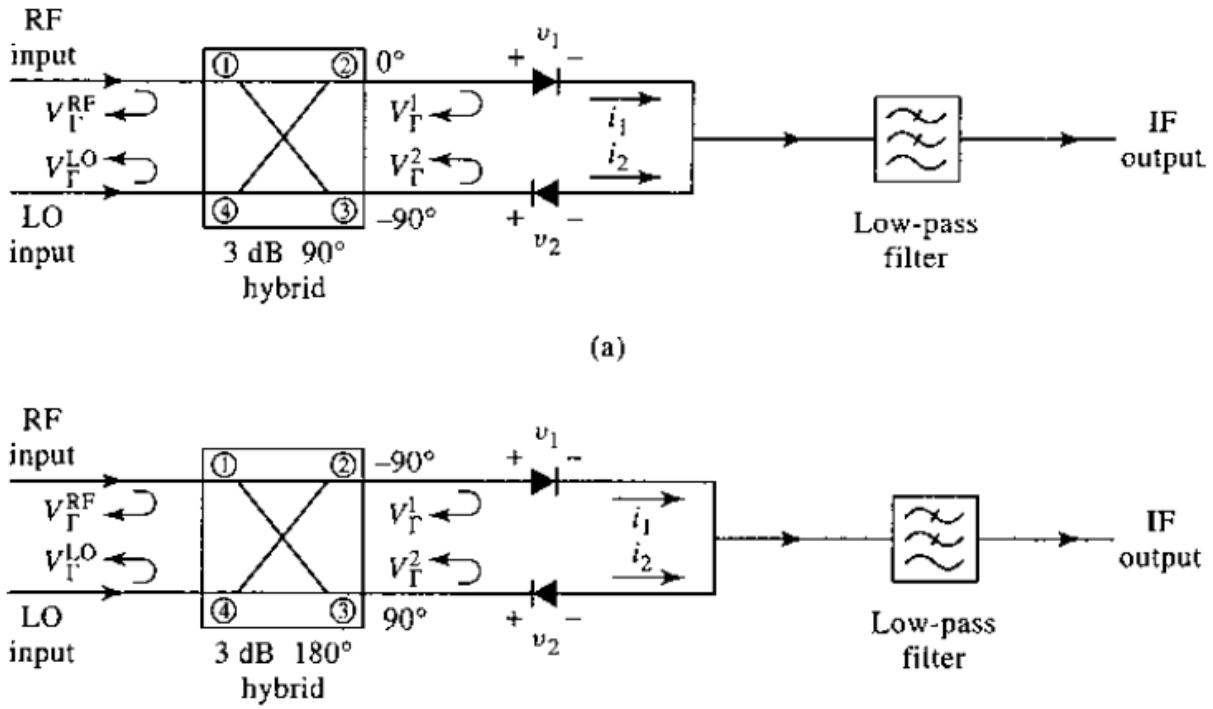


Figure 4

Figure 4 shows the basic configuration, with either a 90^o hybrid (Figure 4a), or a 180^o hybrid (Figure 4b).

A balanced mixer using a 90^o hybrid junction will ideally lead to a perfect input match at the RF port over a wide frequency range, while the use of a 180^o hybrid will ideally lead to perfect RF-LO isolation over a wide frequency range.

In addition, both mixers will reject all even-order intermodulation products.

Microwave Oscillator Design

Transistor Oscillator

In a transistor oscillator, a negative resistance one-port network is effectively created by terminating a potentially unstable transistor with an impedance designed to drive the device in an unstable region.

The circuit model of a transistor oscillator is shown in Figure 1.

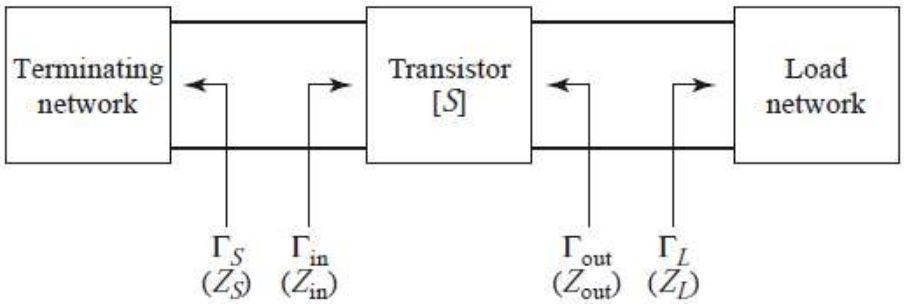


Figure 1

In this circuit, the RF output port is part of the load network on the output side of the transistor, but it is also possible to use the terminating network to the left of the transistor as the output port.
In the case of an amplifier, we preferred a device with a high degree of stability – ideally, an unconditionally stable device.
For an oscillator, we require a device with a high degree of instability.
Typically, common source or common gate FET configurations are used (common emitter or common base for bipolar junction devices), often with positive feedback to enhance the instability of the device.
After the transistor configuration is selected, the output stability circle can be drawn in the Γ_L plane, and Γ_L selected to produce a large value of negative resistance at the input to the transistor.
Then the terminating impedance $Z_S = R_S + jX_S$ can be chosen to match Z_{in}.
Because such a design often relies on the small-signal scattering parameters, and because R_{in} will become less negative as the oscillator power builds up, it is often necessary to choose R_S so that $R_S + R_{in} < 0$.
Otherwise, oscillation may cease if increasing RF power increases R_{in} to the point where $R_S + R_{in} > 0$.
In practice, a value of

$$R_s = -\frac{R_{in}}{3} \quad \longrightarrow (1a)$$

is often used.
The reactive part of Z_S is chosen to resonate the circuit,

$$X_s = -X_{in} \quad \longrightarrow (1b)$$

When oscillation occurs between the termination network and the transistor, oscillation will simultaneously occur at the output port, which we can show as follows.
For steady-state oscillation at the input port, we must have $\Gamma_S\Gamma_{in} = 1$.
Then we have

$$\frac{1}{\Gamma_S} = \Gamma_{in} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} = \frac{S_{11} - \Delta\Gamma_L}{1 - S_{22}\Gamma_L} \quad \longrightarrow (2)$$

where $\Delta = S_{11}\,S_{22} - S_{12}S_{21}$. Solving for Γ_L gives

$$\Gamma_L = \frac{1 - S_{11}\Gamma_S}{S_{22} - \Delta\Gamma_S} \quad \longrightarrow (3)$$

Also we have

$$\Gamma_{out} = S_{22} + \frac{S_{12}S_{21}\Gamma_S}{1 - S_{11}\Gamma_S} = \frac{S_{22} - \Delta\Gamma_S}{1 - S_{11}\Gamma_S} \quad \longrightarrow (4)$$

which shows that $\Gamma_L\Gamma_{out} = 1$, and hence $Z_L = -Z_{out}$.
Thus, the condition for oscillation at the load network is satisfied.
Note that it is preferable to use the large-signal scattering parameters of the transistor in the above development.

MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

EC8791 Embedded and Real Time Systems

Semester - 07

Notes

MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

Vision

To excel in providing value based education in the field of Electronics and Communication Engineering, keeping in pace with the latest technical developments through commendable research, to raise the intellectual competence to match global standards and to make significant contributions to the society upholding the ethical standards.

Mission

- ✓ To deliver Quality Technical Education, with an equal emphasis on theoretical and practical aspects.
- ✓ To provide state of the art infrastructure for the students and faculty to upgrade their skills and knowledge.
- ✓ To create an open and conducive environment for faculty and students to carry out research and excel in their field of specialization.
- ✓ To focus especially on innovation and development of technologies that is sustainable and inclusive, and thus benefits all sections of the society.
- ✓ To establish a strong Industry Academic Collaboration for teaching and research, that could foster entrepreneurship and innovation in knowledge exchange.
- ✓ To produce quality Engineers who uphold and advance the integrity, honour and dignity of the engineering.

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

1. To provide the students with a strong foundation in the required sciences in order to pursue studies in Electronics and Communication Engineering.
2. To gain adequate knowledge to become good professional in electronic and communication engineering associated industries, higher education and research.
3. To develop attitude in lifelong learning, applying and adapting new ideas and technologies as their field evolves.
4. To prepare students to critically analyze existing literature in an area of specialization and ethically develop innovative and research oriented methodologies to solve the problems identified.
5. To inculcate in the students a professional and ethical attitude and an ability to visualize the engineering issues in a broader social context.

PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO1: Design, develop and analyze electronic systems through application of relevant electronics, mathematics and engineering principles.

PSO2: Design, develop and analyze communication systems through application of fundamentals from communication principles, signal processing, and RF System Design & Electromagnetics.

PSO3: Adapt to emerging electronics and communication technologies and develop innovative solutions for existing and newer problems.

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

EC8791 EMBEDDED AND REAL TIME SYSTEMS L T P C 3 0 0 3

Objectives:

- To Understand the concept of embedded system design and analysis.
- To learn the architecture of ARM processor.
- To learn the Programming of ARM processor
- To Expose the basic concepts of embedded programming.
- To Learn real time operating systems

UNIT I INTRODUCTION TO EMBEDDED SYSTEM DESIGN

Complex systems and microprocessors– Embedded system design process –Design example: Model train controller- Design methodologies- Design flows - Requirement Analysis – Specifications-System analysis and architecture design – Quality Assurance techniques - Designing with computing platforms – consumer electronics architecture – platform-level performance analysis.

UNIT II ARM PROCESSOR AND PERIPHERALS

ARM Architecture Versions – ARM Architecture – Instruction Set – Stacks and Subroutines – Features of the LPC 214X Family – Peripherals – The Timer Unit – Pulse Width Modulation Unit – UART – Block Diagram of ARM9 and ARM Cortex M3 MCU.

UNIT III EMBEDDED PROGRAMMING

Components for embedded programs- Models of programs- Assembly, linking and loading – compilation techniques- Program level performance analysis – Software performance optimization – Program level energy and power analysis and optimization – Analysis and optimization of program size- Program validation and testing.

UNIT IV REAL TIME SYSTEMS

Structure of a Real Time System — Estimating program run times – Task Assignment and Scheduling – Fault Tolerance Techniques – Reliability, Evaluation – Clock Synchronisation..

UNIT V PROCESSES AND OPERATING SYSTEMS

Introduction – Multiple tasks and multiple processes – Multirate systems- Preemptive realtime operating systems- Priority based scheduling- Interprocess communication mechanisms – Evaluating operating system performance- power optimization strategies for processes – Example Real time operating systems-POSIX-Windows CE. - Distributed embedded systems – MPSoCs and shared memory multiprocessors. – Design Example - Audio player, Engine control unit – Video accelerator.

TOTAL : 45 PERIODS

Course Outcomes:

At the end of the course, the student will be able to:

- **Summarize** Architecture and programming of ARM processor.
- **Applying** the concepts of embedded systems and its features.
- **Analyze** various Real Time Operating system is used in Embedded System.
- **Design** the flow &Techniques to develop Software for embedded system networks.
- **Analyze** Real-time applications using embedded System Products.

TEXT BOOK:

1. Marilyn Wolf, "Computers as Components - Principles of Embedded Computing System Design", Third Edition "Morgan Kaufmann Publisher (An imprint from Elsevier), 2012. (UNIT I, II, III, V)
2. Jane W.S.Liu,‖ Real Time Systems‖, Pearson Education, Third Indian Reprint, 2003.(UNIT IV)

REFERENCES:

1. Lyla B.Das, ―Embedded Systems : An Integrated Approach‖ Pearson Education, 2013.
2. Jonathan W.Valvano, "Embedded Microcomputer Systems Real Time Interfacing", Third Edition Cengage Learning, 2012.
3. David. E. Simon, "An Embedded Software Primer", 1st Edition, Fifth Impression, Addison-Wesley Professional, 2007.
4. Raymond J.A. Buhr, Donald L.Bailey, "An Introduction to Real-Time Systems- From Design to Networking with C/C++", Prentice Hall, 1999.
5. C.M. Krishna, Kang G. Shin, "Real-Time Systems", International Editions, Mc Graw Hill 1997
6. K.V.K.K.Prasad, "Embedded Real-Time Systems: Concepts, Design & Programming", Dream Tech Press, 2005.
7. Sriram V Iyer, Pankaj Gupta, "Embedded Real Time Systems Programming", Tata Mc Graw Hill, 2004.

EC8791 -EMBEDDED AND REAL TIME SYSTEMS

1. Wayne Wolf, “Computers as Components – Principles of Embedded Computing System Design”, Third Edition “Morgan Kaufmann Publisher (An imprint from Elsevier), 2012.

UNIT I

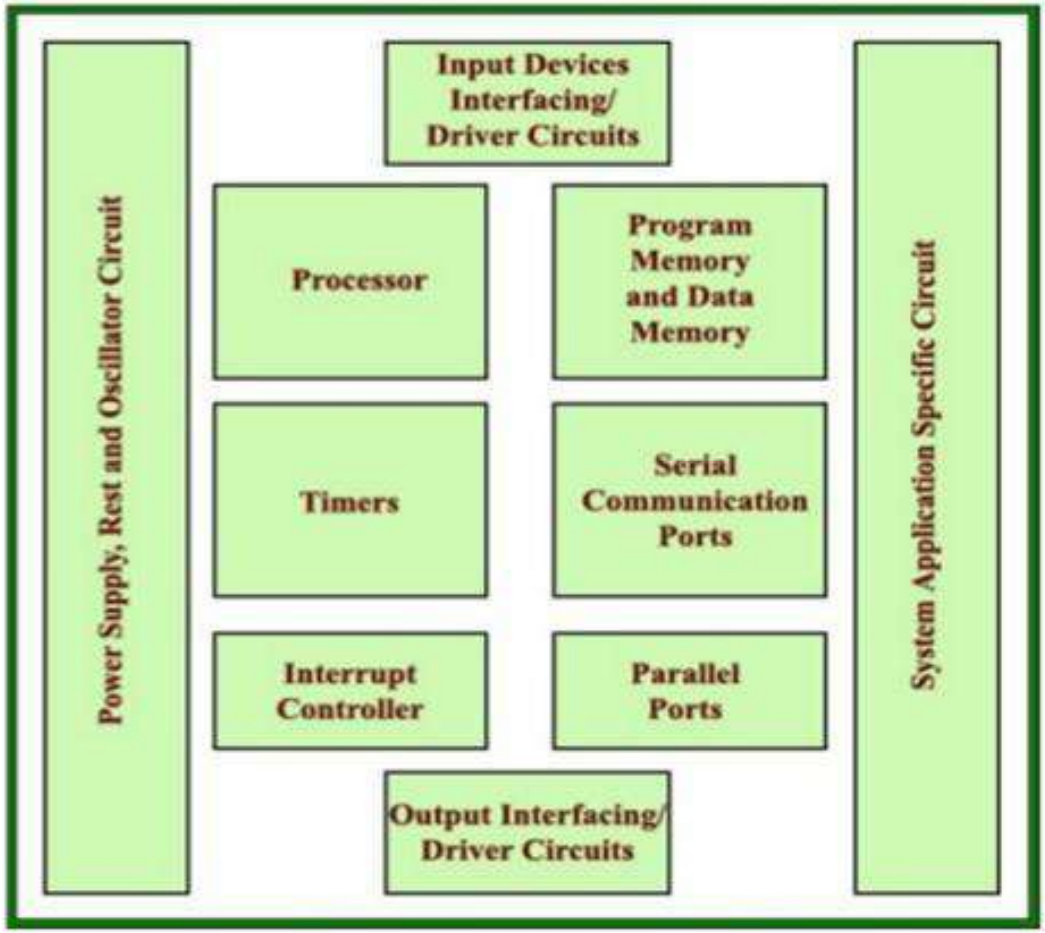
INTRODUCTION TO EMBEDDEDSYSTEM DESIGN

Complex systems and microprocessors– Embedded system design process –Design example: Model train controller- Design methodologies- Design flows - Requirement Analysis – Specifications-System analysis and architecture design – Quality Assurance techniques - Designing with computing platforms – consumer electronics architecture – platform-level performance analysis.

Introduction-Embedded Systems

- An Embedded system is an electronic system that has a software and is embedded in computer hardware.
- It is a system which has collection of components used to execute a task according to a program or commands given to it.
- Examples →Microwave ovens, Washing machine, Telephone answering machine system, Elevator controller system, Printers, Automobiles, Cameras, etc.

EMBEDDED SYSTEM HARDWARE



Components of Embedded system

- Microprocessor
- Memory Unit(RAM,ROM)
- Input unit(Keyboard,mouse,scanner)
- Output unit(pinters,video monitor)
- Networking unit(Ethernet card)
- I/O units(modem)

Real Time Operating System-RTOS

- Real-Time Operating System (**RTOS**) is an operating system (OS) intended to serve real-time applications that process data as it comes in, typically without buffer delays.
- It schedules their working and execution by following a plan to control the latencies and to meet the dead lines.
- Modeling and evaluation of a real-time scheduling system concern is on the analysis of the algorithm capability to meet a process deadline.
- A deadline is defined as the time required for a task to be processed.

Classification of Embedded system

1. Small scale Embedded system→(8/16bit microcontroller)
2. Medium Scale Embedded system→ →(16/32bit microcontroller, more tools like simulator, debugger)
3. Sophisticated Embedded system →(configurable processor and PAL)

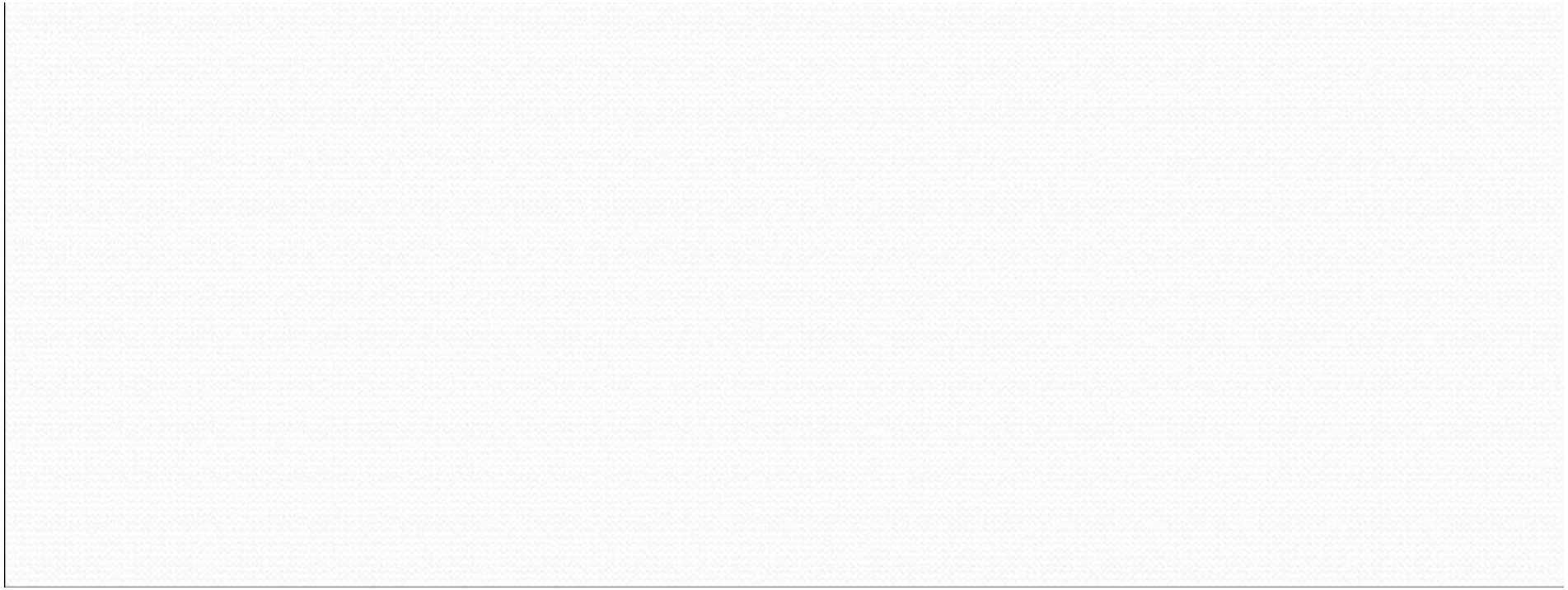
Embedded designer-skills

- Designer has a knowledge in the followings field,
- Microcontrollers, Data comm., motors, sensors, measurements ,C programming, RTOS programming.

1) COMPLEX SYSTEMS AND MICROPROCESSORS

Embedded(+)computer system

- Embedded system is a complex system
- It is any device that includes a programmable computer but is not itself intended to be a general-purpose computer.



History of Embedded computer system

- Computers have been embedded into applications since the earliest days of computing.
- In 1940s and 1950s→Whirlwind, designed a first computer to support real-time operation for controlling an aircraft simulator.
- In 1970s→ The first microprocessor(Intel 4004) was designed for an embedded application (Calculator), provided basic arithmetic functions.
- In 1972s→ The first handheld calculator (HP-35) was to perform transcendental functions , so it used several chips to implement the CPU, rather than a single-chip microprocessor.
- Designer faced critical problems to design a digital circuits to perform operations like trigonometric functions using calculator.
- But ,Automobile designers started making use of the microprocessor for to control the engine by determining when spark plugs fire, controlling the fuel/air mixture

Levels of Microprocessor

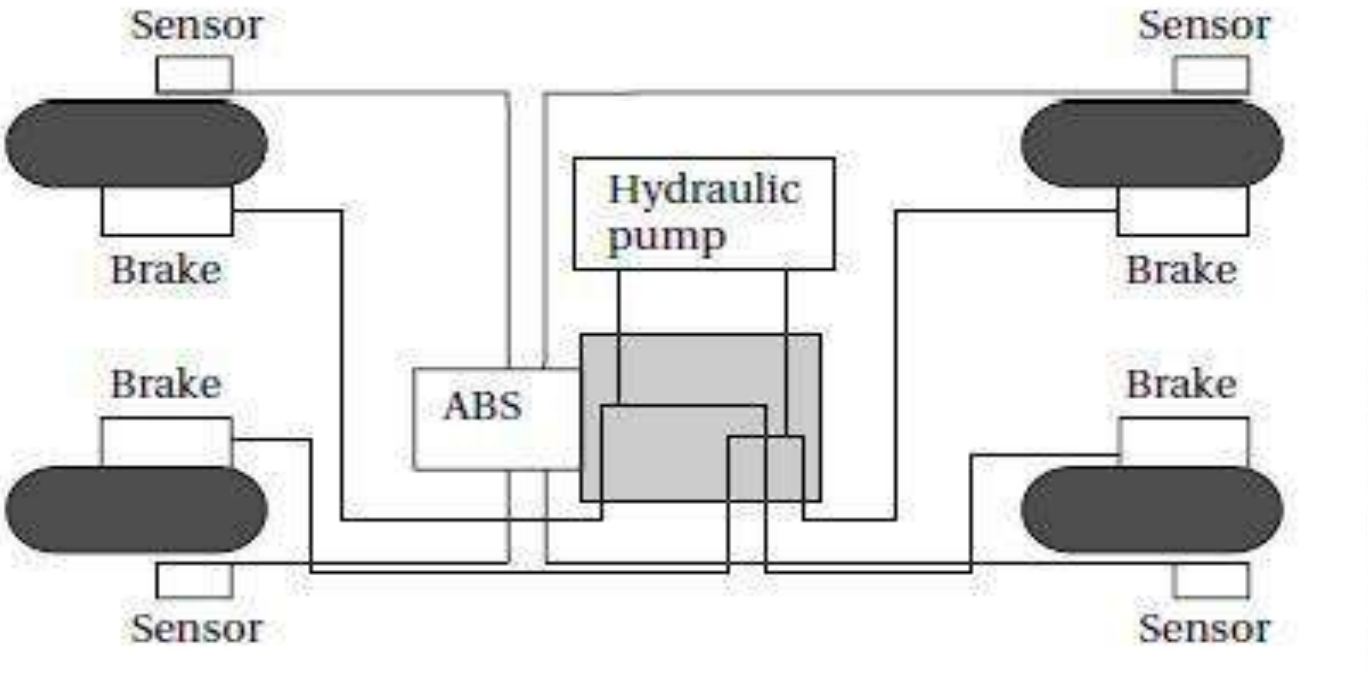
1. 8-bit ***microcontroller*** → for low-cost applications and includes on-board memory and I/O devices.
2. 16-bit microcontroller → used for more sophisticated applications that may require either longer word lengths or off-chip I/O and memory.
3. 32-bit ***RISC*** microprocessor →offers very high performance for computation-intensive applications.

Microprocessor Uses/Applications

- Microwave oven has at least one microprocessor to control oven operation
- Thermostat systems, which change the temperature level at various times during the day
- The modern camera is a prime example of the powerful features that can be added under microprocessor control.
- Digital television makes extensive use of embedded processors.

Embedded Computing Applications

- Ex→BMW 850i Brake and Stability Control System
- The BMW 850i was introduced with a sophisticated system for controlling the wheels of the car.
- Which uses An antilock brake system (ABS) and An automatic stability control (ASC +T) system.



1. An antilock brake system (ABS)

- Reduces skidding by pumping the brakes.
- It is used to temporarily release the brake on a wheel when it rotates too slowly—when a wheel stops turning, the car starts skidding and becomes hard to control.
- It sits between the hydraulic pump, which provides power to the brakes.
- It uses sensors on each wheel to measure the speed of the wheel.
- The wheel speeds are used by the ABS system to determine how to vary the hydraulic fluid pressure to prevent the wheels from skidding.

2. An automatic stability control (ASC +T) system

- It is used to control the engine power and the brake to improve the car's stability during maneuvers.
- It controls four different systems: throttle, ignition timing, differential brake, and (on automatic transmission cars) gear shifting.
- It can be turned off by the driver, which can be important when operating with tire snow chains.
- It has control unit has two microprocessors , one of which concentrates on logic-relevant components and the other on performance-specific components.
- The ABS and ASC+ T must clearly communicate because the ASC+ T interacts with the brake system.

Characteristics of Embedded Computing Applications

1. Complex algorithms-The microprocessor that controls an automobile engine must perform complicated filtering functions to optimize the performance of the car while minimizing pollution and fuel utilization.
2. User interface-The moving maps in Global Positioning System (GPS) navigation are good examples of user interfaces.
3. Real time-Embedded computing systems have to perform in real time—if the data is not ready by a certain deadline,the system breaks. In some cases, failure to meet a deadline or missing a deadline does not create safety problems but does create unhappy customers
4. Multirate-Multimedia applications are examples of ***multirate*** behavior. The audio and video portions of a multimedia stream run at very different rates, but they must remain closely synchronized. Failure to meet a deadline on either the audio or video portions spoils the perception of the entire presentation.

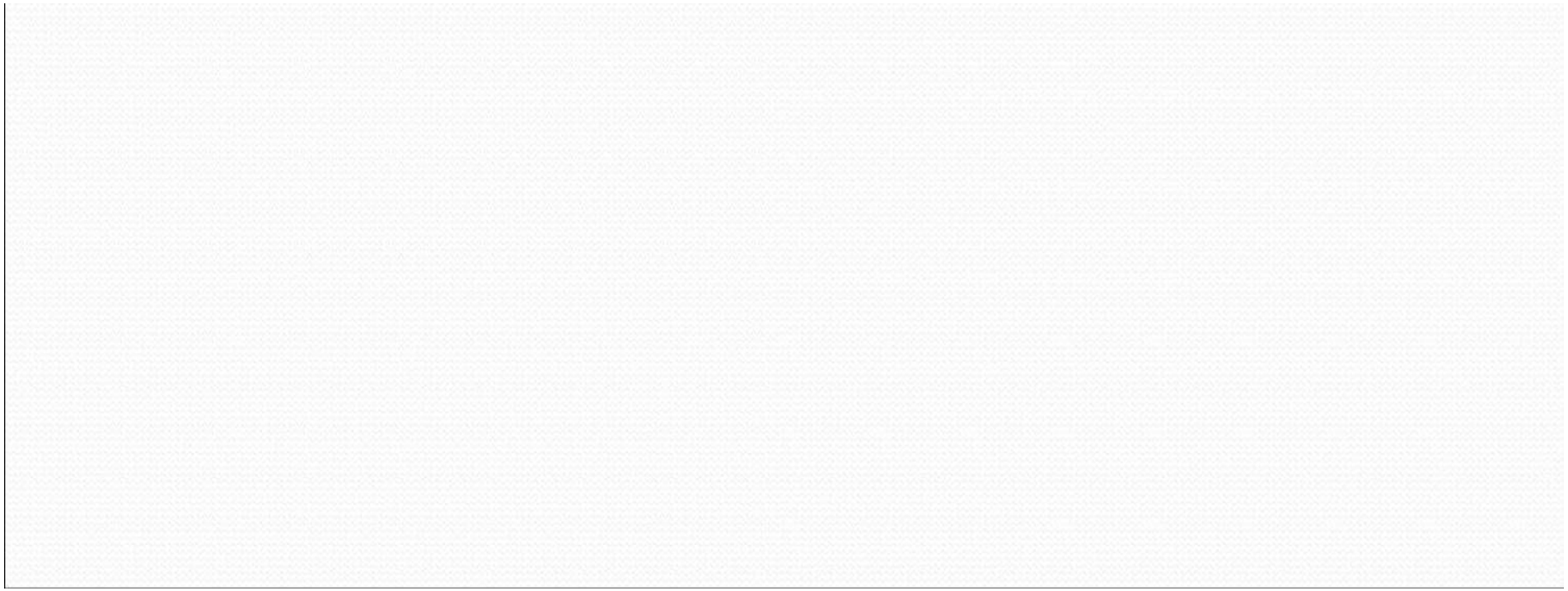
5. Manufacturing cost- It is depends on the type of microprocessor used, the amount of memory required, and the types of I/O devices.
6. Power and energy-Power consumption directly affects the cost of the hardware, since a larger power supply may be necessary.
7. Energy consumption →affects battery life, which is important in many applications, as well as heat consumption, which can be important even in desktop applications.

Why Use Microprocessors?

- Microprocessors are a very efficient way to implement digital systems.
- It make it easier to design families of products with various feature at different price points
- It can be extended to provide new features to keep up with rapidly changing markets.
- It executes program very efficiently
- It make their CPU run very fast
- Implementing several function on a single processor

Why not use PCs for all embedded computing?

- Real time performance is very less in PC because of different architecture.
- It increases the complexity and price of components due to broad mix of computing requirements.



Challenges in Embedded Computing System Design

1. How much hardware do we need?

- To meet performance deadlines and manufacturing cost constraints, the choice of Hardware is important.
- Too much hardware and it becomes too expensive.

2. How do we meet deadlines?

- To speed up the hardware so that the program runs faster. But the system more expensive.
- It is also entirely possible that increasing the CPU clock rate may not make enough difference to execution time, since the program's speed may be limited by the memory system.

3. How do we minimize power consumption?

- In battery-powered applications, power consumption is extremely important.
- In non-battery applications, excessive power consumption can increase heat dissipation.
- Careful design is required to slow down the noncritical parts of the machine for power consumption while still meeting necessary performance goals.

4) How do we design for upgradability?

- The hardware platform may be used over several product generations,or for several different versions ,able to add features by changing software.

  Complex testing: Run a real machine in order to generate the proper data.

- Testing of an embedded computer from the machine in which it is embedded.

  Limited observability and controllability→*No keyboard and screens,* in real-time applications we may not be able to easily stop the system to see what is going on inside and to affect the system's operation.

4.3) Restricted development environments:

- We generally compile code on one type of machine, such as a PC, and download it onto the embedded system.
- To debug the code, we must usually rely on programs that run on the PC or workstation and then look inside the embedded system.

Performance in Embedded Computing

- Embedded system designers have to set their goal —their program must meet its ***deadline.***

Performance Analysis

1. CPU: The CPU clearly influences the behavior of the program, particularly when the CPU is a pipelined processor with a cache.
2. Platform: The platform includes the bus and I/O devices. The platform components that surround the CPU are responsible for feeding the CPU and can dramatically affect its performance.
3. Program: Programs are very large and the CPU sees only a small window of the program at a time. We must consider the structure of the entire program to determine its overall behavior.
4. Task: We generally run several programs simultaneously on a CPU, creating a multitasking system. The tasks interact with each other in ways that have profound implications for performance.
5. Multiprocessor: Many embedded systems have more than one processor—they may include multiple programmable CPUs as well as accelerators. Once again, the interaction between these processors adds yet more complexity to the analysis of overall system performance.

2)EMBEDDED SYSTEM DESIGN PROCESS

Design process has two objectives as follows.

1. It will give us an introduction to the various steps in embedded system design.
2. Design methodology

I. Design to ensure that we have done everything we need to do, such as optimizing performance or performing functional tests.

II. It allows us to develop computer-aided design tools.

III. A design methodology makes it much easier for members of a design team to communicate.

Levels of abstraction in the design process.

1)Requirements

•It can be classified in to functional or nonfunctional

1.1)Functional Requirements

•Gather an informal description from the customers.

•Refine the requirements into a specification that contains enough information to design the system architecture.

•Ex:Sample Requirements form

•Name→Giving a name to the project

Purpose→Brief one- or two-line description of what the system is supposed to do.

•Inputs& Outputs →Analog electronic signals? Digital data? Mechanical inputs?

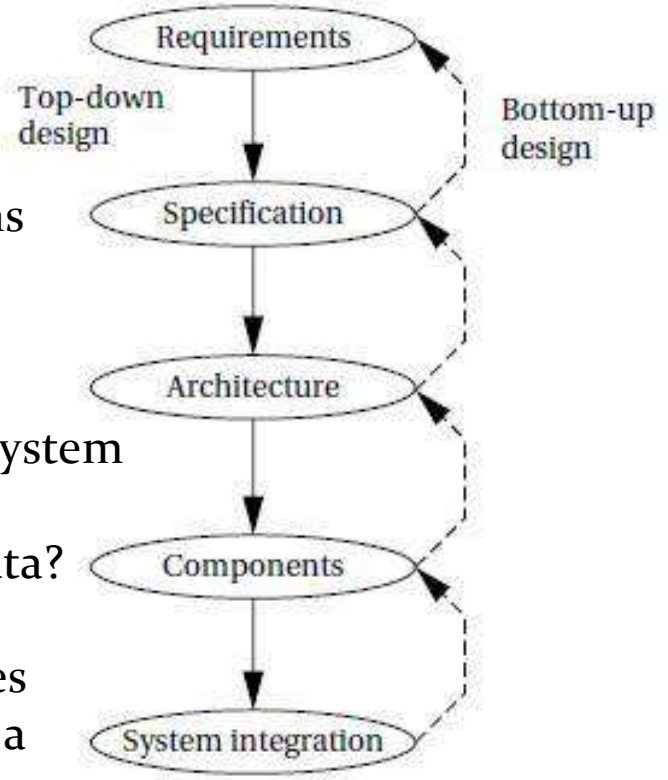
•Functions→ detailed description of what the system does

Performance→ computations must be performed within a certain time frame

•Manufacturing cost→ cost of the hardware components.

Power→ how much power the system can consume

Physical size and weight→ indication of the physical size of the system



- 1.2) Non-Functional Requirements
- Performance→ depends upon approximate time to perform a user-level function and also operation must be completed within deadline.
- Cost→Manufacturing cost includes the cost of components and assembly.
- Nonrecurring engineering (NRE) costs include the personnel and other costs of designing the
- system
- Physical Size and Weight→The final system can vary depending upon the application.
- Power Consumption→Power can be specified in the requirements stage in terms of battery life.

2)SPECIFICATION

- The specification must be carefully written so that it accurately reflects the customer's requirements.
- It can be clearly followed during design.

3) Architecture Design

- The architecture is a plan for the overall structure of the system.
- It is in the form block diagram that shows a major operation and data flow.

4) Designing Hardware and Software Components

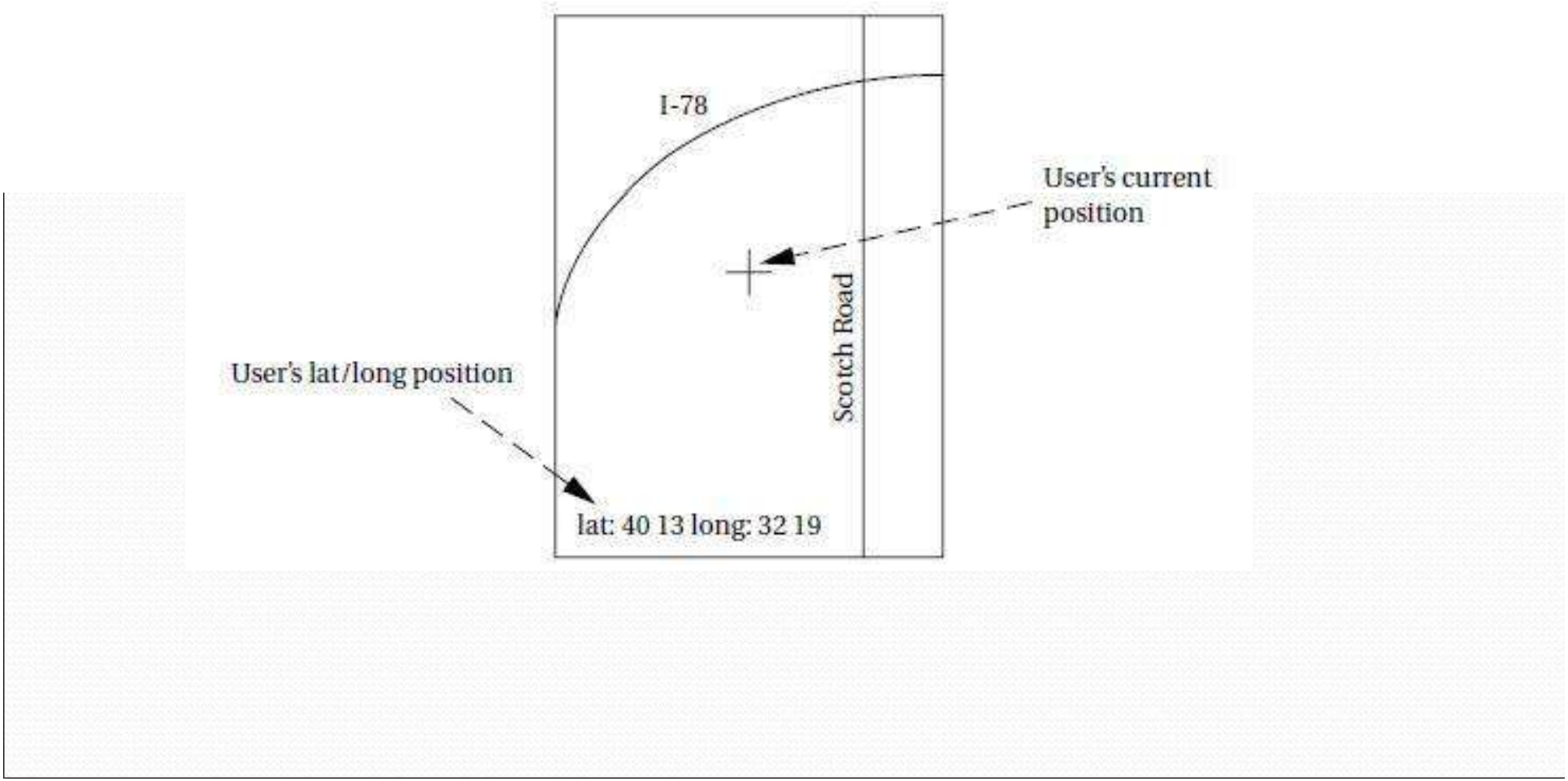
- The architectural description tells us what components we need include both hardware—FPGAs, boards & software modules

5)System Integration

- Only after the components are built, putting them together and seeing a working system.
- Bugs are found during system integration, and good planning can help us find the bugs quickly.

Embedded system Design Example

- GPS moving map



Design Process Steps

1. Requirements analysis of a GPS moving map

- The moving map is a handheld device that displays for the user a map of the terrain around the user's current position.
- The map display changes as the user and the map device change position.
- The moving map obtains its position from the GPS, a satellite-based navigation system.

Name	GPS moving map
Purpose	Consumer-grade moving map for driving use
Inputs	Power button, two control buttons
Outputs	Back-lit LCD display 400 600
Functions	Uses 5-receiver GPS system; three user-selectable resolutions;always displays current latitude and longitude
Performance	Updates screen within 0.25 seconds upon movement
Manufacturing cost	$30
Power	100mW
Physical size and weight	No more than 2"X 6, " *12 ounces*

Design Process Steps

2) Functionality→This system is designed for highway driving and similar uses. The system should show major roads and other landmarks available in standard topographic databases.

3) User interface→The screen should have at least 400X600 pixel resolution. The device should be controlled by no more than 3 buttons.
→A menu system should pop up on the screen when buttons are pressed to allow the user to make selections to control the system.

4) Performance→ The map should scroll smoothly.
→Upon power-up, a display should take no more than 1sec to appear.
→The system should be able to verify its position and display the current map within 15 s.

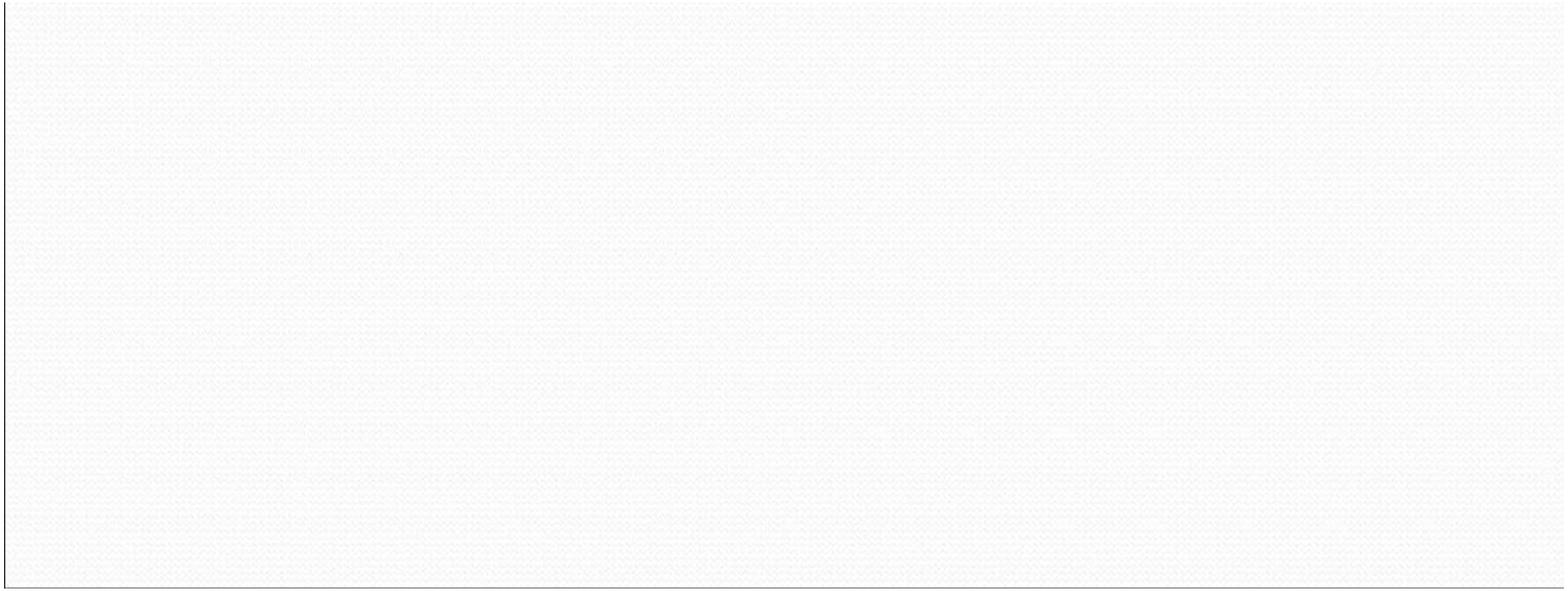
5) Cost→ The selling cost of the unit should be no more than $100.

6) Physical size and weight→The device should fit comfortably in the palm of the hand.

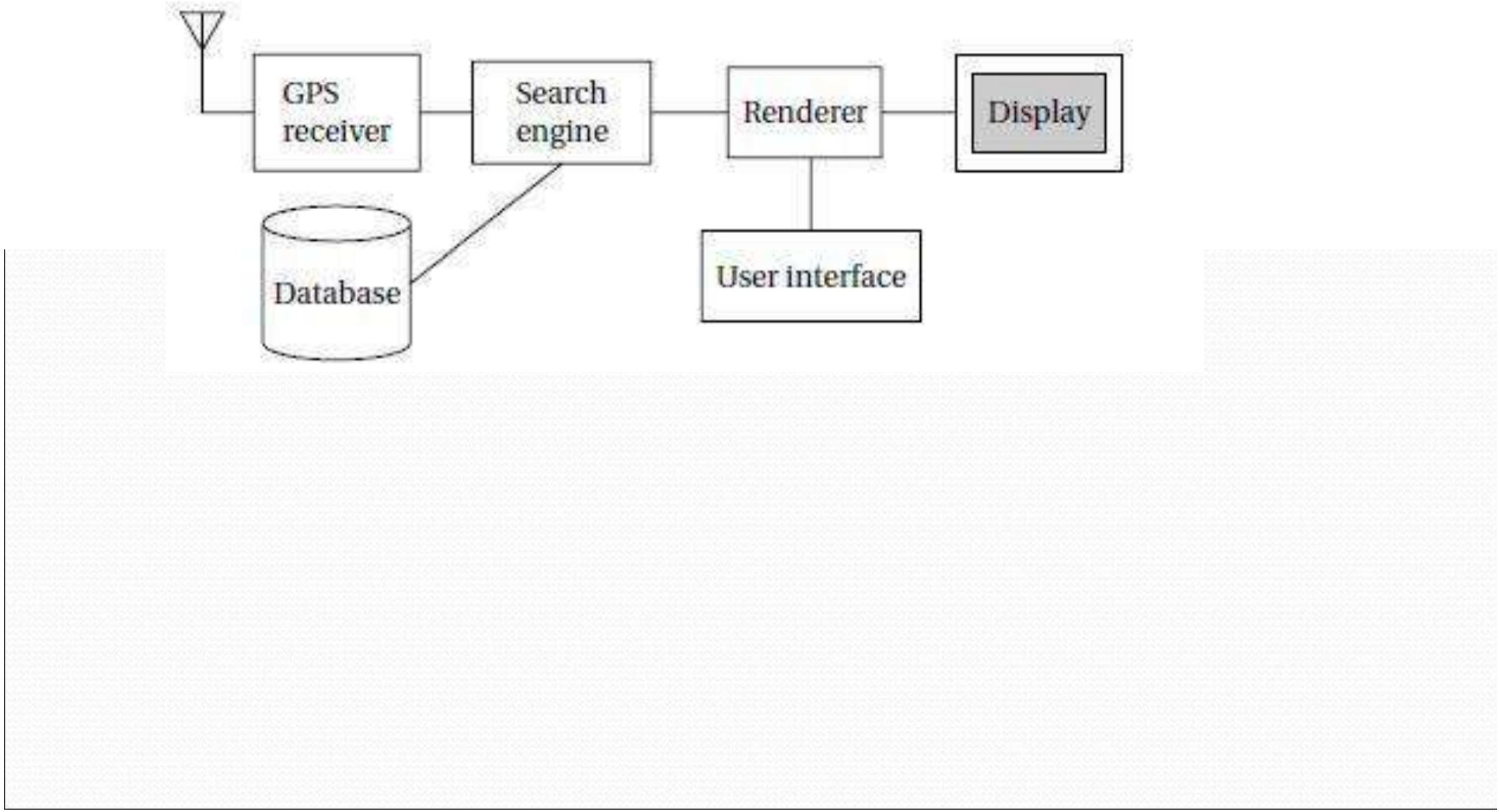
7) Power consumption→ The device run for at least 8 hrs on 4 AA batteries.

8) specification

1. Data received from the GPS satellite constellation.
2. Map data.
3. User interface.
4. Operations that must be performed to satisfy customer requests.
5. Background actions required to keep the system running, such as operating the GPS receiver.

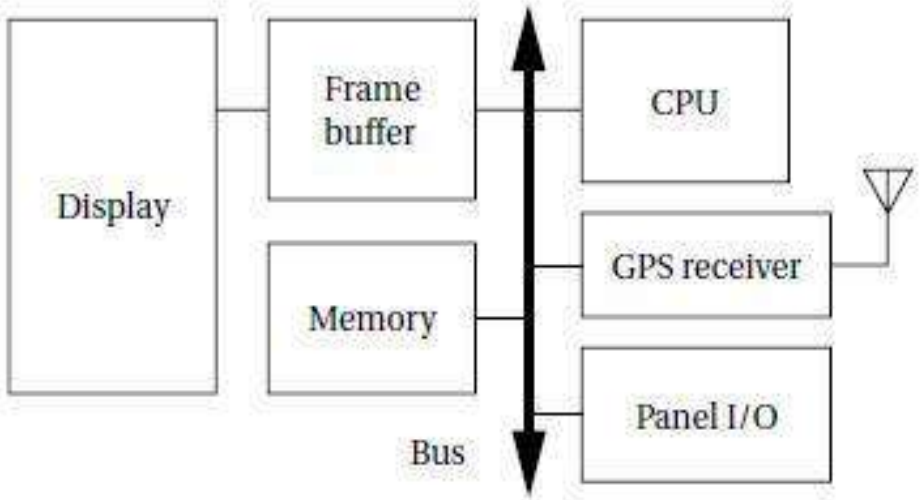


Block Diagram



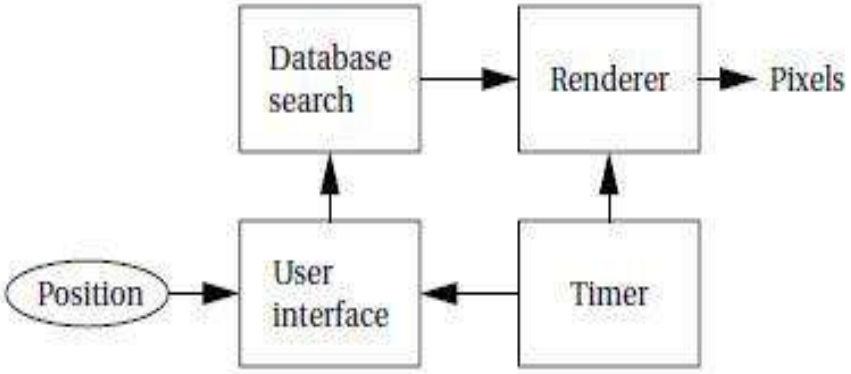
Hardware architecture

•one central CPU surrounded by memory and I/O devices.
• It used two memories: a frame buffer for the pixels to be displayed and a separate program/data memory for general use by the CPU.



Software architecture

•Timer to control when we read the buttons on the user interface and render data onto the screen.
•Units in the software block diagram will be executed in the hardware block diagram and when operations will be performed in time.



3)FORMALISM FOR SYSTEM DESIGN

- UML(Unified Modeling Language) is an object-oriented modeling language→ used to capture all these design tasks.
- It encourages the design to be described as a number of interacting objects, rather than blocks of code.
- objects will correspond to real pieces of software or hardware in the system.
- It allows a system to be described in a way that closely models real-world objects and their interactions.

Classification of descriptor

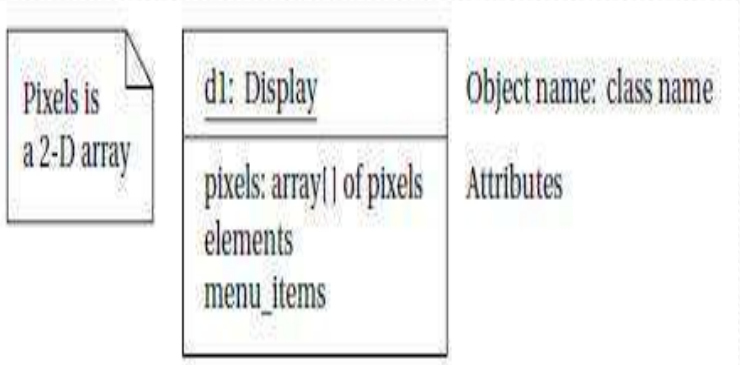
3.1)Structural Description

3.2)Behavioral Description

Structural Description

- It gives basic components of the system and designers can learn how to describe these components in terms of object.

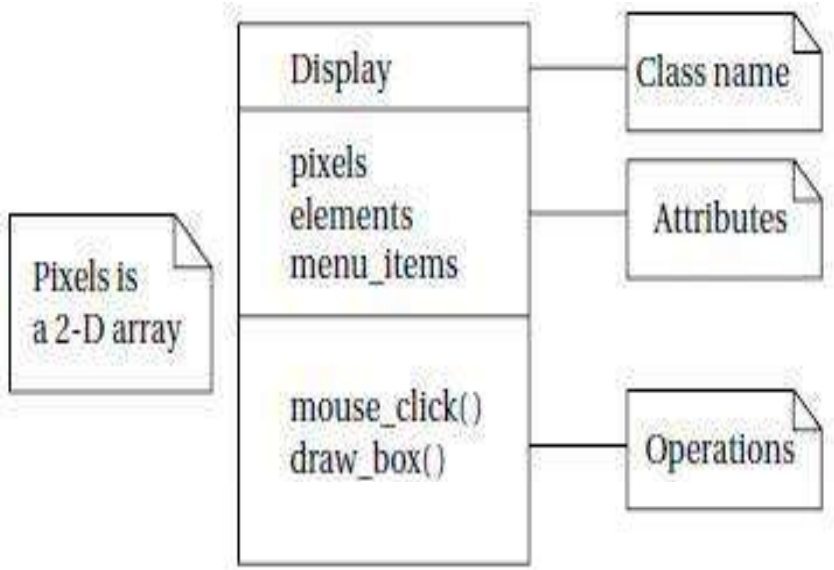
  OBJECT in UML NOTATION
- *An object* includes a set of attributes that define its internal state.
- An object describing a display (CRT screen) is shown in UML notation in Figure.
- The object has a **unique name**, and a member of a **class.**
- The name is underlined to show that this is a description of an object and not of a class.
- The text in the folded-corner page icon is a **note.**



An object in UML notation

CLASS IN UML NOTATION

- All objects derived from the same class have the same characteristics, but attributes may have different values.
- It also defines the **operations** that determine how the object interacts with the rest of the world.
- It defines both the interface for a particular type of object and that object's implementation.

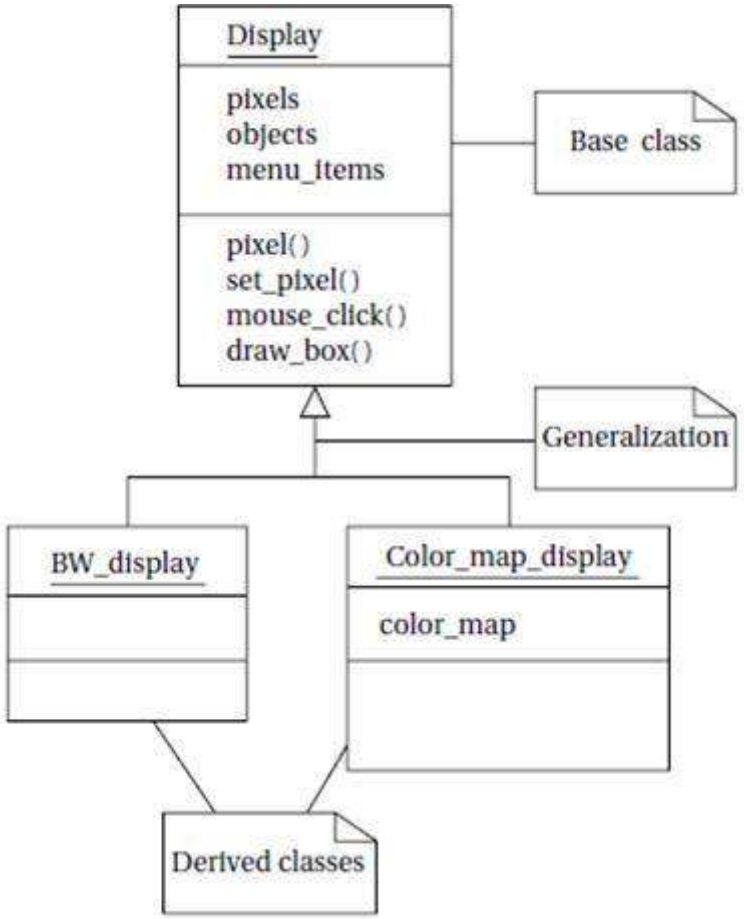


Relationships between objects and classes

1. Association →occurs between objects that communicate with each other but have no ownership relationship between them.
2. Aggregation→ describes a complex object made of smaller objects.
3. Composition →It is a type of aggregation in which the owner does not allow access to the component objects.
4. Generalization →allows us to define one class in terms of another.

Derived classes as a form of generalization in UML

•A derived class is defined to include all the attributes of its base class.
• Display is the base class and BW display and color map display are the two derived classes.
•BW display represents black and white display.



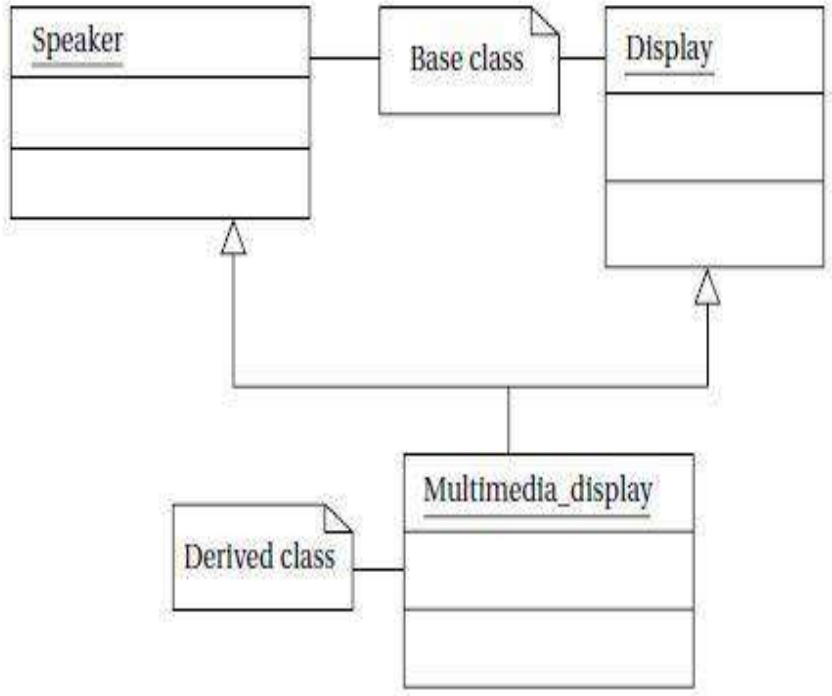
ultiple inheritance in UML

•UML allows to define multiple inheritance, in which a class is derived from more than one base class.

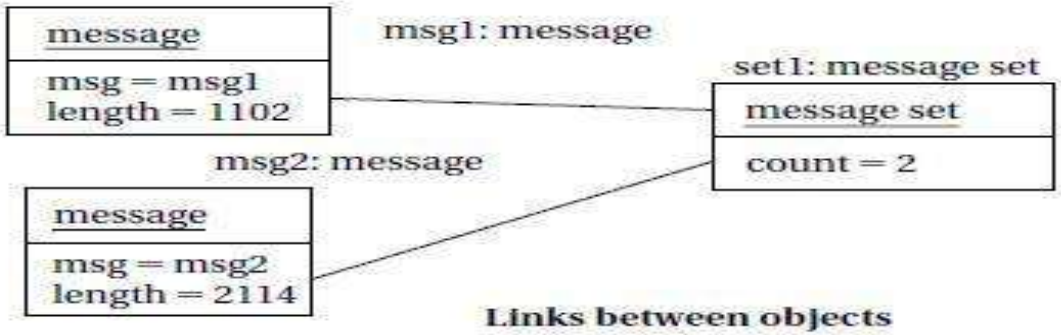
•Multimedia display class by combining the Display class with a Speaker class for sound.

•The derived class inherits all the attributes and operations of both its base classes, Display and Speaker.

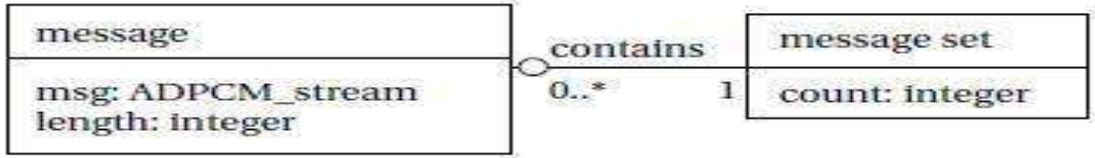
•



Links and Association



Links between objects

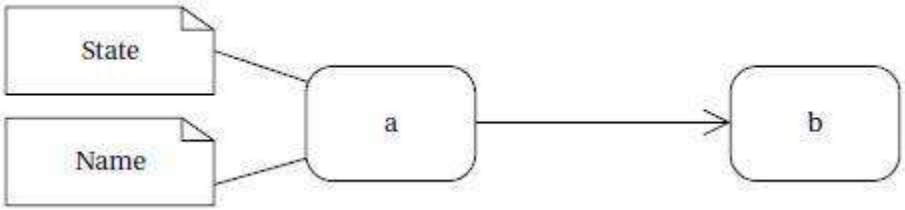


Association between classes

- A link describes a relationship between objects and association is to link as class is to object.
- Links used to make to stand associations capture type information about these links.
- The association is drawn as a line between the two labeled with the name of the association, namely, *contains*.

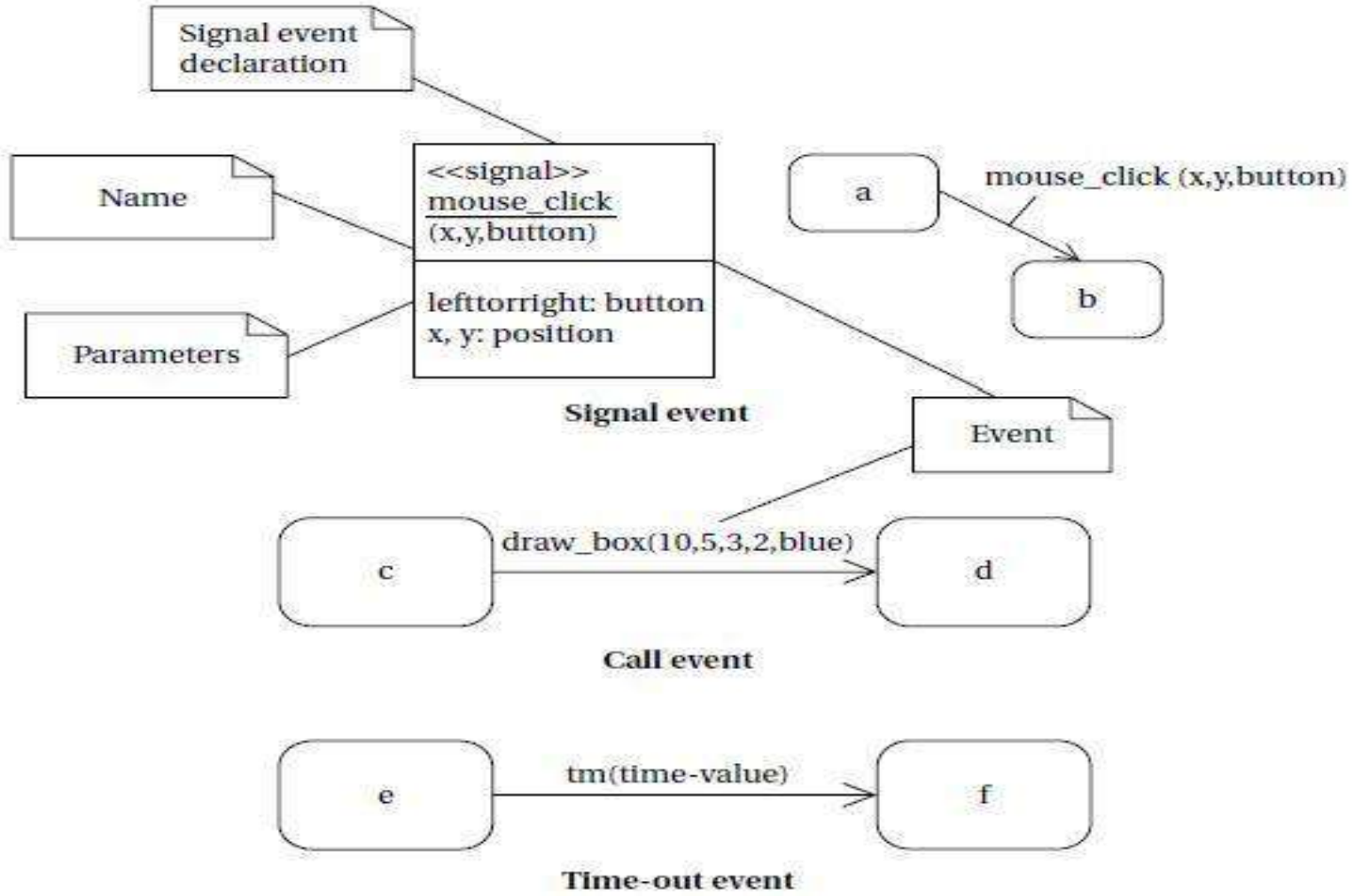
Behavioral Description

- Behavior of an operation is specified by a state machine.



- These state machines will not rely on the operation of a clock.
- Changes from one state to another are triggered by the occurrence of events.
- The event may generated from the outside or inside of the system.

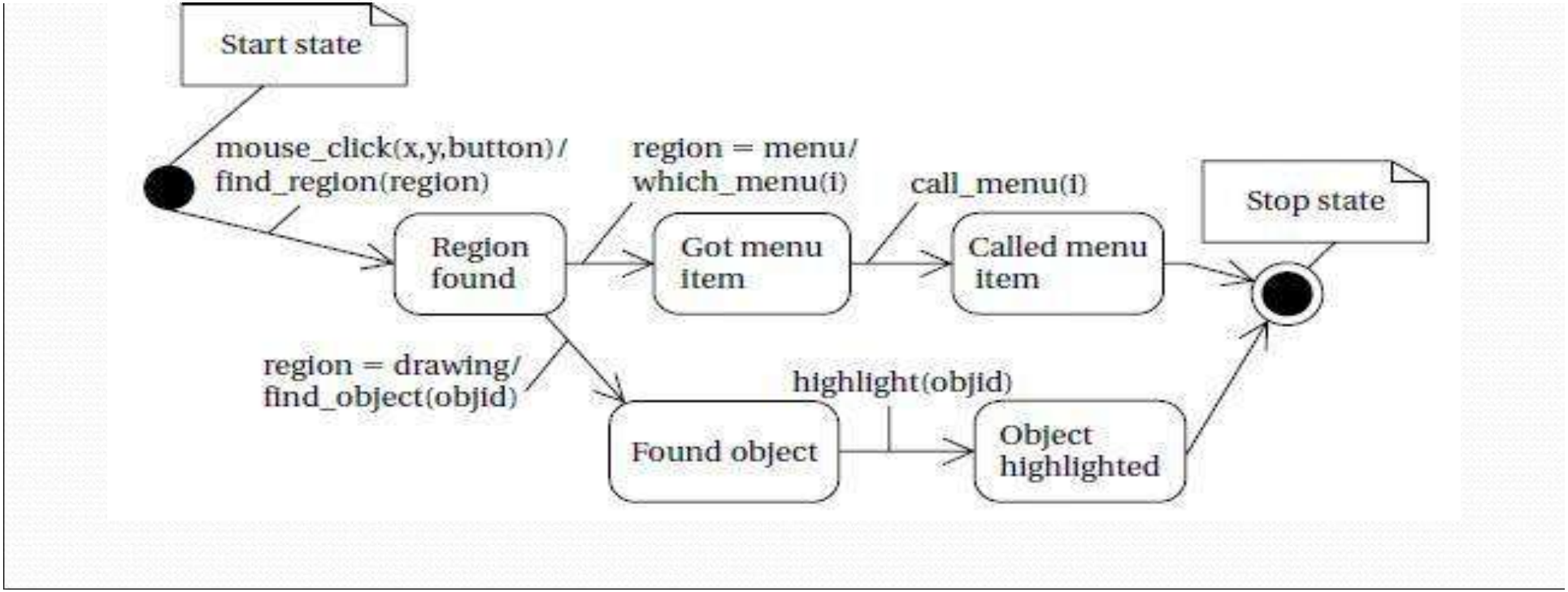
Signal, call, and time-out events in UML.



- Signal →is an asynchronous occurrence.
- It is defined in UML by an object that is labeled as a <<signal>>.
- Signal may have parameters that are passed to the signal's receiver.
- Call event→ follows the model of a procedure call in a programming language.
- Time-out event→ causes the machine to leave a state after a certain amount of time.
- The label tm(time-value) on the edge gives the amount of time after which the transition occurs.
- It is implemented with an external timer.

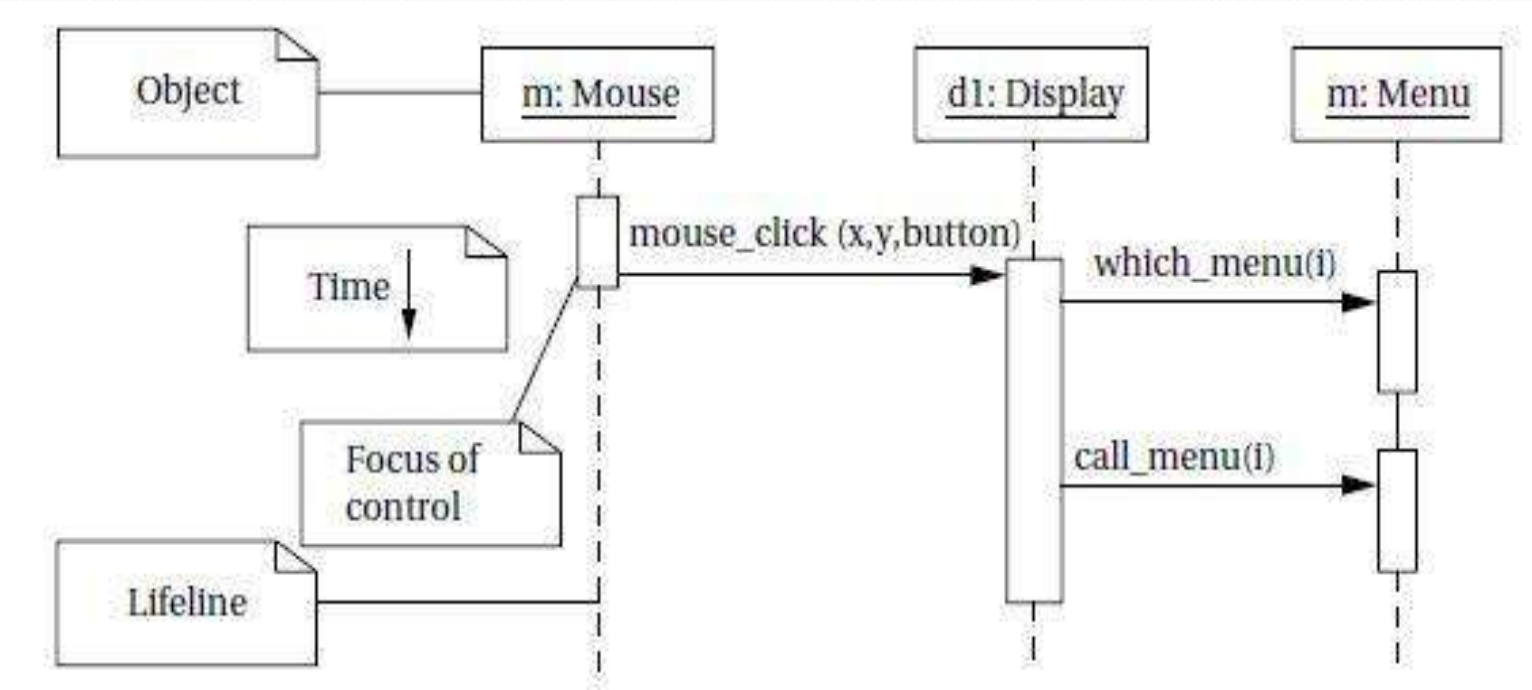
State Machine specification in UML

- The start and stop states are special states which organize the flow of the state machine.
- The states in the state machine represent different operations.
- Conditional transitions out of states based on inputs or results of some computation.
- An unconditional transition to the next state.



Sequence diagram in UML

- Sequence diagram is similar to a hardware timing diagram, although the time flows vertically in a sequence diagram, whereas time typically flows horizontally in a timing diagram.
- It is designed to show particular choice of events—it is not convenient for showing a number of mutually exclusive possibilities.



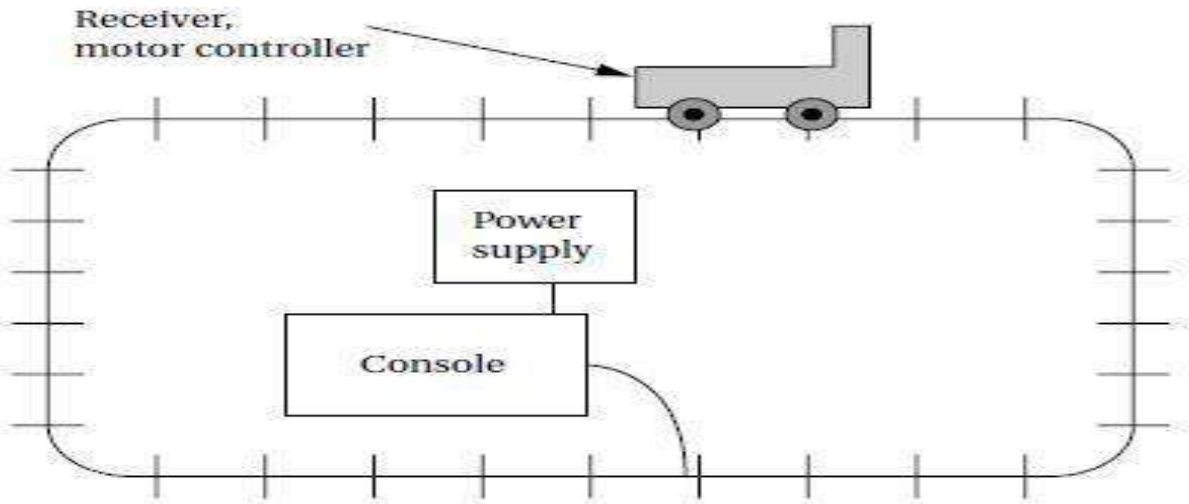
4) Design:Model Train Controller

- In order to learn how to use UML to model systems→ specify a simple system (Ex: model train controller)
- The user sends messages to the train with a control box attached to the tracks.
- The control box may have controls such as a throttle, emergency stop button, and so on.
- The train Rx its electrical power from the two rails of the track.

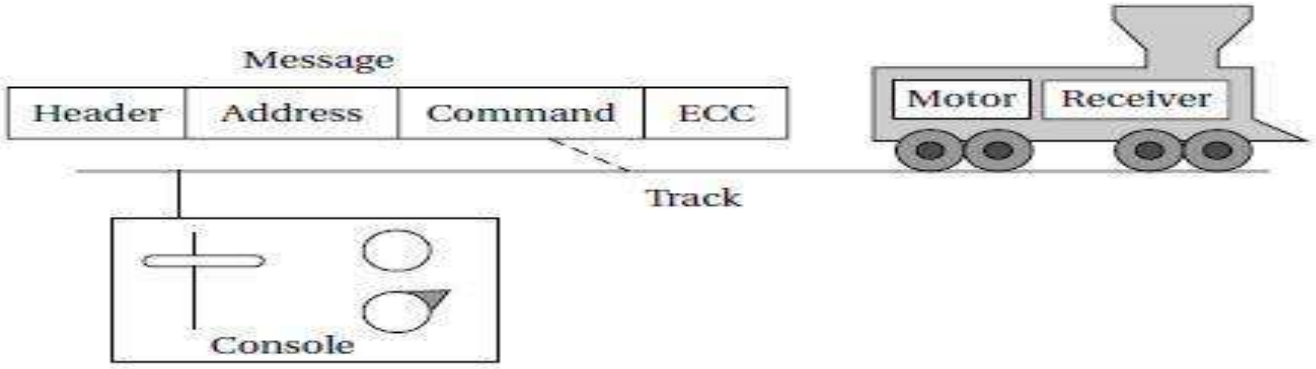
CONSOLE

- Each packet includes an address so that the console can control several trains on the same track.
- The packet also includes an error correction code (ECC) to guard against transmission errors.
- This is a one-way communication system—the model train cannot send commands back to the user.

Model Train Control system



System setup



Signaling the train

REQUIREMENTS

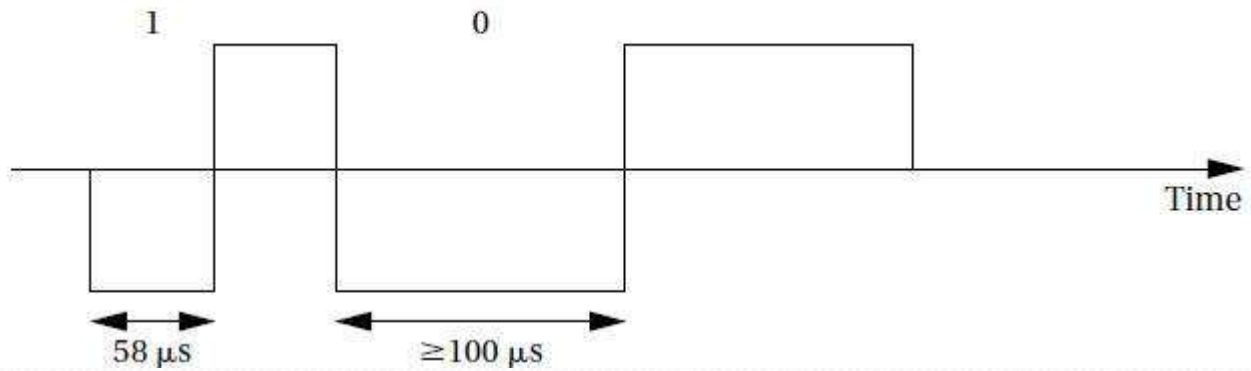
- The console shall be able to control up to eight trains on a single track.
- The speed of each train controllable by a throttle to at least 63 different levels in each direction (forward and reverse).
- There shall be an inertia control→ to adjust the speed of train.
- There shall be an emergency stop button.
- An error detection scheme will be used to transmit messages.

Requirements:Chart Format

Name	Model train controller
Purpose	Control speed of up to eight model trains
Inputs	Throttle, inertia setting, emergency stop, train number
Outputs	Train control signals
Functions	Set engine speed based upon inertia settings; respond
Performance	Can update train speed at least 10 times per second
Manufacturing cost	$50
Power	10W
Physical size and weight	Console should be comfortable for two hands,approximatesize of standard keyboard;
weight	2 pounds

Digital Command Control (DCC)

- Standard S-9.1→ how bits are encoded on the rails for transmission.
- Standard S-9.2→ defines the packets that carry information.
- The signal encoding system should not interfere with power transmission
- Data signal should not change the DC value of the rails.
- Bits are encoded in the time between transitions.
- Bit 0 is at least 100 s while bit 1 is nominally 58 s.



Packet Formation in DCC

- The basic packet format is given by $PSA(sD) + E$
- P → preamble, which is a sequence of at least 10 1 bits.
- S → packet start bit. It is a 0 bit.
- A →address is 8 bits long. The addresses 00000000, 1111110, and 1111111 are reserved.
- s→ data byte start bit, which, like the packet start bit, is a 0.
- D →data byte includes 8 bits. A data byte may contain an address, instruction, data, or error correction information.
- E →packet end bit, which is a 1 bit.

Baseline packet

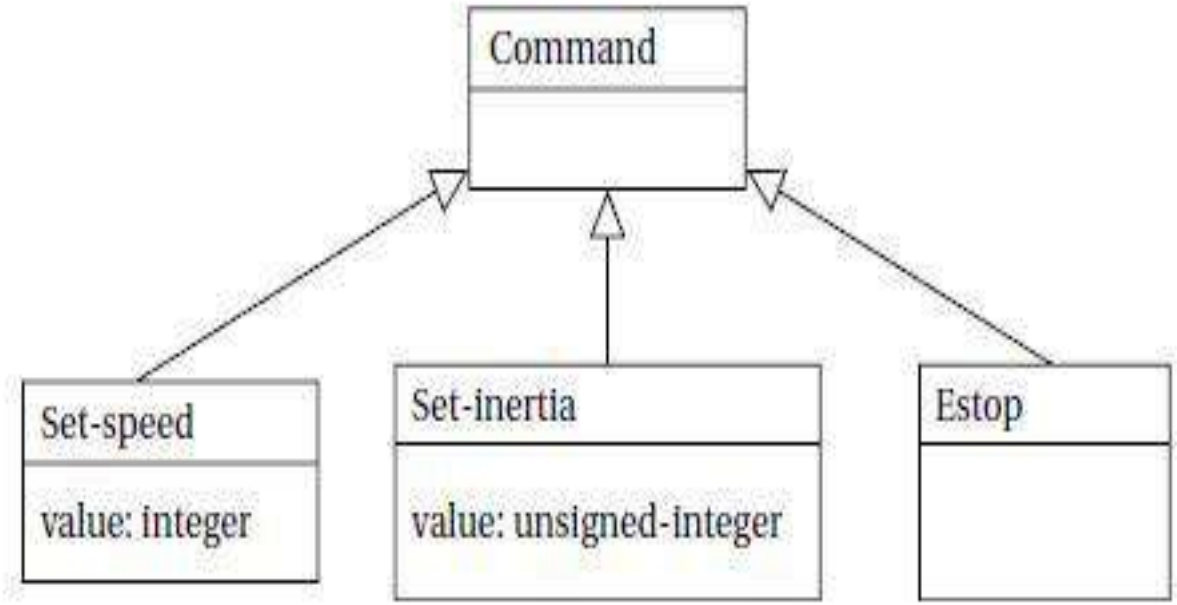
- The minimum packet that must be accepted by all DCC implementations.
- It has three data bytes.
- Address data byte→ gives the intended receiver of the packet
- Instruction data byte→ provides a basic instruction
- Error correction data byte→ is used to detect and correct transmission errors.

Date byte

- Bits 0–3→ provide a 4-bit speed value.
- Bit 4 →has an additional speed bit.
- Bit 5 →gives direction, with 1 for forward and 0 for reverse.
- Bits 6-7 are set at 01 → provides speed and direction.

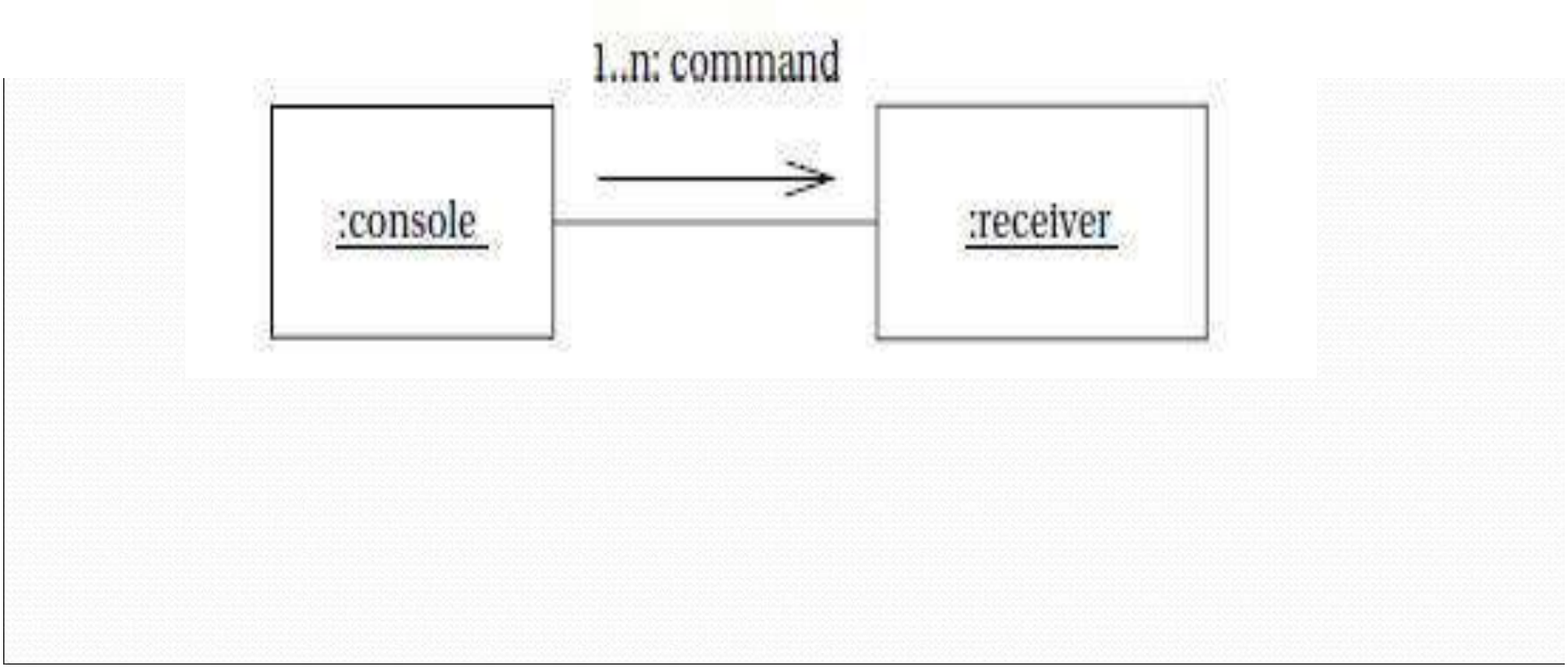
Conceptual Specification

- Conceptual specification allows us to understand the system a little better.
- A train control system turns commands into packets.
- A command comes from the command unit while a packet is transmitted over the rails.
- Commands and packets may not be generated in a 1-to-1 ratio

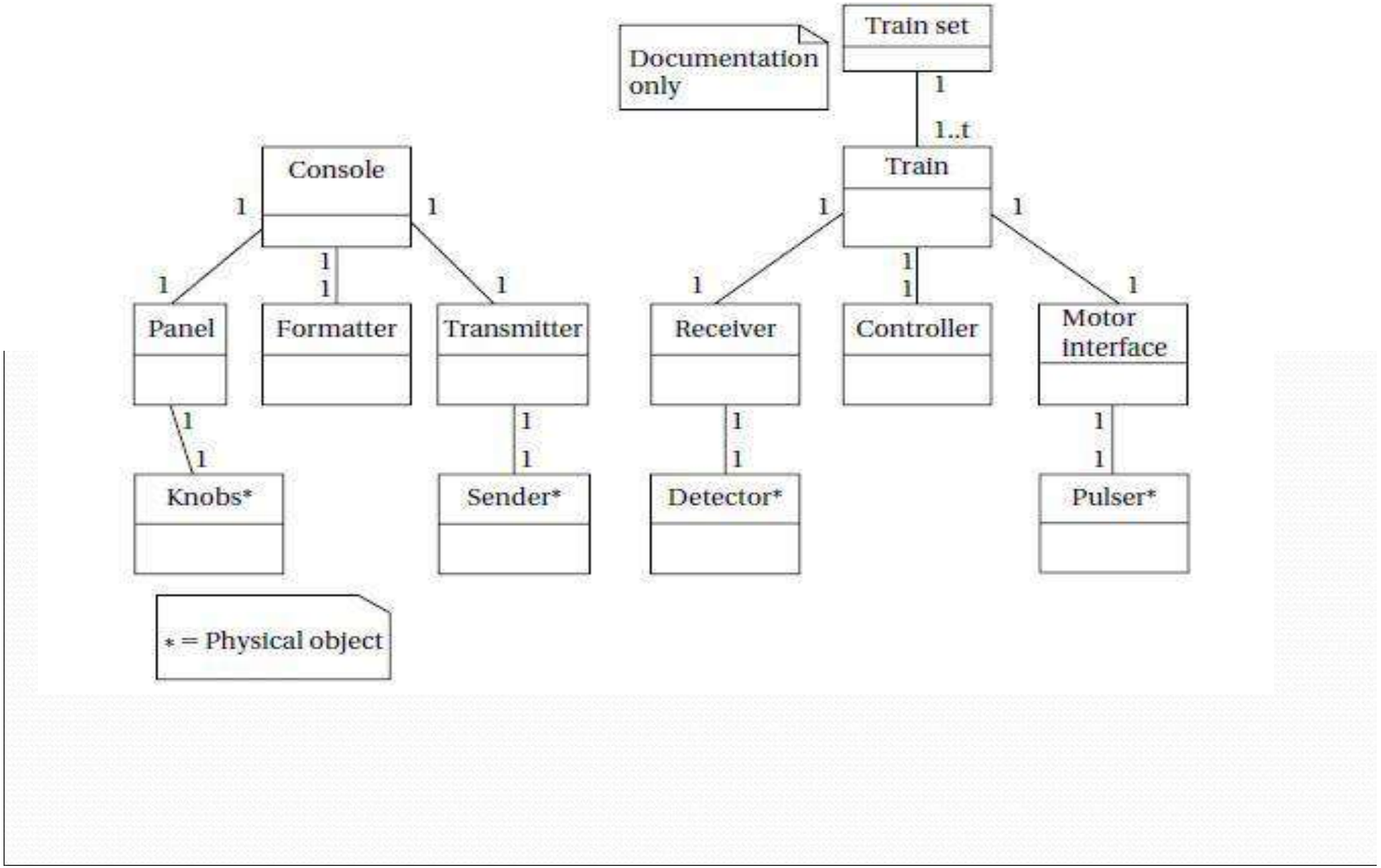


UML collaboration diagram for train controller system

- The command unit and receiver are each represented by objects.
- The command unit sends a sequence of packets to the train's receiver, as illustrated by the arrow messages as 1..n.
- Those messages are of course carried over the track.



UML class diagram for the train controller

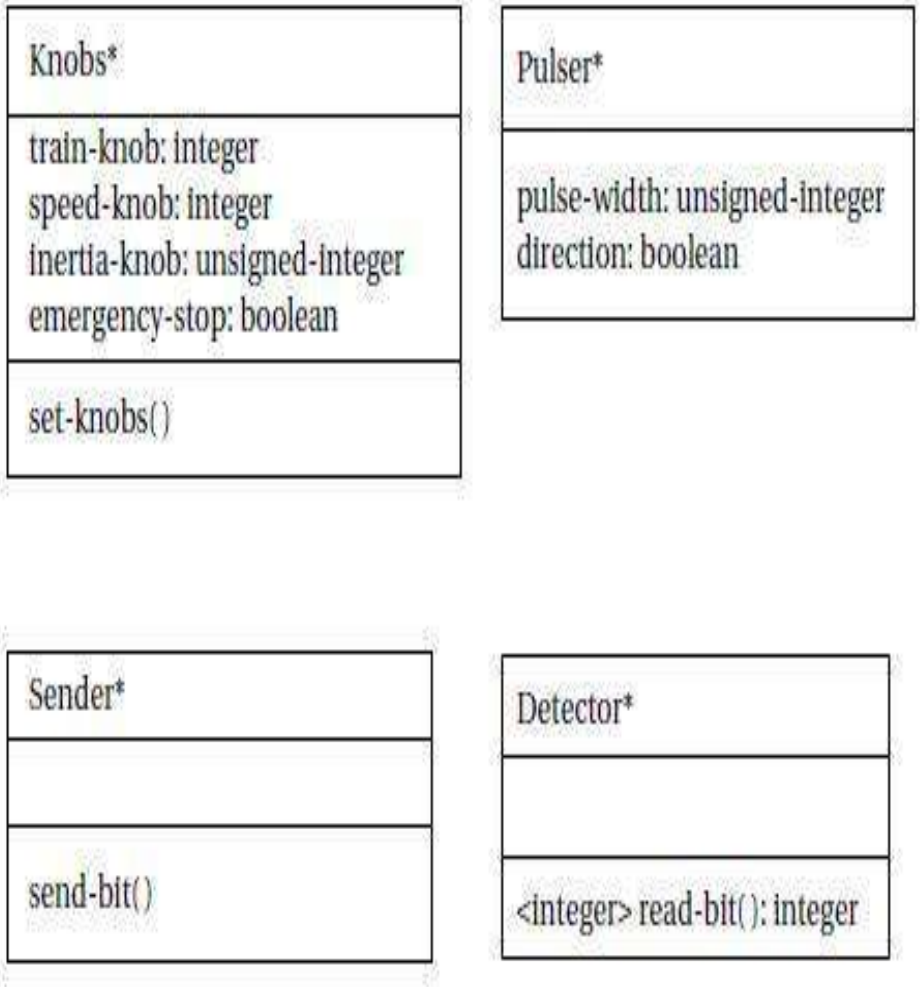


Basic characteristics of UML classes

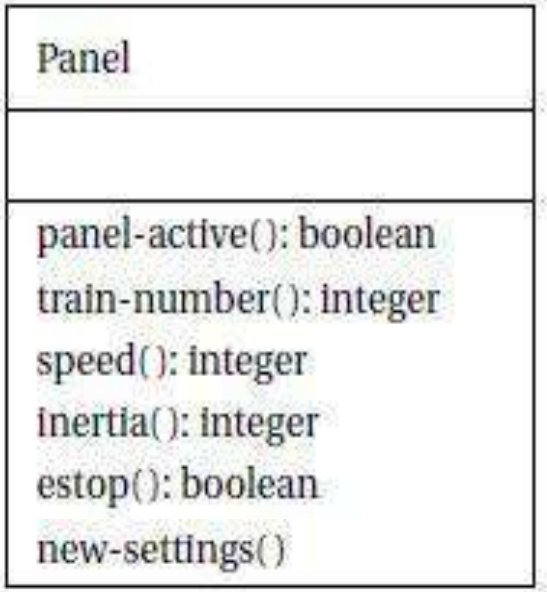
- *Console class→ describes the command unit's front panel, which contains* the analog knobs and hardware to interface to the digital parts of the system.
- *Formatter class→ includes behaviors that know how to read the panel* knobs and creates a bit stream for the required message.
- *Transmitter class →interfaces to analog electronics to send the message* along the track
- *Knobs*→ describes the actual analog knobs, buttons, and levers on the control* panel.
- *Sender*→ describes the analog electronics that send bits along the track.*
- *Receiver class→ knows how to turn the analog signals on the track into* digital form.
- *Controller class→ includes behaviors that interpret the commands and* figures out how to control the motor.
- *Motor interface class→ defines how to generate the analog signals required* to control the motor.
- *Detector* →detects analog signals on the track and converts them into digital* form.
- *Pulser* →turns digital commands into the analog signals required to control the* motor speed.

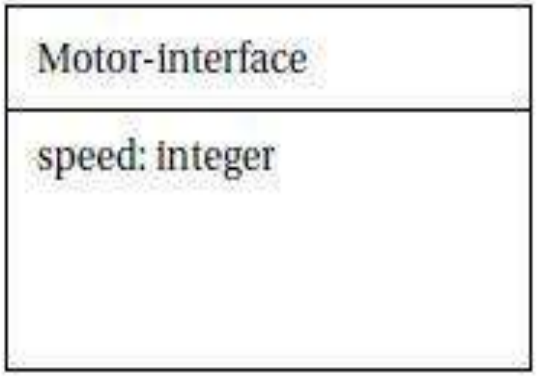
Detailed Specification

•The Panel has three knobs
• train number (which train is currently being controlled).
•speed (which can be positive or negative), and inertia.
•It also has one button for emergency-stop.
•When we change the train number setting, to reset the other controls to the proper values for that train.
• so that the previous train's control settings are not used to change the current train's settings.



Class diagram for panel

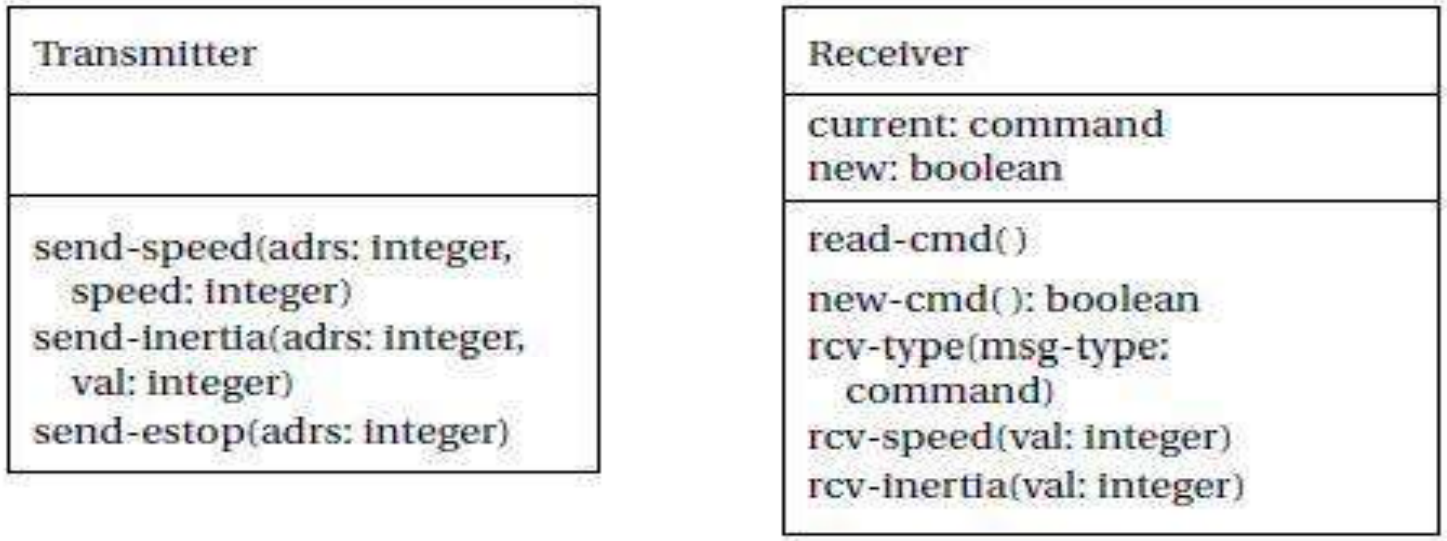




- The *Panel class* defines a behavior for each of the controls on the panel.
- The *new-settings behavior uses the set-knobs behavior of the Knobs**
- Change the knobs settings whenever the train number setting is changed.
- The *Motor-interface defines an attribute for speed that can be set by other classes.*

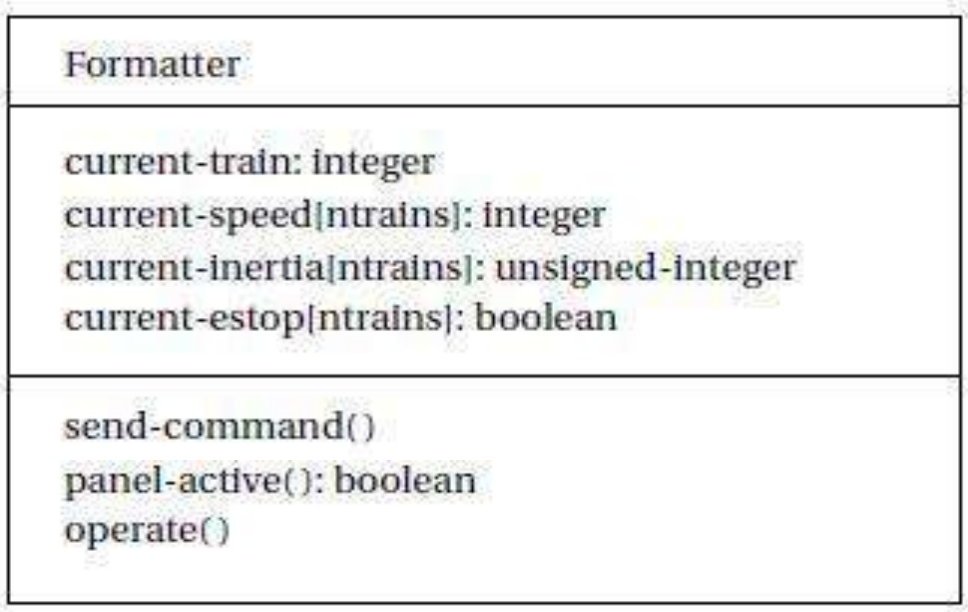
.

Class diagram for the Transmitter and Receiver



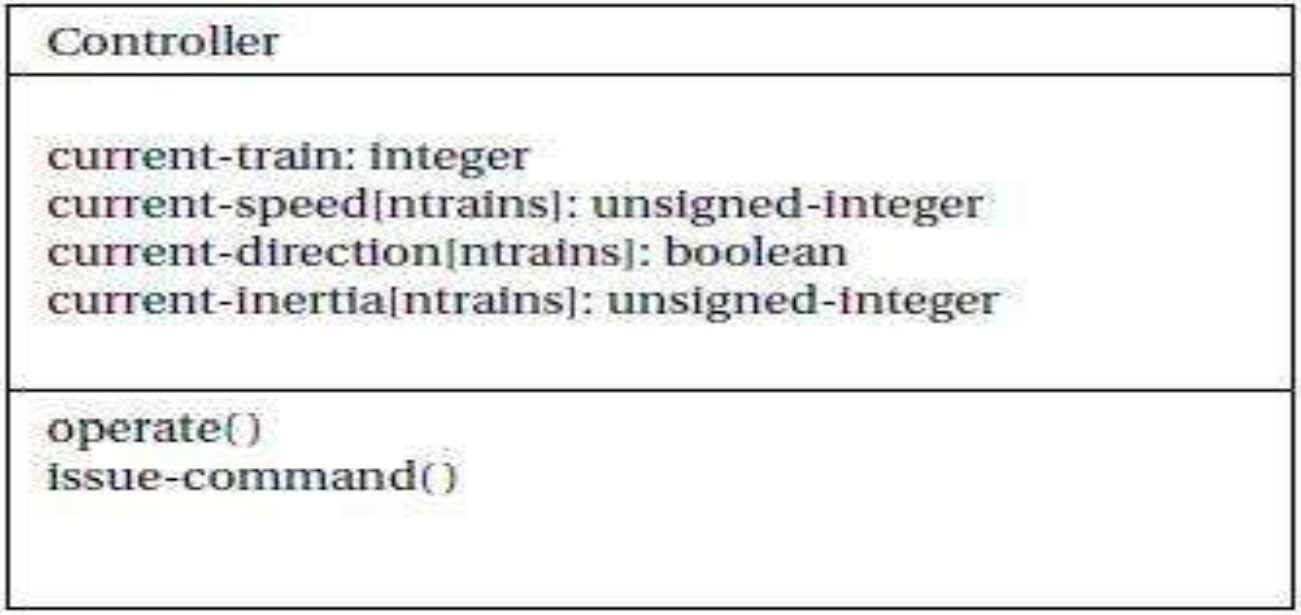
- They provide the software interface to the physical devices that send and receive bits along the track.
- The Transmitter provides a behavior message that can be sent
- The Receiver class provides a read-cmd behavior to read a message off the tracks.

Class diagram for Formatter



- The formatter holds the current control settings for all of the trains.
- The *send-command* serves as the interface to the transmitter.
- The *operate function performs the* basic actions for the object.
- The *panel-active behavior* returns true whenever the panel's values do not correspond to the current values

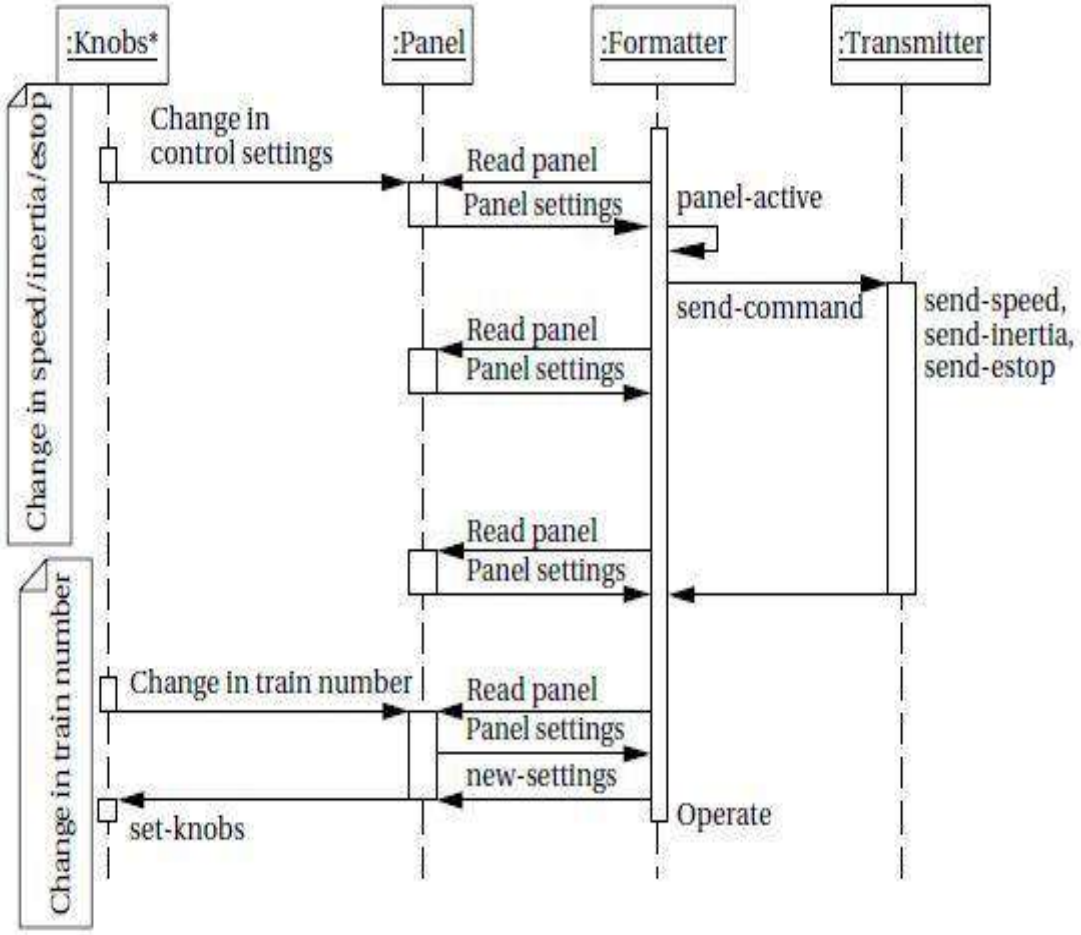
Class diagram for Controller



- The *Controller's operate behavior must execute* several behaviors to determine the nature of the message.
- Once the speed command has been parsed, it must send a sequence of commands to the motor to smoothly change the train's speed.

Sequences diagram for transmitting a control input

- Sequence diagram specify the interface between more than one classes.
- Its detailed operations and what ways its going to operate



DESIGN METHODOLOGIES

- Design of Embedded system is not an easy task.
- The main goal of a design process is to create a product that does something useful.
- Typical specifications for a product are functionality , manufacturing cost, performance and power consumption.

Design process has several important goals as follows

Time-to-market

- Customers always want new features.
- The product that comes out first can win the market, even setting customer preferences for future generations of the product.

Design cost

- Consumer products are very cost sensitive, and it is distinct from manufacturing cost.
- Design costs can dominate manufacturing costs.
- Design costs can also be important for high-volume consumer devices when time-to-market pressures cause teams to swell in size.

Quality

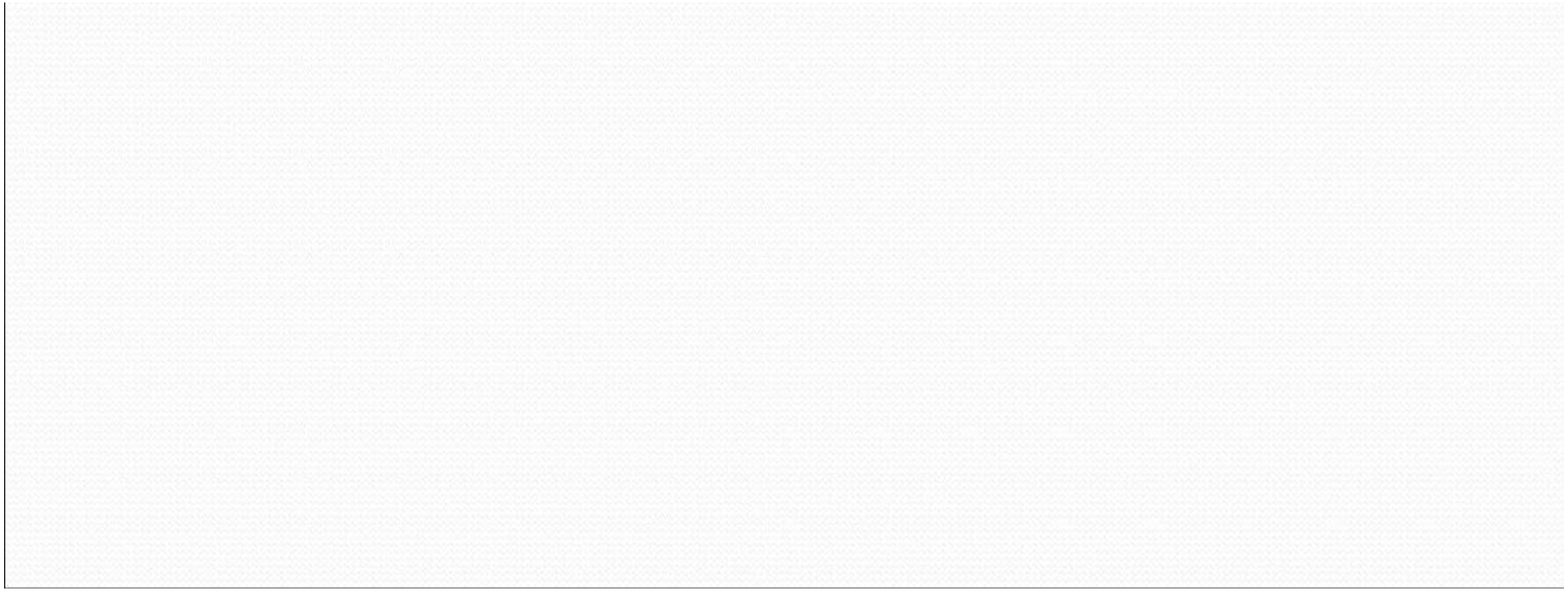
- Customers want their products fast and cheap.
- Correctness, reliability, and usability must be explicitly addressed from the beginning of the design job to obtain a high-quality product at the end

Design flows

- A design flow is a sequence of steps to be followed during a design.
- Some of the steps can be performed by tools and other steps can be performed by hand.

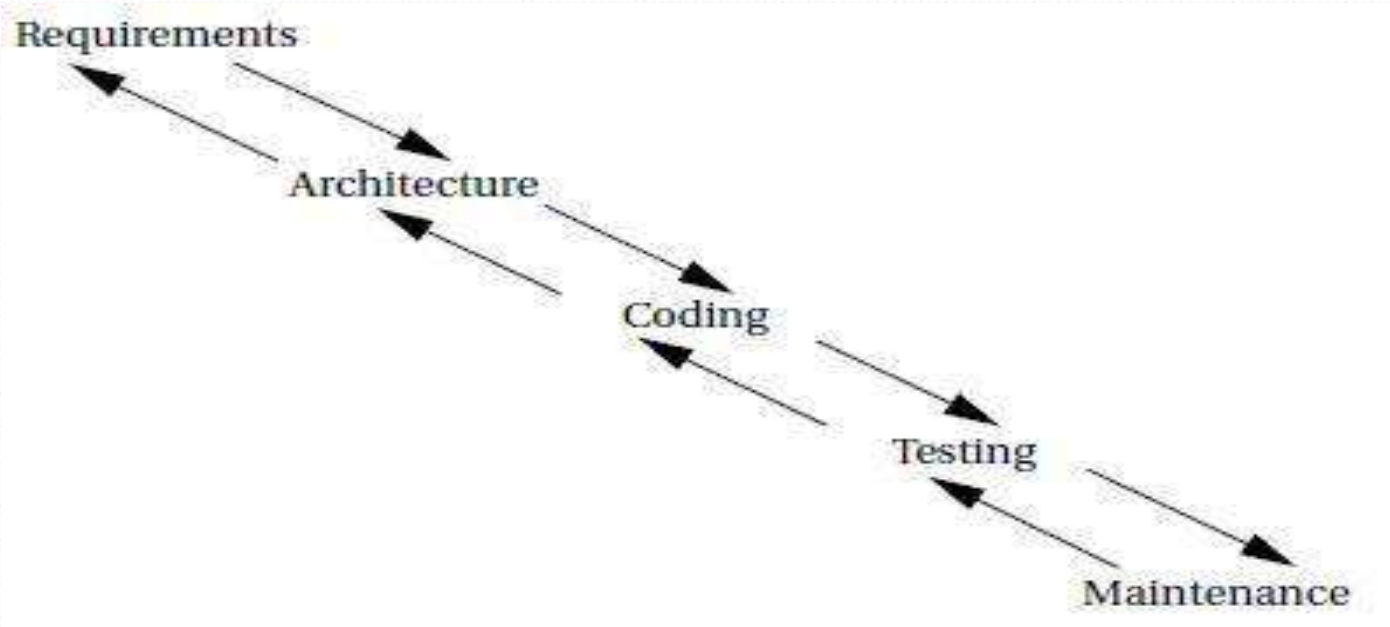
Types of Software development models

1. Waterfall model
2. Spiral model
3. Successive refinement development model
4. Hierarchical design model



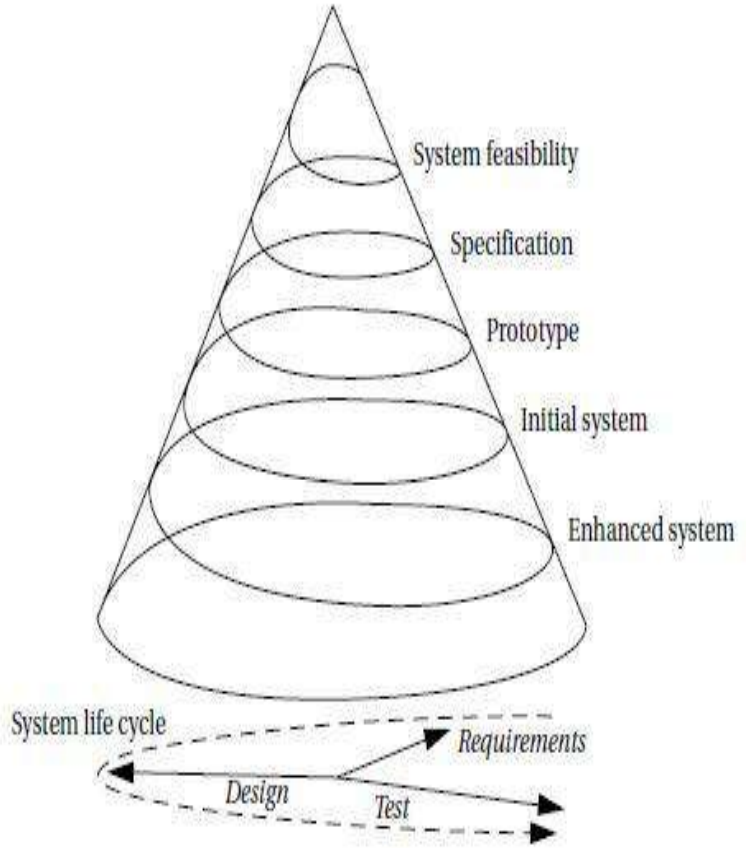
Waterfall model

- The waterfall development model consists of five major phases.
- Requirements analysis→ determines the basic characteristics of the system.
- Architecture design→It decomposes the functionality into major components
- Coding→It implements the pieces and integrates them.
- Testing→It detemines bugs.
- Maintenance→ It entails deployment in the field, bug fixes,and upgrades.
- The waterfall model makes work flow information from higher levels of abstraction to more detailed design steps.



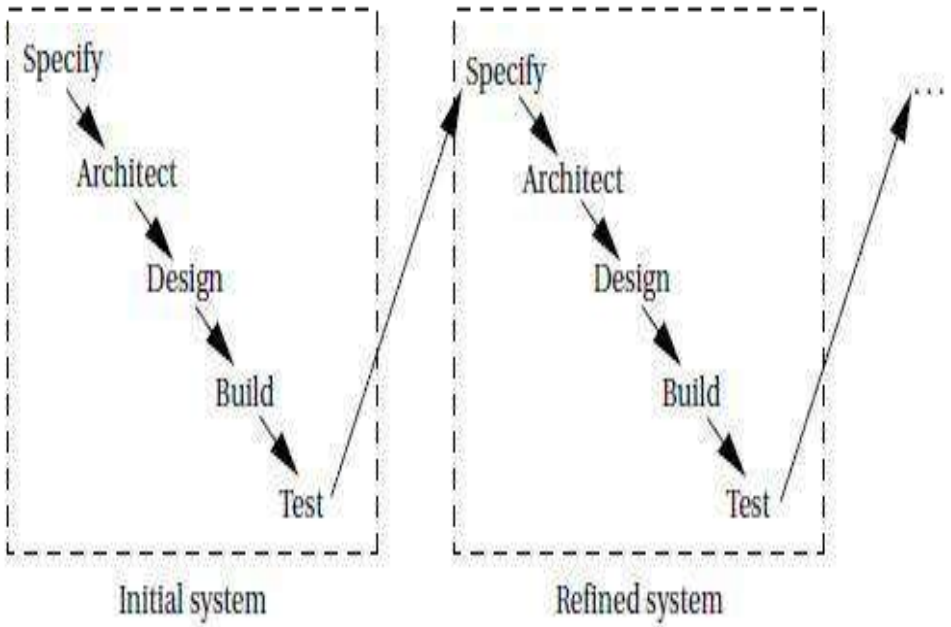
Spiral model

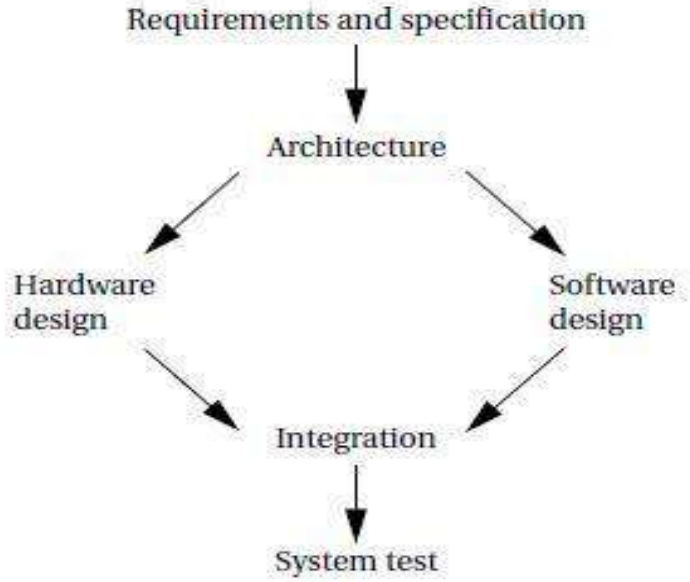
- The spiral model assumes that several versions of the system will be built.
- Each level of design, the designers go through requirements,construction,and testing phases.
- At later stages when more complete versions of the system are constructed.
- Each phase requires more work, widening the design spiral.
- The first cycles at the top of the spiral are very small and short.
- The final cycles at the spiral's bottom add detail learned from the earlier cycles of the spiral.
- The spiral model is more realistic than the waterfall model because multiple iterations needed to complete a design.
- But too many spirals may take long time required for design.

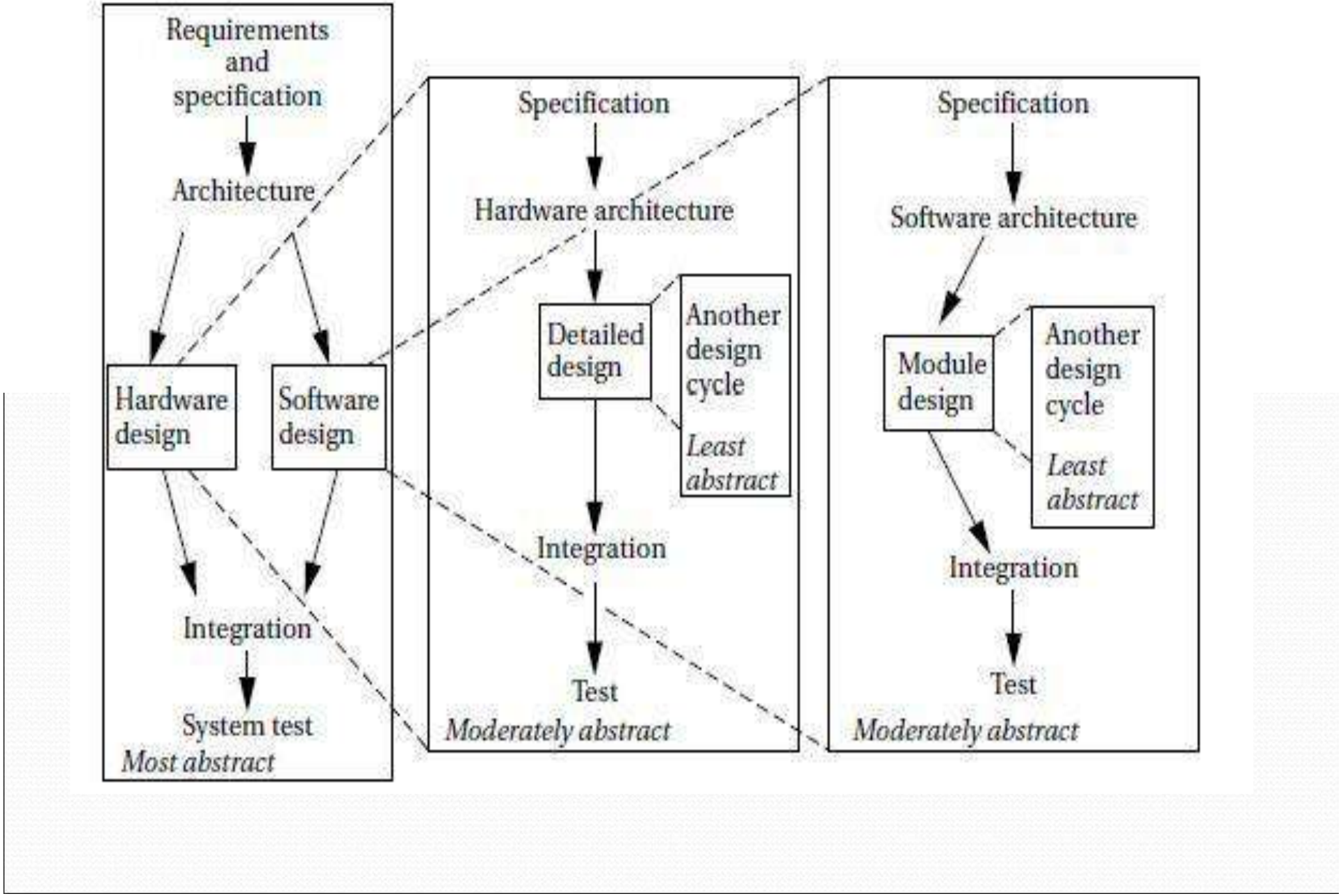


Successive refinement design model

- In this approach, the system is built several times.
- A first system is used as a rough prototype.
- Embedded computing systems are involved the design of hardware/software project.
- Front-end activities→ are specification and architecture and also includes hardware and software aspects.
- Back-end activities → includes integration and testing.
- Middle activities→includes hardware and software development.






Requirements and specification
Architecture
Hardware design
Software design
Integration
System test
Most abstract
Specification
Hardware architecture
Detailed design
Another design cycle
Least abstract
Integration
Test
Moderately abstract
Specification
Software architecture
Module design
Another design cycle
Least abstract
Integration
Test
Moderately abstract

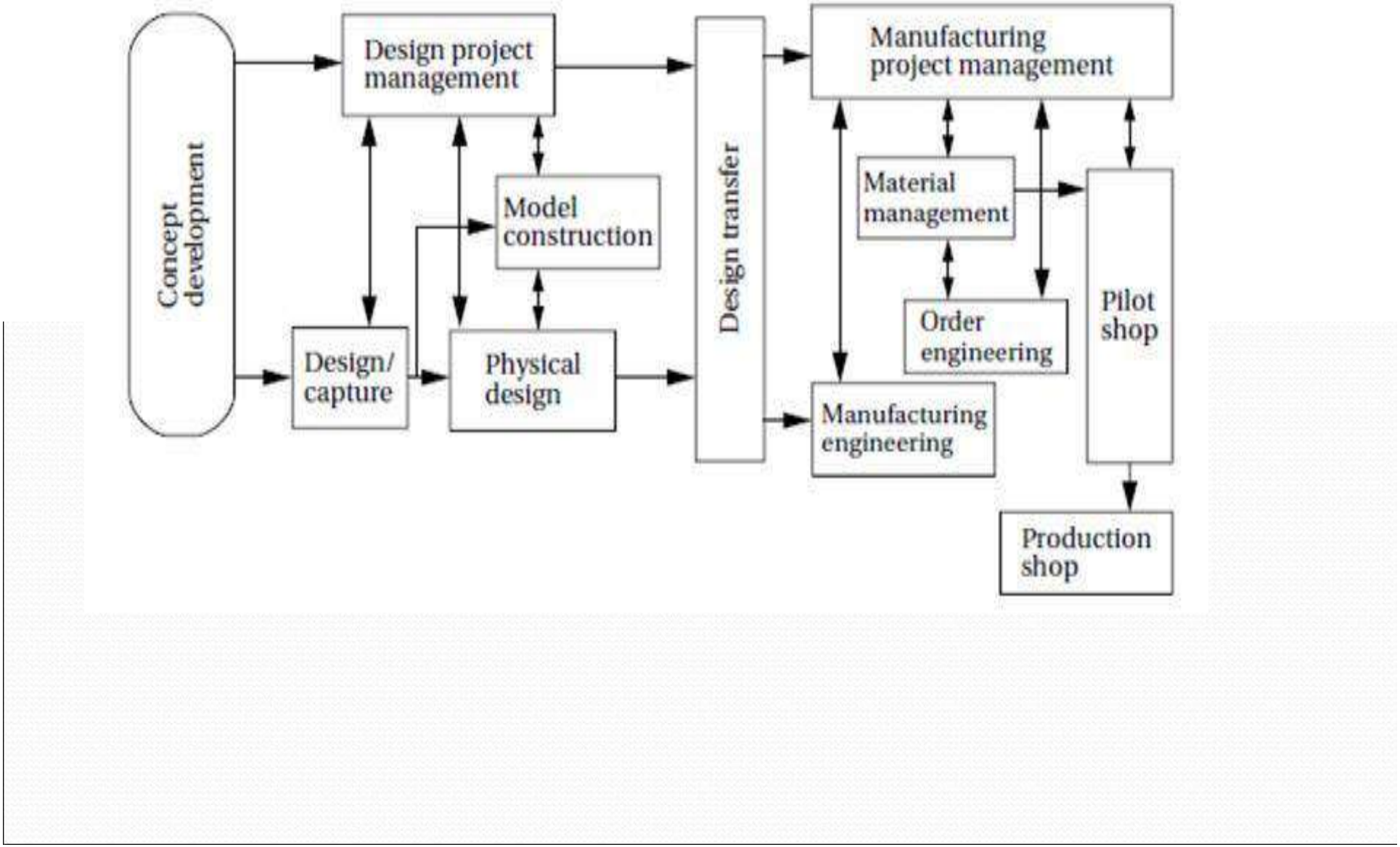
Concurrent engineering

- Reduced design time is an important goal for concurrent engineering.
- It eliminate "over-the-wall" design steps, one designer performs an isolated task and then throws the result to the next designer.

Concurrent engineering efforts are comprised of several elements.

- Cross-functional teams →include members from various disciplines (manufacturing, hardware and software design, marketing)
- Concurrent product→ realize the process activities .
- Designing various subsystems simultaneously, is reducing design time.
- Integrated project management→ensures that someone is responsible for the entire project.
- Early and continual supplier→ make the best use of suppliers' capabilities.
- Early and continual customer →ensure that the product meets customers' needs.

Concurrent Engineering Applied to Telephone Systems



1. **Benchmarking→** They compared themselves to competitors and found that it took them 30% longer to introduce a new product than their best competitors.
2. **Breakthrough improvement.**

- Increased partnership between design and manufacturing.
- Continued existence of the basic organization of design labs and manufacturing.
- Support of managers at least two levels above the working level.

3. **Characterization of the current process.**

- Too many design and manufacturing tasks were performed sequentially.

4. **Create the target process→** The core team created a model for the new development process.
5. **Verify the new process→** test the new process.
6. **Implement across the product line→**This activity required training of personnel, documentation of the new standards and procedures, and improvements to information systems.
7. **Measure results and improve→P**erformance of the new design was measured.

REQUIREMENTS ANALYSIS

- Requirements→It is a informal descriptions of what the customer wants.
- A functional requirement→ states what the system must do.
- A nonfunctional requirement→It can be physical size, cost, power consumption, design time, reliability, and so on.

Requirements of tests

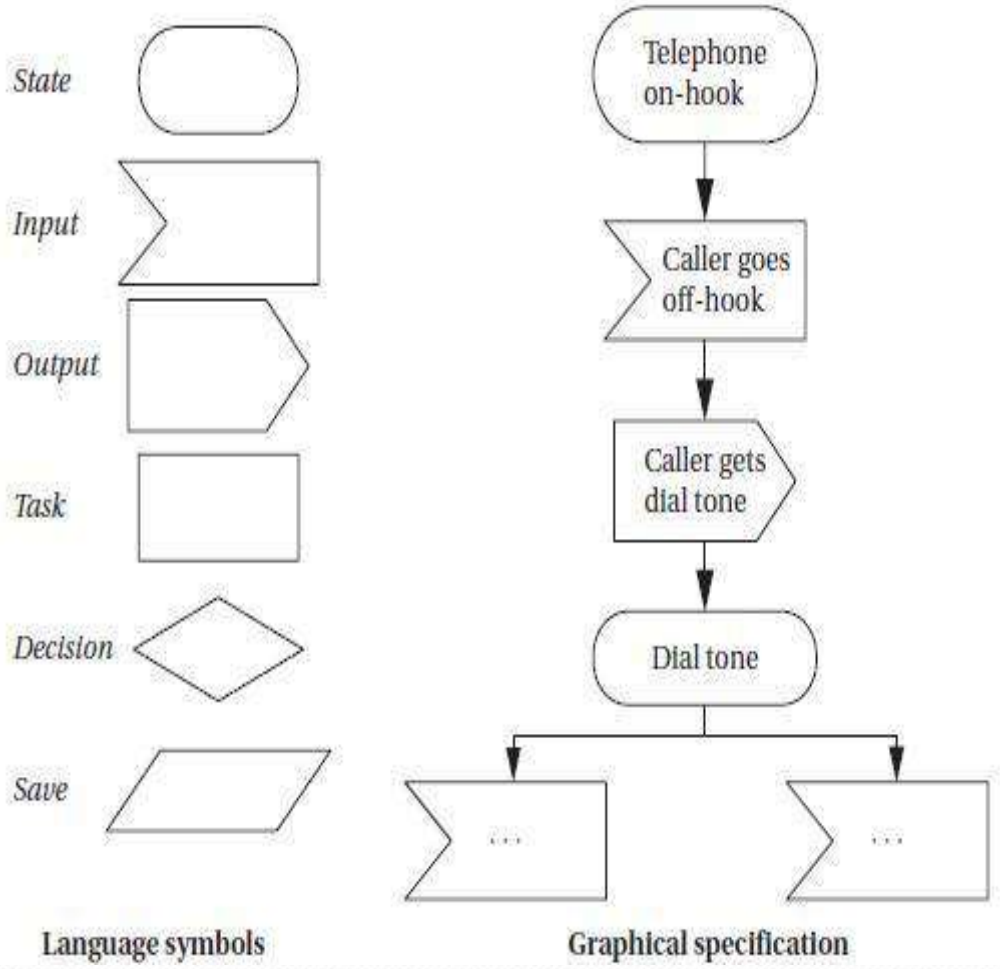
- Correctness→Requirements should not mistakenly describe what the customer wants.
- Unambiguousness→Requirements document should be clear and have only one plain language interpretation.
- Completeness→ Requirements all should be included.
- Verifiability→ cost-effective way to ensure that each requirement is satisfied in the final product.
- Consistency→One requirement should not contradict another requirement.
- Modifiability→The requirements document should be structured so that it can be modified to meet changing requirements without losing consistency.
- Traceability→Able to trace forward /backward from the requirements.

SPECIFICATIONS

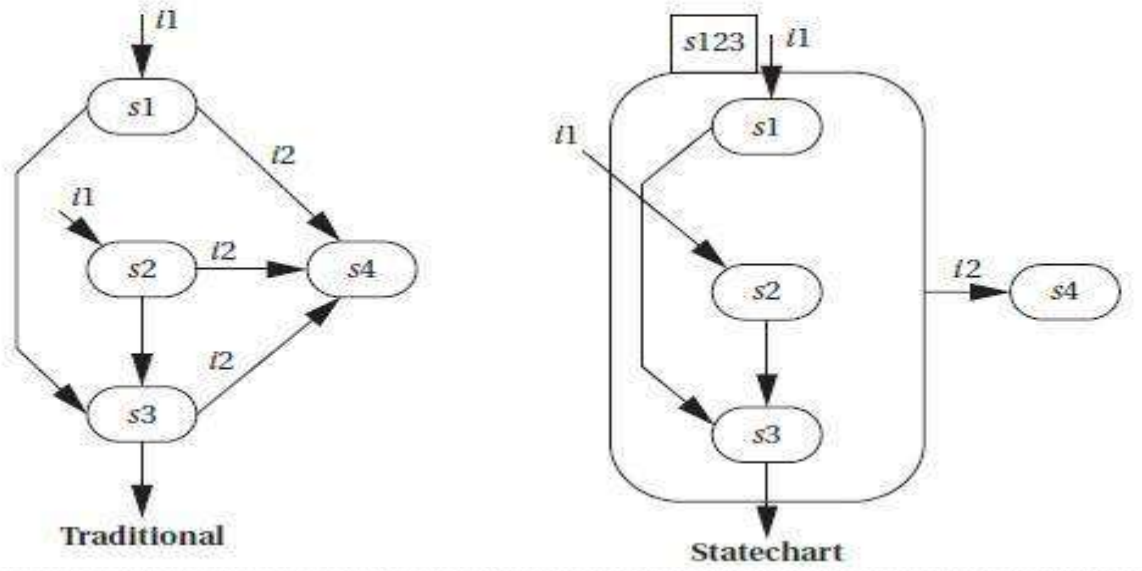
- Specifications→It is a detailed descriptions of the system that can be used to create the architecture.

Control-oriented specification languages

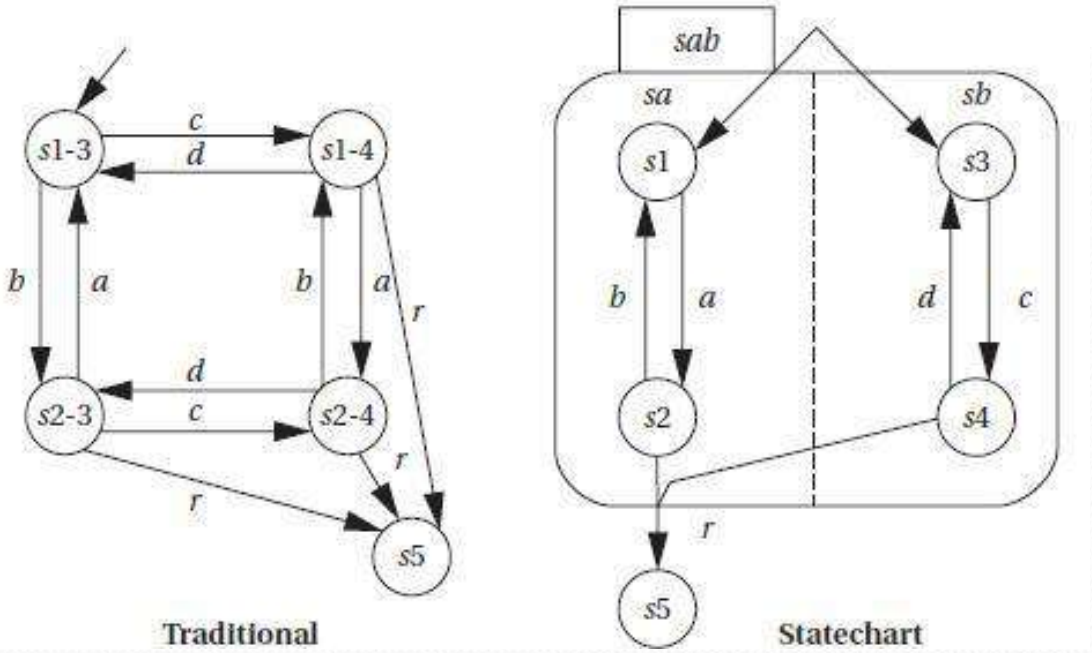
- SDL specifications include states, actions, and both conditional and unconditional transitions between states.
- SDL is an event-oriented state machine model.
- State chart has some important concepts.
- State charts allow states to be grouped together to show common functionality.



- ## Basic groupings(OR)
- State machine specifies that the machine goes to state s4 from any of s1, s2, or s3 when they receive the input i2.
- The State chart denotes this commonality by drawing an OR state around s1, s2, and s3 .
- Single transition out of the OR state s123 specifies that the machine goes to s4 when it receives the i2 input while in any state included in s123.
- Multiple ways to get into s123 (via s1 or s2), and transitions between states within the OR state (from s1 to s3 or s2 to s3).
- The OR state is simply a tool for specifying some of the transitions relating to these states.



- ## Basic groupings(AND)
- In the State chart, the AND state sab is decomposed into two components, sa and sb.
- When the machine enters the AND state, it simultaneously inhabits the state s1 of component sa and the state s3 of component sb.
- When it enters sab, the complete state of the machine requires examining both sa and sb.
- State s1-3 in the State chart machine having its sa component in s1 and its sb component in s3.
- When exit from cluster states go to s5 only when in the traditional specification, we are in state s2-4 and receive input r.



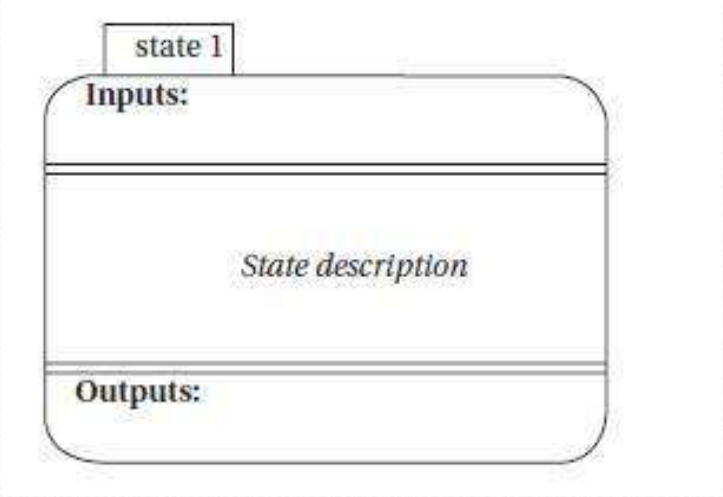
Advanced specifications

- It ensure the correctness and safety of this system.

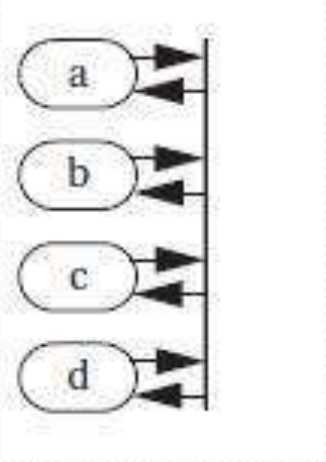
Ex→ Traffic Alert and Collision Avoidance System(TCAS)

- It is a collision avoidance system for aircraft.
- TCAS unit in an aircraft keeps track of the position of other nearby aircraft.
- It uses pre-recorded voice ("DESCEND!) commands for mid-air collision.
- TCAS makes sophisticated decisions in real time and is clearly safety critical.
- It must detect as many potential collision events as possible .
- It must generate a few false alarms ,at extreme maneuvers in potentially dangerous.

TCAS-II specification(RSML Language)

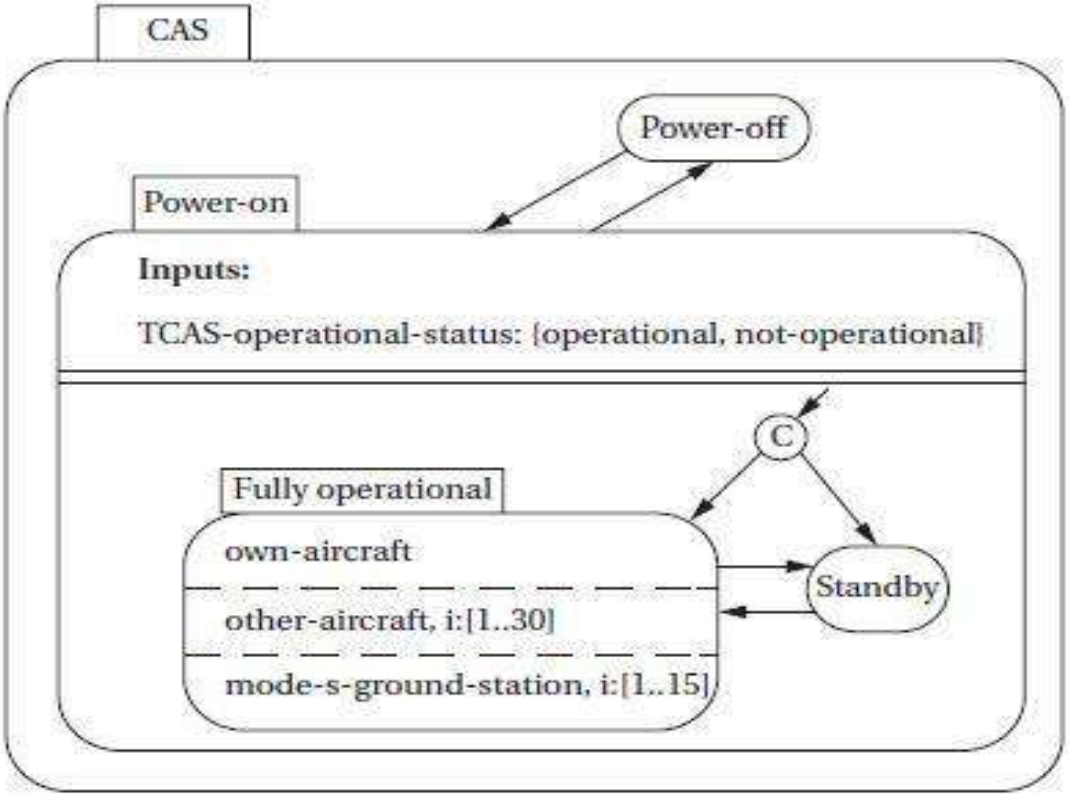


Transition states



Collision Avoidance system

- The system has Power-off and Power-on states .
- In the power on state, the system may be in Standby or Fully operational mode.
- In the Fully operational mode, three components are operating in parallel, as specified by the AND state.
- The own aircraft subsystem to keep track of up to 30 other aircraft.
- Subsystem to keep track of up to 15 Mode S ground stations, which provide radar information.



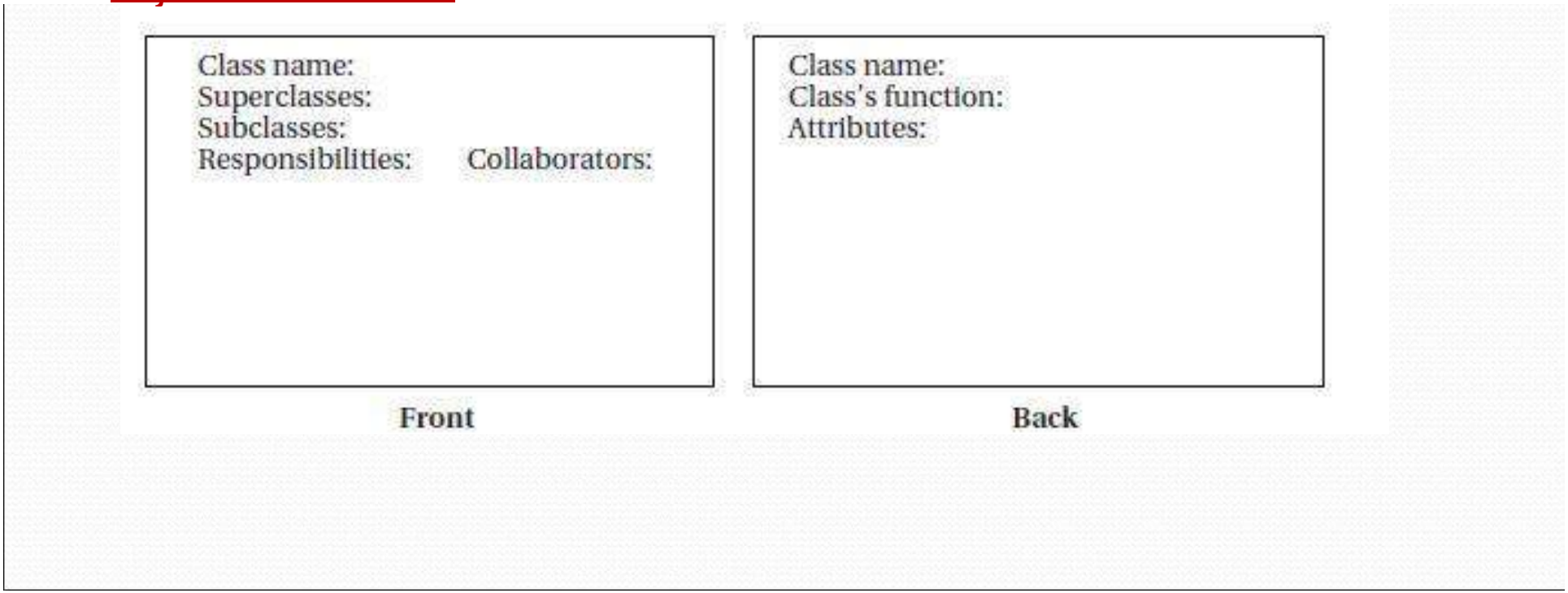
SYSTEM ANALYSIS AND ARCHITECTURE DESIGN

- The CRC card methodology analyze and understanding the overall structure of a complex system.

CRC cards

- Classes define the logical groupings of data and functionality.
- Responsibilities describe what the classes do.
- Collaborators are the other classes with which a given class works.
- It has space to write down the class name, its responsibilities and collaborators, and other information.

Layout of CRC card



- A class may represent a real-world object of the system.
- A class has both an internal state and a functional interface.
- The functional interface describes the class's capabilities.
- The responsibility set is describing that functional interface.
- The collaborators of a class are simply the classes that it talks or calls upon to help it do its work.
- CRC card Analysis Process

1. Develop an initial list of classes→Write down the class name and functions of it.
2. Write an initial list of responsibilities and collaborators.
3. Create some usage scenarios→describe what the system does.
4. Walk through the scenarios→Each person on the team represents one or more classes.
5. Refine the classes, responsibilities, and collaborators→ make changes to the CRC cards.
6. Add class relationships→ subclass and super-class can be added to the cards.

Ex:Elevator system

1. One passenger requests a car on a floor, gets in the car when it arrives, requests another floor, and gets out when the car reaches that floor.
2. One passenger requests a car on a floor, gets in the car when it arrives, and requests the floor that the car is currently on.
3. A second passenger requests a car while another passenger is riding in the elevator.
4. Two people push floor buttons on different floors at the same time.
5. Two people push car control buttons in different cars at the same time.

Class	Responsibilities	Collaborators
Elevator car*	Moves up and down	Car control, car sensor, car control sender
Passenger*	Pushes floor control and car control buttons	Floor control, car control
Floor control*	Transmits floor requests	Passenger, floor control reader
Car control*	Transmits car requests	Passenger, car control reader
Car sensor*	Senses car position	Scheduler
Car state	Records current position of car	Scheduler, car sensor
Floor control reader	Interface between floor control and rest of system	Floor control, scheduler
Car control reader	Interface between car control and rest of system	Car control, scheduler
Car control sender	Interface between scheduler and car	Scheduler, elevator car
Scheduler	Sends commands to cars based upon requests	Floor control reader, car control reader, car control sender, car state

Capability Maturity Model (CMM)

- It is used to measuring the quality of an organization's software development.

1. Initial→A poorly organized process, with very few well-defined processes. Success of a project depends on the efforts of individuals, not the organization itself.
2. Repeatable→ provides basic tracking mechanisms to understand cost, scheduling .
3. Defined→The management and engineering processes are documented and standardized.
4. Managed→detailed measurements of the development process and product quality.
5. Optimizing→feedback from detailed measurements is used to continually improve the organization's processes.

- Verifying the specification→Discovering bugs early is crucial because it prevents bugs from being released to customers, minimizes design costs, and reduces design time.
- Validation of specifications→creating the requirements, including correctness, completeness, consistency, and so on

Design reviews

- The review leader coordinates the pre-meeting activities, the design review itself, and the post-meeting follow-up.
- The reviewer records the minutes of the meeting so that designers and others know which problems need to be fixed.
- The review audience studies the component.

DESIGNING WITH COMPUTING PLATFORM

System Architecture

- The architecture of an embedded computing system includes both hardware and software elements

HARDWARE

- ***CPU*** →The choice of the CPU is one of the most important, but it can be considered the software that will execute on the machine.
- ***Bus*** →The choice of a bus is closely tied to that of a CPU,bus can handle the traffic.
- ***Memory***→Selection depends total size and speed of the memory will play a large part in determining system performance.
- **Input and output devices**→ Dependig upon the system requirements

SOFTWARE

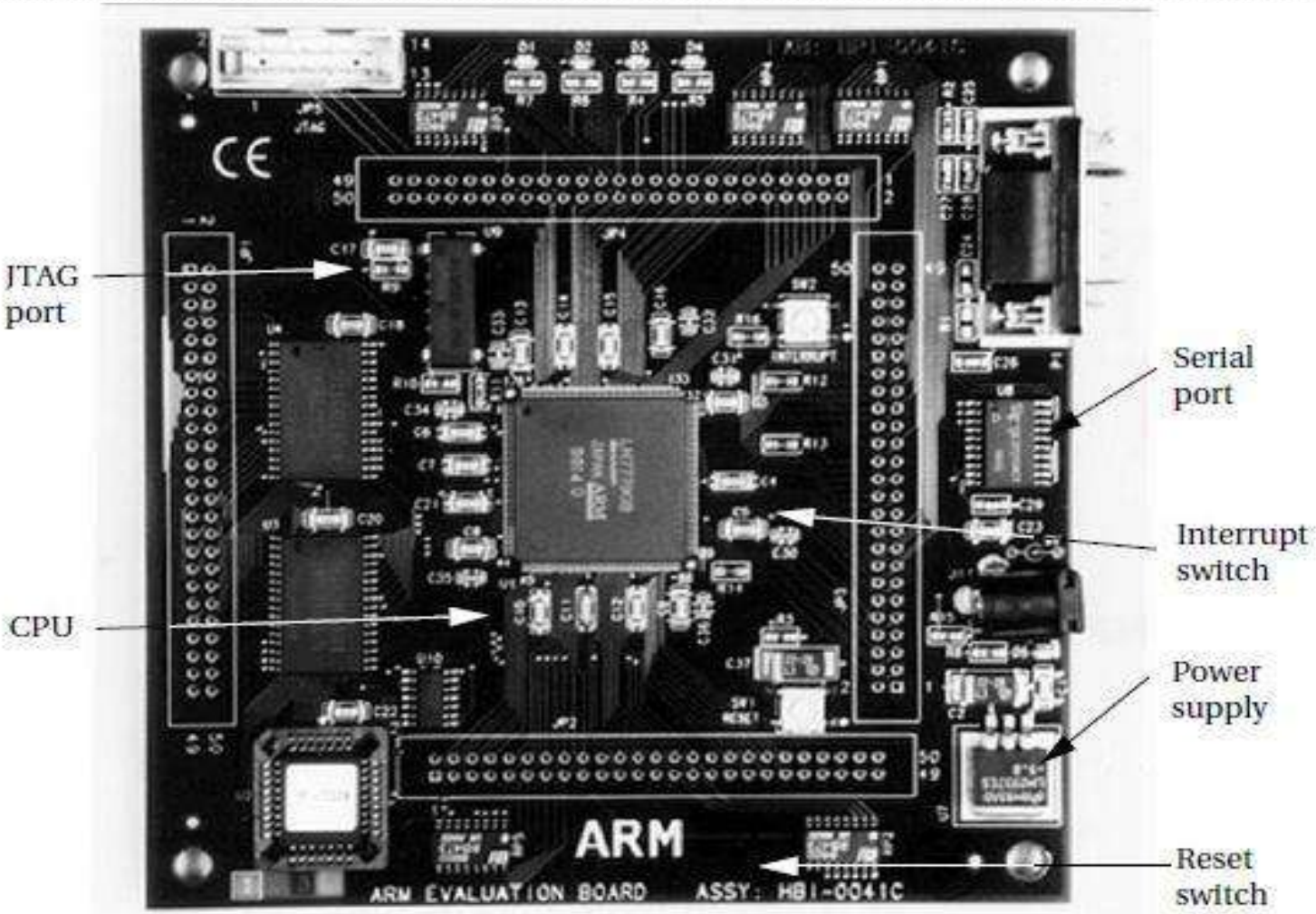
Run Time components

- It is a critical part of the platform.
- An operating system is required to control CPU and its multiple processes .
- A file system is used in many embedded systems to organize internal data and interface with other systems

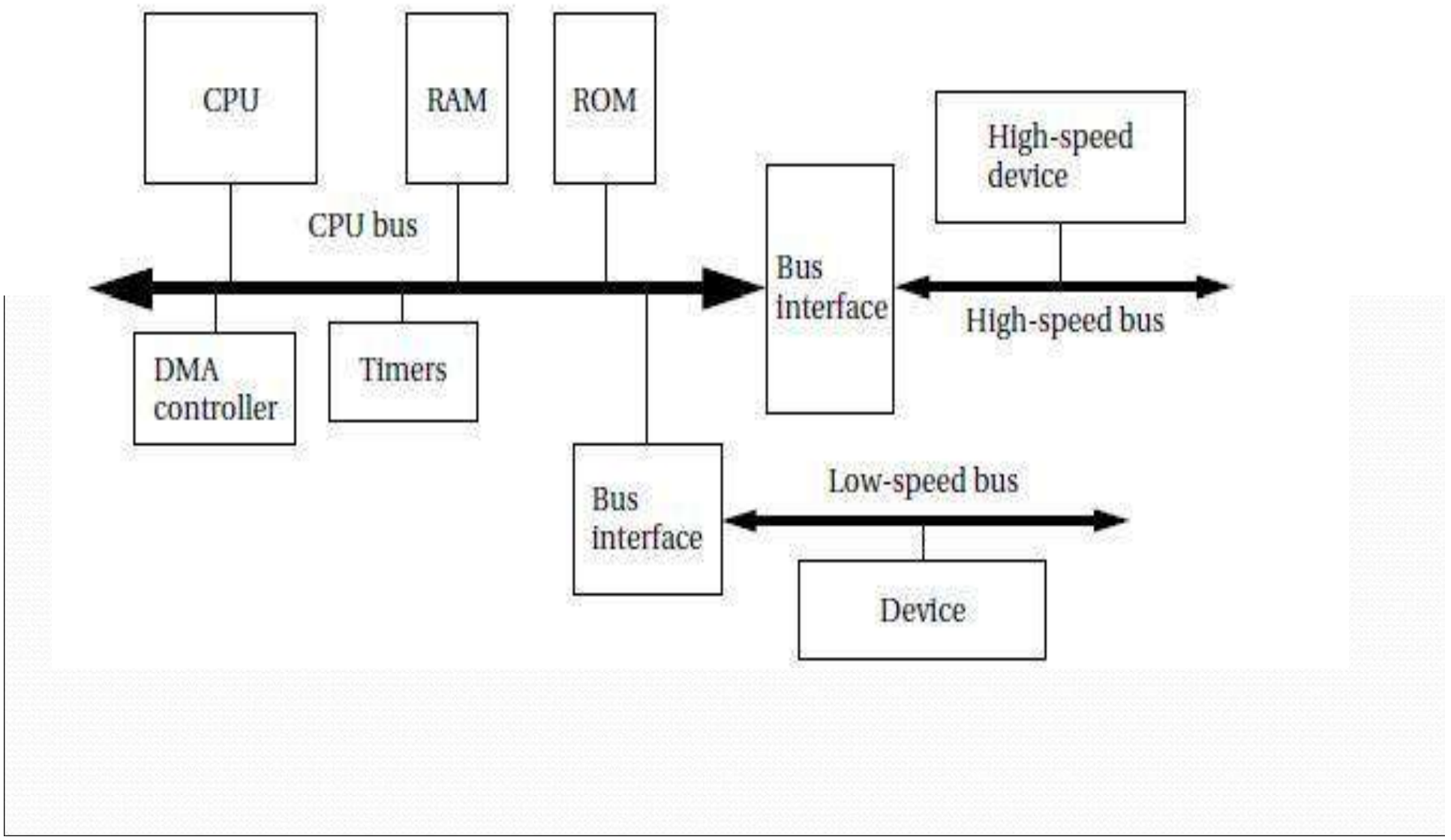
Support components

- It is a complex hardware platform.
- Without proper code development and operating system, the hardware itself is useless.

ARM evaluation board



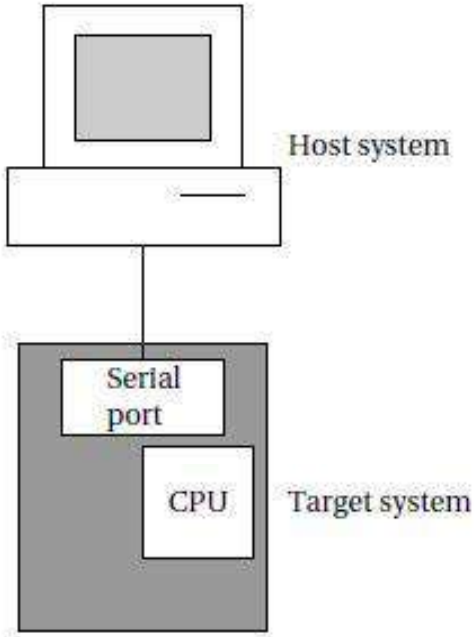
1.10.2)The PC as a Platform



- CPU →provides basic computational facilities.
- RAM→ is used for program storage.
- ROM→holds the boot program.
- DMA→controller provides DMA capabilities.
- Timers→ used by the operating system for a variety of purposes.
- High-speed bus→connected to the CPU bus through a bridge, allows fast devices to communicate with the rest of the system.
- low-speed bus→ provides an inexpensive way to connect simpler devices.

Development Environments

- Development process → used to make a complete design of the system.
- It guides the developers how to design a system .
- An embedded computing system has CPU ,memory, I/O devices.
- Development of embedded system have both hardware& software.
- The software development on a PC or workstation known as a host.



•The host and target are frequently connected by a USB link.
•The target must include a small amount of software to talk to the host system.

Functions of Host system

- Load programs into the target
- Start and stop program execution on the target
- Examine memory and CPU registers.

Cross-Compiler

- Compiler→ kind of software that translate one form of pgm to another form of pgm.
- Cross Compiler→is a compiler that runs on one type of machine but generates
- code for another
- After compilation, the executable code is downloaded to the embedded system by a serial link.
- A PC or workstation offers a programming environment .
- But one problem with this approach emerges when debugging code talks to I/O devices.
- Testbench program→ can be built to help debug the embedded code.
- It may also take the output values and compare them against expected values.

Debugging Techniques

- It is the process of checking errors and correcting those errors.
- It can be done by compiling and executing the code on a PC or workstation.
- It can be performed by both H/W and S/W sides.

Software debugging tools

1. Serial Port tool

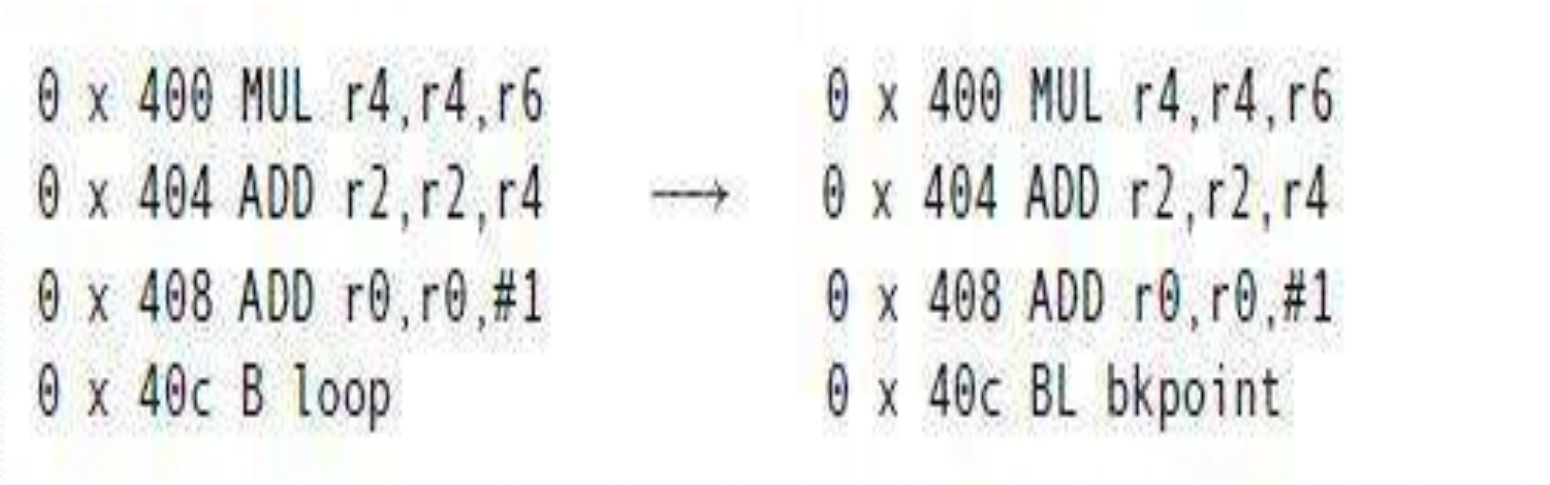
- It will perform the debugging process from the initial state of embedded system design.
- It can be used not only for development debugging but also for diagnosing problems in the field.

2. Breakpoints tool

- user to specify an address at which the program's execution is to break.
- When the PC reaches that address, control is returned to the monitor program.
- From the monitor program, the user can examine and/or modify CPU registers, after which execution can be continued.

▪Breakpoint is a location in memory at which a program stops executing and returns to the debugging tool or monitor program.

To establish a breakpoint at location 0x40c in some ARM code, replaced the branch (B) instruction with a subroutine call (BL) to the breakpoint handling routine



- Hardware debugging tools
- Hardware can be deployed to give a clearer view on what is happening when the system is running.

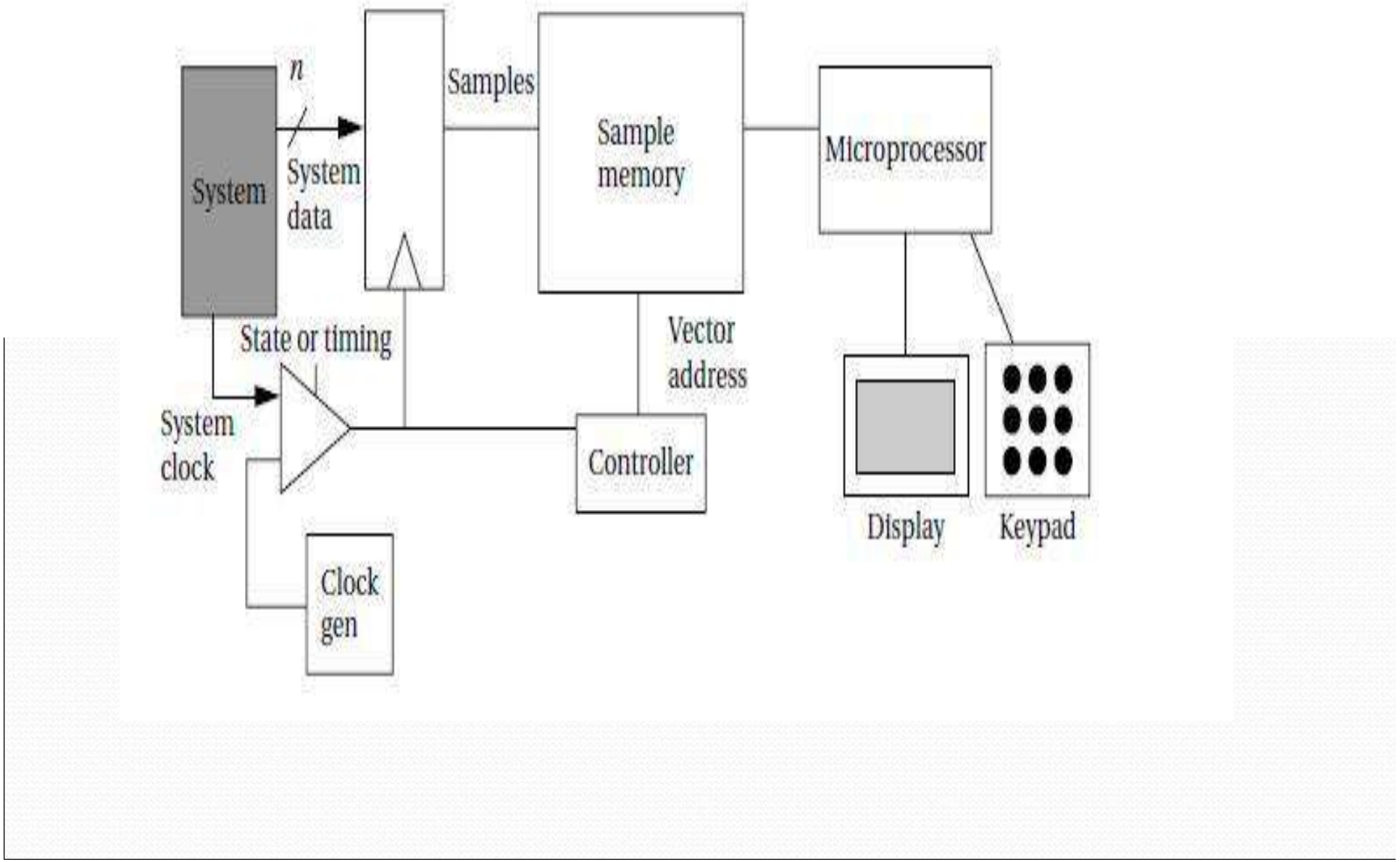
1. Microprocessor In-Circuit Emulator (ICE)

- It is a specialized hardware tool that can help debug software in a working embedded system.
- In-circuit emulator is a special version of the microprocessor that allows its internal registers to be read out when it is stopped

2. Logic Analyer

- The analyzer can sample many different signals simultaneously but can display only 0, 1, or changing values for each.
- The logic analyzer records the values on the signals into an internal memory and then displays the results on a display once the memory is full.

Architecture of a logic analyzer



Data modes of logic analyzer

State modes

- State mode represent different ways of sampling the values.
- It uses the own clock to control sampling
- It samples each signal only one per clock cycle.
- It has less memory to store a given number of system clock.

Timing modes

- Timing mode uses an internal clock that is fast enough to take several samples per clock period in a typical system.

1.10.5)Debugging Challenges

- Logical errors in software can be hard to track down and it will create many problems in real time code.
- Real-time programs are required to finish their work within a certain amount of time.
- Run time pgm run too long, they can create very unexpected behavior.
- Missing of Deadline makes debugging process as difficult.

Consumer Electronic Architecture

- Consumer electronic refers to any device containing an electronic circuit board that is intended for eneryday use by individuals.
- Eg→TV,cameras,digital cameras,calculators,DVDs,audio devices,smart phones etc..,

1.11..1)Functional Requirements

1. *Multimedia*
- The media may be audio, still images, or video.
- These multimedia objects are generally stored in compressed form and must be uncompressed to be played .
- Eg→ multimedia compression standards (MP3,Dolby Digital(TM))
- audio; JPEG for still images; MPEG-2, MPEG-4, H.264, etc. for video.

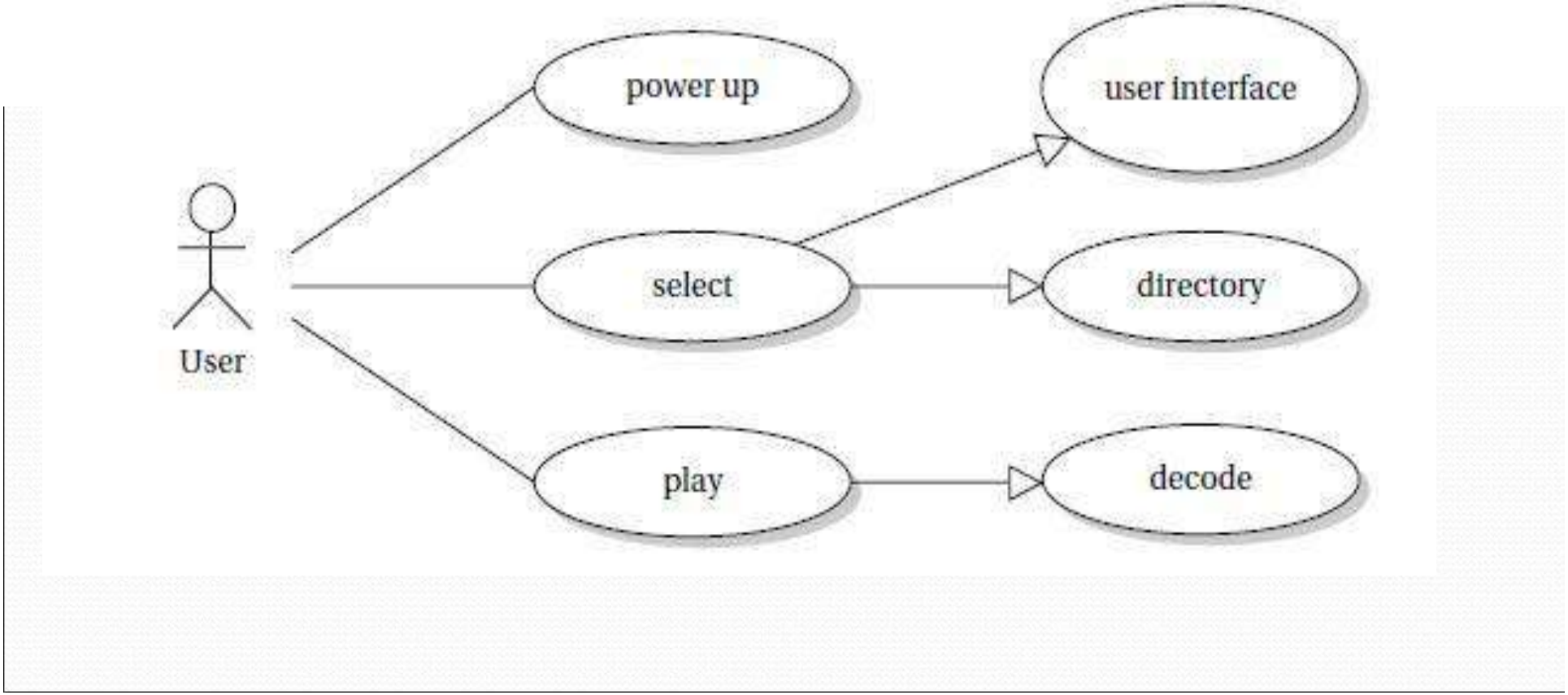
2. *Data storage and management*→ People want to select what multimedia objects they save or play, data storage goes hand-in-hand with multimedia capture and display. Many devices provide PC-compatible file systems so that data can be shared more easily.
3. *Communications*→ Communications may be relatively simple, such as a USB
- and another is Ethernet port or a cellular telephone link.

Non-Functional Requirements

- Many devices are battery-operated, which means that they must operate under strict energy budgets.
- Battery(75mW) →support not only the processors but also the display, radio, etc.
- Consumer electronics must also be very inexpensive but provide very high performance.

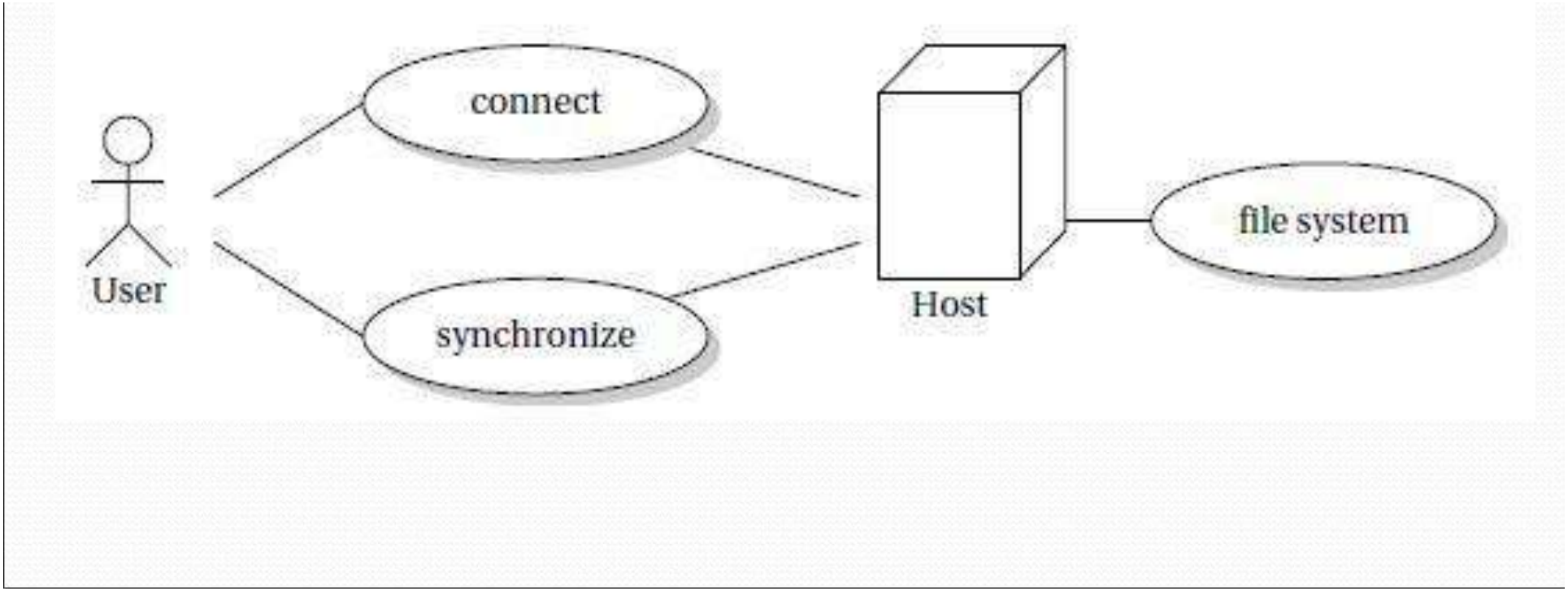
Use case for playing multimedia

- use case for selecting and playing a multimedia object (audio clip, a picture,etc.).
- Selecting an object makes use of both the user interface and the file system.
- Playing also makes use of the file system as well as the decoding subsystem and I/O subsystem.

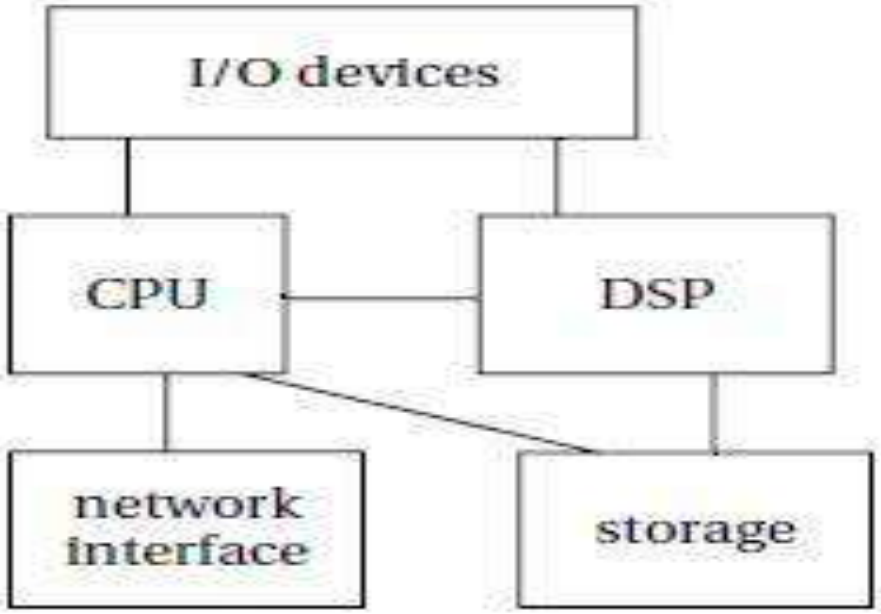


Use case of synchronizing with a host system

- use case for connecting to a client.
- The connection may be either over a local connection like USB or over the Internet.
- Some operations may be performed locally on the client device
- most of the work is done on the host system while the connection is established



Functional architecture of Consumer Electronics Device(CED)



- It is a two-processor architecture.
- If more computation is required, more DSPs and CPUs may be added.
- The RISC-CPU runs the operating system, runs the user interface, maintains the file system, etc.
- DSP→ it is a programmable one, which performs signal processing.
- Operating system→ runs on the CPU must maintain processes and the file system.
- Depending on the complexity of the device, the operating system may not need to create tasks dynamically.
- If all tasks can be created using initialization code, the operating system can be made smaller and simpler.

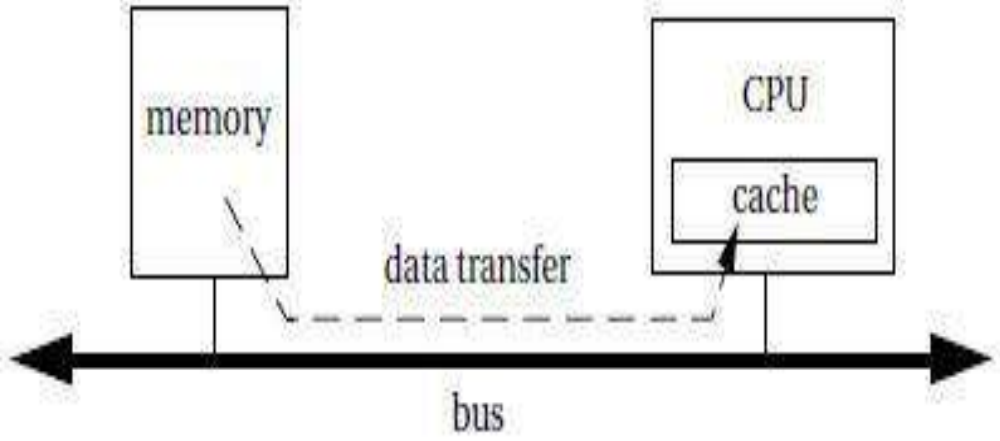
1.11.4 Flash File Systems

- Many consumer electronics devices use flash memory for mass storage.
- Flash memory is a type of semiconductor memory ,unlike DRAM or SRAM, provides permanent storage.
- Values are stored in the flash memory cell as electric charge using a specialized capacitor that can store the charge for years.
- The file system of a device is typically shared with a PC.
- Standard file system→has two layers.bottom layer handles physical reads and writes on the storage device and the top layer provides a logical view of the file system.
- Flash file system→imposes an intermediate layer that allows the logical-to-physical mapping of files to be changed.

Platform-Level Performance Analysis

- System-Level Performance involves much more than the CPU.
- To move data from memory to the CPU to process it. To get the data from memory to the CPU we must.

1. Read from the memory.
2. Transfer over the bus to the cache.
3. Transfer from the cache to the CPU.



- The performance of the system based on Bandwidth of the system.
- We can increase bandwidth in two ways:

1) By increasing the clock rate of the bus
2) By increasing the amount of data transferred per clock cycle.

For example, bus to carry four bytes or 32 bits per transfer, we would reduce the transfer time to 0.058 s. If we also increase the bus clock rate to 2 MHz, then we would reduce the transfer time to 0.029 s ,which is within our time budget for the transfer.

$$t=TP$$

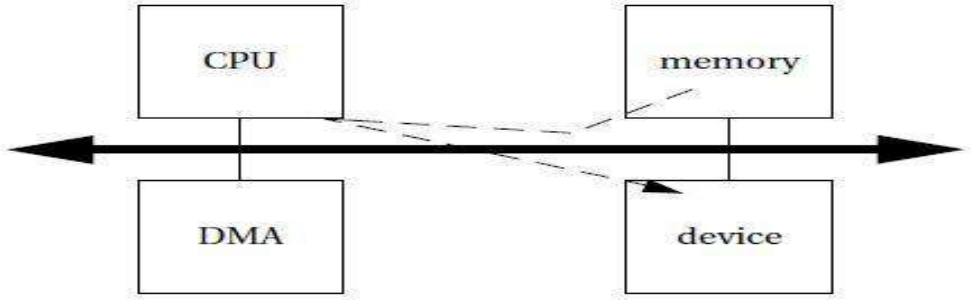
t→bus cycle counts

T→bus cycles.

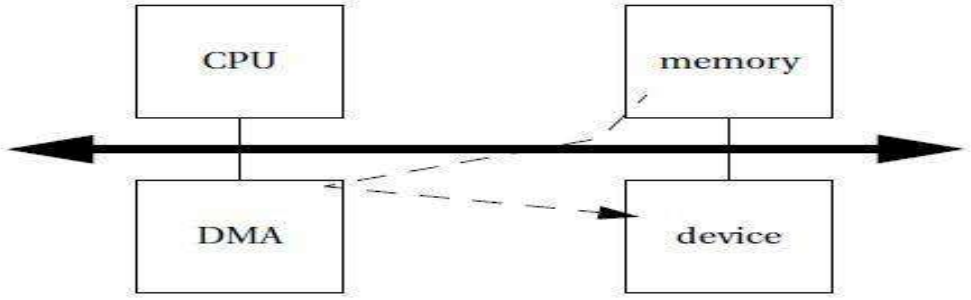
p→bus clock period

Parallelism

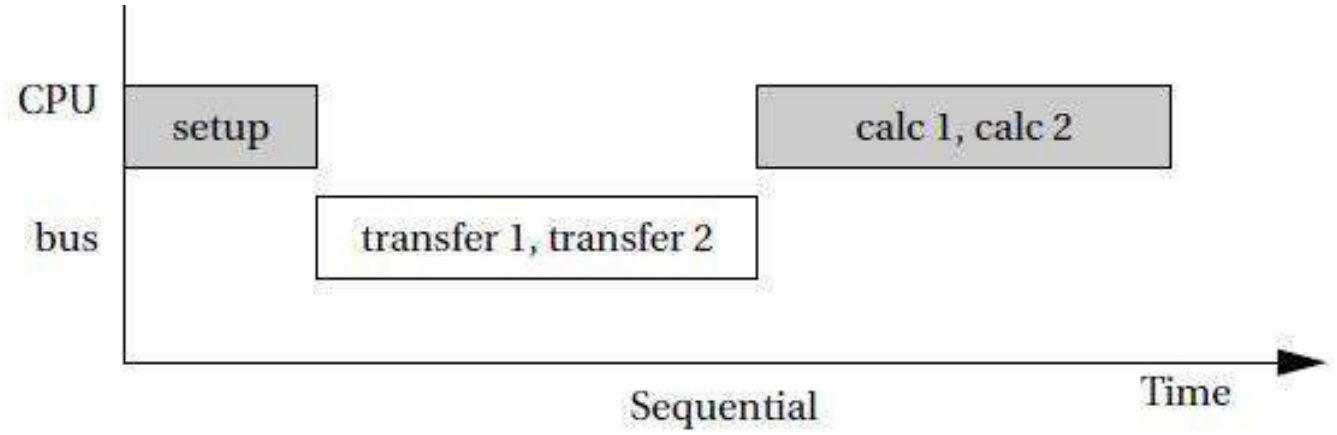
- Direct memory access is a example of parallelism.
- DMA was designed to off-load memory transfers from the CPU.
- The CPU can do other useful work while the DMA transfer is running.

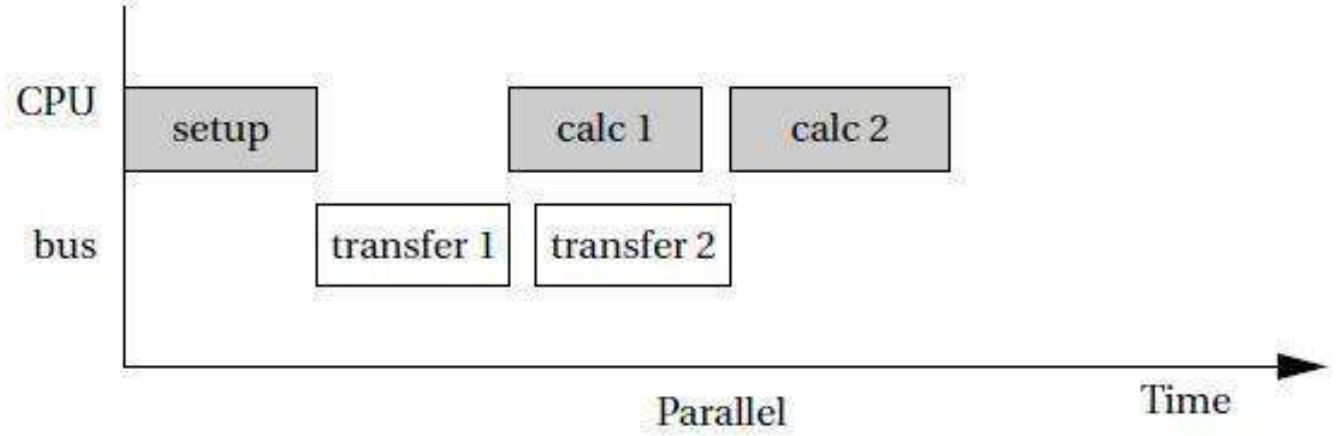


transfer without DMA



transfer without DMA


CPU
setup
calc 1, calc 2
bus
transfer 1, transfer 2
Time
Sequential


CPU
setup
calc 1
calc 2
bus
transfer 1
transfer 2
Time
Parallel

UNIT II

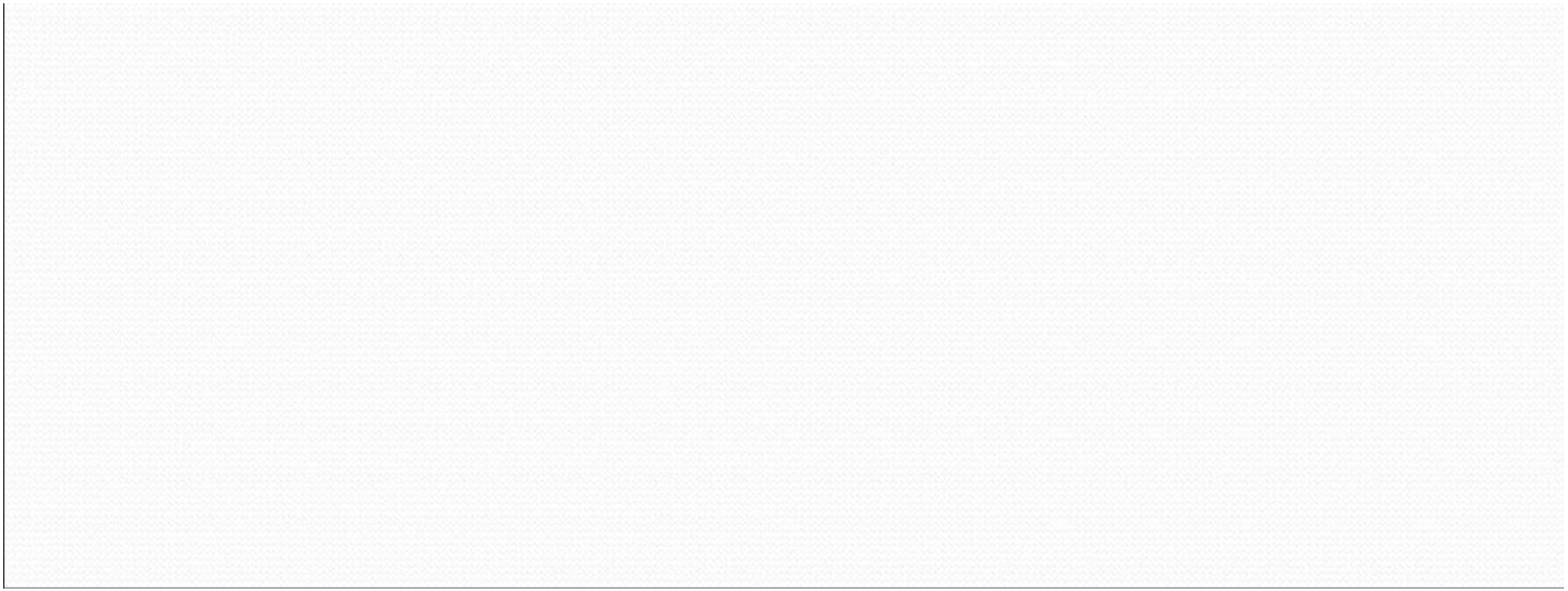
ARM PROCESSOR AND PERIPHERALS

ARM Architecture Versions – ARM Architecture – Instruction Set – Stacks and Subroutines – Features of the LPC 214X Family – Peripherals – The Timer Unit – Pulse Width Modulation Unit – UART – Block Diagram of ARM9 and ARM Cortex M3 MCU.

ARM Architecture Versions

- The ARM processor is a Reduced Instruction Set Computer (RISC).
- The first ARM processor was developed at Acorn Computers Limited, of Cambridge, England, between October 1983 and April 1985. It is very simple architecture.
- At that time, and until the formation of Advanced RISC Machines Limited (which later was renamed simply ARM Limited) in 1990, ARM stood for Acorn RISC Machine

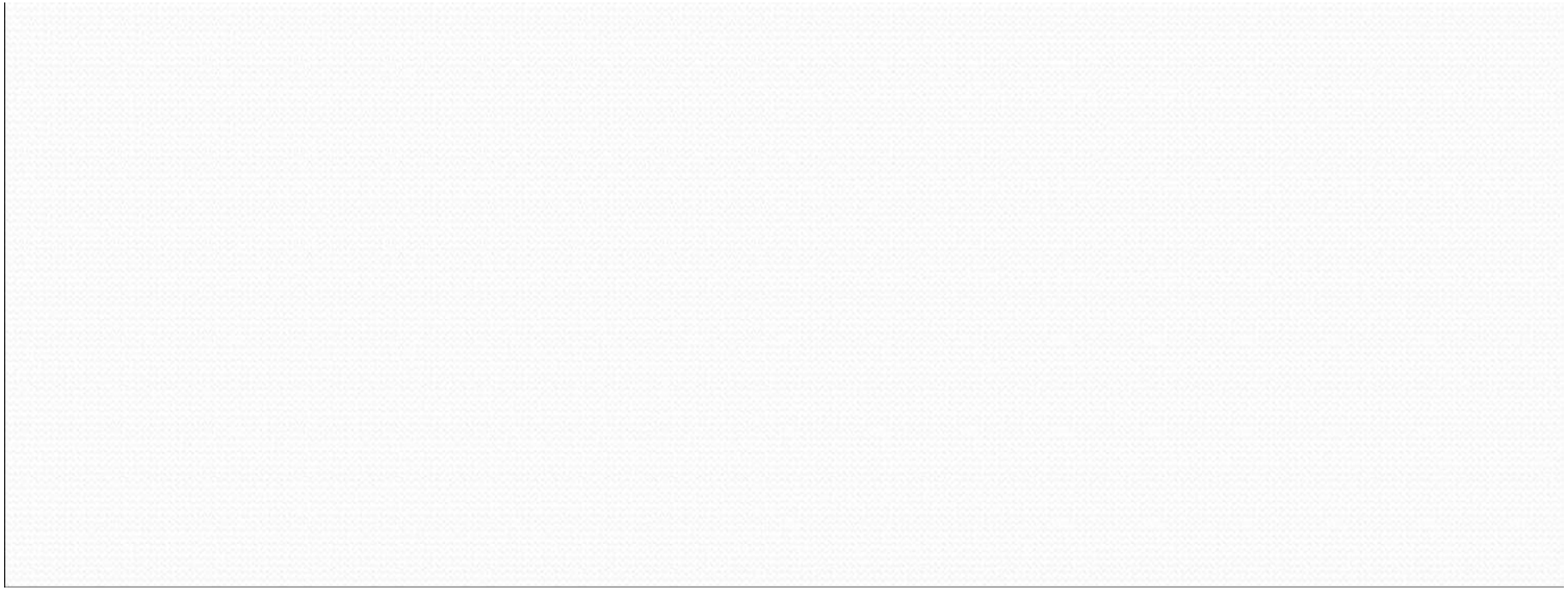
- Second, both ARM ISA and pipeline design are aimed to minimize the energy consumption.
- Third, the ARM architecture is highly modular only mandatory component of ARM processor is the integer pipeline, others are optional. This gives more flexibility in application dependent architecture



Revision	Example core implementation	ISA Enhancement
ARM v1	ARM1	•First ARM Processor •26bit addressing
ARMv2	ARM2	•32bit multiplier •32bit coprocessor support
ARMv2a	ARM3	•On chip cache •Atomic swap instruction •Coprocessor 15 for cache management
ARMv3	ARM6 and ARM7DI	•32 bit addressing •Separate cpsr (current Program status register)and spsr (Saved program status register) •New modes undefined instruction and abort •MMU support(Memory Management Unit)

ARMv3M	ARM7M	Signed and un signed long multiply instruction
ARMv4	Strong ARM	•load store instructions for signed half words/bytes •Reserve SWI(software interrupt) space fro architecturally define operations. •26 bit addressing mode no longer supported
ARMv4T	ARM7TDMI and ARM9T	•Thumb
ARMV5TE	ARM9E AND ARM10E	•Superset of ARM •Enhanced multiply instructions •Extra DSP type instruction •Faster multiply instruction

ARM V5tej	ARM7EJ & ARM926EJ	Java Acceleration
ARMv6	ARM11	•Improved Multiprocessor instructions •Unaligned and Mixed endian data handling



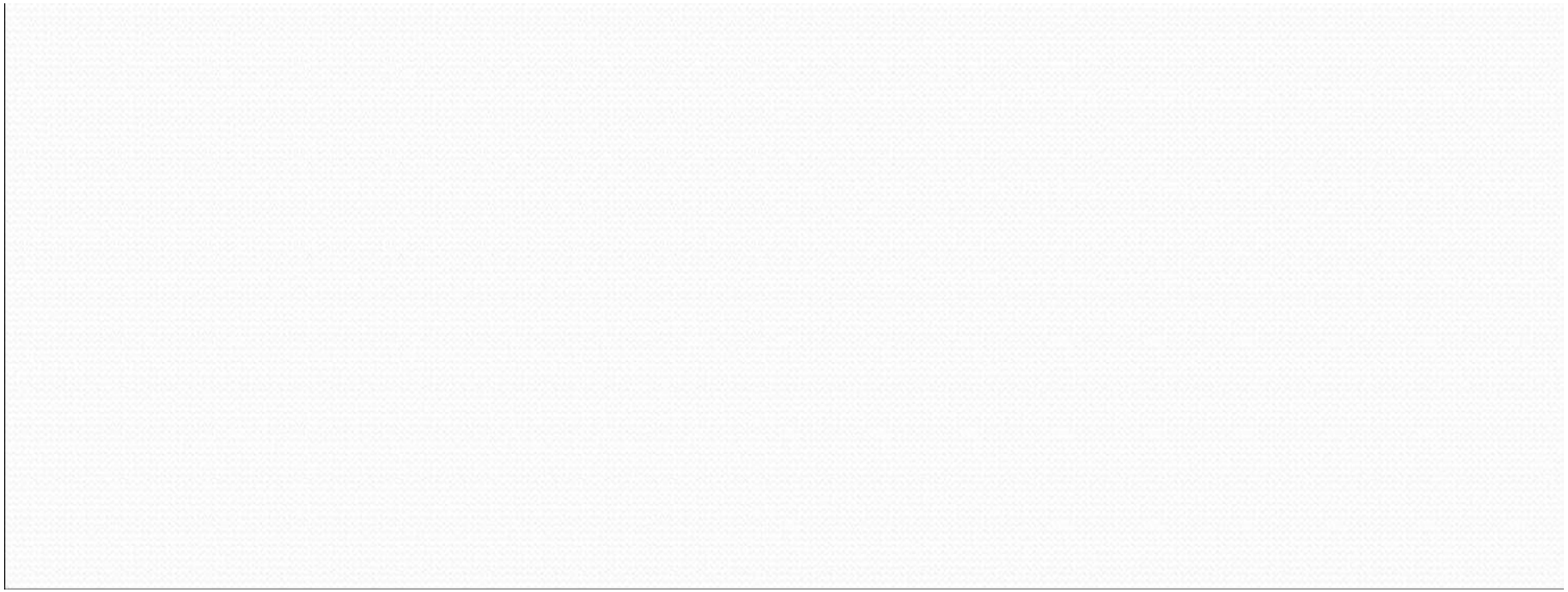
ARM processor Features

Terms	Extention
X	Family or series
Y	Memory Management
Z	Cache
T	16bit thumb decoder
D	Jtag Debugger
M	Fast multiplier
I	Embedded In circuit Emulator
E	Enhanced Instruction for DSP
J	Jazelle
F	Vector floating point unit
S	Synthesizable version

ARM 7family

- ARM7 core has a von neumann style architecture
- ARM7 TDMI is first processor introduced in 1995 by ARM
- It provide a very good performance to power ratio
- ARM7TDMI-S has the synthesizable
- ARM720T is the most fexible member of ARM7 family because it include MMU. MMU handle both platforms Linux and windows
- It having unified 8k cache and vector table are relocated depend on the priority

- ARM7EJS processor, also synthesizable. Its having five stage pipeline and execute ARMv5TEJ instruction
- This version only support java acceleration.



ARM9 family

- The ARM9 family was announced in 1997
- ARM9 has five stage pipeline and high clock frequencies
- Memory have been redesign Harvard architecture
- ARM9 process includes cache and MMU
- Operating system requiring virtual memory support
- ETM (Embedded Trace Macrocell) which allows a developer to trace instruction and data execution in real time operation. So that debugging is done during the critical time segments.
-

- ARM946E-S include TCM, cache and MPU. The size of the TCM and cache are configurable
- The processor is designed for the embedded control application that require deterministic real time response
- ARM926EJ-S synthesizable processor core, announced in 2000
- It a java enable device such as 3G phones and personal digital assistant

RM10 FAMILY

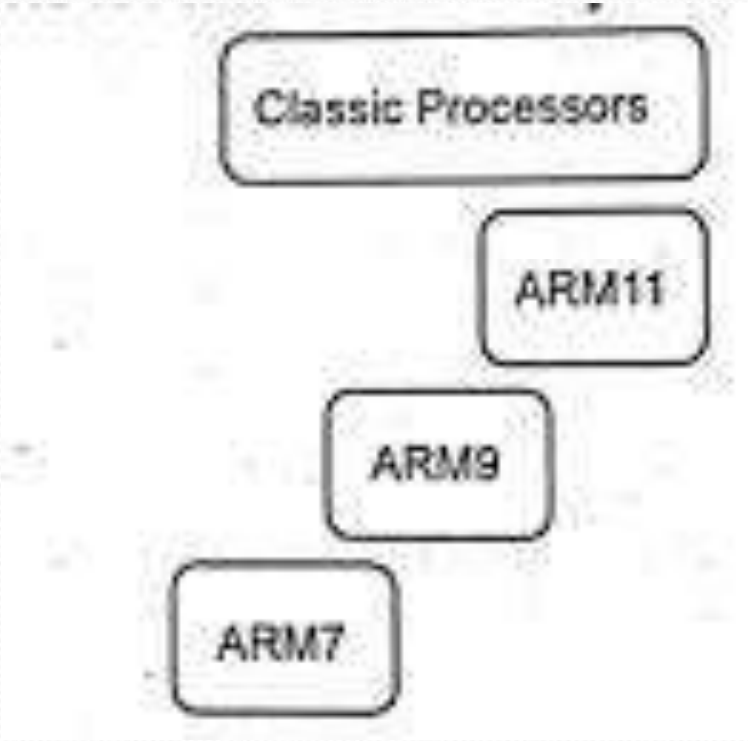
- The ARM10 announced in 1997 was designed for performance
- It extended version of 6 stage pipeline
- Vector floating point unit which adds a seventh stage to the ARM10 pipeline
- VFP combined with IEEE 754.1985 floating point
- ARM1020 E it includes E instruction. it having cache, VFP and MMU
- ARM1026EJ-S is similar to ARM926EJ-S . But ARM10 is flexible when compare to ARM9

ARM11

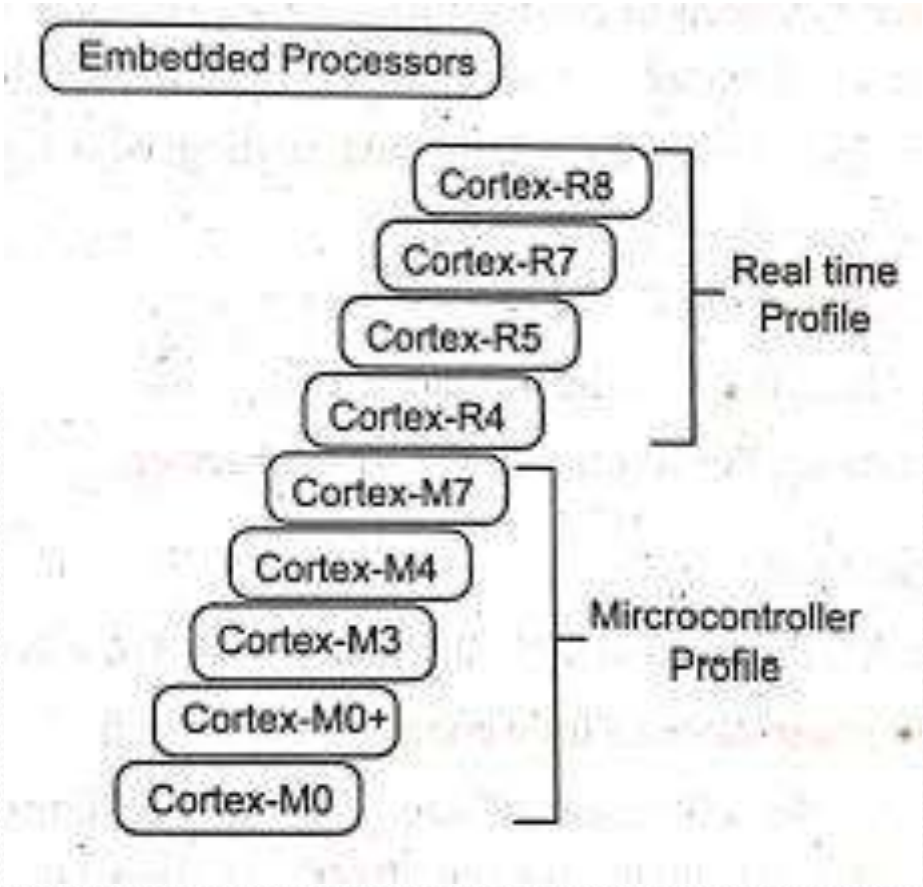
- ARM1136J-S, announced in 2003 was designed for high performance and power efficient applications
- ARM1136J-S was the first processor to execute architecture ARMv6 instructions
- It has eight pipeline stages with load and store arithmetic pipeline.
- ARMv6 instruction are single instruction with multiple data extensions for media processing.

- 2.2 ARM PROCESSORS
- ARM Processor can be divided into three types
  - ARM classic processor
  - ARM Embedded Processor
  - ARM Application processor

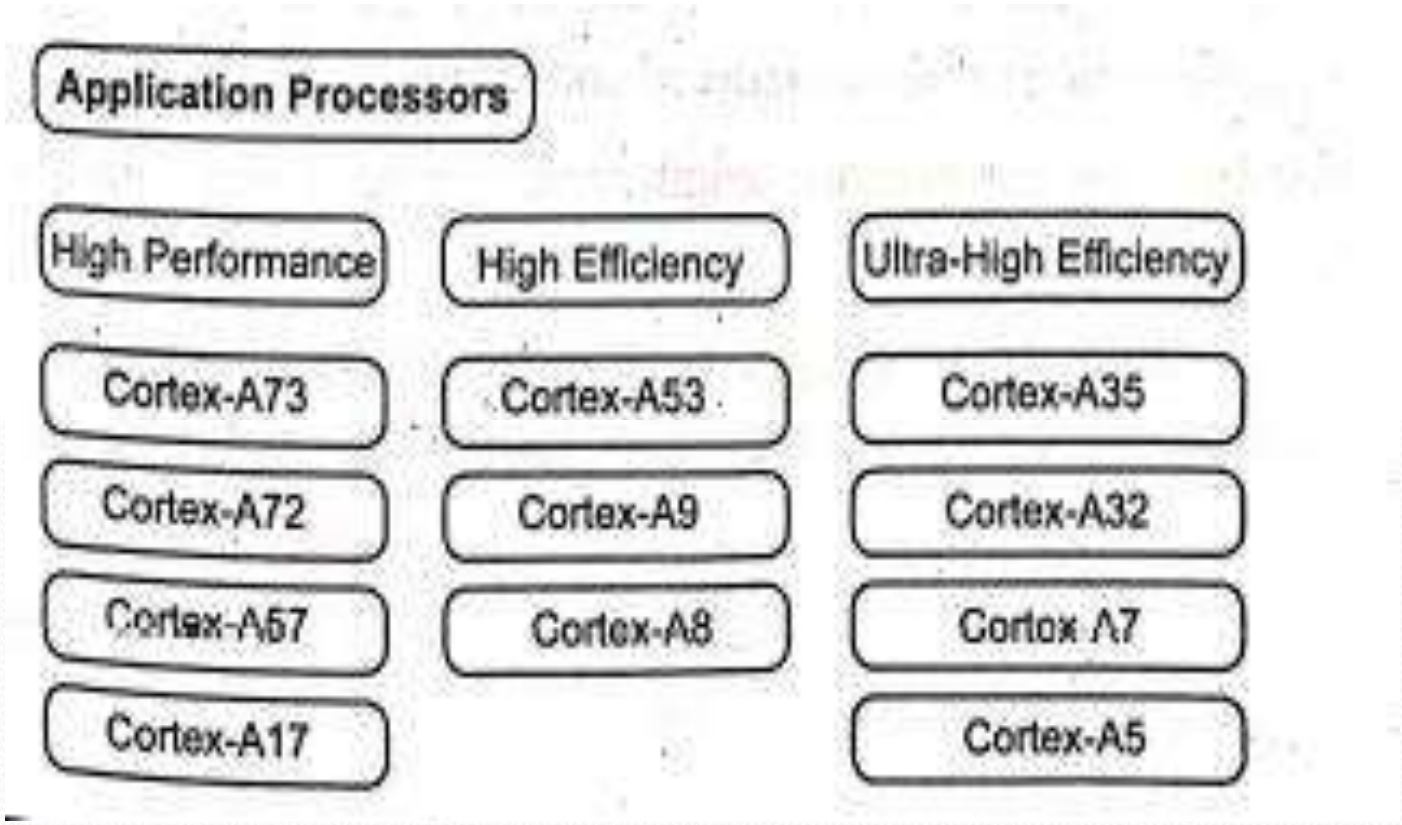
ARM classic processor



ARM Embedded Processor



ARM Application processor

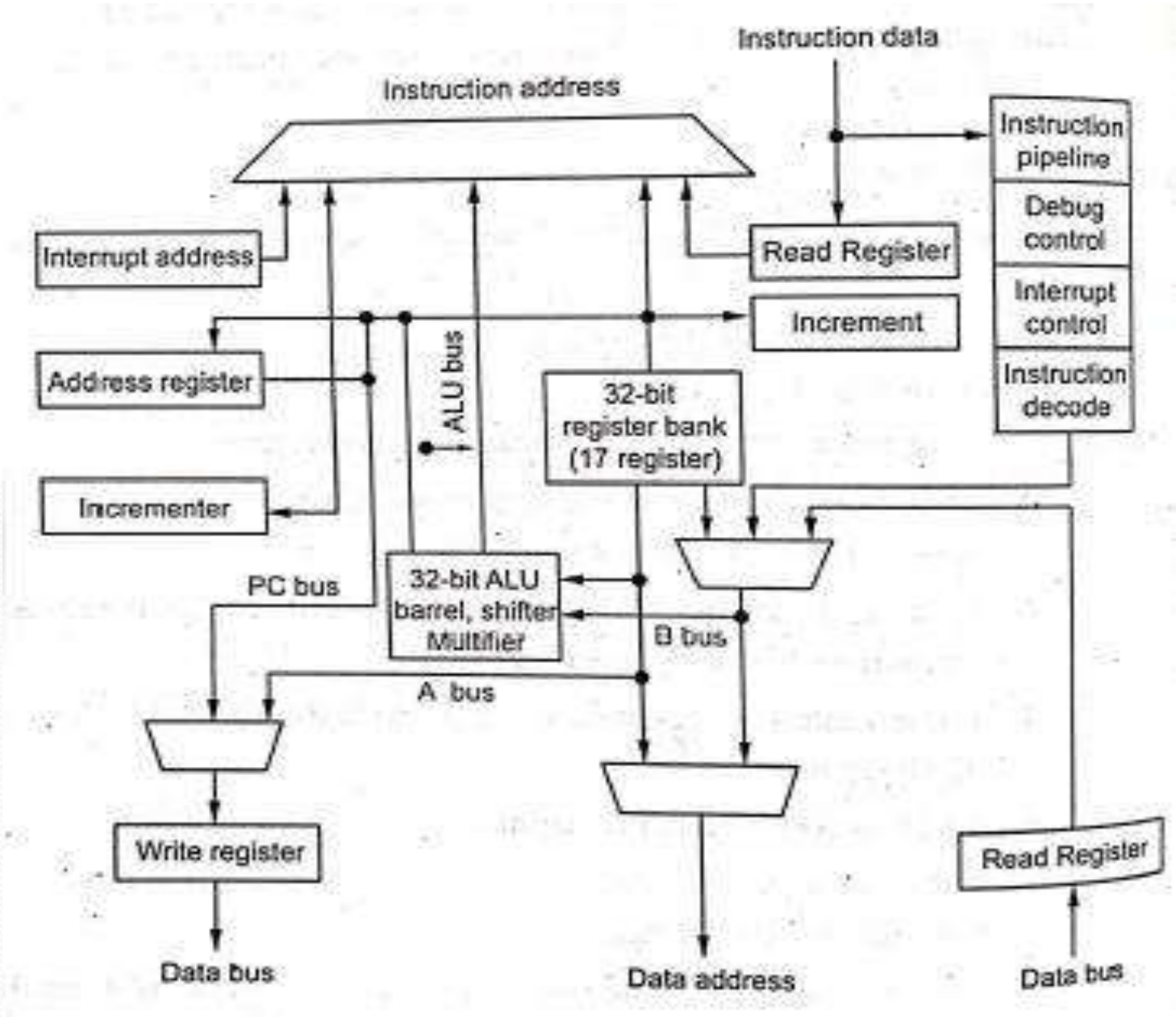


ARM ARCHITECTURE

- The architecture has evolved over time, and starting with cortex series of cores, three profiles are,
- Application Profile Cortex- A series
- Real time profile- Cortex- R series
- Microcontroller profile-Cortex –M series

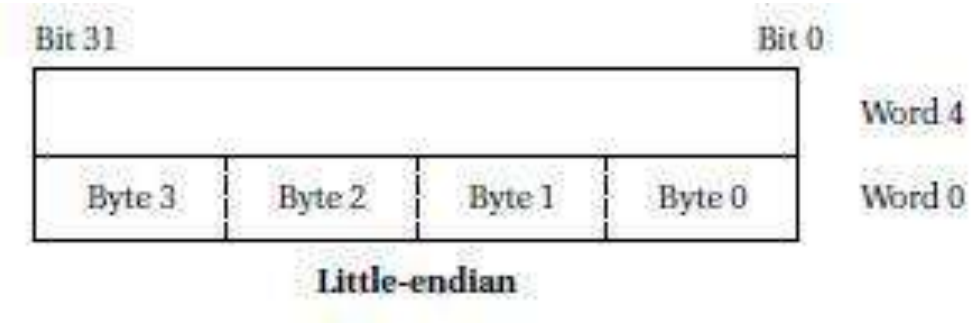
Arm Features

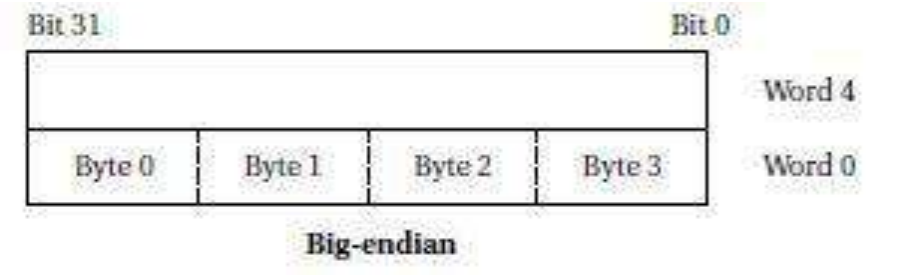
- A load-store architecture,
- Fixed-length 32-bit instructions
- 3-Address instruction formats.


Instruction data
Instruction address
Instruction pipeline
Debug control
Interrupt control
Instruction decode
Interrupt address
Read Register
Increment
Address register
ALU bus
32-bit register bank (17 register)
Incrementer
PC bus
32-bit ALU barrel, shifter Multiplier
B bus
A bus
Write register
Read Register
Data bus
Data address
Data bus

- It has 32 bit architecture but it supports to 16bit and 8 bit data types also
- A wide choice of development tools and simulation models for leading EDA (Electronic Design Automation) environments and excellent debug support
- ARM uses a Intelligent Memory Manager (IEM). It implements advanced algorithms to optimally balance processor workload and power consumption.IEM work with operating system and mobile OS
- ARM uses AHB (AMBA Advanced High performance Bus) interface. AMBA is open source specification for on chip interconnection
-

Byte organizations with an ARM word





ARM ARCHITECTURE

- ARM core is functional units connected by data buses.
- Arrow represents the flow of data.
- Lines represent buses.
- Boxes represent either operation unit or storage area
- Design of ARM is simple and Programmer's design.
- Power Saving design module.
- Flexible design for different application with simple changes
- Instruction Pipeline and Read Data Register are 32 bit

- ARM instructions have two registers:
  - Rm, Rn- source register
  - Rd-destination register.
- Address bus line A(31:0) and data in lines DATA (31:0) to store the data into the register.
- Address Register holds the address of next instruction / data to be fetched
- Address Incrementer the address register value to appropriate amount to point the next instruction/ data
- It contains 31 Register bank, each register are 32 bit registers and also contains 6 status registers each of 32 bits

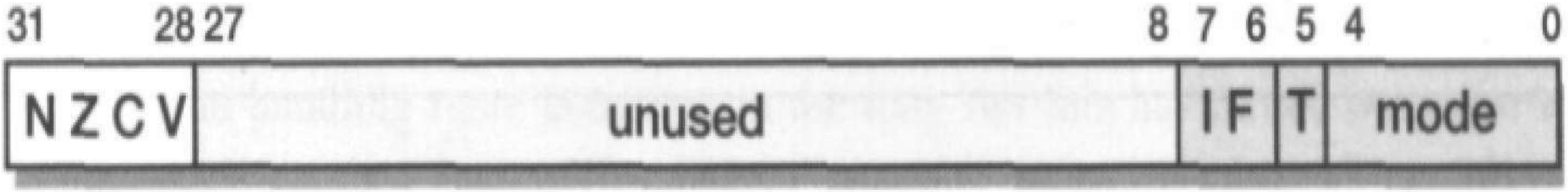
CPU Modes of ARM:

- **User mode:** It is used for programs and applications. It is a only non privileged mode.
- **System Mode:** It is a special version of user mode. It allows the full read write access to the CPSR.
- **Supervisor Mode:** it is privileged mode it enters whenever the processor get reset or SWI instruction is executed. In this mode OS kernel operates in.
- **Abort Mode:** It occurs when there is a failed attempt to access the memory. This mode is entered when prefetch abort and data abort exception occurs.

- **Undefined mode:** it is used when the processor encountered an instruction that is undefined or not supported by the implementation. It is a privileged mode.
- **Interrupt Mode :** It is a privileged mode. When the processor accepts the IRQ it occurs.
- **Fast Interrupt Mode :** It is a privileged mode. When the processor accepts the IRQ it occurs.
- **HYP Mode:** This mode introduced in the ARMV-7A fir cortex- A15 processor to providing hardware virtualization support.

The Current Program Status Register (CPSR)

- It gives the status of ALU result for every execution
- The CPSR is used in user-level programs to store the condition code bits.
- Example, to record the result of a comparison operation and to control whether or not a conditional branch is taken



- **N: Negative**; the last ALU operation which changed the flags produced a **negative result**
- **Z: Zero**; the last ALU operation which changed the flags produced a **zero result** (every bit of the 32-bit result was zero).
- **C: Carry**; the last ALU operation which changed the flags **generated a carry-out**, either as a result of an arithmetic operation in the ALU or from the shifter.
- **V: oVerflow**; the last arithmetic ALU operation which changed the flags **generated an overflow** into the sign bit.

ARM Data Instruction

For arithmetic	ADD	Add
	ADC	Add with carry
	SUB	Subtract
	SBC	Subtract with carry
	RSB	Reverse subtract

	RSC	Reverse subtract with carry
	MUL	Multiply
	MLA	Multiply and accumulate
For logical	AND	Bit wise and
	ORR	Bit wise or
	EOR	Bit wise exclusive or
	BIC	Bit clear
For shift/rotate	LSL	Logical shift left
	LSR	Logical shift right
	ASL	Arithmetic shift left
	ASR	Arithmetic shift right
	ROR	Rotate right
	RRX	Rotate right extended with C
ARM Comparison Instructions	CMP	Compare
	CMN	Negated compare
	TST	Bit wise test
	TEQ	Bit wise negated test
ARM Move Instructions	MOV	Move
	MVN	Move negated
ARM Load Store Instructions and	LDR	Load
	STR	Store
	LDRSH	Load Half Word

Pseudo Operations	STRH	Store Half Word
	LDRSH	Load Half Word Signed
	LDRB	Load byte
	STRB	Store byte
	ADR	Set register to address

- **Example Program:**
- int a,b,c, X;
- X= a+b-c;

```
ADR r4, a     : get address for a
LDR r0, [r4]  : get value of a
ADR r4, b     : get address for b, reusing r4
LDR r1, [r4]  : load value of b
ADD r3, r0, r1 : set intermediate result for X to a + b
ADR , r4, c   : get address for C
LDR ,r2, [r4] : get value of C
SUB r3, r3, r2 : complete computation of X
ADR r4, X     : get address for X
STR r3, [r4]  : store X at proper location
```

ARM INSTRUCTION SET

Types of instruction set

- Data Processing Instructions
- Branch Instructions
- Load Store Instructions
- Software interrupt Instructions
- Program Status Register Instructions

DATA PROCESSING INSTRUCTIONS:

- Move instruction
- Arithmetic instruction
- Logical instruction
- Comparison instruction
- Multiply instruction

Move instruction

- MOV operand2
- MVN NOT operand2
- MOVS – Update In Status Reg

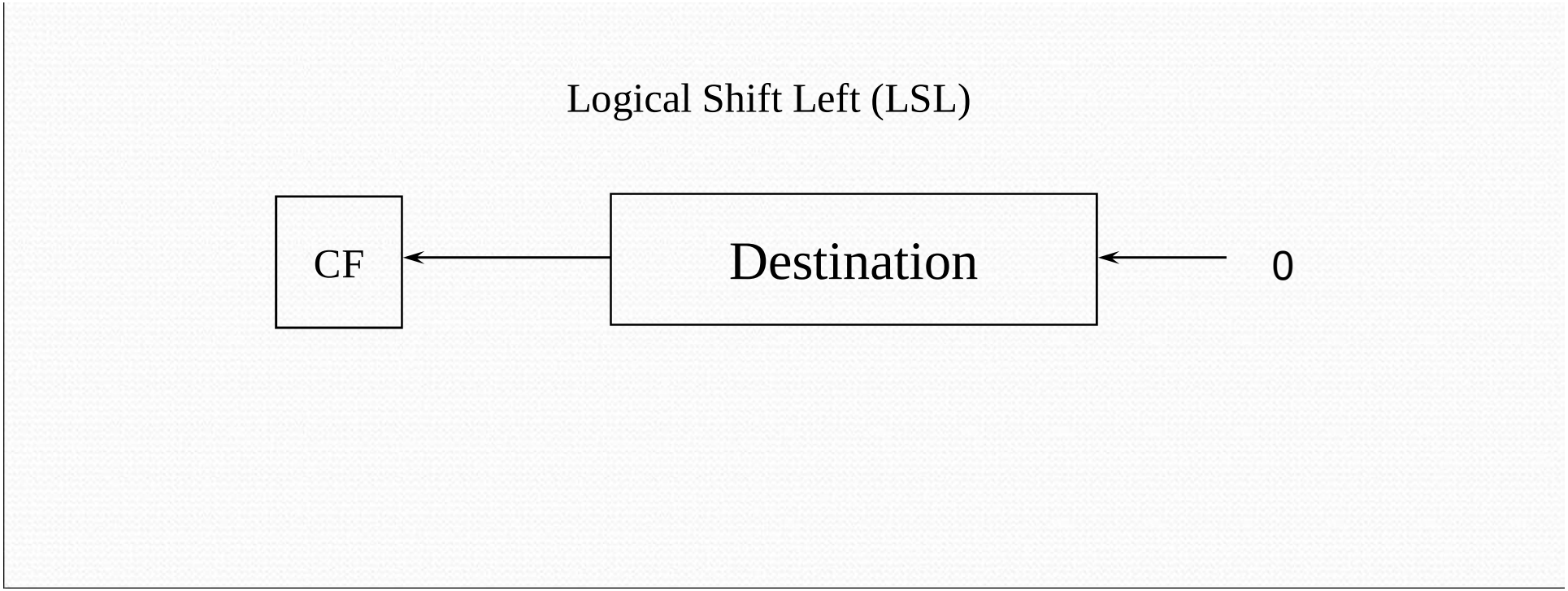
- Syntax:
  - <Operation>{<cond>}{S} Rd, Operand2
- Examples:
  - MOV r0, r1
  - MOVS r2, #10

The Barrel Shifter

- The ARM doesn't have actual shift instructions.
- Instead it has a barrel shifter which provides a mechanism to carry out shifts as part of other instructions.
- **Barrel Shifter - Left Shift**
- Shifts left by the specified amount (multiplies by powers of two)
- e.g.

  LSL #5 = multiply by 32

Barrel Shifter - Left Shift

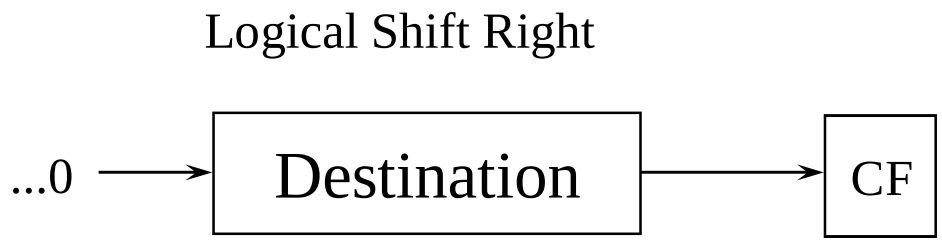


Barrel Shifter - Right Shifts

Logical Shift Right

- Shifts right by the specified amount (divides by powers of two) e.g.

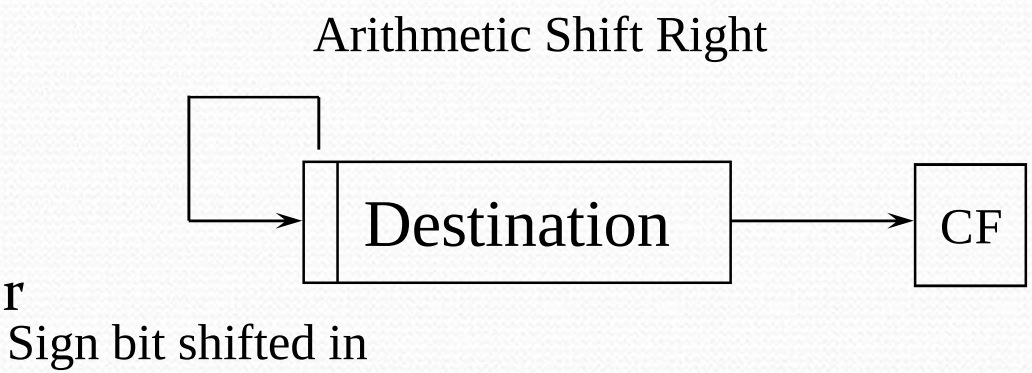
  LSR #5 = divide by 32



Arithmetic Shift Right

- Shifts right (divides by powers of two) and preserves the sign bit, for 2's complement operations. e.g.

  ASR #5 = divide by 32

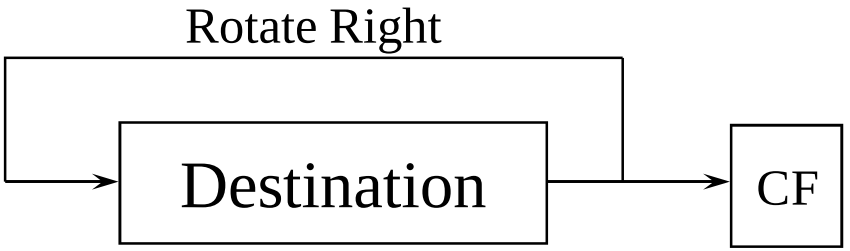


Barrel Shifter - Rotations

Rotate Right (ROR)

- Similar to an ASR but the bits wrap around as they leave the LSB and appear as the MSB.

  e.g. ROR #5

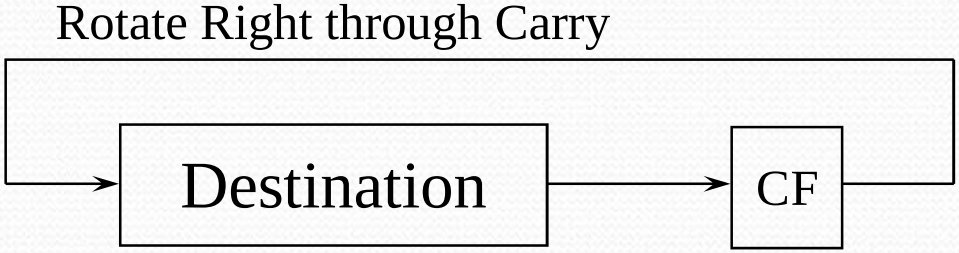


- Note the last bit rotated is also used as the Carry Out.

Rotate Right Extended (RRX)

- This operation uses the CPSR C flag as a 33rd bit.
- Rotates right by 1 bit.

  Encoded as ROR #0.



Arithmetic instruction

ADC	add two 32-bit values and carry	Rd = Rn + N + carry
ADD	add two 32-bit values	Rd = Rn + N
RSB	reverse subtract of two 32-bit values	Rd = N – Rn
RSC	reverse subtract with carry of two 32-bit values	Rd = N – Rn – !(carry flag)
SBC	subtract with carry of two 32-bit values	Rd = Rn – N – !(carry flag)
SUB	subtract two 32-bit values	Rd= Rn – N

example

- ADD r0, r1, r2
  - R0 = R1 + R2
- SUB r5, r3, #10
  - R5 = R3 – 10
- RSB r2, r5, #0xFF00
  - R2 = 0xFF00 – R5

Logical instruction

AND	Logical bit wise AND of two 32-bit values	Rd = Rn & N
ORR	logical bitwise OR of two 32-bit values	Rd = Rn | N
EOR	logical exclusive OR of two 32-bit values	Rd = Rn ^ N
BIC	logical bit clear (AND NOT)	Rd = Rn & ~ N

Comparison instruction

CMN	compare negated	flags set as a result of Rn + N
CMP	compare	flags set as a result of Rn – N
TEQ	test for equality of two 32-bit values	flags set as a result of Rn ^ N
TST	test bits of a 32-bit value	flags set as a result of Rn & N

Multiply instruction

MLA	multiply and accumulate	Rd = (Rm * Rs) + Rn
MUL	multiply	Rd = Rm * Rs
SMLAL	signed multiply accumulate long	[RdHi, RdLo] = [RdHi, RdLo] + (Rm * Rs)
SMULL	signed multiply long	[RdHi, RdLo] = (Rm * Rs)
UMLAL	unsigned multiply accumulate long	[RdHi, RdLo] = [RdHi, RdLo] + (Rm * Rs)
UMULL	unsigned multiply long	[RdHi, RdLo] + (Rm * Rs)

2. Branch Instructions

B	branch	pc = label
BL	branch with link	pc = label lr = address of the next instruction after the BL
BX	branch exchange	pc = Rm & Oxfffffffe, T = Rm & 1
BLX	branch exchange with link	pc = label, T = 1 pc = Rm & Oxfffffffe, T = Rm & 1 lr = address of the enxt instruction after the BLX

3.Load Store Instructions

Single register transfer

LDR	load word into a register	Rd < - mem32[address]
STR	save byte or word from a register	Rd - > mem32[address]
LORB	load byte into a register	Rd < - mem8[address]
STRB	save byte from a register	Rd - > mem8[address]

LDRH	load halfword into a register	Rd < - mem16[address]
STRH	save halfword into a register	Rd - > mem16[address]

LDRSB	load signed byte into a register	Rd < - Sign Extend (mem8[address])
LDRSH	load signed halfword into a register	Rd < - Sign Extend (mem16[address])

Single register load store addressing mode

Index Method	Data	Base address register	Example
Preindex with writeback	mem[base + offset]	base + offset	LDR r0, [r1, #4]!
Preindex	mem[base + offset]	not updated	LDR r0, [r1, #4]
Postindex	mem[base]	base + offset	LDR r0, [r1], #4

Multiple Register Transfer

LDM	load multiple registers	{Rd}*N < - mem 32 [start address + 4*N] optional Rn updated
STM	save multiple registers	{Rd}*N -> mem32[start address + 4*N] optional Rn updated

Swap instruction

SWP	swap a word between memory and a register	tmp = mem32[Rn] mem32[Rn] = Rm Rd = tmp
SWPB	swap a byte between memory and a register	tmp = mem8[Rn] mem8[Rn] = Rm Rd = tmp

Software interrupt Instructions

SWI	software interrupt	lr_svc = address of instruction following the SWI spsr_svc = cpsr pc=vectors + 0 × 8 cpsr mode = SVC cpsrI = 1 (mask IRQ interrupts)

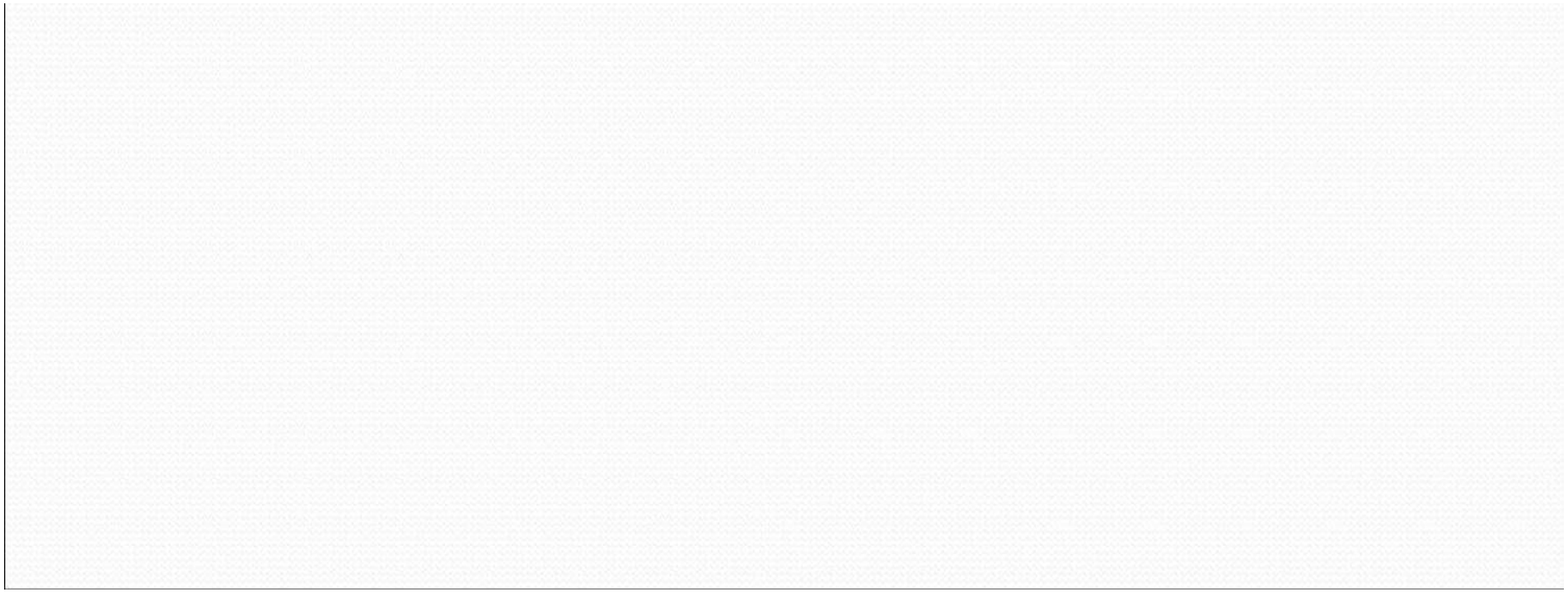
Program Status Register Instructions

MRS	copy program status register to a general-purpose register	Rd= psr
MSR	move a general-purpose register to a program status register	psr[field] = Rm
MSR	move an immediate value to a program status register	psr[field] = immediate

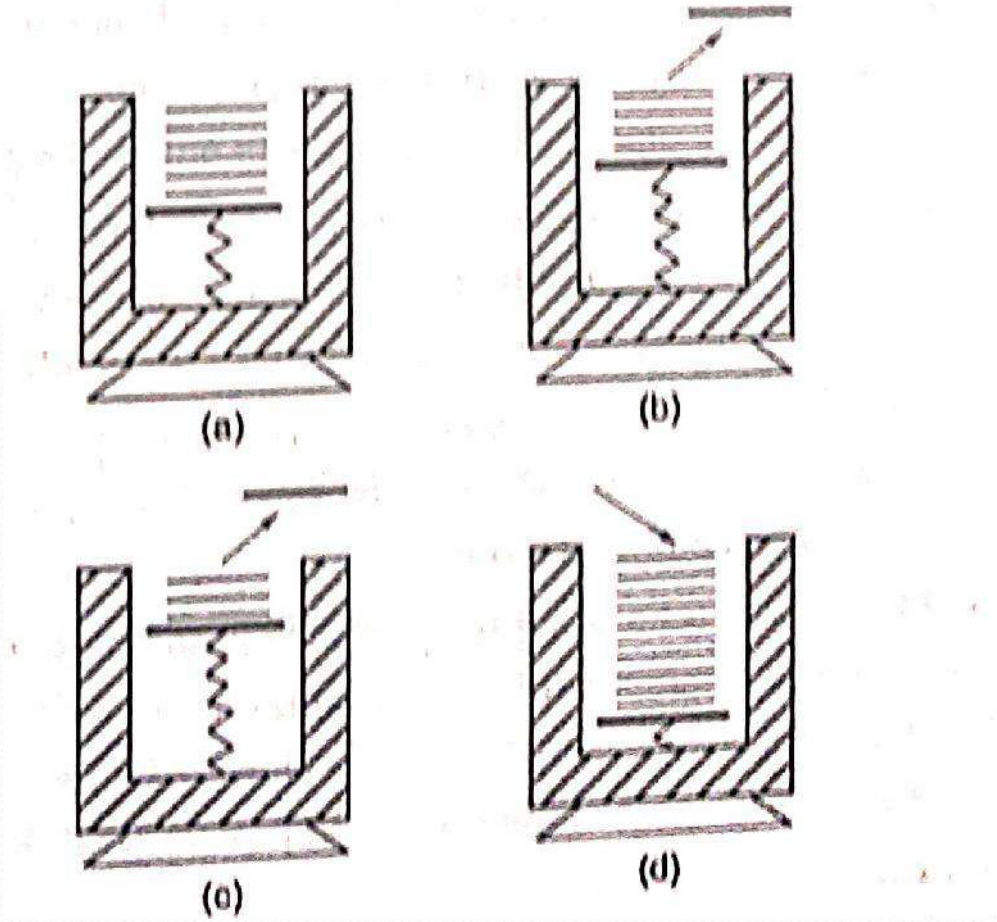
Coprocessor Instruction

CDP	coprocessor data processing – perform an operation in a coprocessor
MRC MCR	coprocessor register transfer – move data to/from coprocessor registers
LDC STC	coprocessor memory transfer – load and store blocks of memory to/from a coprocessor

Stack and subroutine

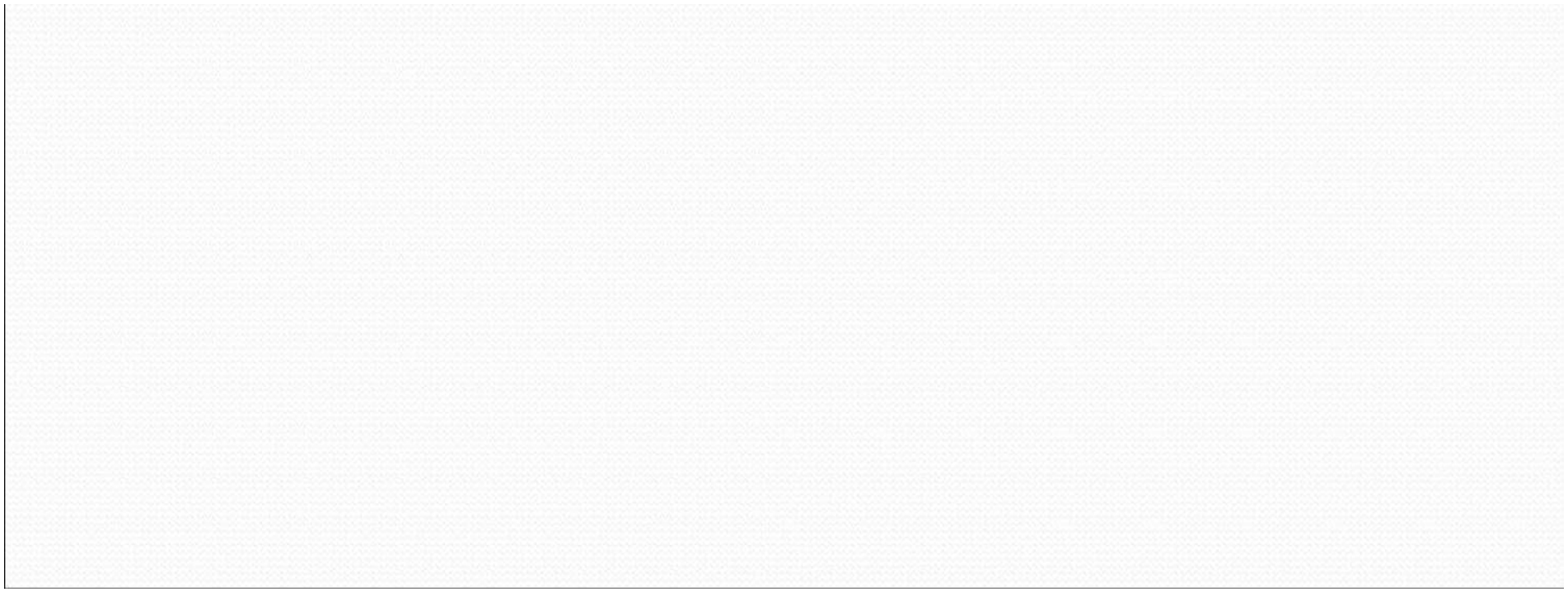


Stack and subroutine



```
        BL      subrout         ; Main program
        .
        .
subrout PUSH    {R0}            ; subroutine
        PUSH    {R1}
        PUSH    {R2}
        PUSH    {R3}
        .
        .   SUBROUTINE INSTRUCTIONS GO HERE
        .
        .
        .
        POP     {R3}
        POP     {R2}
        POP     {R1}
        POP     {R0}
```

- Calling A subroutine
- Parameter passing
- Software delay



2.6 Features of the LPC 214x family

- The LPC2148 is a 16 bit or 32 bit ARM7 family based microcontroller and available in a small LQFP64 package.
- ISP (in system programming) or IAP (in application programming) using on-chip boot loader software.
- On-chip static RAM is 8 kB-40 kB, on-chip flash memory is 32 kB-512 kB, the wide interface is 128 bit, or accelerator allows 60 MHz high-speed operation.
- It takes 400 milliseconds time for erasing the data in full chip and 1 millisecond time for 256 bytes of programming.

- Embedded Trace interfaces and Embedded ICE RT offers real-time debugging with high-speed tracing of instruction execution and on-chip Real Monitor software.
- It has 2 kB of endpoint RAM and USB 2.0 full speed device controller. Furthermore, this microcontroller offers 8kB on-chip RAM nearby to USB with DMA.
- One or two 10-bit ADCs offer 6 or 14 analogs i/ps with low conversion time as 2.44 µs/ channel.
- Only 10 bit DAC offers changeable analog o/p.
- External event counter/32 bit timers-2, PWM unit, & watchdog.
- Low power RTC (real time clock) & 32 kHz clock input.

- Several serial interfaces like two 16C550 UARTs, two I2C-buses with 400 kbit/s speed.5 volts tolerant quick general purpose Input/output pins in a small LQFP64 package.
- Outside interrupt pins-21.60 MHz of utmost CPU CLK-clock obtainable from the programmable-on-chip phase locked loop by resolving time is 100 µs.
- The incorporated oscillator on the chip will work by an exterior crystal that ranges from 1 MHz-25 MHz
- The modes for power-conserving mainly comprise idle & power down.
- For extra power optimization, there are individual enable or disable of peripheral functions and peripheral CLK scaling.

- **2.7 PERIPHERALS:**

- Embedded systems that interacts with the outside world, needs some peripheral device. A peripheral device performs input and output functions for the chip by connecting to other devices or sensors that are off chip.
- Each peripheral device performs one function from outside of chip. Peripheral range is from simple serial communication to complex 802.11 wireless devices.
- All ARM peripherals are memory mapped. It has set of addressed registers. This address registers used to select the exact peripheral device address

Controllers-Specialized peripherals for higher level functionality. Its two types are,

- Memory controllers.
- Interrupt controllers.

Memory controllers:

- Connect different types of memory to the processor bus.
- On- power-up a memory controller is configured in hardware to allow the certain memory devices to be active.
- Some memory devices must be set up by software.

Interrupt controllers:

- When a peripheral device requires a attention it raises the interrupt to the processor.
- The interrupt controller provides the programmable governing policy that allows the software to determine which peripheral device can interrupt the processor at specific time. This is done by bits in the interrupt controller register.

Two types of interrupt controllers for ARM:

- The Standard interrupt controller.
- The Vector interrupt controller (VIC).

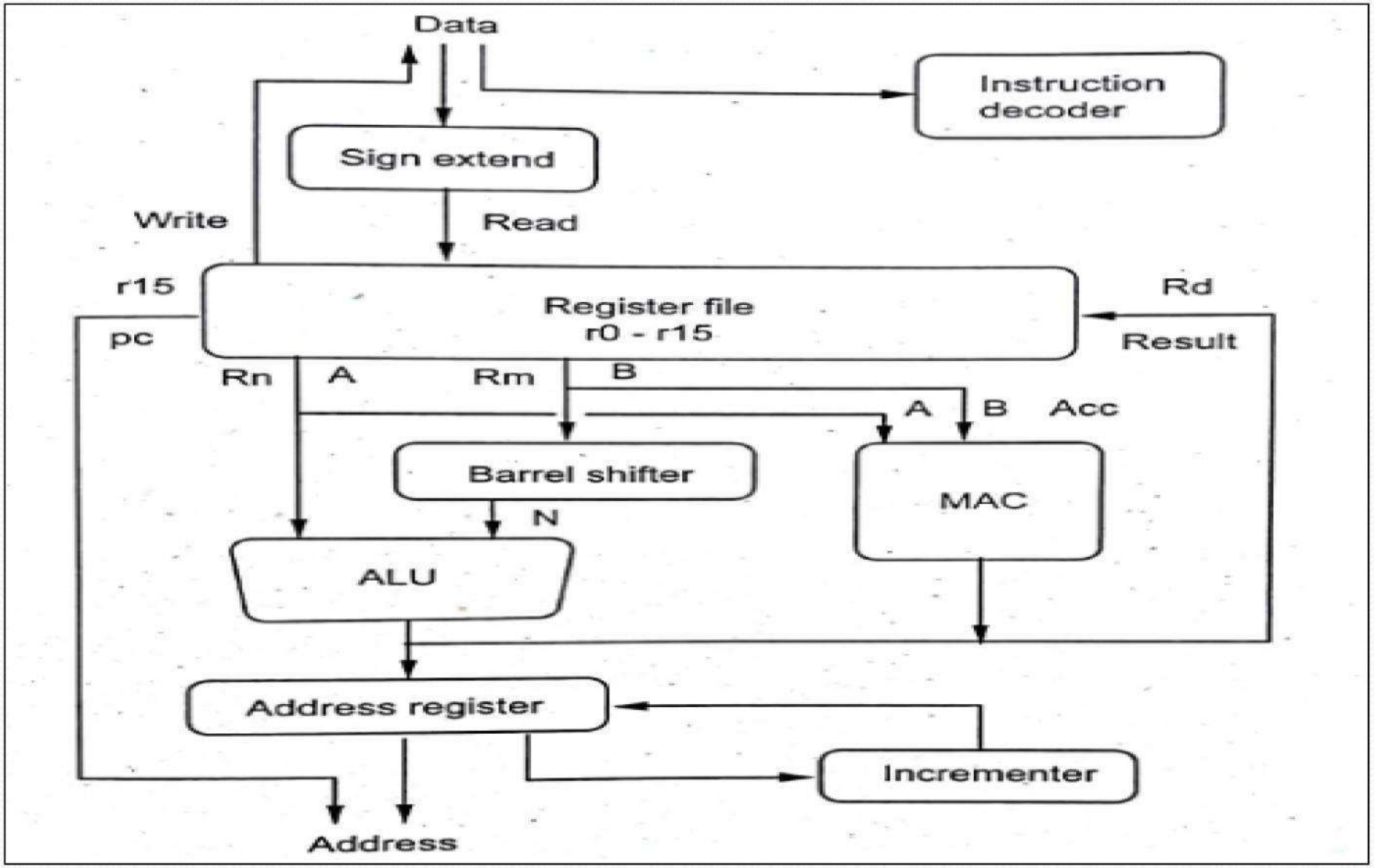
The Standard interrupt controller:

- It sends the interrupt signal to the processor core, when an external device requests servicing.
- It can be programmed to ignore or mask other individual device or set of devices.
- The interrupt handler determines which device requires to servicing by reading a device bitmap register in the interrupt controller.

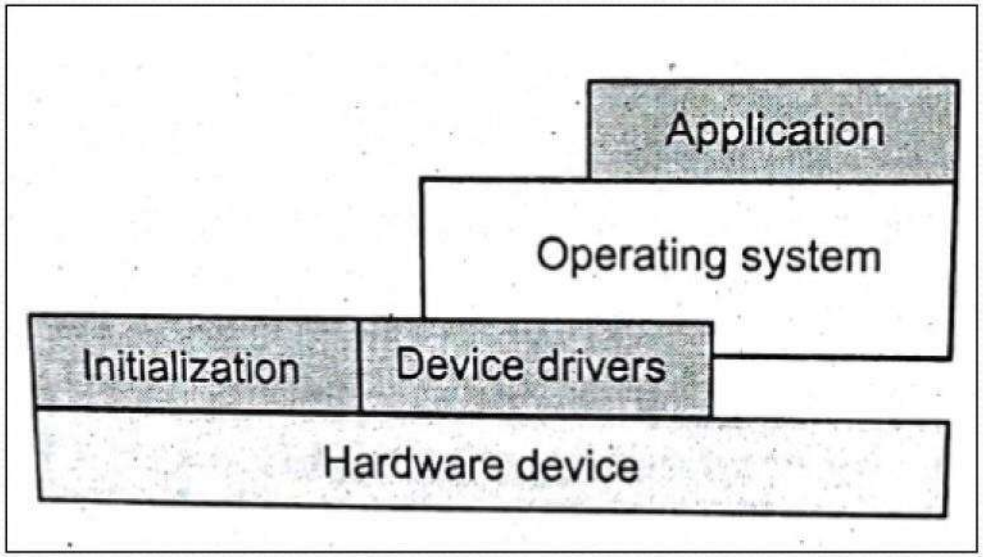
The Vector interrupt controller (VIC):

- It is powerful than Standard interrupt controller. It has prioritizes interrupts. So determination of which device caused the interrupt is simple.
- The VIC only allows an interrupt signal to the core if the new higher priority came than currently executing interrupt.

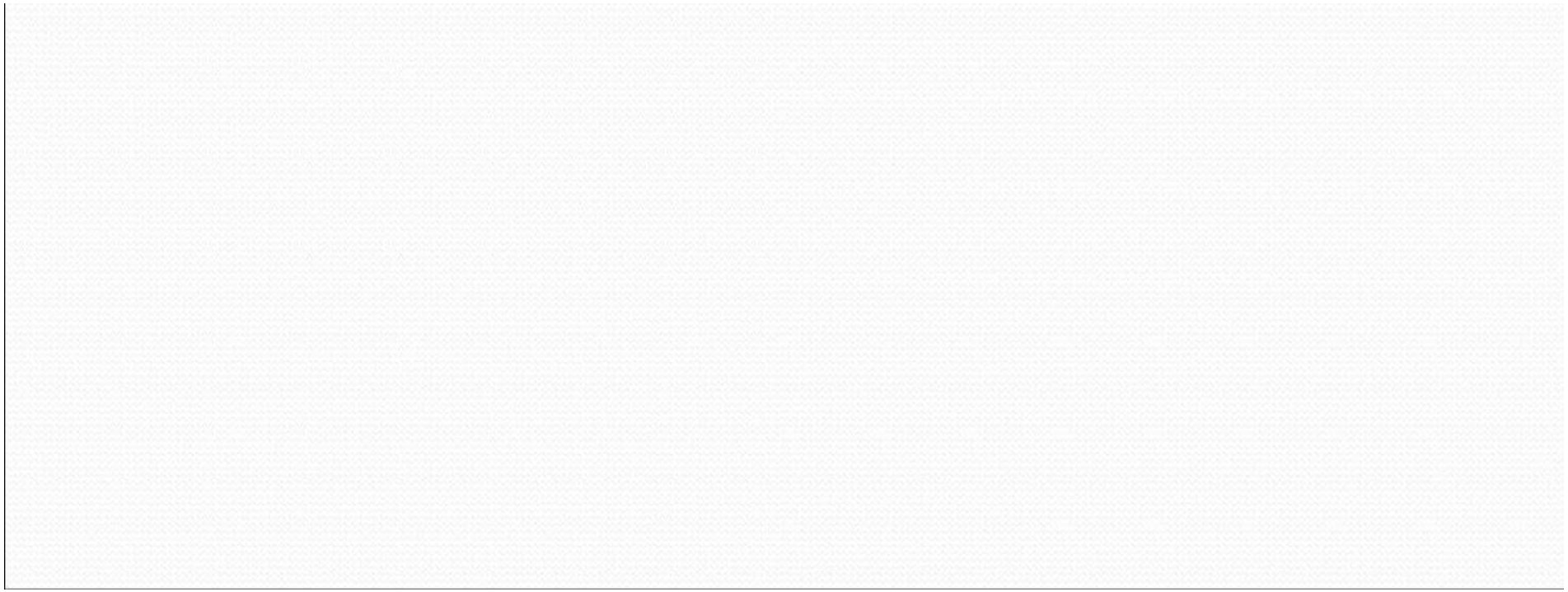
The ARM core data flow model:



Software abstraction layers executing on hardware



The Timer Unit

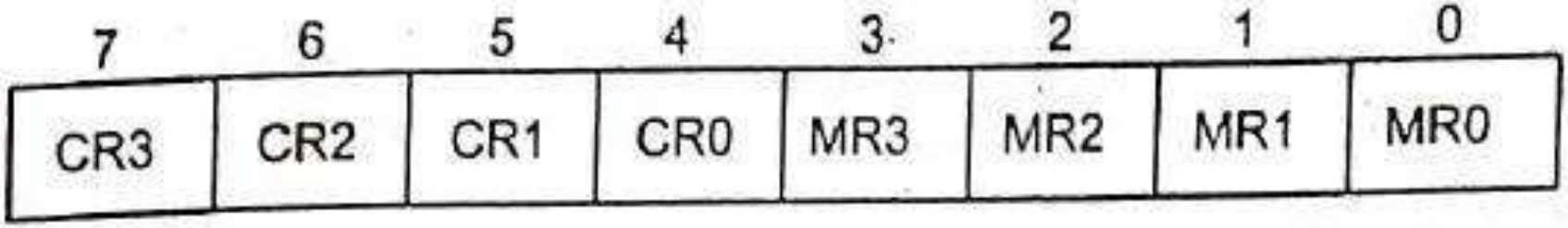


Register Associated with timer in LPC2148

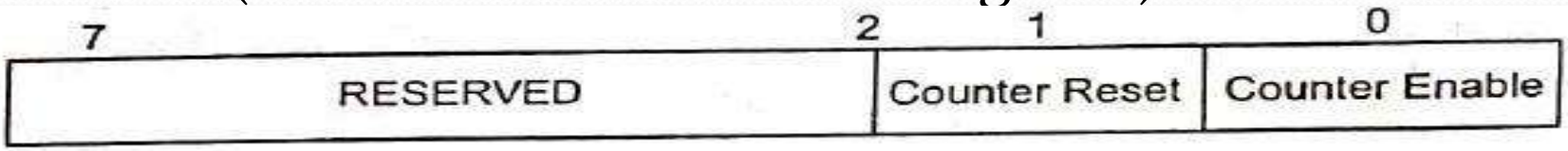
- Prescale register (PR)
- Prescaler Counter register (PC)
- Timer counter register(TC)
- Ttimer control register(TCR)
- Conter control register(CTCR)
- Match control Register (MCR)
- Interrupt Register(IR)

- Timer 0 register

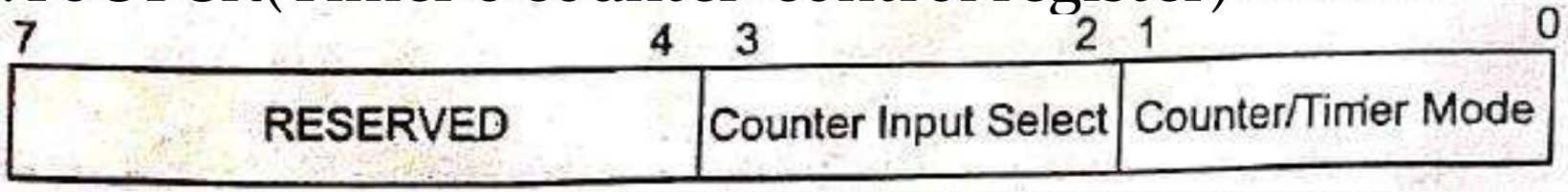
1. T0IR(Timer 0 interrupt Register)



2.T0TCR(Timer 0 Timer Control Register)



3.T0CTCR(Timer 0 counter control register)



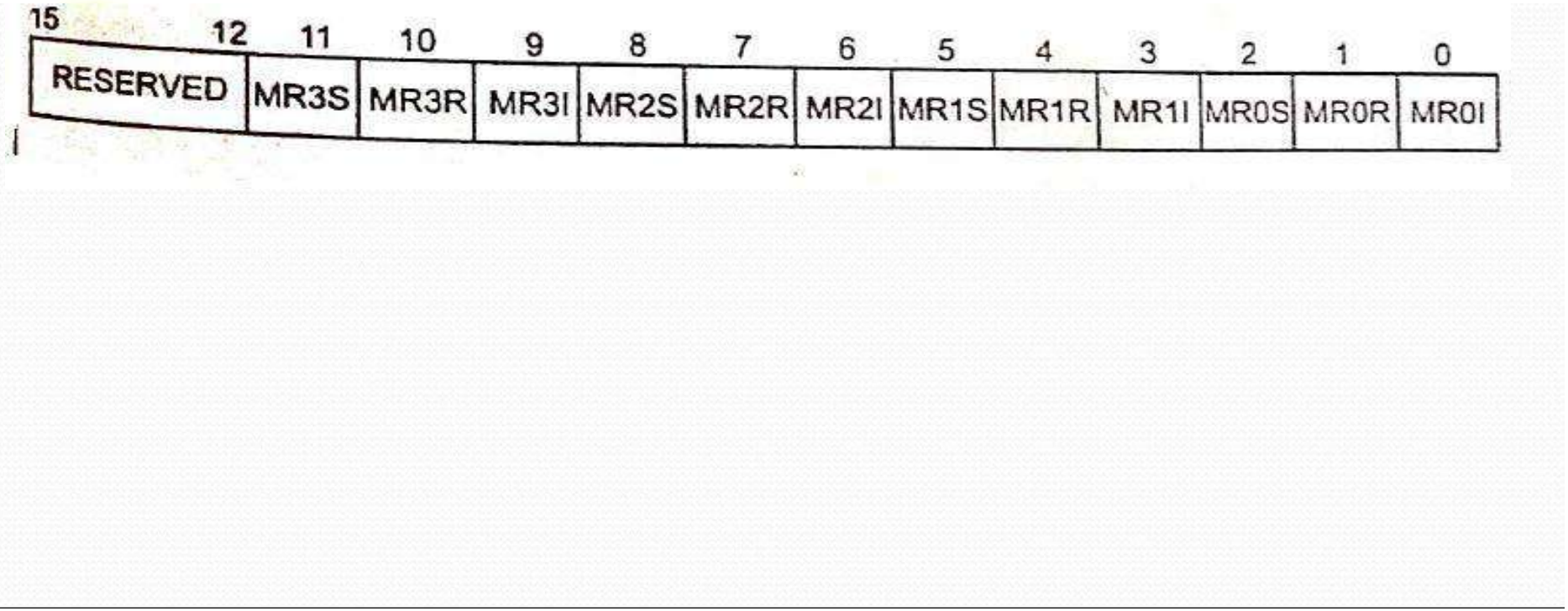
4. ToTC(Timer 0 Timer Counter)

5.ToPR(Timer 0 Prescale Register)

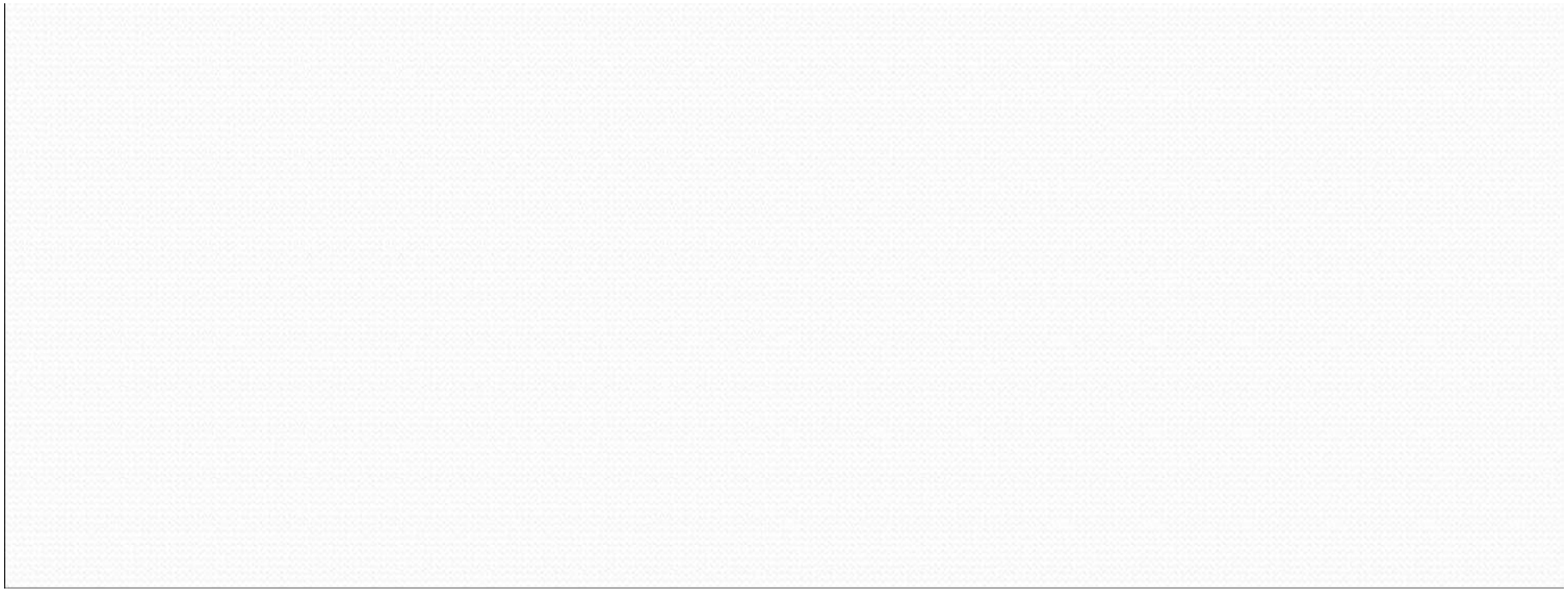
6.ToPC(Timer 0 prescale counter register)

7.ToMR0-ToMR3(Timero Match Register)

8.ToMCR(Timero Match Control Register)

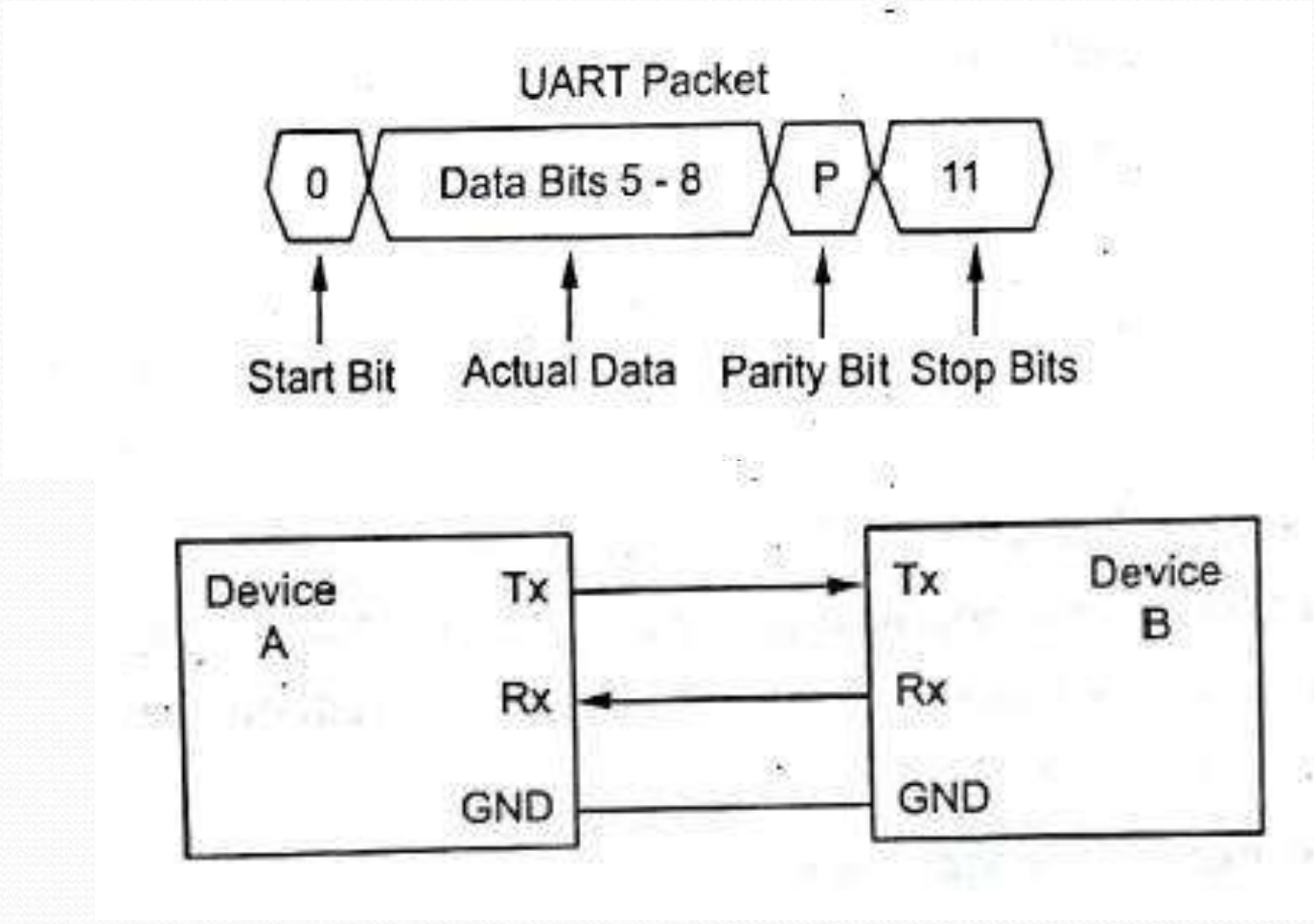


UART



UART

- Universal Asynchronous Receiver/Transmitter



- UART in LPC2148 ARM 7 Micro controller

UART0		UART1	
TXD0	P0.0	TXD1	P0.8
RXD0	P0.1	RXD1	P0.9

Register Associated with UART in LPC2148

- UART0 Receiver Buffer Register(U0RBR)
- UART0 Transmit Holding Register(U0THR)
- UART0 Divisor Latch Register (U0DLL and U0DLM)

  Determine the baud rate generator (U0DLL / U0DLM). (0x00:00x01)
- UART0 Fractional divider register (U0FDR)
  - It is used for prescale for the baud rate
  - Both Multiply and Division can be done in prescale
  - Bit 0 – 3 used for prescale divisor value for baud rate
  - Bit 4 -7 used multiplier value

- UART0 Interrupt Enable Register(U0IER)
  - 0 bit- RBR (Receiver buffer Register)interrupt
  - 1 bit- Interrupt enable register
  - 2 bit- Rx line status register
  - 8 bit – End of auto baud rate interrupt
  - 9 bit- auto baud time out interrupt
- **U0LCR (UART0 Line Control Register)**
- **Bit 1:0 - Word Length Select**

  00 = 5-bit character length
  01 = 6-bit character length
  10 = 7-bit character length
  11 = 8-bit character length

- **Bit 2 - Number of Stop Bits**
  0 = 1 stop bit
  1 = 2 stop bits
- **Bit 3 - Parity Enable**
  0 = Disable parity generation and checking
  1 = Enable parity generation and checking
- **00Bit 5:4 - Parity Select**
  00 = Odd Parity
  01 = Even Parity
  10 = Forced "1" Stick Parity
  11 = Forced "0" Stick Parity
- **Bit 6 - Break Control**
  0= Disable break transmission
  1 = Enable break transmission
- **Bit 7 - Divisor Latch Access Bit (DLAB)**
  0 = Disable access to Divisor Latches
  1 = Enable access to Divisor Latches

U0LSR (UART0 Line Status Register)

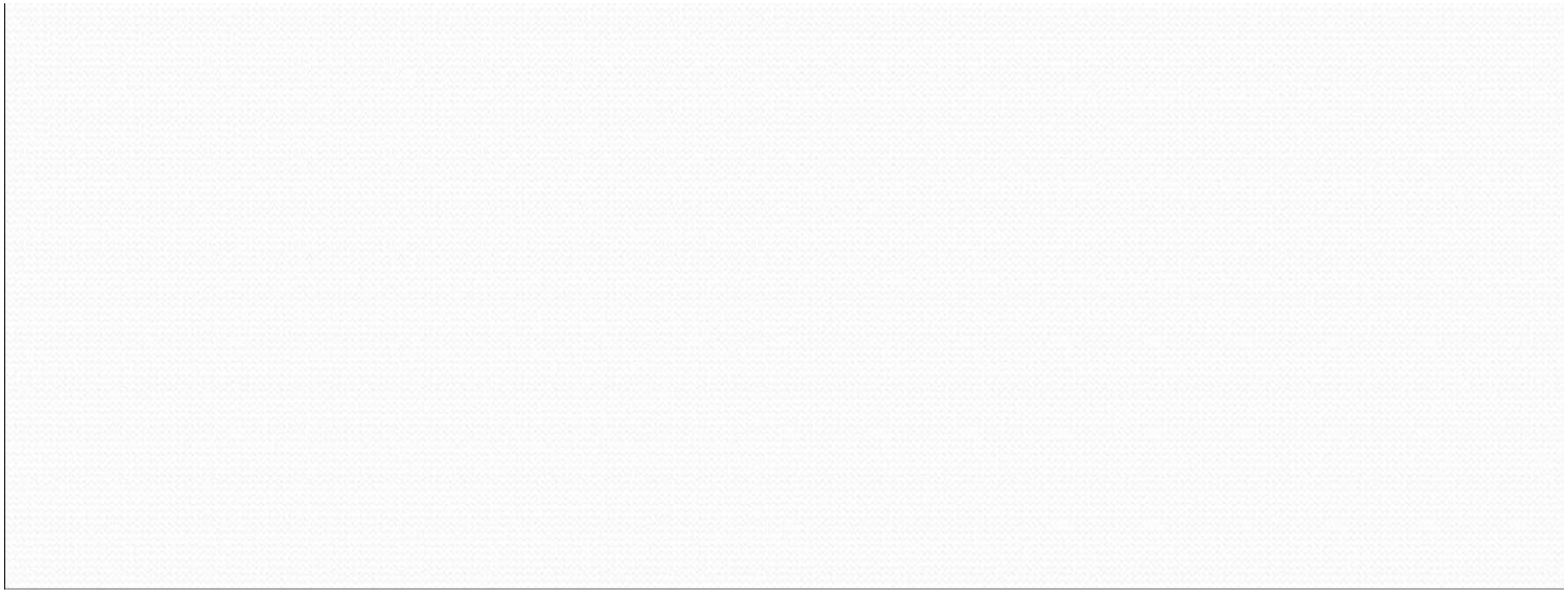
- It provides status information on UART0 RX and TX blocks.
- **Bit 0 - Receiver Data Ready**
  0 = U0RBR is empty
  1 = U0RBR contains valid data
- **Bit 1 - Overrun Error**
  0 = Overrun error status inactive
  1 = Overrun error status active
  This bit is cleared when U0LSR is read.
- **Bit 2 - Parity Error**
  0 = Parity error status inactive
  1 = Parity error status active
  This bit is cleared when U0LSR is read.

- **Bit 3 - Framing Error**
  0 = Framing error status inactive
  1 = Framing error status active
  This bit is cleared when U0LSR is read.
- **Bit 4 - Break Interrupt**
  0 = Break interrupt status inactive
  1 = Break interrupt status active
  This bit is cleared when U0LSR is read.
- **Bit 5 - Transmitter Holding Register Empty**
  0 = U0THR has valid data
  1 = U0THR empty
- **Bit 6 - Transmitter Empty**
  0 = U0THR and/or U0TSR contains valid data
  1 = U0THR and U0TSR empty
- **Bit 7 - Error in RX FIFO (RXFE)**
  0 = U0RBR contains no UART0 RX errors
  1 = U0RBR contains at least one UART0 RX error
  This bit is cleared when U0LSR is read

U0TER (UART0 Transmit Enable Register)

- The U0TER enables implementation of software flow control. When TXEn=1, UART0 transmitter will keep sending data as long as they are available. As soon as TXEn becomes 0, UART0 transmission will stop.
- Software implementing software-handshaking can clear this bit when it receives an XOFF character (DC3). Software can set this bit again when it receives an XON (DC1) character.
- **Bit 7 : TXEN**
  0 = Transmission disabled
  1 = Transmission enabled
- If this bit is cleared to 0 while a character is being sent, the transmission of that character is completed, but no further characters are sent until this bit is set again

Block Diagram of ARM9



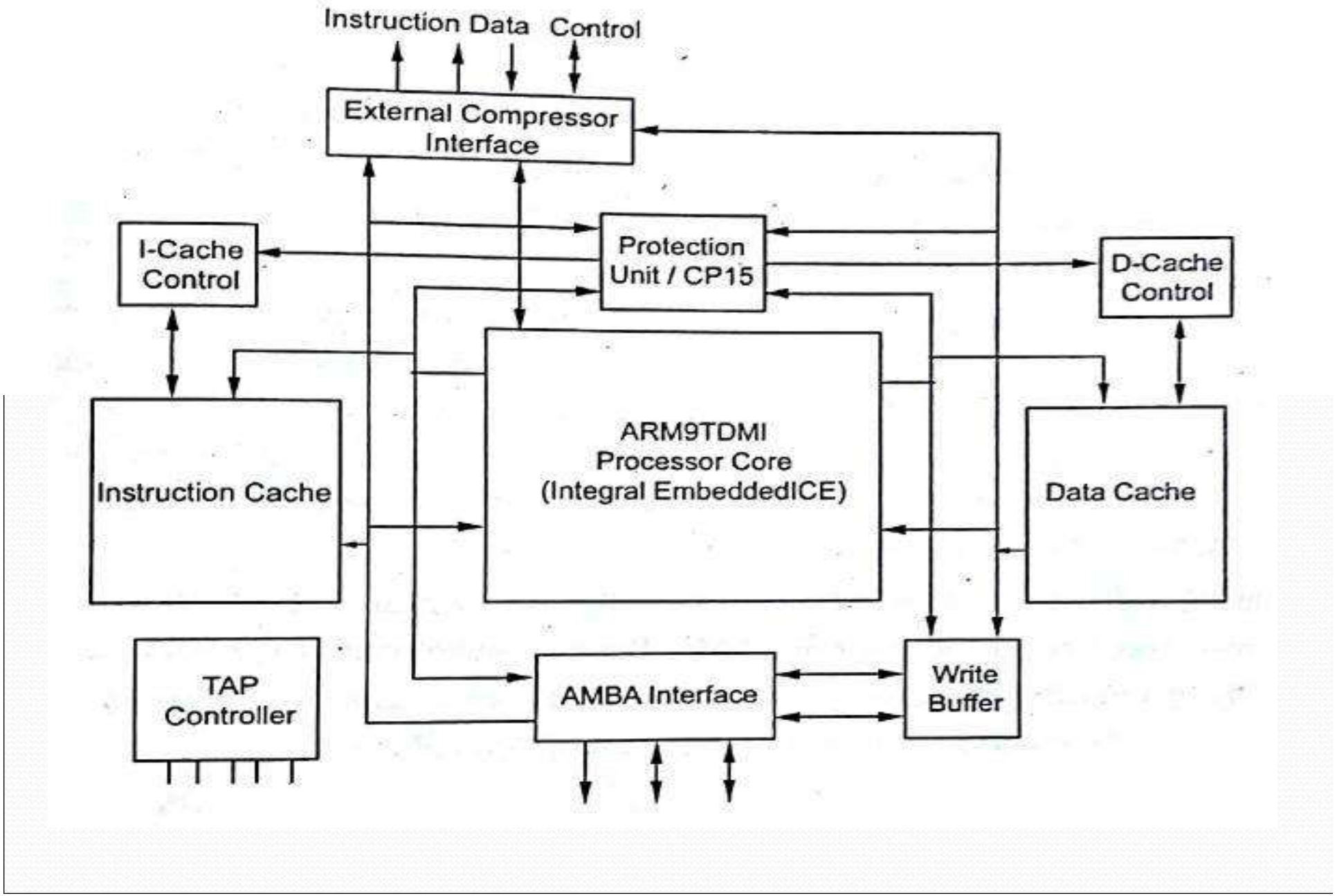
2.12.1. FEATURES OF ARM9

1. **Pipeline Depth:** 5 stage (Fetch, Decode, Execute, Decode, Write)
2. **Operating frequency:** 150 MHz
3. **Power Consumption:** 0.19 mW/MHz
4. **MIPS/MHz:** 1.1
5. **Architecture used:** Harvard
6. **MMU/MPU:** Present
7. **Cache Memory:** Present (separate 16k/8k)
8. **ARM/ Thumb Instruction:** Support both
9. **ISA (Instruction Set Architecture):** V5T(ARM926EJ-S)
10. 31 (32-Bit size) Registers
11. 32-bit ALU & Barrel Shifter
12. Enhanced 32- bit MAC block
13. Memory Controller

Memory operations are controlled by MMU or MPU

1. **MMU:**
   - Provides Virtual Memory Support
   - Fast Context Switching Extensions
2. **MPU:**
   - Enables memory protection & bounding
   - Sand – boxing of applications

14. Flexible Cache Design (sizes can be 4KB to 128KB)
15. Flexible Core Design
16. DSP Enhancements: (very important)
17. Single cycle 32 ×16 multiplier Implementation
18. Speed up all the multiply instructions
19. New 32 ×16 & 16 ×16 multiply instructions
20. Allows independent access to 16 bit halves of registers


Instruction Data Control
External Compressor Interface
I-Cache Control
Protection Unit / CP15
D-Cache Control
Instruction Cache
ARM9TDMI Processor Core (Integral EmbeddedICE)
Data Cache
TAP Controller
AMBA Interface
Write Buffer

ARM9TDMI

- ARM920T with 16 KB each of I/D cache and an MMU
- ARM922T with 8 KB each of I/D cache and an MMU
- ARM940T with cache and a Memory Protection Unit (MPU)

ARM940T Cached Processor

- Firstly, it allows the processor to operate at its maximum frequency since memory accesses are to the local, high performance cache.
- Secondly, since main memory is accessed infrequently, system power is reduced. Also, the main memory system may now be used for other tasks, such as DMA, while the processor is executing from its caches.

Comparision between ARM9TDMI and ARM7TDMI

Instruction	% taken	% Skipped	ARM7TDMI	ARM9TDMI
Data processing	49	4	1	1
Data processing with PC	3	0	3	3
Branch/Branch with link	11	4	3	3
Load register	14	1	3	1-2
Store register	8	1	2	1
Load multiple registers	1	0	7	5
Store multiple registers	2	0	7	6
CPI			1.9	1.5

Pipeline Process

...maximum operating frequency.

ARM7TDMI Pipeline Operation

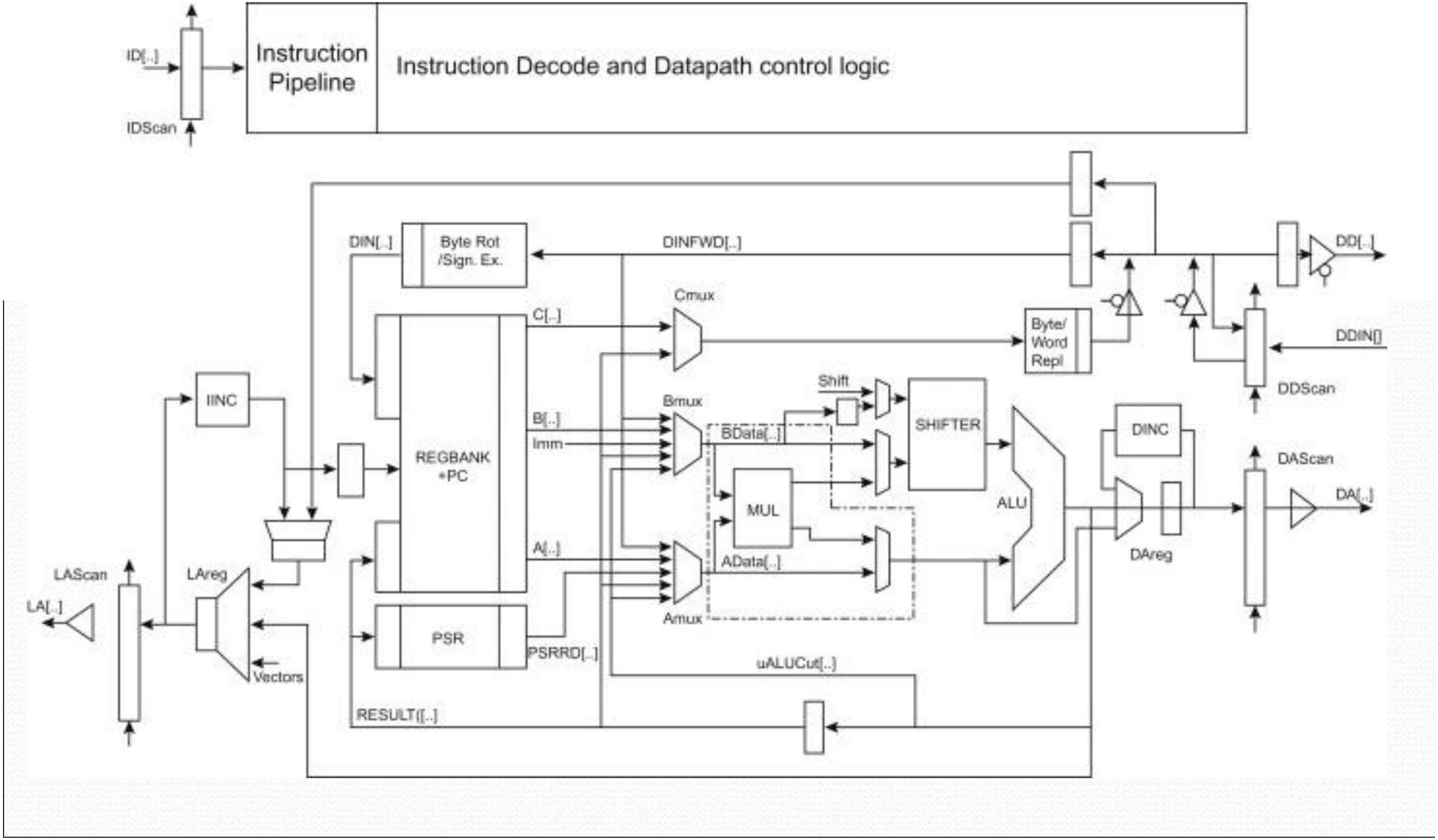
Fetch	Decode	Execute
Instruction Fetch	Convert Thumb to ARM Main Decode Register Address Decode	Register Read Shifter ALU Writeback

ARM9TDMI Pipeline Operation

Fetch	Decode	Execute	Memory	Writeback
Instruction Fetch	Reg. Address Decode Register Read Thumb Decode Reg. Address Decode Register Read	Shifter ALU	Memory Data access	ALU Result and / or Load data Writeback

Fig. 2.30 ARM7TDMI...

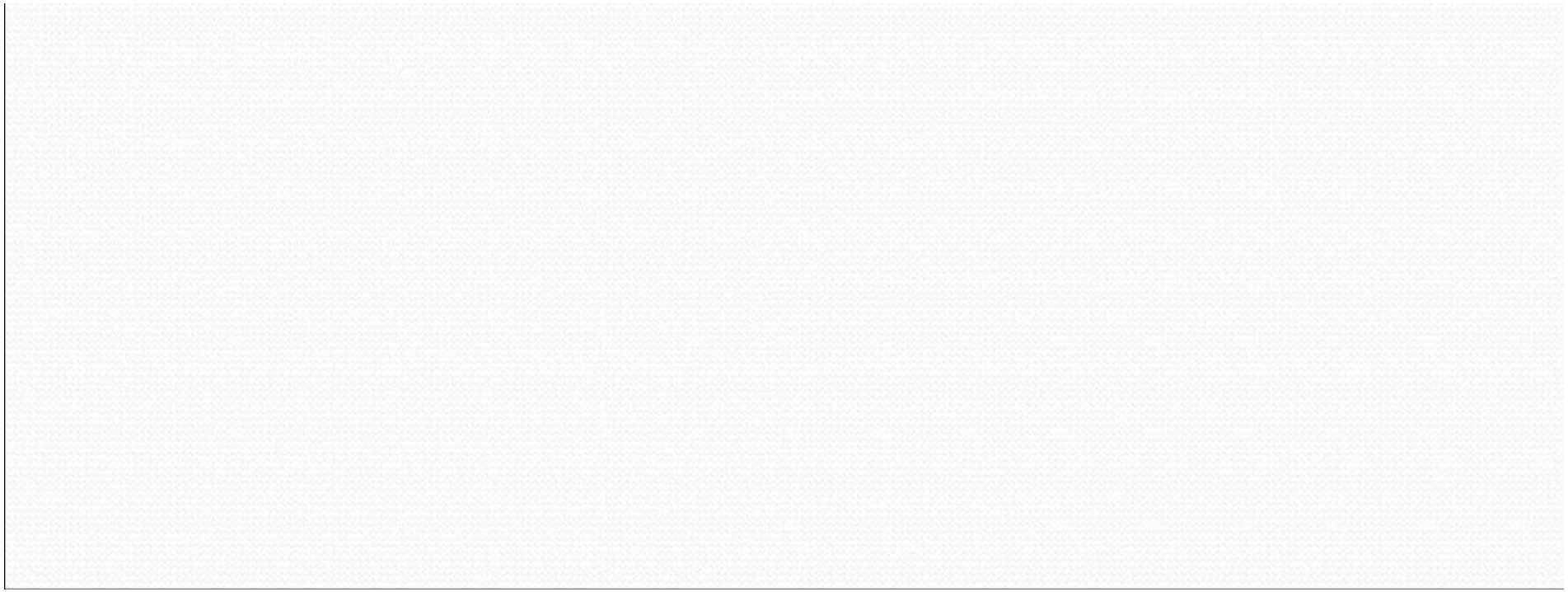
DATA FLOW


ID[..]
IDScan
Instruction Pipeline
Instruction Decode and Datapath control logic
DIN[..]
Byte Rot /Sign. Ex.
DINFWD[..]
DD[..]
Cmux
C[..]
Byte/ Word Repl
DDIN[]
DDScan
IINC
Bmux
Shift
B[..]
Imm
BData[..]
SHIFTER
DINC
DAScan
REGBANK *PC
MUL
ALU
DA[..]
A[..]
AData[..]
DAreg
LAScan
LAreg
LA[..]
Amux
PSR
PSRRD[..]
uALUCut[..]
Vectors
RESULT[..]

COMPARISION SUMMARY

	ARM7TDMI	ARM9TDMI
Area (mm^2,0.35 μm)	2.2	4.15
Transistor count	74k	112k
Pipeline stages	3	5
CPI	1.9	1.5
MIPS/MHz	0.9	1.1
Typical Max Clock rate (0.35 μm)	60	120
Power (mW/MHz @ 3.0V)	1.5	1.8

2.9 Pulse Width Modulation(PWM)



$$\text{Duty Cycle (In \%)} = \frac{T_{on}}{T_{on} + T_{off}} \times 100$$

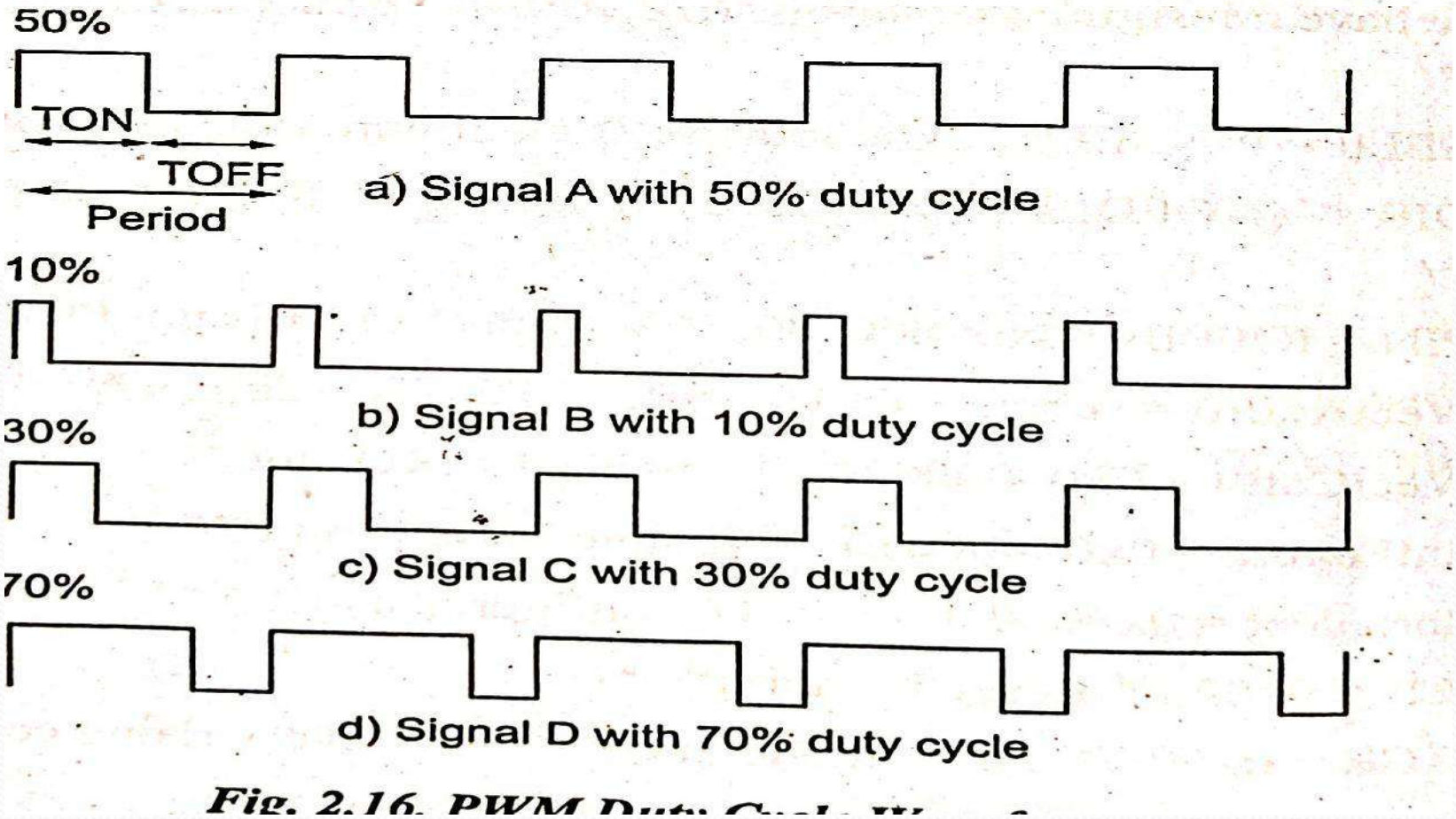
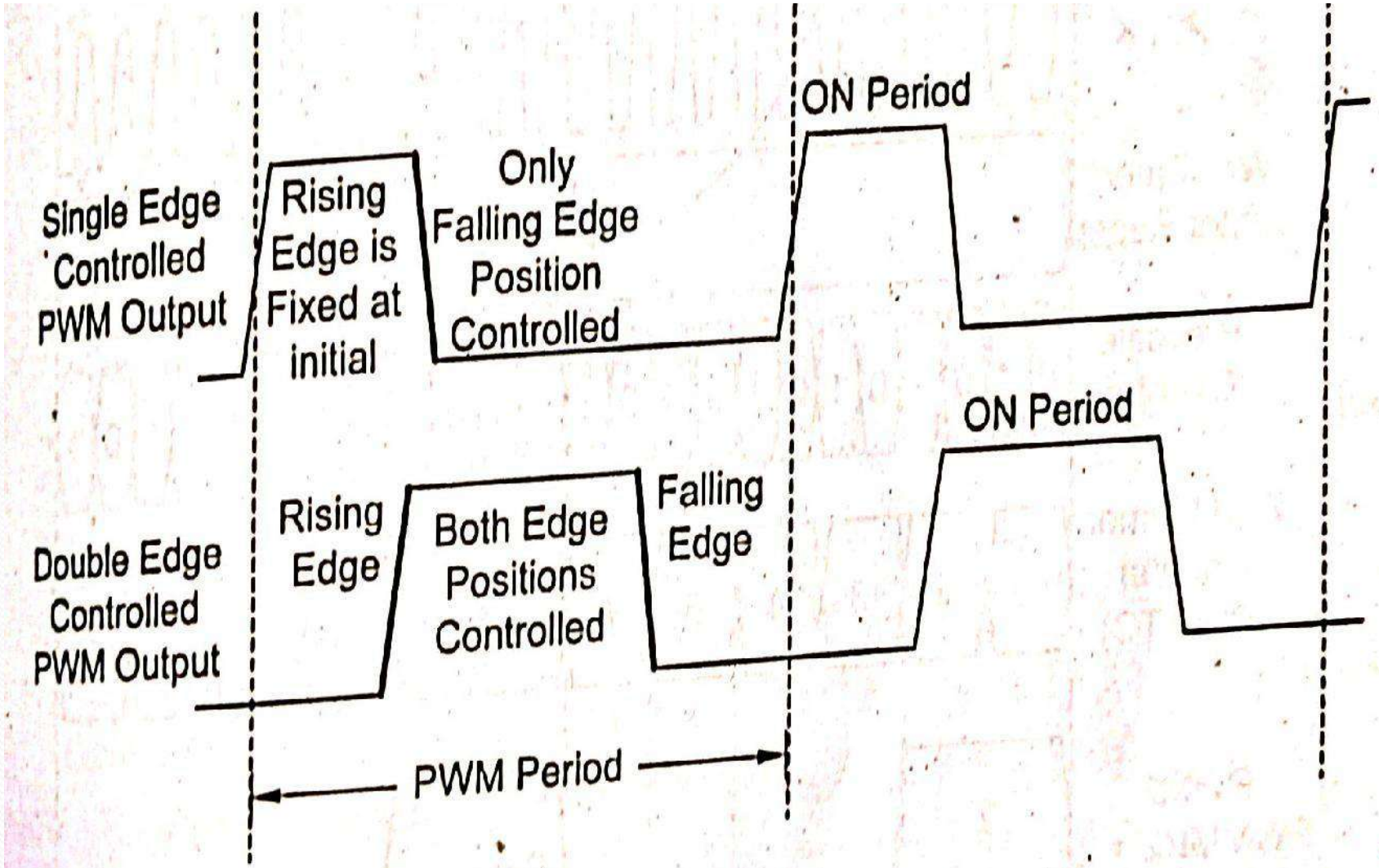


Fig. 2.16. PWM Duty Cycle Wa...

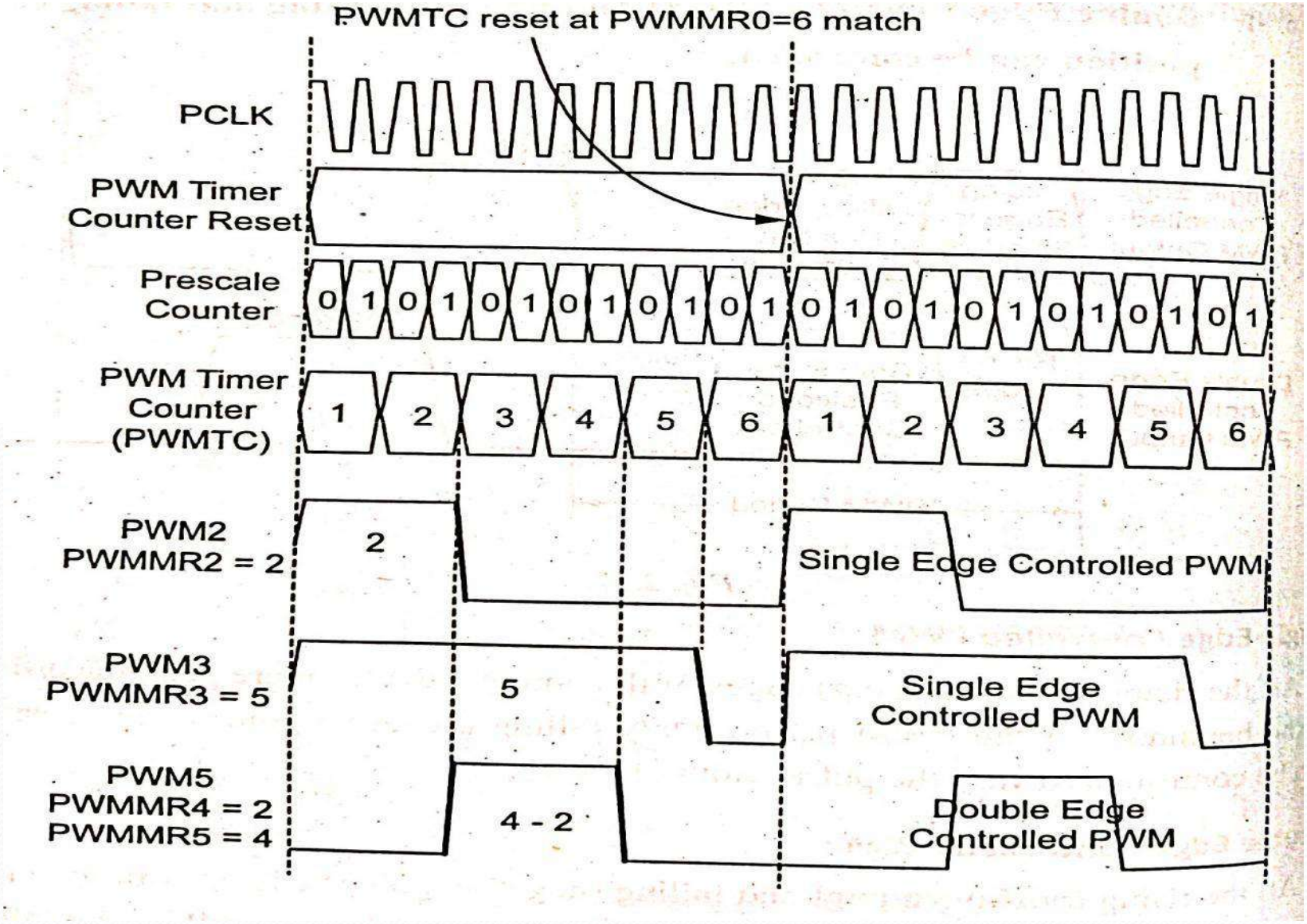

Single Edge
Controlled
PWM Output
Rising
Edge is
Fixed at
initial
Only
Falling Edge
Position
Controlled
ON Period
Double Edge
Controlled
PWM Output
Rising
Edge
Both Edge
Positions
Controlled
Falling
Edge
ON Period
PWM Period

LPC 2148

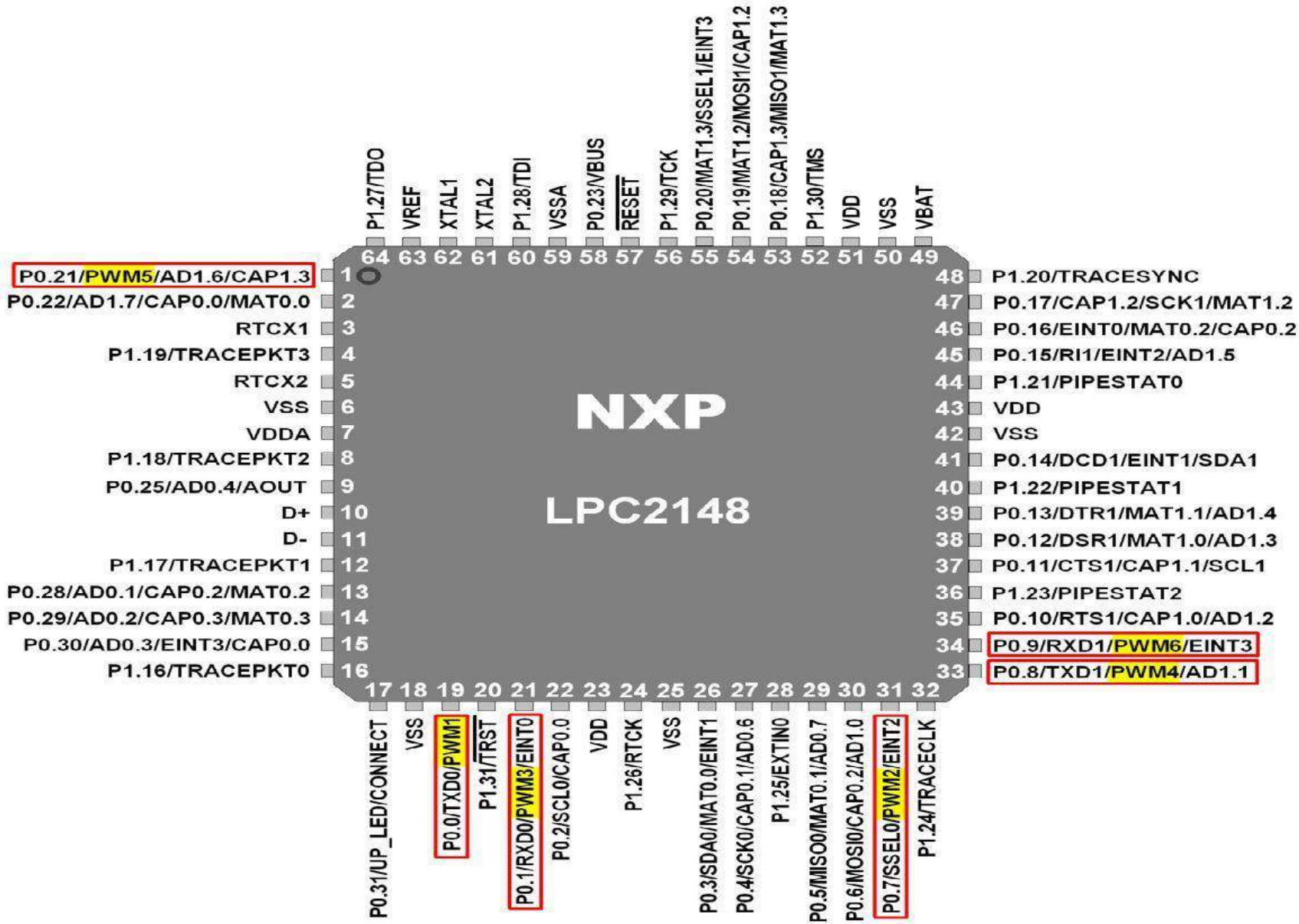
- It consist of 32 timer /counter ie PWMTC
- Counter count the cycles of peripheral clock(PCLK)
- It having 32bit prescale register (PWMPR)
- It having 7 matching register (PWMR0-PWMR06)
- 6 different pwm signal in single edge controlled pwm

or 3 different pwm signal in double edge controlled pwm

- Match register will match and then it will reset the timer/counter or stop.

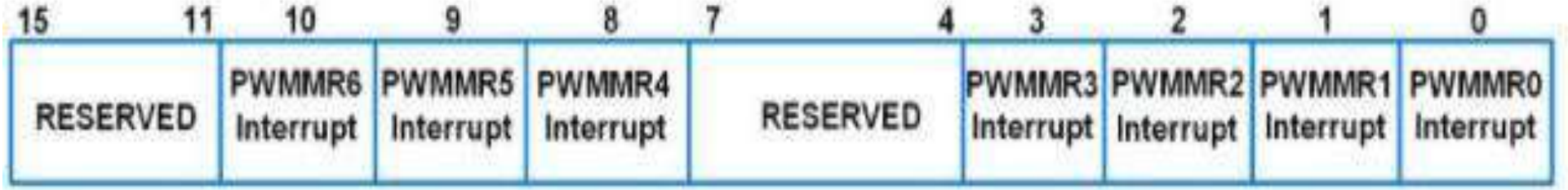

PWMTC reset at PWMMR0=6 match
PCLK
PWM Timer Counter Reset
Prescale Counter
0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1
PWM Timer Counter (PWMTC)
1 2 3 4 5 6 1 2 3 4 5 6
PWM2
PWMMR2 = 2
2
Single Edge Controlled PWM
PWM3
PWMMR3 = 5
5
Single Edge Controlled PWM
PWM5
PWMMR4 = 2
PWMMR5 = 4
4 - 2
Double Edge Controlled PWM

PWM Channel	Single Edge Controlled		Double Edge Controlled	
	Set by	Reset by	Set by	Reset by
1	Match 0	Match 1	Match 0	Match 1
2	Match 0	Match 2	Match 1	Match 2
3	Match 0	Match 3	Match 2	Match 3
4	Match 0	Match 4	Match 3	Match 4
5	Match 0	Match 5	Match 4	Match 5
6	Match 0	Match 6	Match 5	Match 6


NXP
LPC2148
1 P0.21/PWM5/AD1.6/CAP1.3
2 P0.22/AD1.7/CAP0.0/MAT0.0
3 RTCX1
4 P1.19/TRACEPKT3
5 RTCX2
6 VSS
7 VDDA
8 P1.18/TRACEPKT2
9 P0.25/AD0.4/AOUT
10 D+
11 D-
12 P1.17/TRACEPKT1
13 P0.28/AD0.1/CAP0.2/MAT0.2
14 P0.29/AD0.2/CAP0.3/MAT0.3
15 P0.30/AD0.3/EINT3/CAP0.0
16 P1.16/TRACEPKT0
17 P0.31/UP_LED/CONNECT
18 VSS
19 P0.0/TXD0/PWM1
20 P1.31/TRST
21 P0.1/RXD0/PWM3/EINT0
22 P0.2/SCL0/CAP0.0
23 VDD
24 P1.26/RTCK
25 VSS
26 P0.3/SDA0/MAT0.0/EINT1
27 P0.4/SCK0/CAP0.1/AD0.6
28 P1.25/EXTIN0
29 P0.5/MISO0/MAT0.1/AD0.7
30 P0.6/MOSI0/CAP0.2/AD1.0
31 P0.7/SSEL0/PWM2/EINT2
32 P1.24/TRACECLK
33 P0.8/TXD1/PWM4/AD1.1
34 P0.9/RXD1/PWM6/EINT3
35 P0.10/RTS1/CAP1.0/AD1.2
36 P1.23/PIPESTAT2
37 P0.11/CTS1/CAP1.1/SCL1
38 P0.12/DSR1/MAT1.0/AD1.3
39 P0.13/DTR1/MAT1.1/AD1.4
40 P1.22/PIPESTAT1
41 P0.14/DCD1/EINT1/SDA1
42 VSS
43 VDD
44 P1.21/PIPESTAT0
45 P0.15/RI1/EINT2/AD1.5
46 P0.16/EINT0/MAT0.2/CAP0.2
47 P0.17/CAP1.2/SCK1/MAT1.2
48 P1.20/TRACESYNC
49 VBAT
50 VSS
51 VDD
52 P1.30/TMS
53 P0.18/CAP1.3/MISO1/MAT1.3
54 P0.19/MAT1.2/MOSI1/CAP1.2
55 P0.20/MAT1.3/SSEL1/EINT3
56 P1.29/TCK
57 RESET
58 P0.23/VBUS
59 VSSA
60 P1.28/TDI
61 XTAL2
62 XTAL1
63 VREF
64 P1.27/TDO

PWM Registers

1. PWMIR (PWM Interrupt Register)



- It has 7 interrupt bits corresponding to the 7 PWM match registers.
- If an interrupt is generated, then the corresponding bit in this register becomes HIGH.
- Otherwise the bit will be LOW.
- Writing a 1 to a bit in this register clears that interrupt.
- Writing a 0 has no effect.

2. PWMTCR (PWM Timer Control Register)

- It is an 8-bit register.
- It is used to control the operation of the PWM Timer Counter.
- **Bit 0 – Counter Enable**
  When 1, PWM Timer Counter and Prescale Counter are enabled.
  When 0, the counters are disabled.
- **Bit 1 – Counter Reset**
  When 1, the PWM Timer Counter and PWM Prescale Counter are synchronously reset on next positive edge of PCLK.
  Counter remains reset until this bit is returned to 0.
- **Bit 3 – PWM Enable**
  This bit always needs to be 1 for PWM operation. Otherwise PWM will operate as a normal timer.
  When 1, PWM mode is enabled and the shadow registers operate along with match registers.
  A write to a match register will have no effect as long as corresponding bit in PWMLER is not set.

3. PWMTC (PWM Timer Counter)

- It is a 32-bit register.
- It is incremented when the PWM Prescale Counter (PWMPC) reaches its terminal count.

4. PWMPR (PWM Prescale Register)

- It is a 32-bit register.
- It holds the maximum value of the Prescale Counter.

5. PWMPC (PWM Prescale Counter)

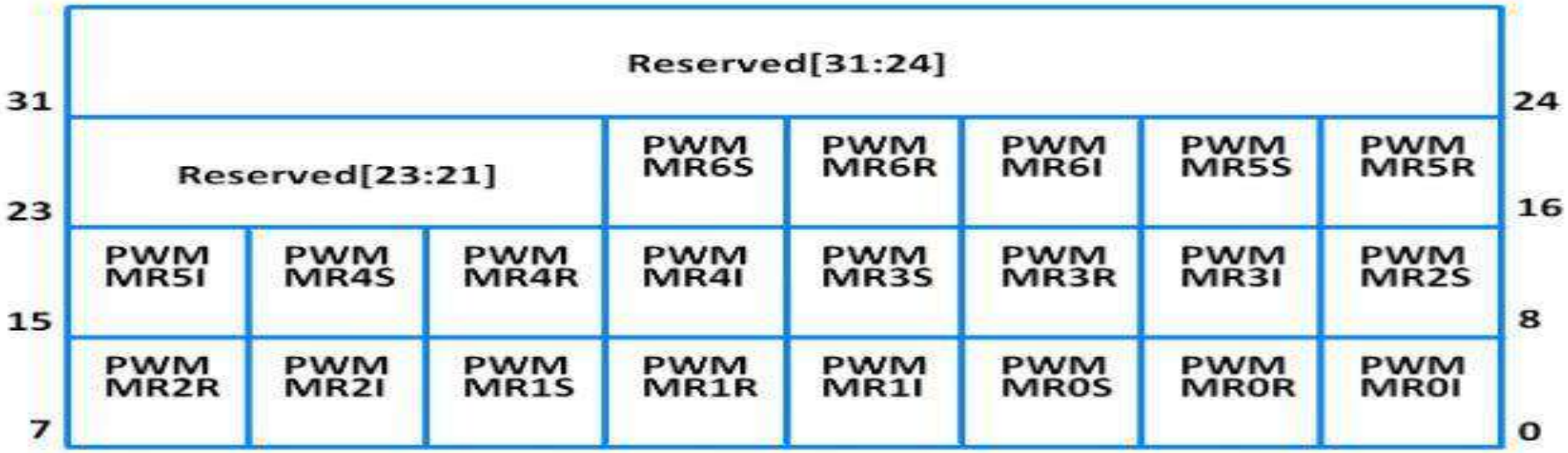
- It is a 32-bit register.
- It controls the division of PCLK by some constant value before it is applied to the PWM Timer Counter.
- It is incremented on every PCLK.
- When it reaches the value in PWM Prescale Register, the PWM Timer Counter is incremented and PWM Prescale Counter is reset on next PCLK.

6. PWMMRo-PWMMR6 (PWM Match Registers)

- These are 32-bit registers.
- The values stored in these registers are continuously compared with the PWM Timer Counter value.
- When the two values are equal, the timer can be reset or stop or an interrupt may be generated.
- The PWMMCR controls what action should be taken on a match.

7. PWMMCR (PWM Match Control Register)

- It is a 32-bit register.
- It controls what action is to be taken on a match between the PWM Match Registers and PWM Timer Counter.



Bit 0 – PWMMR0I (PWM Match register 0 interrupt)

0 = This interrupt is disabled

1 = Interrupt on PWMMR0. An interrupt is generated when PWMMR0 matches the value in PWMTC

Bit 1 – PWMMR0R (PWM Match register 0 reset)

0 = This feature is disabled

1 = Reset on PWMMR0. The PWMTC will be reset if PWMMR0 matches it

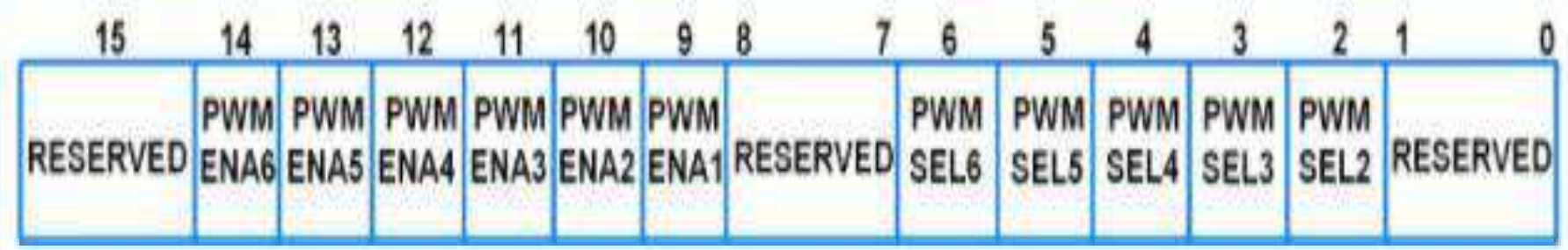
Bit 2 – PWMMR0S (PWM Match register 0 stop)

0 = This feature is disabled

1 = Stop on PWMMR0. The PWMTC and PWMPC is stopped and Counter Enable bit in PWMTCR is set to 0 if PWMMR0 matches PWMTC

PWMMR1, PWMMR2, PWMMR3, PWMMR4, PWMMR5 and PWMMR6 has same function bits (stop, reset, interrupt) as in PWMMR0.

- **Bit 2 – PWMSEL2**
  0 = Single edge controlled mode for PWM2
  1 = Double edge controlled mode for PWM2



- All other PWMSEL bits have similar operation as PWMSEL2 above.
- **Bit 10 – PWMENA2**
  0 = PWM2 output disabled
  1 = PWM2 output enabled
- All other PWMENA bits have similar operation as PWMENA2 above.

9. PWMLER (PWM Latch Enable Register)

- It is an 8-bit register.

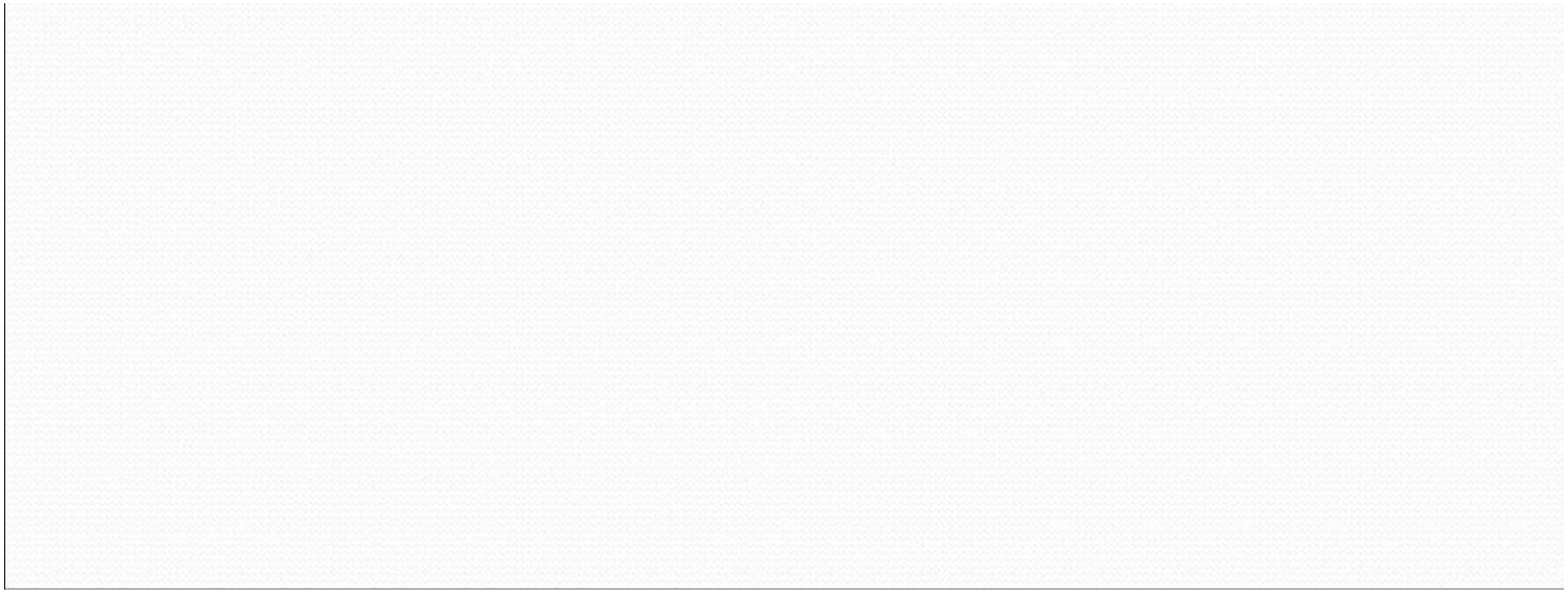
7	6	5	4	3	2	1	0
RESERVED	Enable PWM Match6 Latch	Enable PWM Match5 Latch	Enable PWM Match4 Latch	Enable PWM Match3 Latch	Enable PWM Match2 Latch	Enable PWM Match1 Latch	Enable PWM Match0 Latch

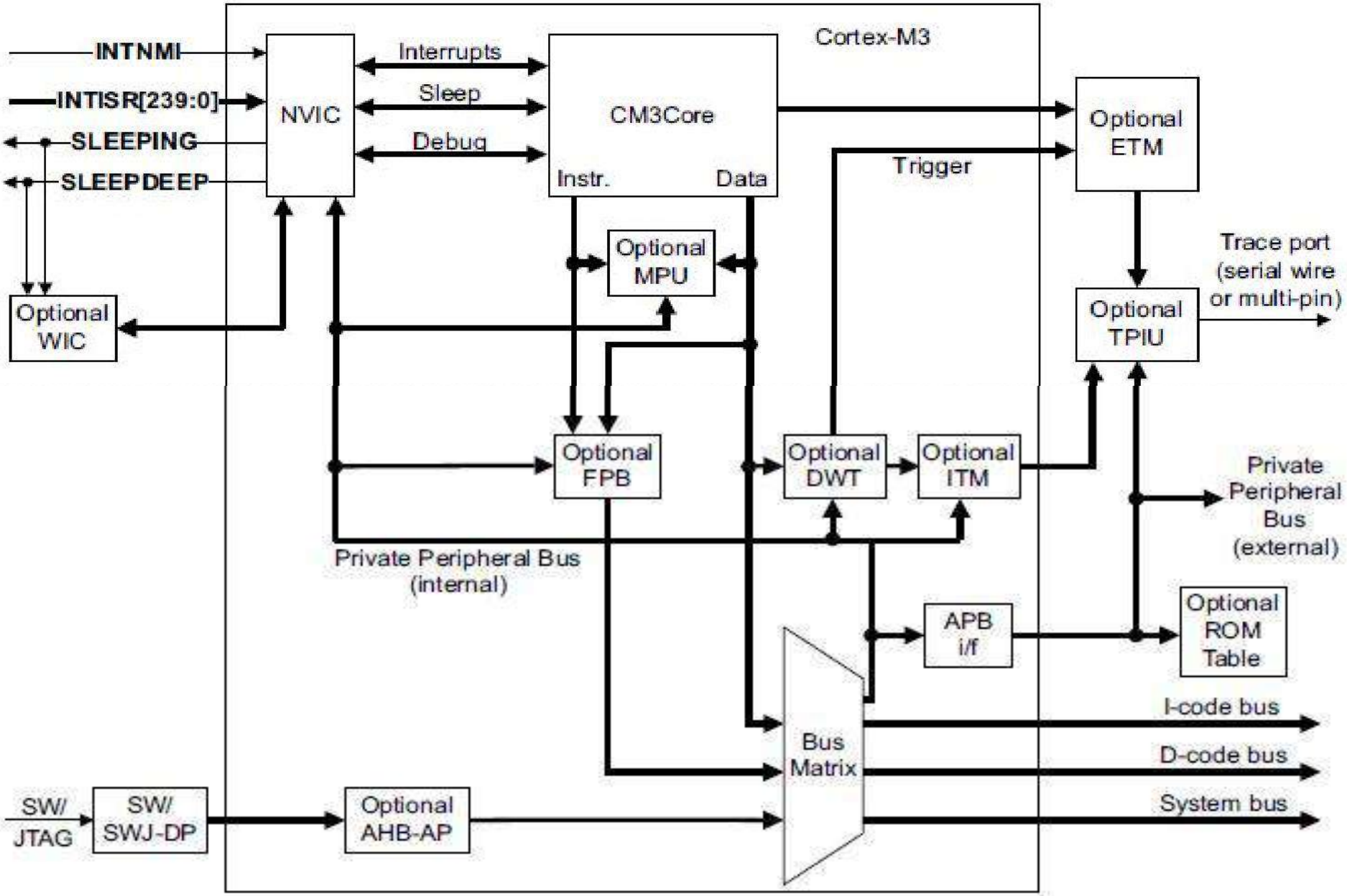
- It is used to control the update of the PWM Match Registers when they are used for PWM generation.
- When a value is written to a PWM Match Register while the timer is in PWM mode, the value is held in the shadow register. The contents of the shadow register are transferred to the PWM Match Register when the timer resets (PWM Match 0 event occurs) and if the corresponding bit in PWMLER is set.
- **Bit 6 – Enable PWM Match 6 Latch**
  Writing a 1 to this bit allows the last written value to PWMMR6 to become effective when timer next is reset by the PWM match event.
- Similar description as that of Bit 6 for the remaining bits.

Steps for PWM generation

- Reset and disable PWM counter using PWMTCR
- Load prescale value according to need of application in the PWMPR
- Load PWMMRo with a value corresponding to the time period of your PWM wave
- Load any one of the remaining six match registers (two of the remaining six match registers for double edge controlled PWM) with the ON duration of the PWM cycle. (PWM will be generated on PWM pin corresponding to the match register you load the value with).
- Load PWMMCR with a value based on the action to be taken in the event of a match between match register and PWM timer counter.
- Enable PWM match latch for the match registers used with the help of PWMLER
- Select the type of PWM wave (single edge or double edge controlled) and which PWMs to be enabled using PWMPCR
- Enable PWM and PWM counter using PWMTCR

Block diagram of ARM CORTEX M3 MCU




INTNMI
INTISR[239:0]
SLEEPING
SLEEPDEEP
NVIC
Interrupts
Sleep
Debug
CM3Core
Instr.
Data
Cortex-M3
Optional ETM
Trigger
Optional MPU
Trace port (serial wire or multi-pin)
Optional WIC
Optional TPIU
Optional FPB
Optional DWT
Optional ITM
Private Peripheral Bus (external)
Private Peripheral Bus (internal)
APB i/f
Optional ROM Table
Bus Matrix
I-code bus
D-code bus
System bus
SW/ JTAG
SW/ SWJ-DP
Optional AHB-AP

- **INTNMI**- Non-maskable interrupt
- **INTISR[239:0]**- External interrupt signals
- **SLEEPING**- Indicates that the Cortex-M3 clock can be stopped.
- **SLEEPDEEP** - Indicates that the Cortex-M3 clock can be stopped
- **WIC** - Wake-up Interrupt Controller
- **NVIC**- Nested Vectored Interrupt Controller
- **ETM**- Embedded Trace Macrocell
- The ETM is an optional debug component that enables reconstruction of program execution. The ETM is designed to be a high-speed, low-power debug tool that only supports instruction trace

- **MPU**- Memory Protection Unit
- The MPU provides full support for:
- protection regions
- overlapping protection regions, with ascending region priority:
- — 7 = highest priority
- — 0 = lowest priority.
- access permissions
- exporting memory attributes to the system.

- **FPB**-Flash Patch and Breakpoint
  - unit to implement breakpoints and code patches.
- ***DWT*** -Data Watchpoint and Trace () unit to implement watchpoints, trigger resources, and system profiling.
- ***ITM****- Instrumentation Trace Macrocell for application-driven trace source that* supports printf style debugging.
- **TPIU**- Trace Port Interface Unit
  - it is an optional component that acts as a bridge between the on-chip trace data from the *Embedded Trace Macrocell (ETM) and the Instrumentation Trace Macrocell*(ITM), with separate IDs, to a data stream, encapsulating IDs where required, that is then captured by a *Trace Port Analyzer (TPA).*

- **SW/SWJ-DP** - SW-DP or SWJ-DP debug port interfaces.
- The debug port provides debug access to all registers and memory in the system, including the processor registers.
- The SW/SWJ-DP might not be present in the production device if no debug functionality is present in the implementation.

UNIT III

EMBEDDED PROGRAMMING

Syllabus

Components for embedded programs- Models of programs- Assembly, linking and loading – compilation techniques- Program level performance analysis – Software performance optimization – Program level energy and power analysis and optimization – Analysis and optimization of program size- Program validation and testing

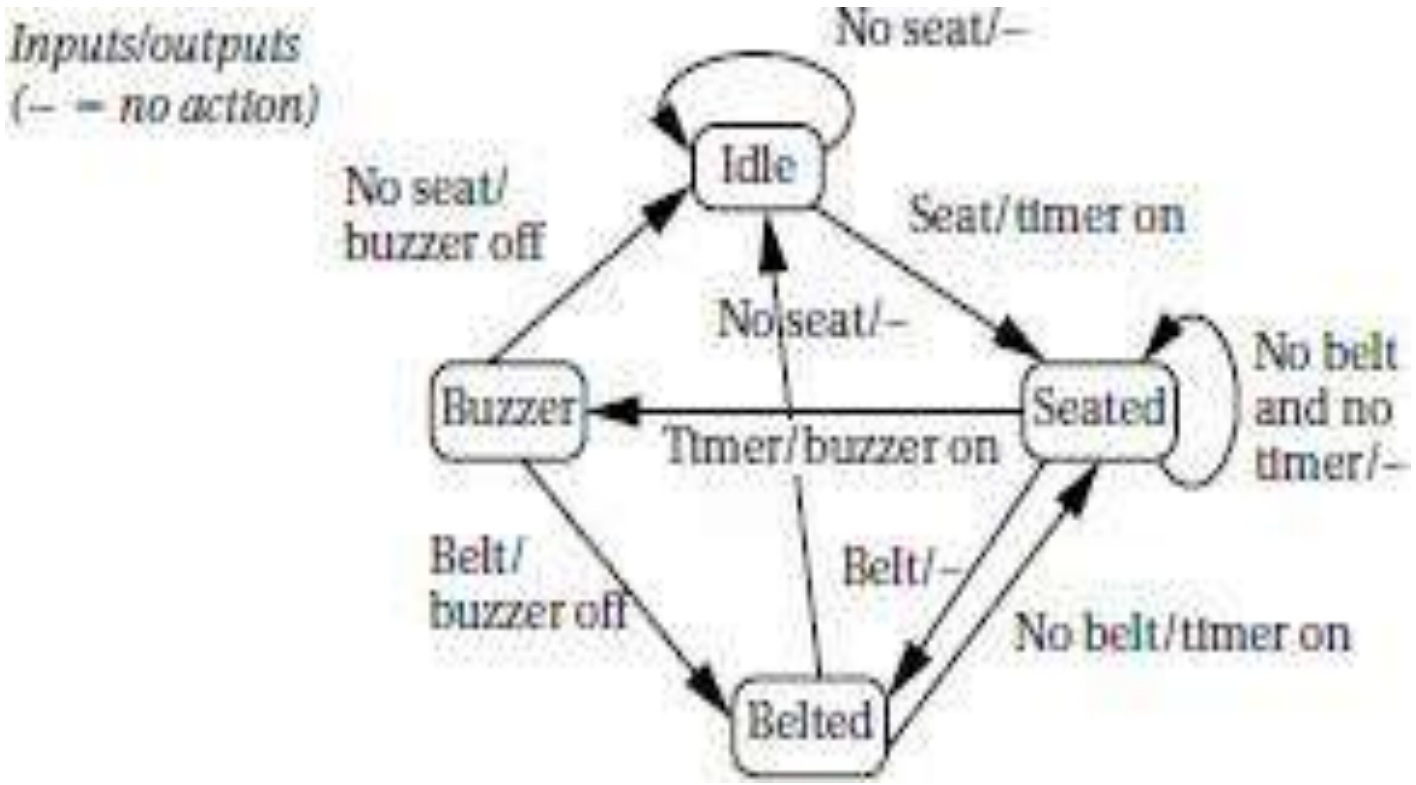
1. COMPONENTS FOR EMBEDDED PROGRAMS

- Embedded components are given by State machine, Circular buffer, and the Queue.

STATE MACHINE

- The reaction of most systems can be characterized in terms of the input received and the current state of the system.
- The finite-state machine style of describing the reactive system's behavior..
- Finite-state machines are usually first encountered in the context of hardware design.

software state machine



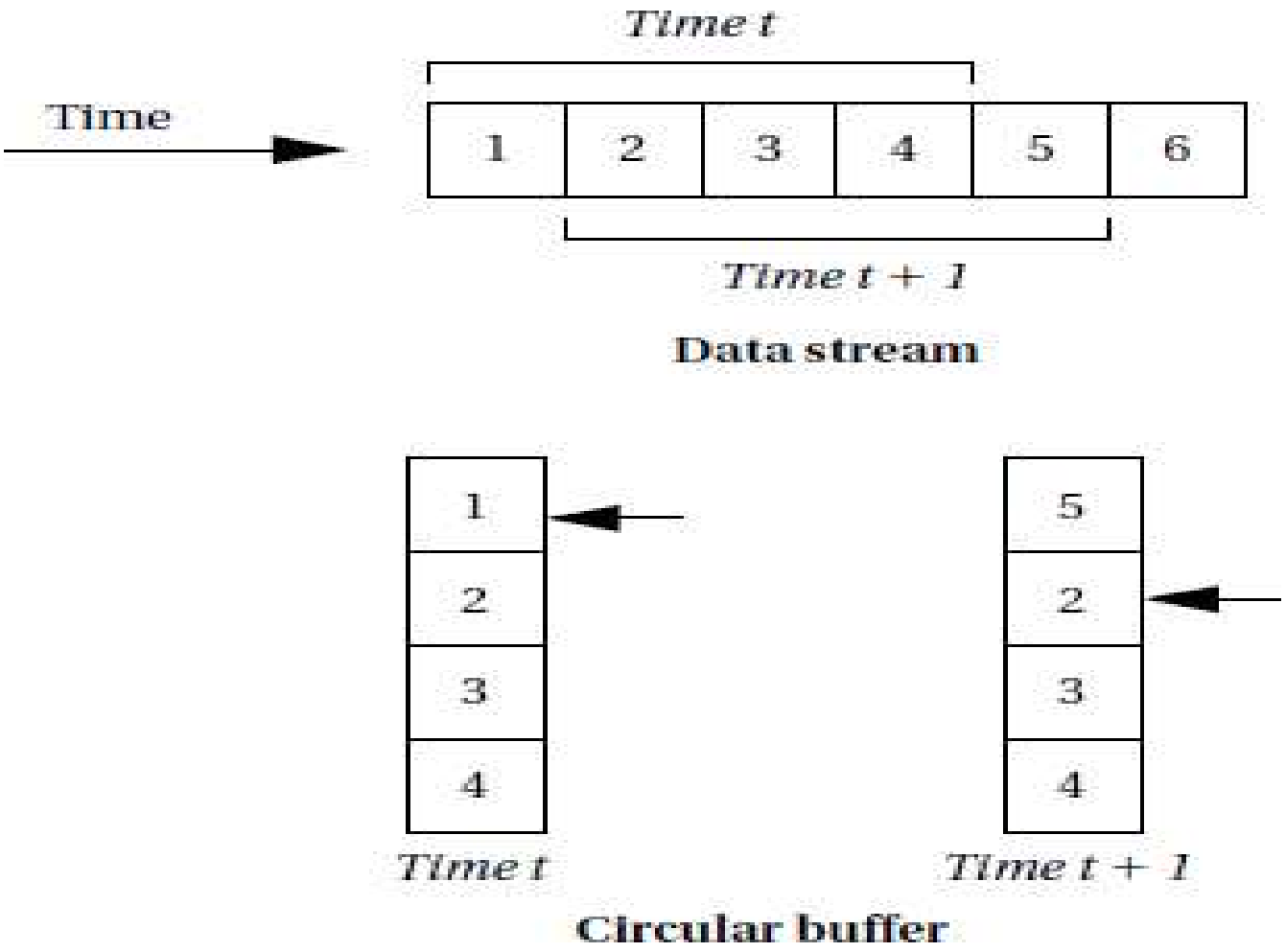
seat, belt, timer

```
#define IDLE 0
#define SEATED 1
#define BELTED 2
#define BUZZER 3
switch (state) { /* check the current state
    */
case IDLE:
if (seat) { state = SEATED;
timer_on = TRUE; }
/* default case is self-loop */
break;
case SEATED:
if (belt) state = BELTED; /* won't hear the
buzzer */
else if (timer) state = BUZZER; /* didn't
    put on
belt in time */
/* default is self-loop */
break;
case BELTED:
if (!seat) state = IDLE; /* person left */
else if (!belt) state = SEATED; /* person
    still
in seat */
break;
case BUZZER:
if (belt) state = BELTED; /* belt is on—
    turn off
buzzer */
else if (!seat) state = IDLE; /* no one in
seat—turn off buzzer */
break;
}
```

Stream-Oriented Programming and Circular Buffers

- The circular buffer is a data structure that handle streaming data in an efficient way.
- Size of the window does not change.
- Fixed-size buffer to hold the current data.
- To avoid constantly copying data within the buffer, move the head of the buffer in time.
- The buffer points to the location at which the next sample will be placed.
- Every time add a sample, automatically overwrite the oldest sample, which is the one that needs to be thrown out.
- When the pointer gets to the end of the buffer, it wraps around to the top.

Circular buffer for streaming data.



1.3 QUEUES

- Queues are also used in signal processing and event processing.
- Queues are used whenever data may arrive and depart at somewhat unpredictable times or when variable amounts of data may arrive.
- A queue is often referred to as an Elastic buffer.

2.MODELS OF PROGRAMS

- Programs are collection of instructions to execute a specified task.
- Models for programs are more general than source code.
- source code can't be used directly because of different type s such as assembly language,C code.
- Single model to describe all of them.
- control/data flow graph (CDFG)→it is the fundamental model for programs

2.1 DATA FLOW GRAPH

- A data flow graph is a model of a program with no conditionals.
- In a high-level programming language, a code segment with no conditionals have only one entry and exit point—is known as a basic block.

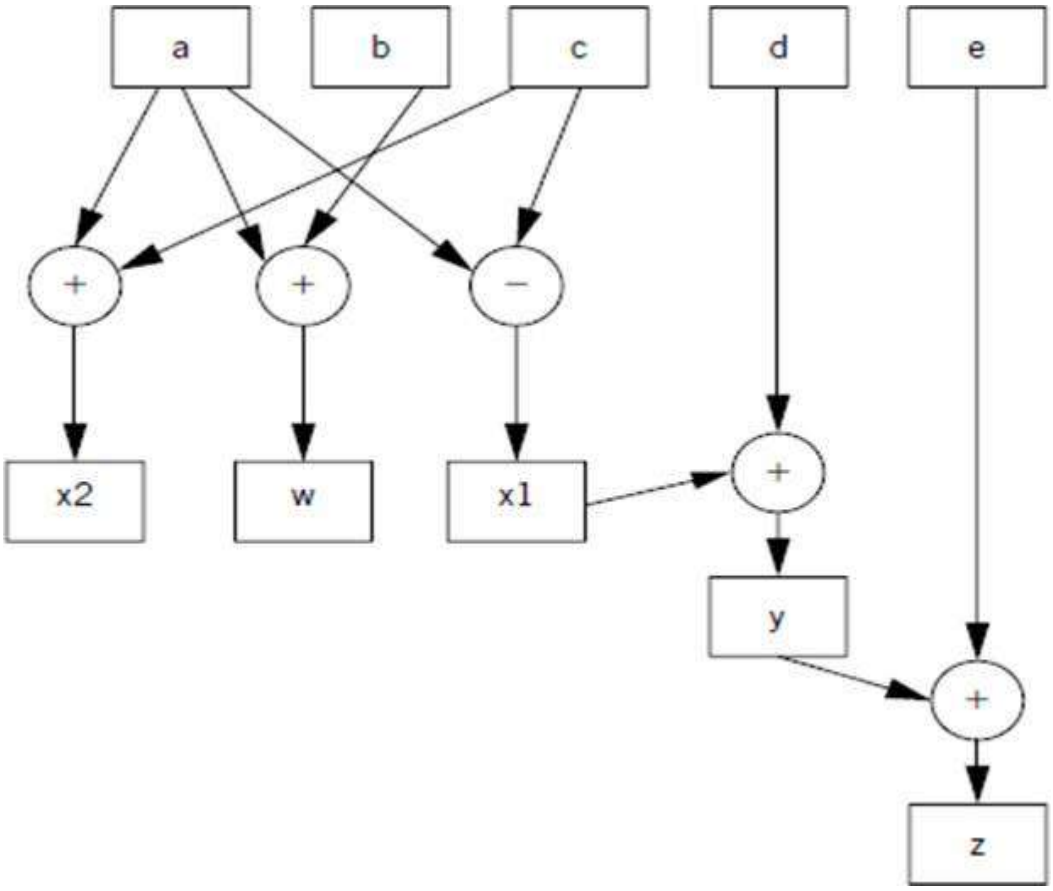
  - A basic block in C

```
w = a + b;
x = a - c;
y = x + d;
x = a + c;
z = y + e;
```

An extended data flow graph for our sample basic block

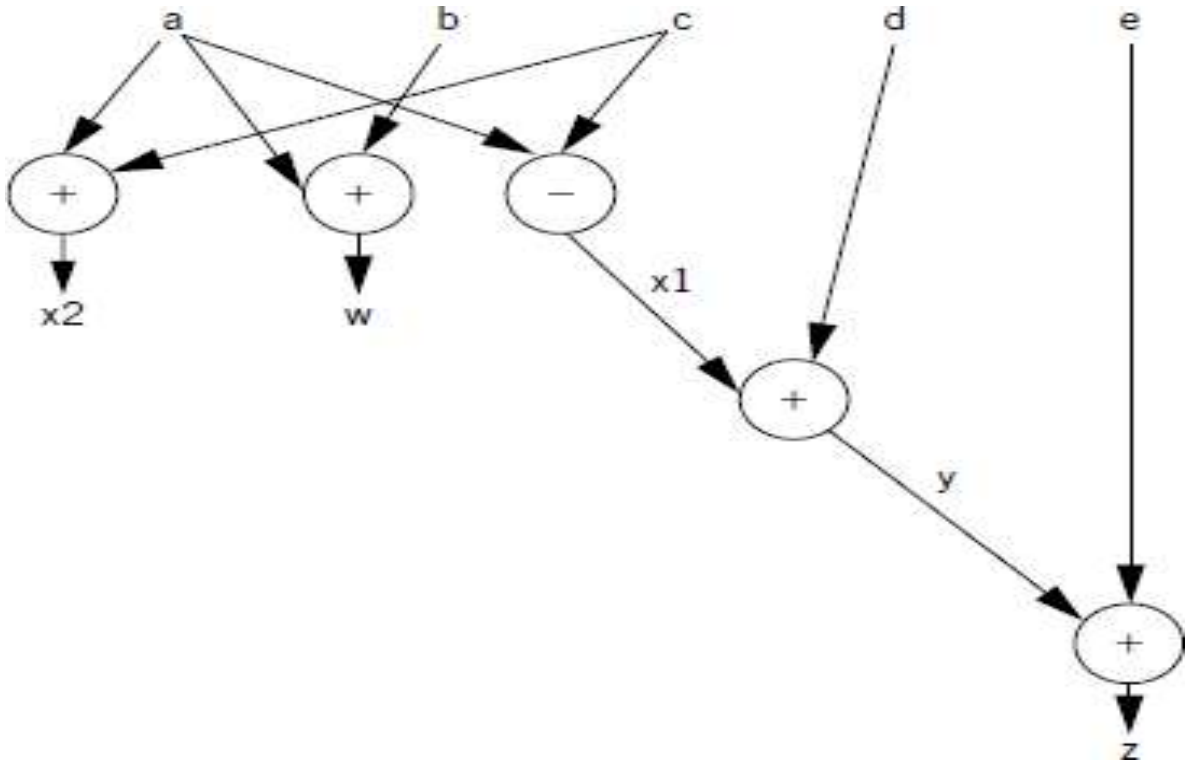
•The basic block in single-assignment form

```
w = a + b;
x1 = a - c;
y = x1 + d;
x2 = a + c;
z = y + e;
```



•Round nodes→denote operators
•Square nodes→denote values.
•The value nodes may be either inputs(a,b) or variables(w,x1).

Standard data flow graph for our sample basic block



2.2. Control/Data Flow Graphs(CDFG)

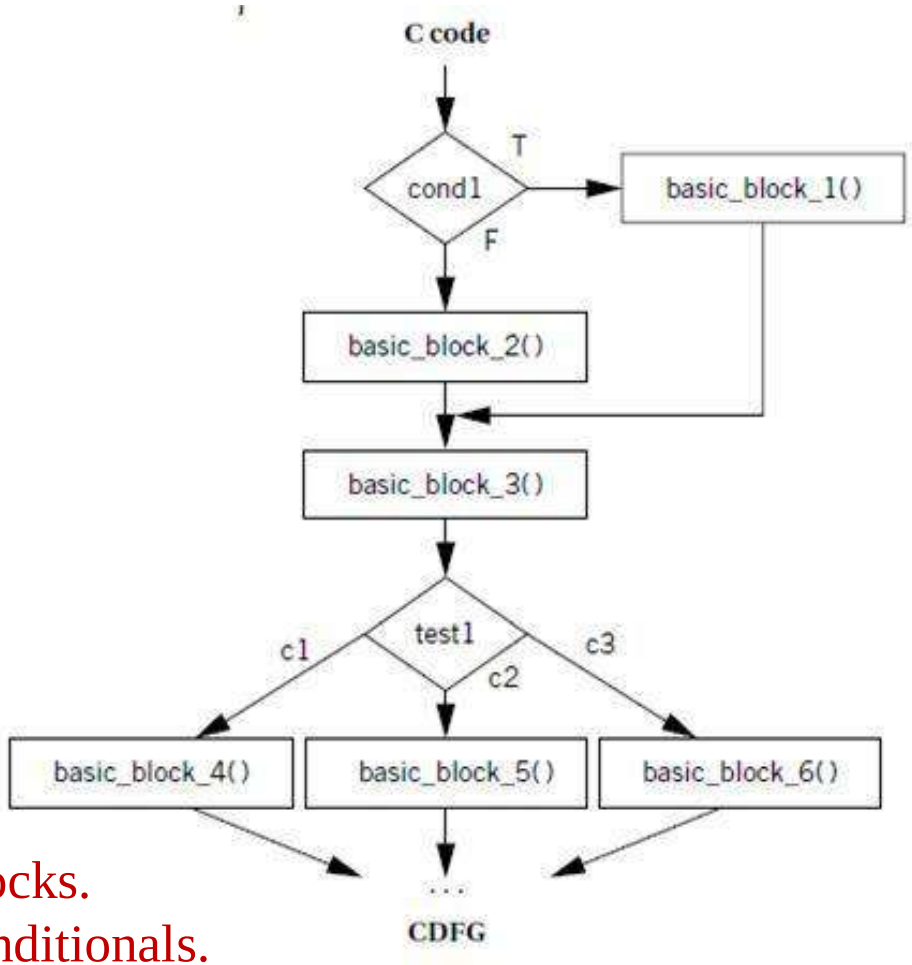
- A CDFG uses a data flow graph as an element,adding constructs to describe control.

CDFG having following two types of nodes.

1. Decision nodes→used to describe the control in a sequential program
- Data flow nodes→ encapsulates a complete data flow graph to represent a data.

C code and

```
if (cond1)
basic_block_1( );
else
basic_block_2();
basic_block_3( );
switch (test1) {
case c1: basic_block_4( ); break;
case c2: basic_block_5( ); break;
case c3: basic_block_6( ): break;
}
```



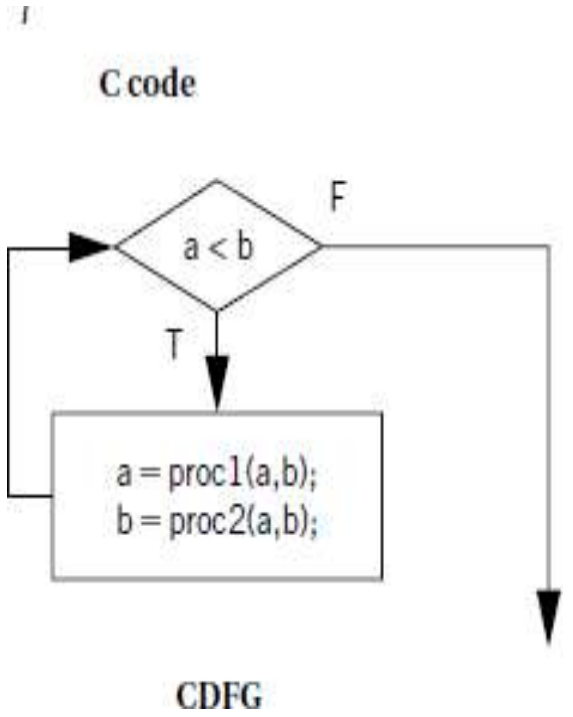
- Rectangular nodes → represent the basic blocks.
- Diamond-shaped nodes → represent the conditionals.
- Label → node's condition
- Edges are labeled with the possible outcomes of evaluating the condition

CDFG for a while loop

```
while (a < b) {
a5proc1(a,b);
b5proc2(a,b);
}
```

CDFG for a while loop

```
while (a < b) {
a5proc1(a,b);
b5proc2(a,b);
}
```

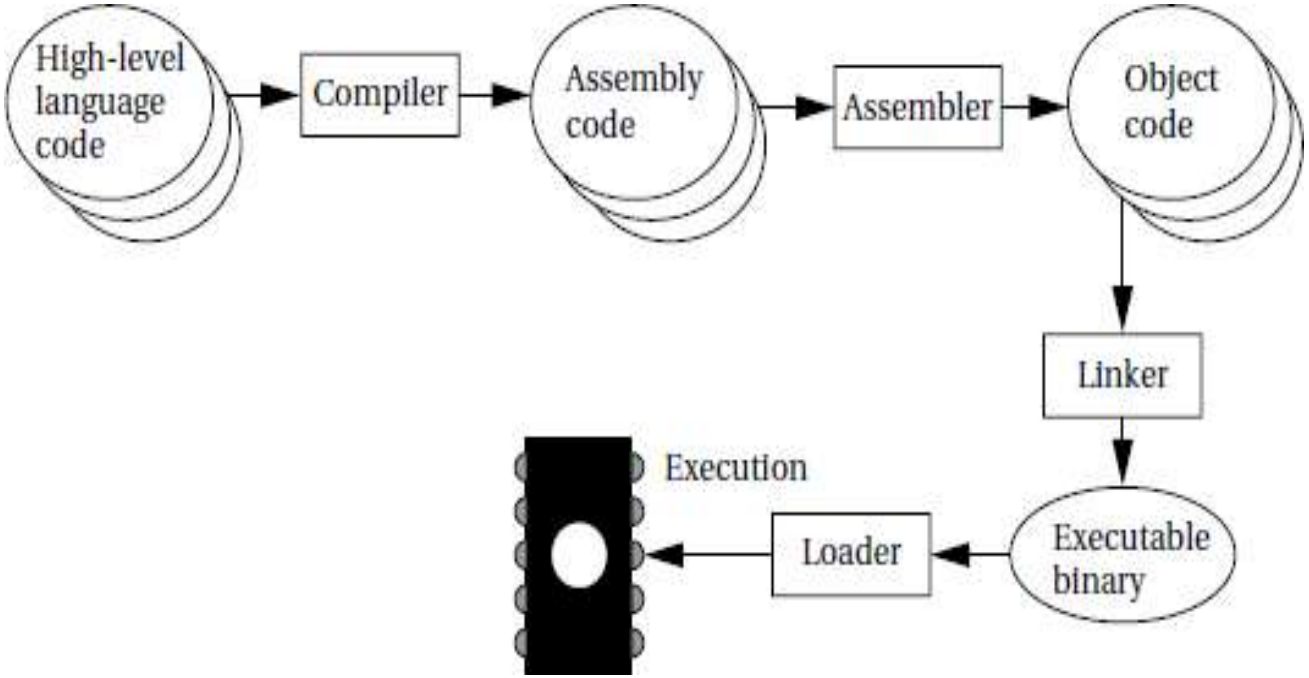


3. ASSEMBLY, LINKING AND LOADING

•Assembly and linking→ last steps in the compilation process

•They convert list of instructions into an image of the program's bits in memory.

•Loading→ puts the program in memory so that it can be executed.



- Compilers → used to create the instruction-level program in to assembly language code.
- Assembler's→ used to translate symbolic assembly language statements into bit-level representations of instructions known as object code and also translating labels into addresses.
- Linker→ determining the addresses of instructions.
- Loader→ load the program into memory for execution.
- Absolute addresses→ Assembler assumes that the starting address of the ALP has been specified by the programmer.
- Relative addresses→ specifying at the start of the file address is to be computed later.

3.1 Assemblers

- Assembler→ Translating assembly code into object code also assembler must translate opcodes and format the bits in each instruction, and translate labels into addresses.
- Labels →it is an abstraction provided by the assembler.
- Labels→know the locations of instructions and data.

Label processing requires making two passes

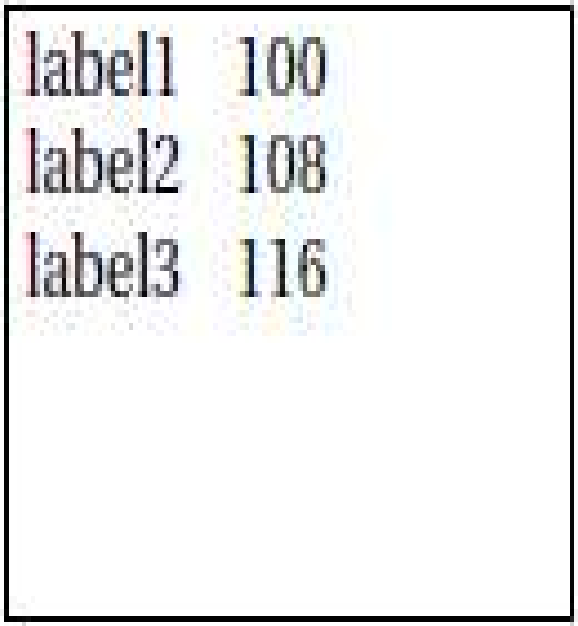
1. first pass scans the code to determine the address of each label.
2. second pass assembles the instructions using the label values computed in the first pass.

EXAMPLE

CODE

```
        ORG 100
label1  ADR r4,c
        LDR r0,[r4]
label2  ADR r4,d
        LDR r1,[r4]
label3  SUB r0,r0,r1
```

SYMBOL TABLE



Symbol table

3.2)LINKING

- A linker allows a program to be stitched together out of several smaller pieces.
- The linker operates on the object files and links between files.
- Some labels will be both defined and used in the same file.
- Other labels will be defined in a single file but used elsewhere .
- The place in the file where a label is defined is known as an entry point.
- The place in the file where the label is used is called an external reference.

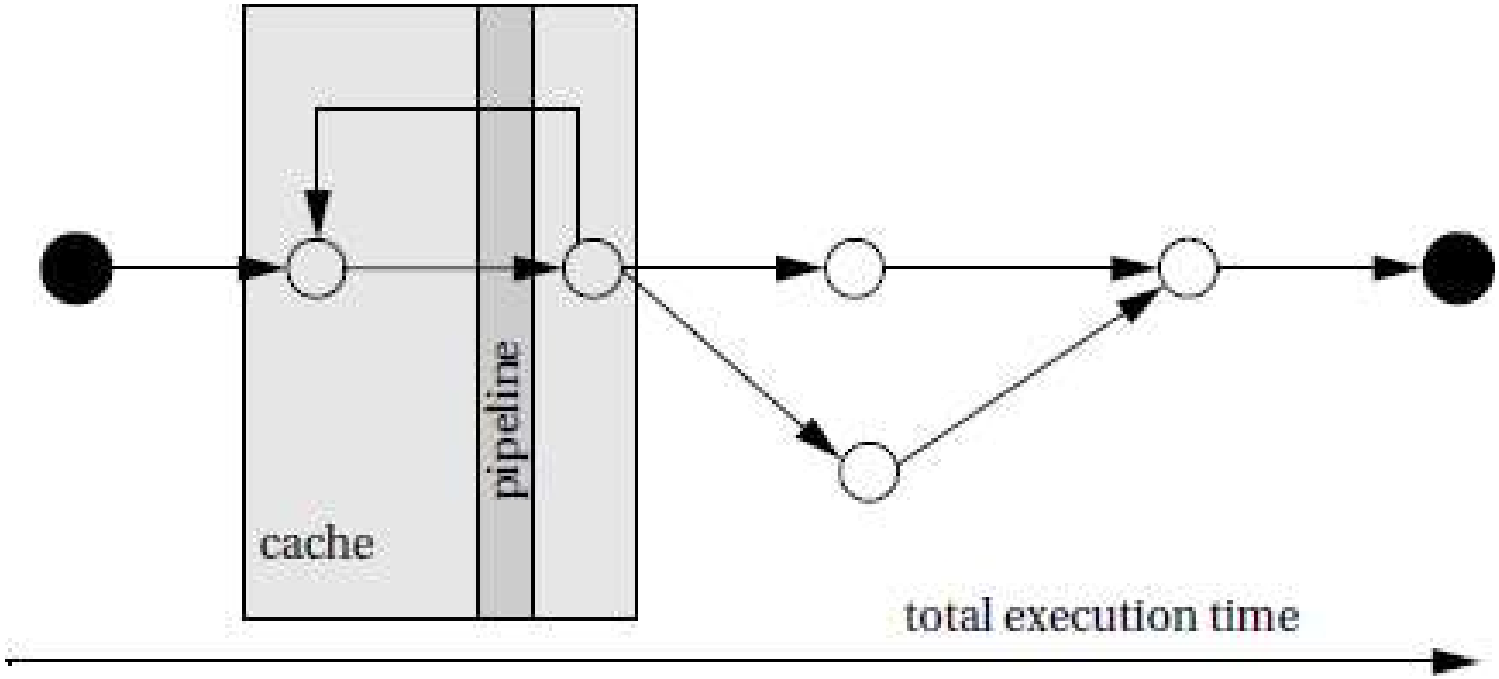
Phases of linker

- First Phase→it determines the address of the start of each object file
- Second Phase→the loader merges all symbol tables from the object files into a single,large table.

4. PROGRAM-LEVEL PERFORMANCE ANALYSIS

- The techniques we use to analyze program execution time are also helpful in analyzing properties such as power consumption.
- The CPU executes the entire program at the rate we desire.
- The execution time of a program often varies with the input data values.
- The cache has a major effect on program performance.
- Cache's behavior depends in part on the data values input to the program.
- The execution time of an instruction in a pipeline depends not only on that instruction but on the instructions around it in the pipeline.

Execution time of a program



Program Performance Measuring techniques

1. Simulator

- It runs on a PC, takes as input an executable for the microprocessor along with input data, and simulates the program.

2. Timer

- It is can be used to measure performance of executing sections of code.
- The length of the program that can be measured is limited by the accuracy of the timer.

3. Logic analyzer

- It is used to measure the start and stop times of a code segment.
- The length of code that can be measured is limited by the size of the logic analyzer's buffer.

4.2)Types of performance Parameters

1. Average-case execution time

- This is the typical execution time we would expect for typical data.

2. Worst-case execution time

- The longest time that the program can spend on any input sequence is clearly important for systems that must meet deadlines.

3. Best-case execution time

- This measure can be important in multi-rate real-time systems.

4.3) Elements of Program Performance

- Execution time =Program path +Instruction timing
- Program path→ It is the sequence of instructions executed by the program.
- Instruction timing→It is determined based on the sequence of instructions traced by the program path.
- Not all instructions take the same amount of time.
- The execution time of an instruction may depend on operand values.

Measurement-Driven Performance Analysis

- To measure the program's performance →need CPU or its simulator .
- Measuring program performance → combination of determination of the execution path and the timing of that path.
- program trace → record of the execution path of a program.

Cycle-Accurate Simulator

- It can determine the exact number of clock cycles required for execution.
- It is built with detailed knowledge of how the processor works .
- It is slower than the processor itself, but a variety of techniques can be used to make them surprisingly fast.
- It has a complete model of the processor, including the cache.
- It can provide information about why the program runs too slowly.

5. SOFTWARE PERFORMANCE OPTIMIZATION

5.1)Loop Optimizations-Loops are important targets for optimization because programs with loops tend to spend a lot of time executing those loops.

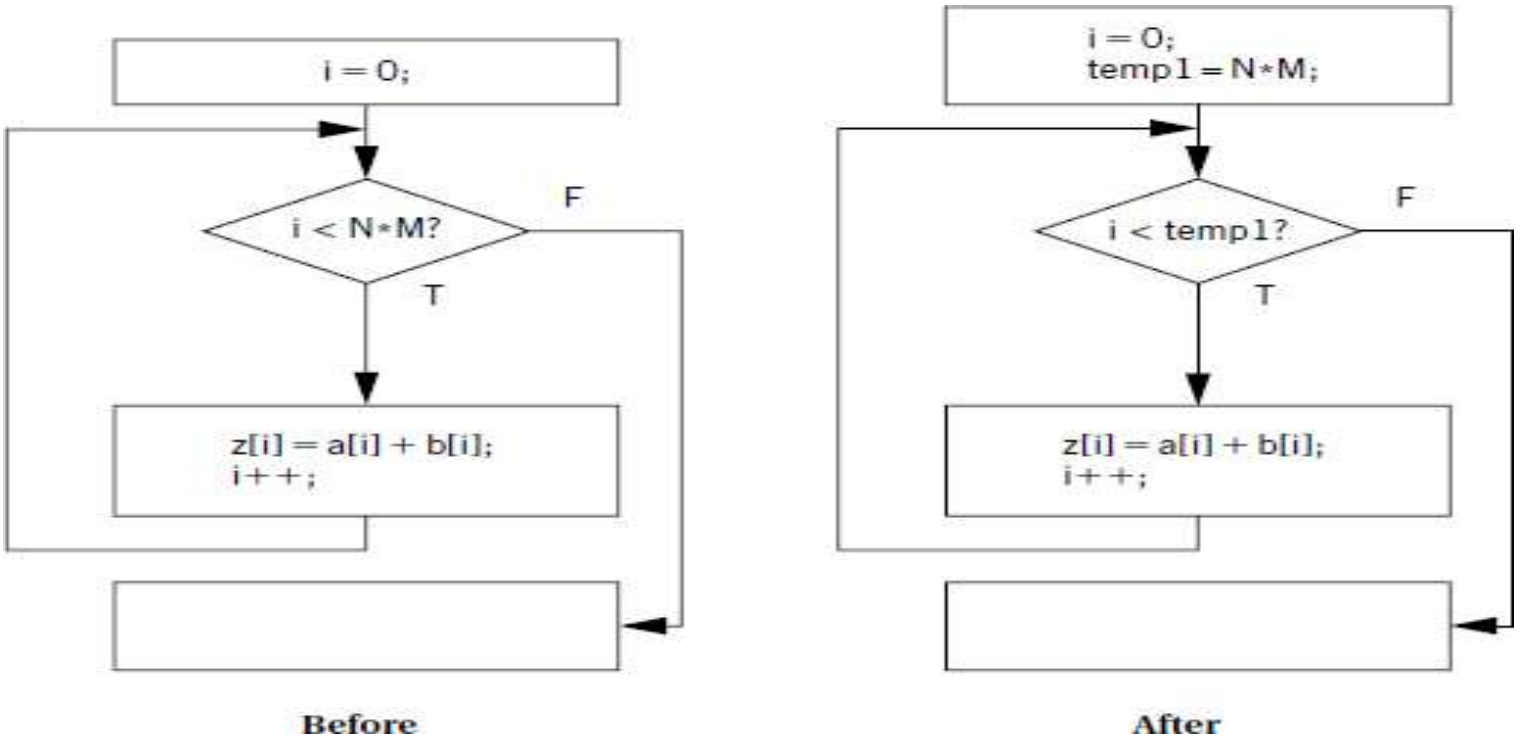
- Code motion
- Induction variable elimination
- Strength reduction

Code motion

- It can move unnecessary code out of a loop.
- If a computation's result does not depend on operations performed in the loop body,thenwe can safely move it out of the loop.

```
for (i = 0; i < N*M; i++)
 {
z[i] = a[i] + b[i];
}
```

Code motion in a loop



•The loop bound computation is performed on every iteration during the loop test, even though the result never changes.

•We can avoid *N X M- 1 unnecessary executions of this statement by moving it before the loop.*

Induction variable elimination

- It is a variable whose value is derived from the loop iteration variable's value.
- The compiler often introduces induction variables to help it implement the loop.
- Properly transformed→ able to eliminate some variables and apply strength reduction to others.
- A nested loop is a good example of the use of induction variables.

```
for (i = 0; i < N; i++)
for (j = 0; j < M; j++)
z[i][j] = b[i][j];
```

- The compiler uses induction variables to help it address the arrays. Let us rewrite the loop in C using induction variables and pointers

```
for (i = 0; i < N; i++)
for (j = 0; j < M; j++) {
zbinduct = i*M + j;
*(zptr + zbinduct) = *(bptr + zbinduct);
}
```

Strength reduction

- It reduce the cost of a loop iteration.

Consider the following assignment

```
y = x * 2;
```

- In integer arithmetic, we can use a left shift rather than a multiplication by 2
- If the shift is faster than the multiply, then perform the substitution.
- This optimization can often be used with induction variables because loops are often indexed with simple expressions.

5.2 Cache Optimizations

- A loop nest is a set of loops, one inside the other.
- Loop nests occur when we process arrays.
- A large body of techniques has been developed for optimizing loop nests.
- Rewriting a loop nest changes the order in which array elements are accessed.
- This can expose new parallelism opportunities that can be exploited by later stages of the compiler, and it can also improve cache performance.

6. PROGRAM-LEVEL ENERGY AND POWER ANALYSIS AND OPTIMIZATION

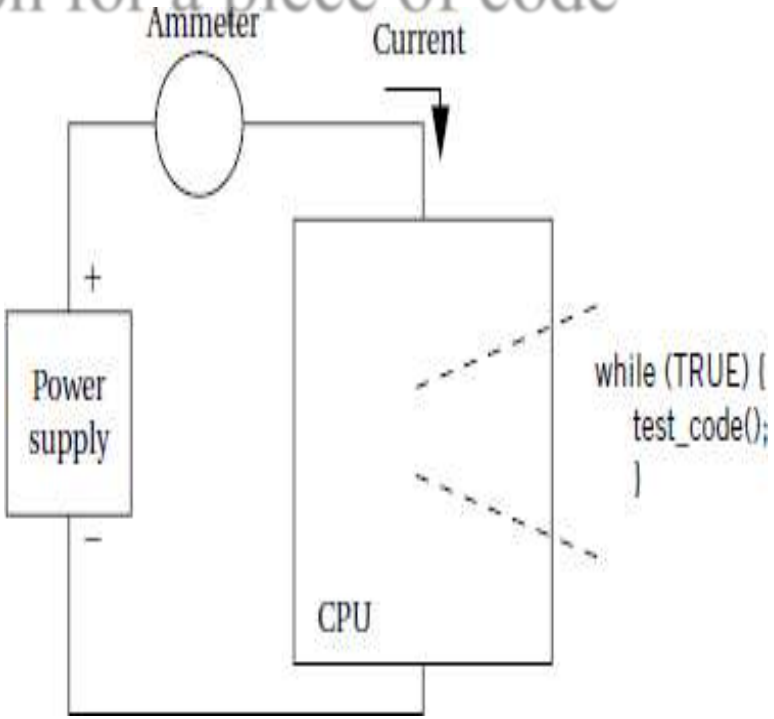
- Power consumption is a important design metric for battery-powered systems.
- It is increasingly important in systems that run off the power grid.
- Fast chips run hot, and controlling power consumption is an important element of increasing reliability and reducing system cost.

Power consumption reduction techniques.

- To replace the algorithms with others that consume less power.
- By optimizing memory accesses ,able to significantly reduce power.
- To turn off the subsystems of CPU, chips in the system, in order to save power.

Measuring energy consumption for a piece of code

- Program's energy consumption→ how much energy the program consumes.
- To measure power consumption for an instruction or a small code fragment.
- It is used to executes the code under test over and over in a loop.
- By measuring the current flowing into the CPU,we are measuring the power consumption of the complete loop, including both the body and other code.
- By separately measuring the power consumption of a loop with no body.
- we can calculate the power consumption of the loop body code as the difference b/w the full loop and the bare loop energy cost of an instruction.



List of the factors contribution for energy consumption of the program.

- Energy consumption varies somewhat from instruction to instruction.
- The sequence of instructions has some influence.
- The opcode and the locations of the operands also matter.

Steps to Improve Energy Consumption

- Try to use registers efficiently(r4)
- Analyze cache behavior to find major cache conflicts.
- Make use of page mode accesses in the memory system whenever possible.
- Moderate loop unrolling eliminates some loop control overhead. when the loop is unrolled too much, power increases.
- Software pipelining reducing the average energy per instruction.
- Eliminating recursive procedure calls where possible saves power by getting rid of function call overhead.
- Tail recursion can often be eliminated, some compilers do this automatically.

7. ANALYSIS AND OPTIMIZATION OF PROGRAM SIZE

- Memory size of a program is determined by the size of its data and instructions.
- Both must be considered to minimize program size.
- Data provide an opportunity to minimizing the size of program.
- Data buffers can be reused at several different points in program, which reduces program size.
- Some times inefficient programs keep several copies of data, identifying and eliminating duplications can lead to significant memory savings.
- Minimizing the size of the instruction text and reducing the number of instructions in a program→ which reduces program size
- Proper instruction selection may reduce code size.
- Special compilation modes produce the program in terms of the dense instruction set.
- Program size of course varies with the type of program, but programs using the dense instruction set are often 70 to 80% of the size of the standard instruction set equivalents.

8. PROGRAM VALIDATION AND TESTING

- Complex systems→ need testing to ensure the working behavior of the systems.
- Software Testing→used to generate a comprehensive set of tests to ensure that our system works properly.
- The testing problem is divided into sub-problems and analyze each sub problem.

Types of testing strategies

1. White/Clear-box Testing→ generate tests ,based on the program structure.
2. Black-box Testing→ generate tests ,without looking at the internal structure of the program.

Clear box testing

- Testing→requires the control/data flow graph of a program's source code.
- To test the program→ exercise both its control and data operations.
- To execute and evaluate the tests→ control the variables in the program and observe the results .

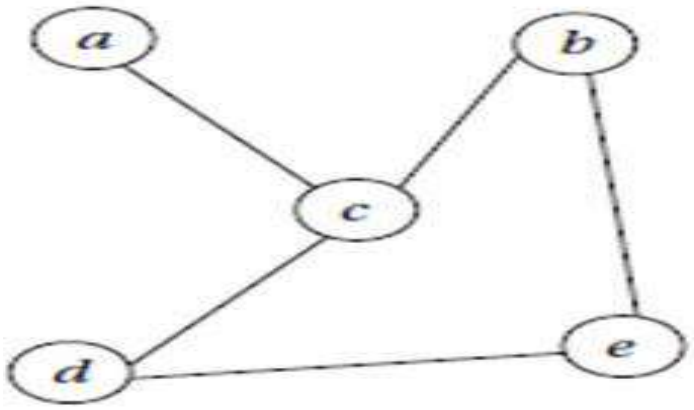
The following three things to be followed during a test

1. Provide the program with inputs for the test.
2. Execute the program to perform the test.
3. Examine the outputs to determine whether the test was successful.

- Execution Path→To test the program by forcing the program to execute along chosen paths. (giving it inputs that it to take the appropriate branches)

Graph Theory

•It help us get a quantitative handle on the different paths required.
•Undirected graph-→form any path through the graph from combinations of basis paths.
•Incidence matrix contains each row and column represents a node.
•1 is entered for each node pair connected by an edge.



Graph

	a	b	c	d	e
a	0	0	1	0	0
b	0	0	1	0	1
c	1	1	0	1	0
d	0	0	1	0	1
e	0	1	0	1	0

Incidence matrix

a	1	0	0	0	0
b	0	1	0	0	0
c	0	0	1	0	0
d	0	0	0	1	0
e	0	0	0	0	1

Basis set

Cyclomatic Complexity

- It is a software metric tool.
- Used to measure the control complexity of a program.

$$M = e - n + 2p.$$

- *e* → *number of edges in the flow graph*
- *n* → *number of nodes in the flow graph*
- *p* → *number of components in the graph*

Types of Clear Box test strategy

1. Branch testing
2. Domain testing
3. Data flow testing

Branch testing

- This strategy requires the true and false branches of a conditional.
- Every simple condition in the conditional's expression to be tested at least once.

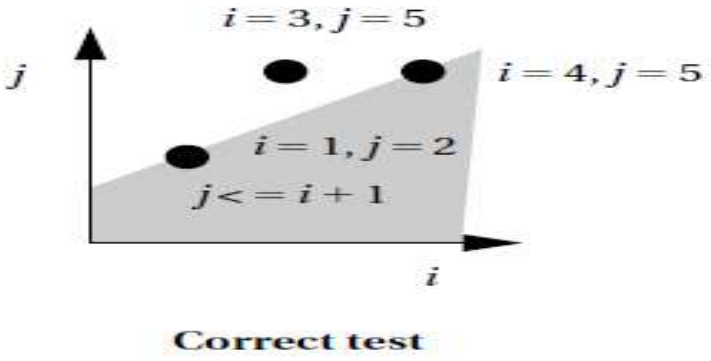
```
if ((x == good_pointer) && (x->field1 == 3))
{ printf("got the value\n"); }
```

The bad code we actually wrote

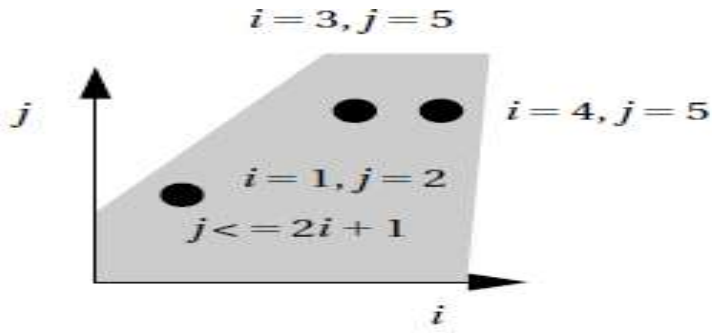
```
if ((x = good_pointer) && (x->field1 == 3))
{ printf("got the value\n"); }
```

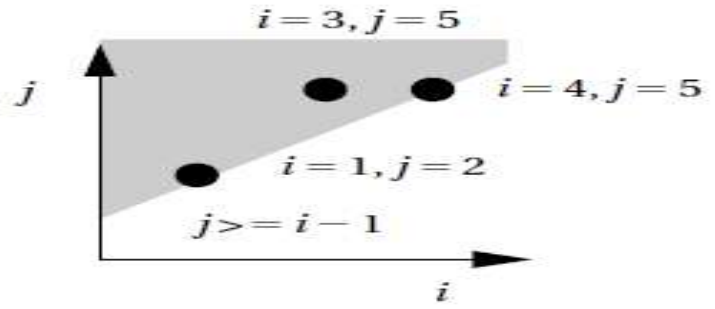
Domain testing

- It concentrates on linear-inequalities.
- The program should use for the test is $j <= i + 1$
- We test the inequality with three test points
- Two on the boundary of the valid region
- Third outside the region but between the *i values of the other* two points.



Correct test

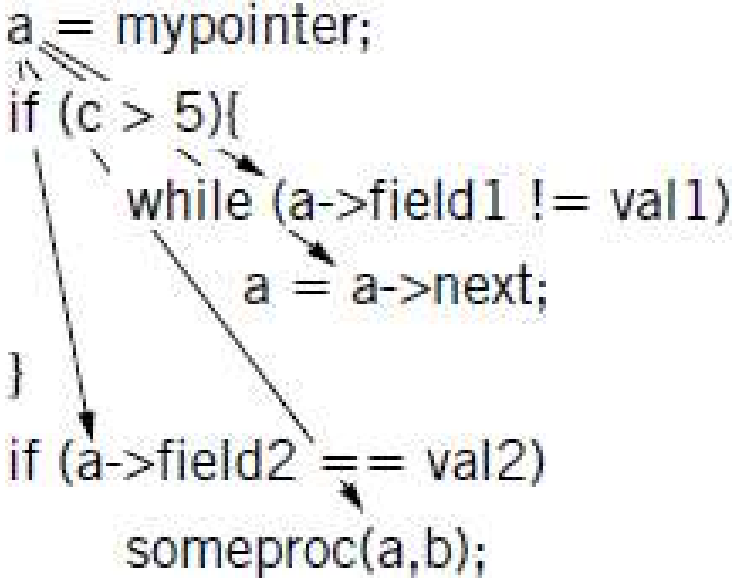




Incorrect tests

Data flow testing

- It use of def-use analysis (definition-use analysis).
- It selects paths that have some relationship to the program's function.
- Compilers→ which use def-use analysis for Optimization.
- A variable's value is defined when an assignment is made to the variable.
- It is used when it appears on the right side of an assignment.



8.2)Block Box Testing

- Black-box tests are generated without knowledge of the code being tested.
- It have a low probability of finding all the bugs in a program.
- We can't test every possible input combination, but some rules help us select reasonable sets of inputs.

1. Random Tests

- Random values are generated with a given inputs.
- The expected values are computed first, and then the test inputs are applied.

2. Regression Tests

- When tests are created during earlier or previous versions of the system.
- Those tests should be saved→ apply to the later versions of the system.
- It simply exercise current version of the code and possibly exercise different bugs.
- In digital signal processing systems→ Signal processing algorithms are implemented to save hardware costs.
- Data sets can be generated for the numerical accuracy of the system.
- These tests can often be generated from the original formulas without reference to the source code.

UNIT IV
REAL TIME SYSTEMS

Structure of a Real Time System - Estimating program run times – Task Assignment and Scheduling – Fault Tolerance Techniques – Reliability, Evaluation – Clock Synchronization.

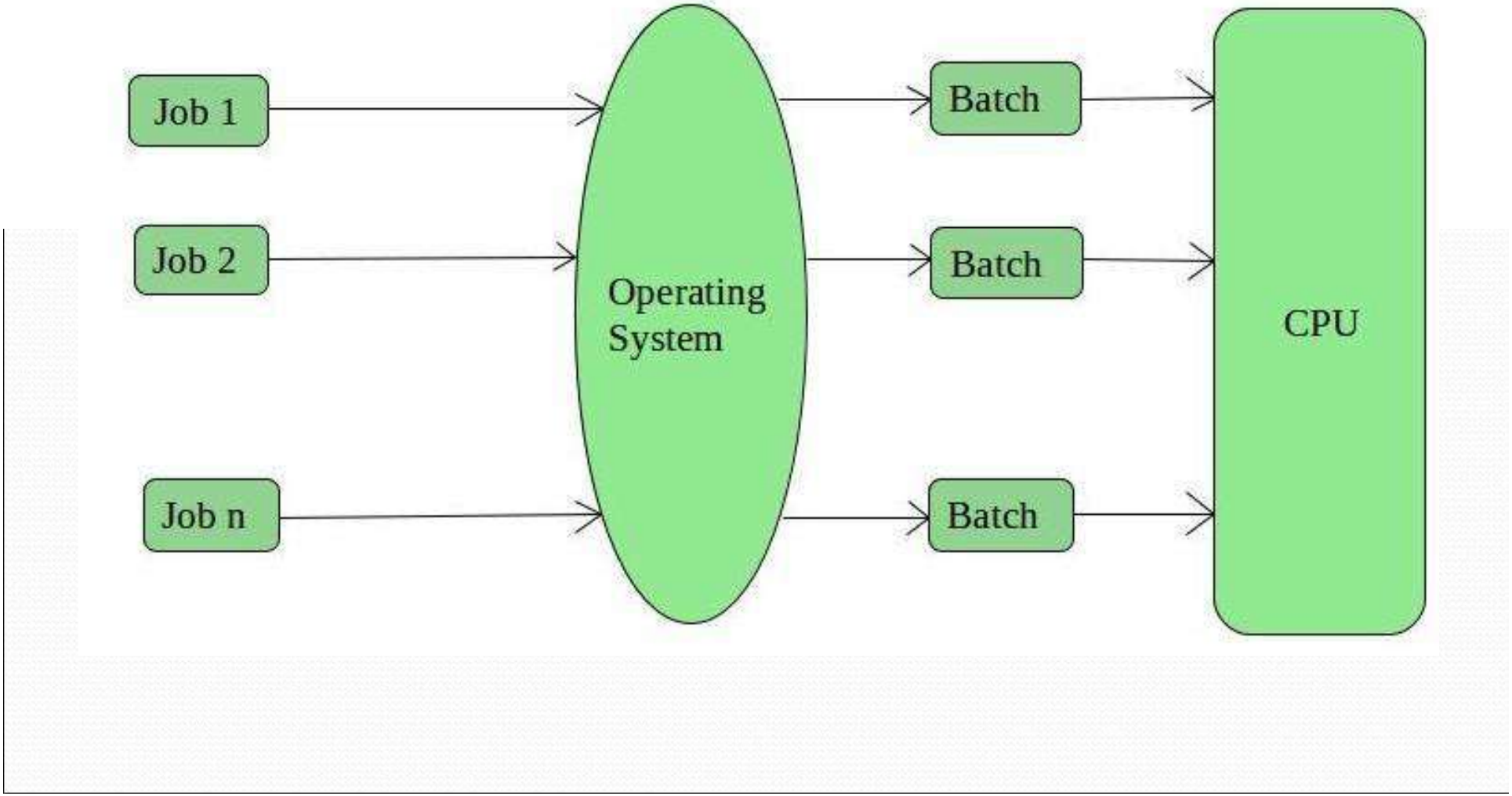
Operating System

- An Operating System performs all the basic tasks like managing file, process, and memory.
- Thus operating system acts as manager of all the resources, i.e. resource manager.
- Thus operating system becomes an interface between user and machine.

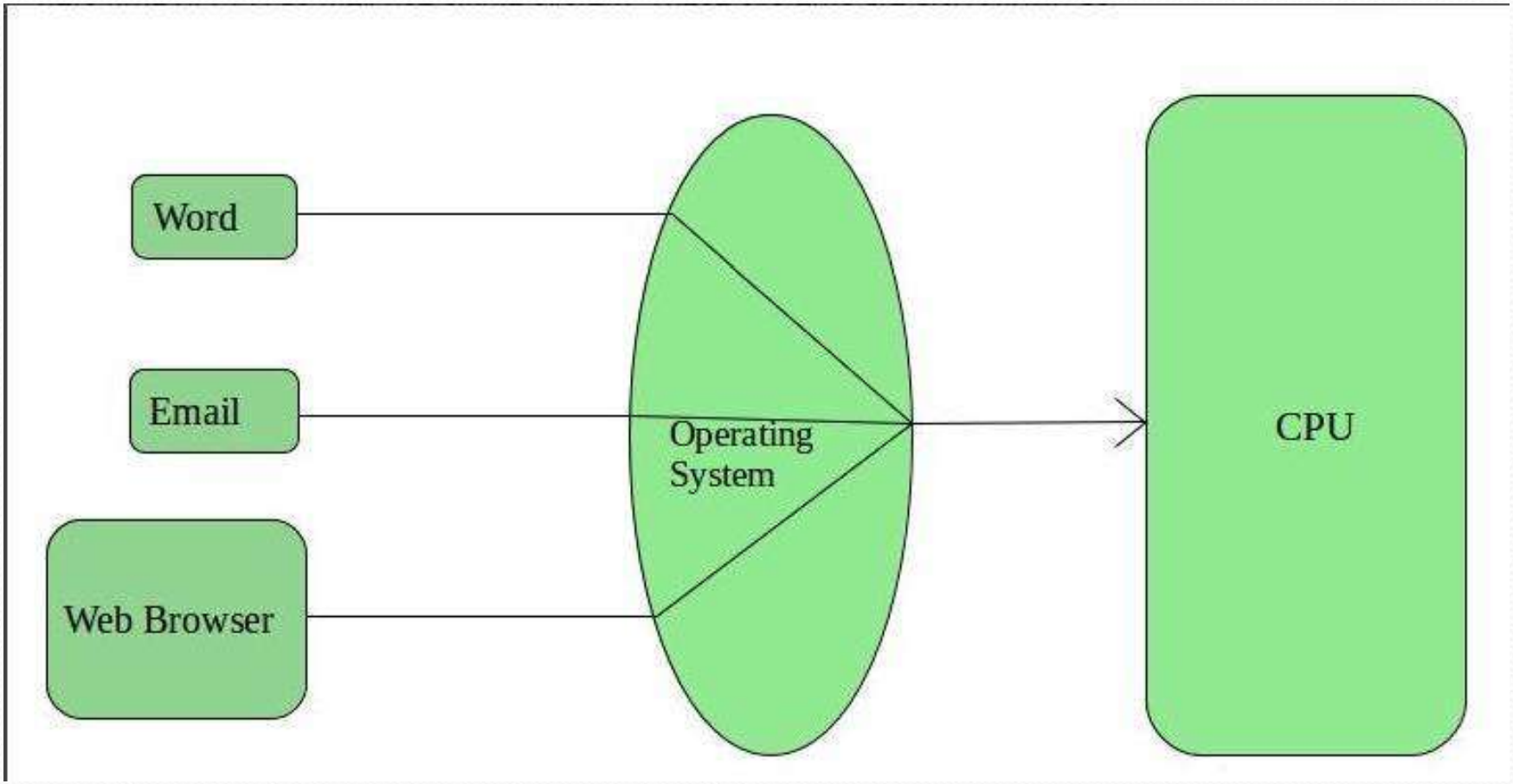
Types of Operating Systems

- **Batch Operating System**
- **Time-Sharing Operating Systems**
- **Distributed Operating System**
- **Network Operating System**
- **Real-Time Operating System**

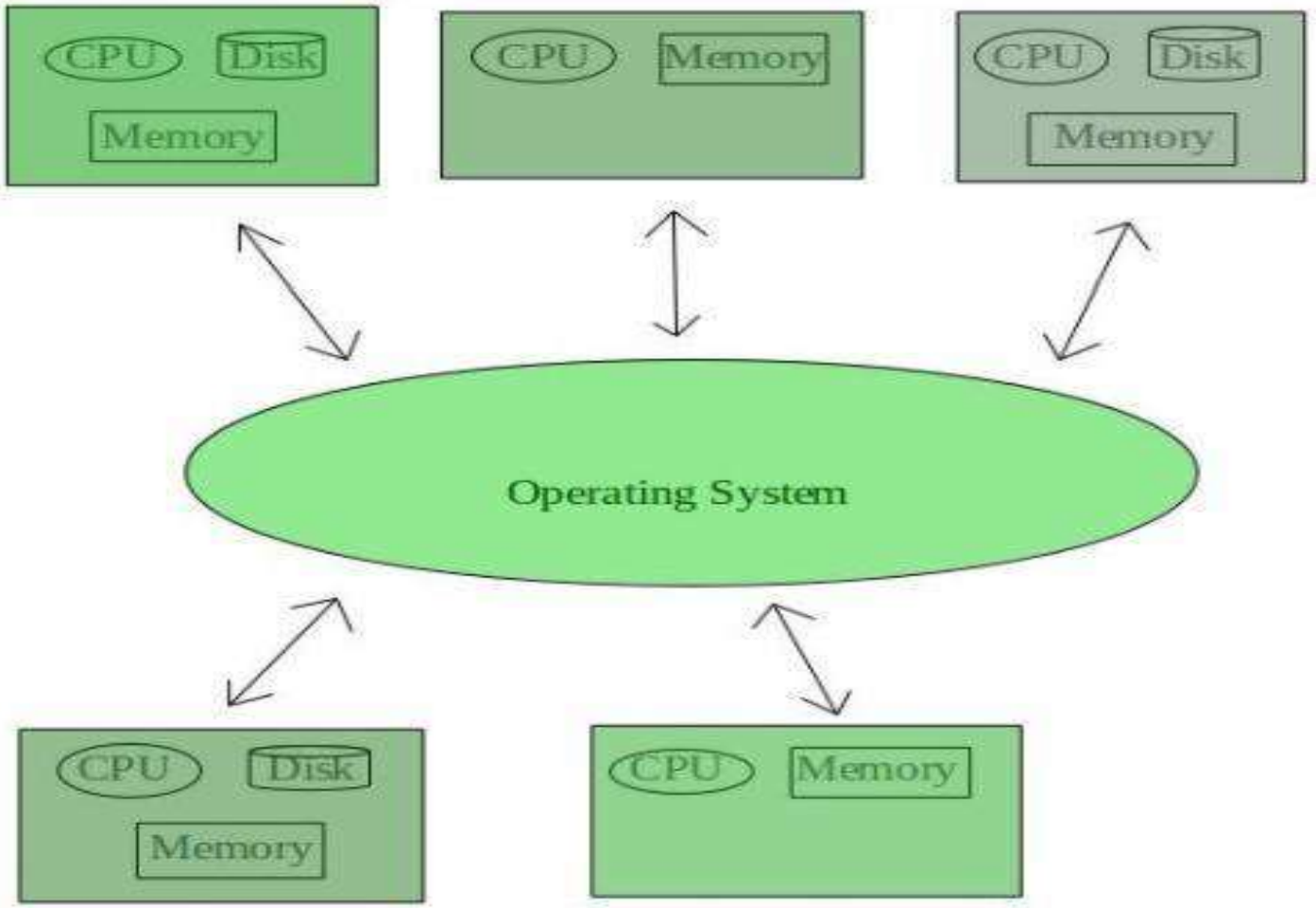
1. Batch Operating System



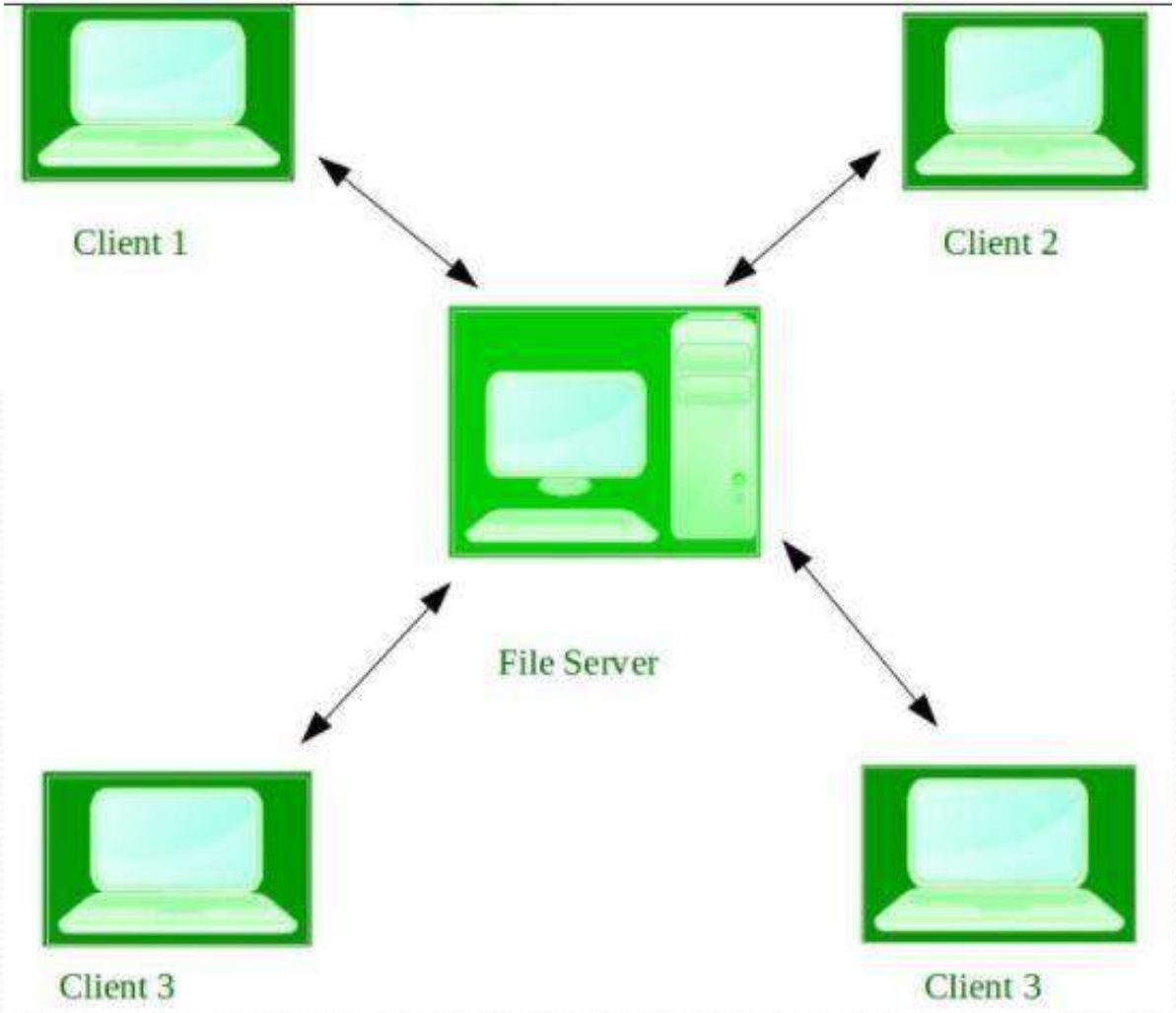
Time-Sharing Operating Systems



Distributed Operating System



Network Operating System



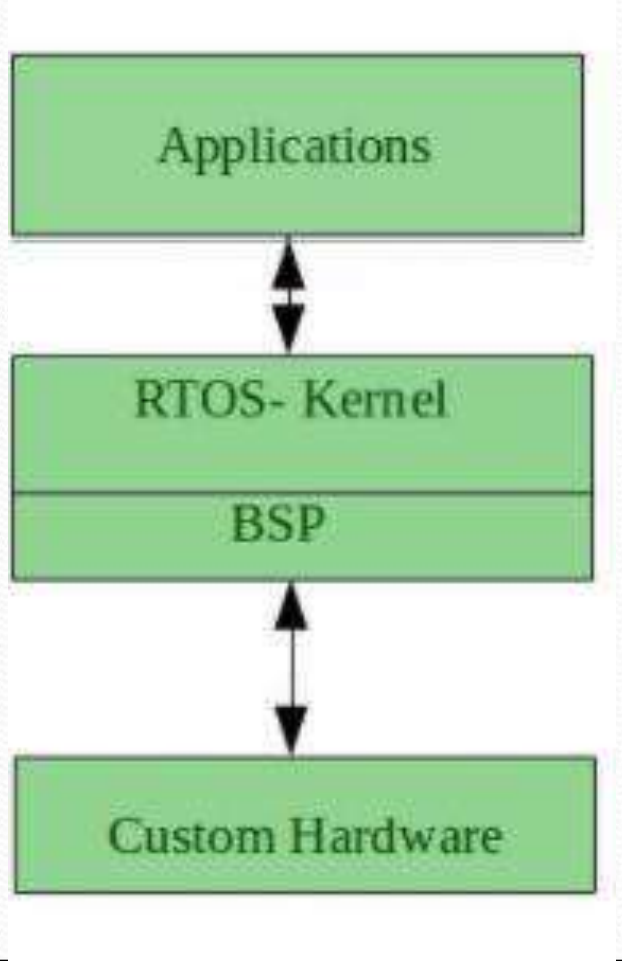
Real-Time Operating System

- These types of OSs serves the real-time systems. The time interval required to process and respond to inputs is very small. This time interval is called **response time**.

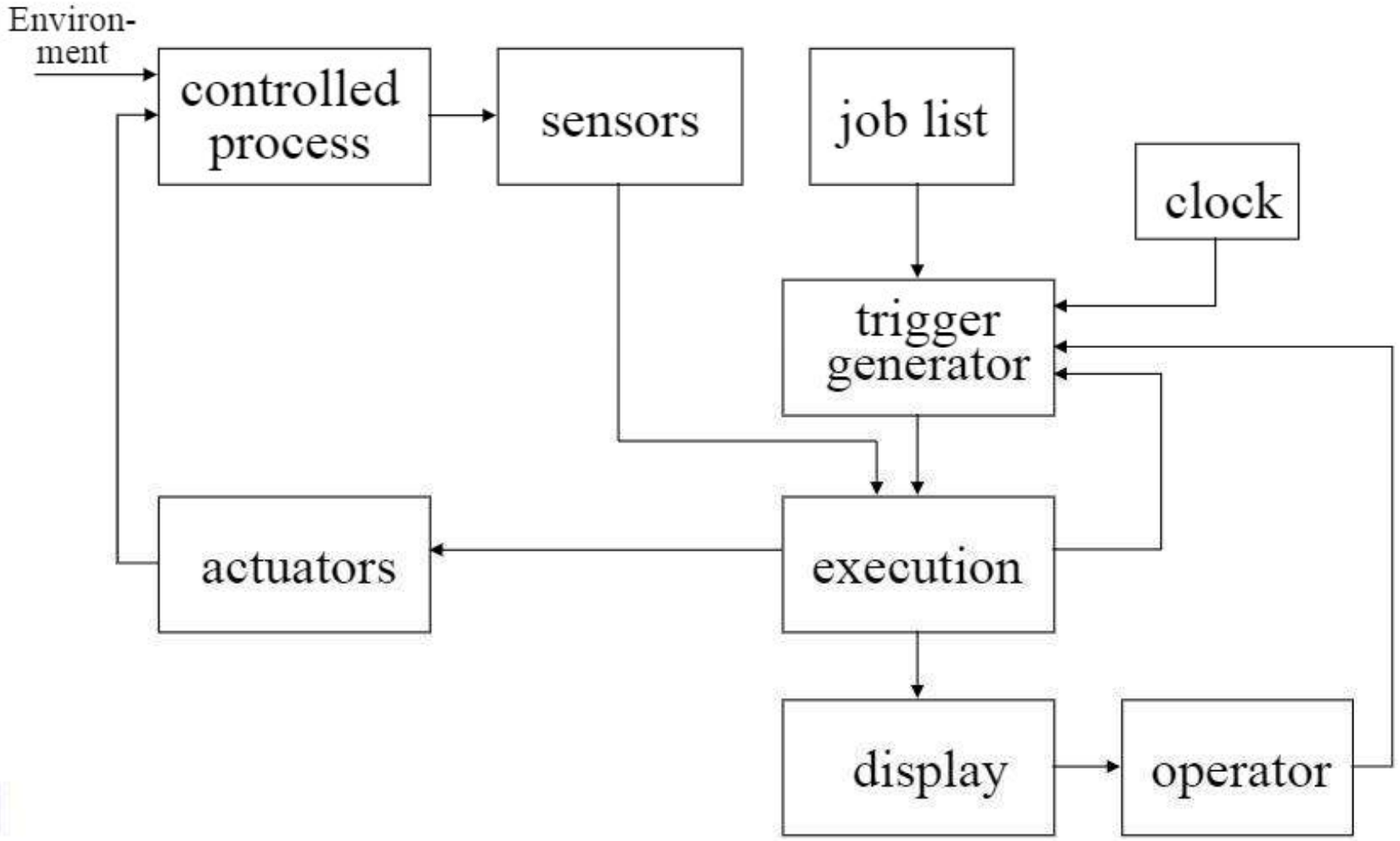
- **Real-time systems** are used when there are time requirements are very strict like missile systems, air traffic control systems, robots etc
- **Two types of Real-Time Operating System which are as follows:**
  - **Hard Real-Time Systems:**
  - **Soft Real-Time Systems:**

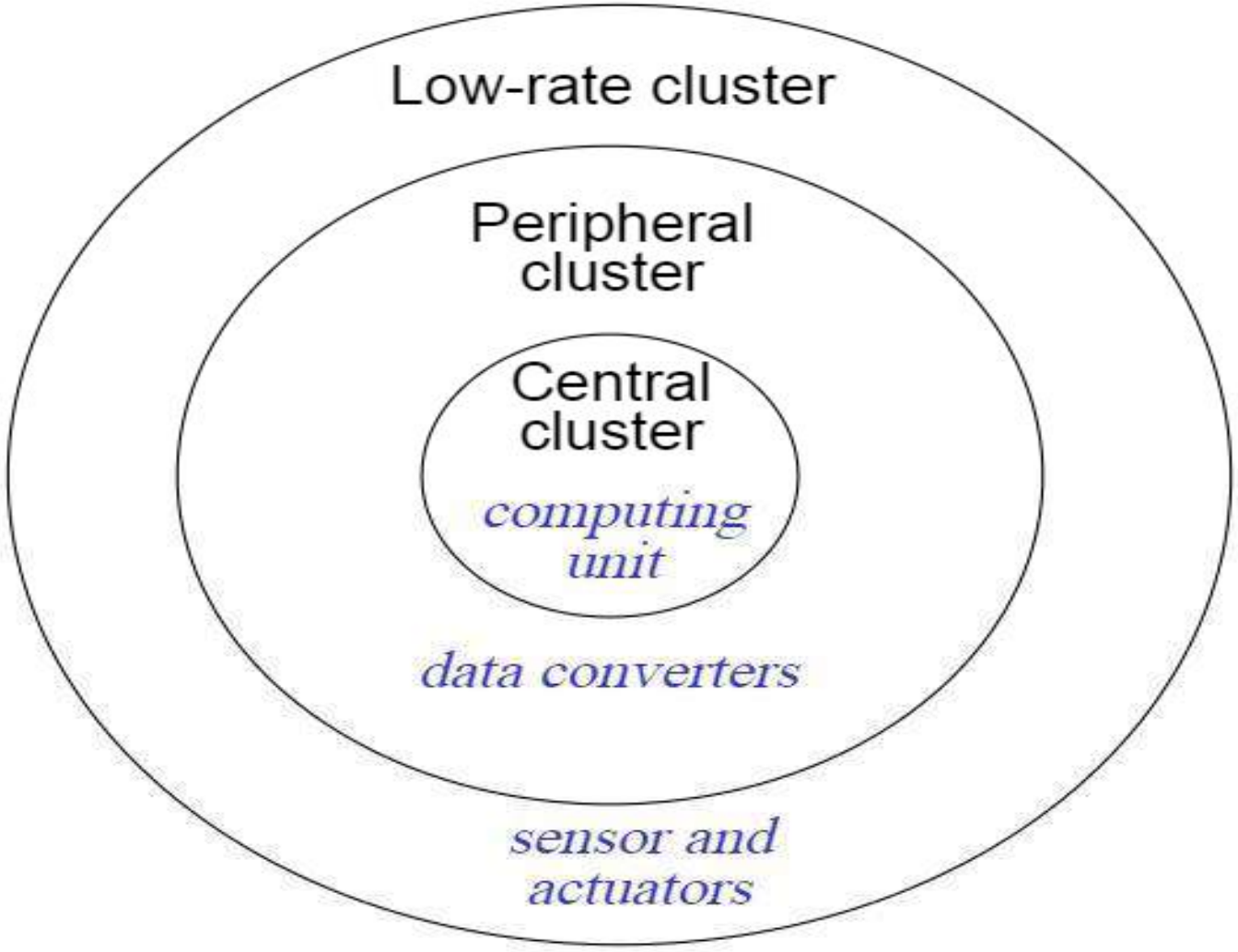
- **Hard Real-Time Systems:**
  These OSs are meant for the applications where time constraints are very strict and even the shortest possible delay is not acceptable.
- These systems are built for saving life like automatic parachutes or air bags which are required to be readily available in case of any accident.
- Virtual memory is almost never found in these systems.

- **Soft Real-Time Systems:**
  These OSs are for applications where for time-constraint is less strict.



Structure of a Real Time System


Environ-
ment
controlled process
sensors
job list
clock
trigger generator
actuators
execution
display
operator


Low-rate cluster
Peripheral cluster
Central cluster
computing unit
data converters
sensor and actuators

4.2 Estimating Program Run Times

- Real time system meet deadlines, it is important to be able to accurately estimate program run times.
- Estimating the executing time of any given program is a very difficult task
- It depend on the following factors
  - Source code
  - Compiler-Mapping should be depend on the compiler used.
  - Machine architecture
  - Operating system

- Analysis of a source code
  - L1: a = b x c;
  - L2: b = d + e;
  - L3: d = a – f;

$$\sum_{}^{3} T_{exec}(L_i)$$

L1.1	Get the address of *c*
L1.2	Load *c*
L1.3	get the address of *b*
L1.4	Load *b*
L1.5	Multiply
L1.6	Store into *a*

- Example 2

```
L4. While (p) do
L5.          Q1;
L6.          Q2;
L7.          Q3;
L8.       End While;
```

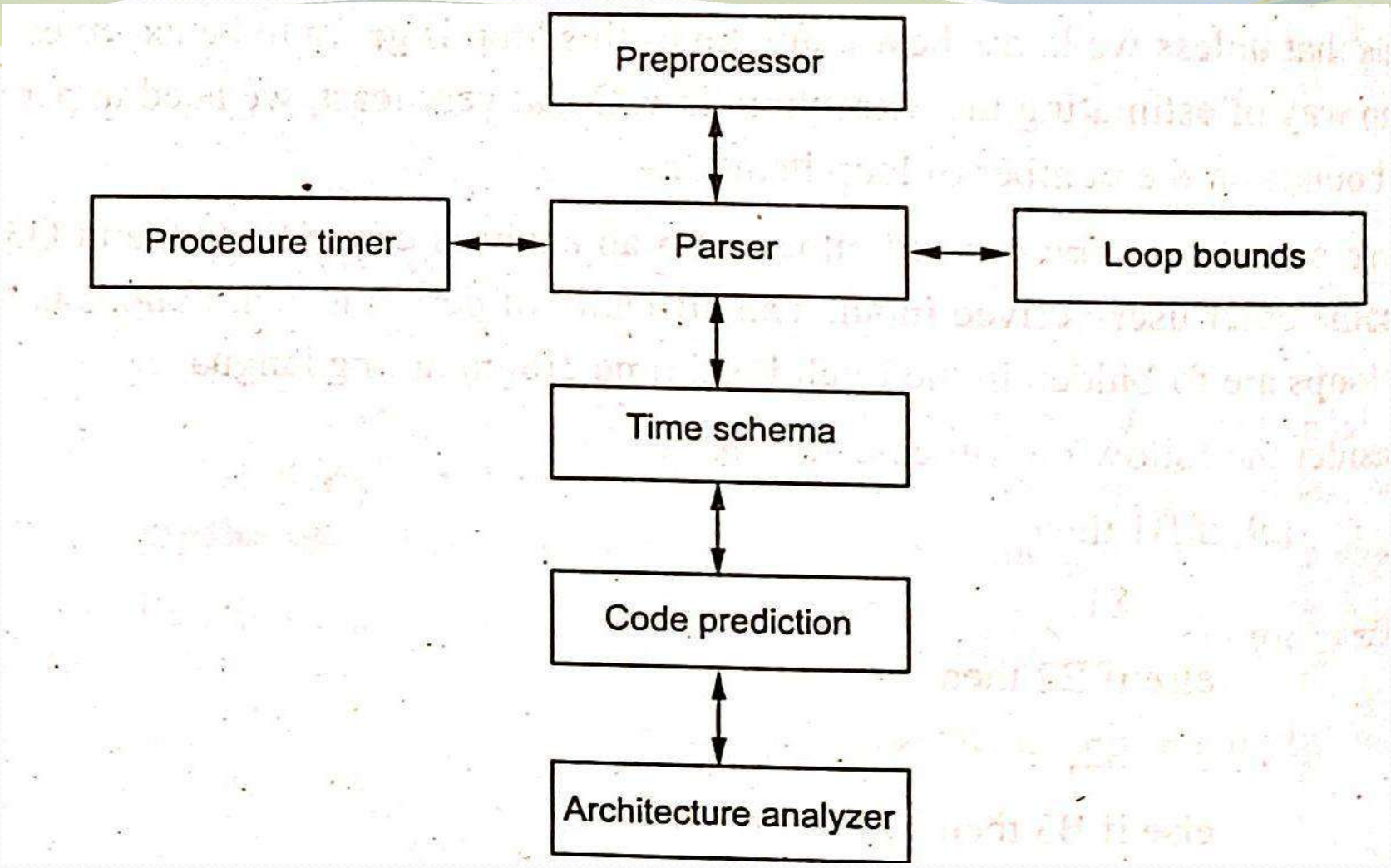
```
L9, if B1 then
        S1;
    else if B2 then
        S2;
    else if B3 then
        S3;
    else
        S4;
    end if;
```

In the case where B1 is true, the execution time is

$$T(B1) + T(S1) + T(JMP)$$

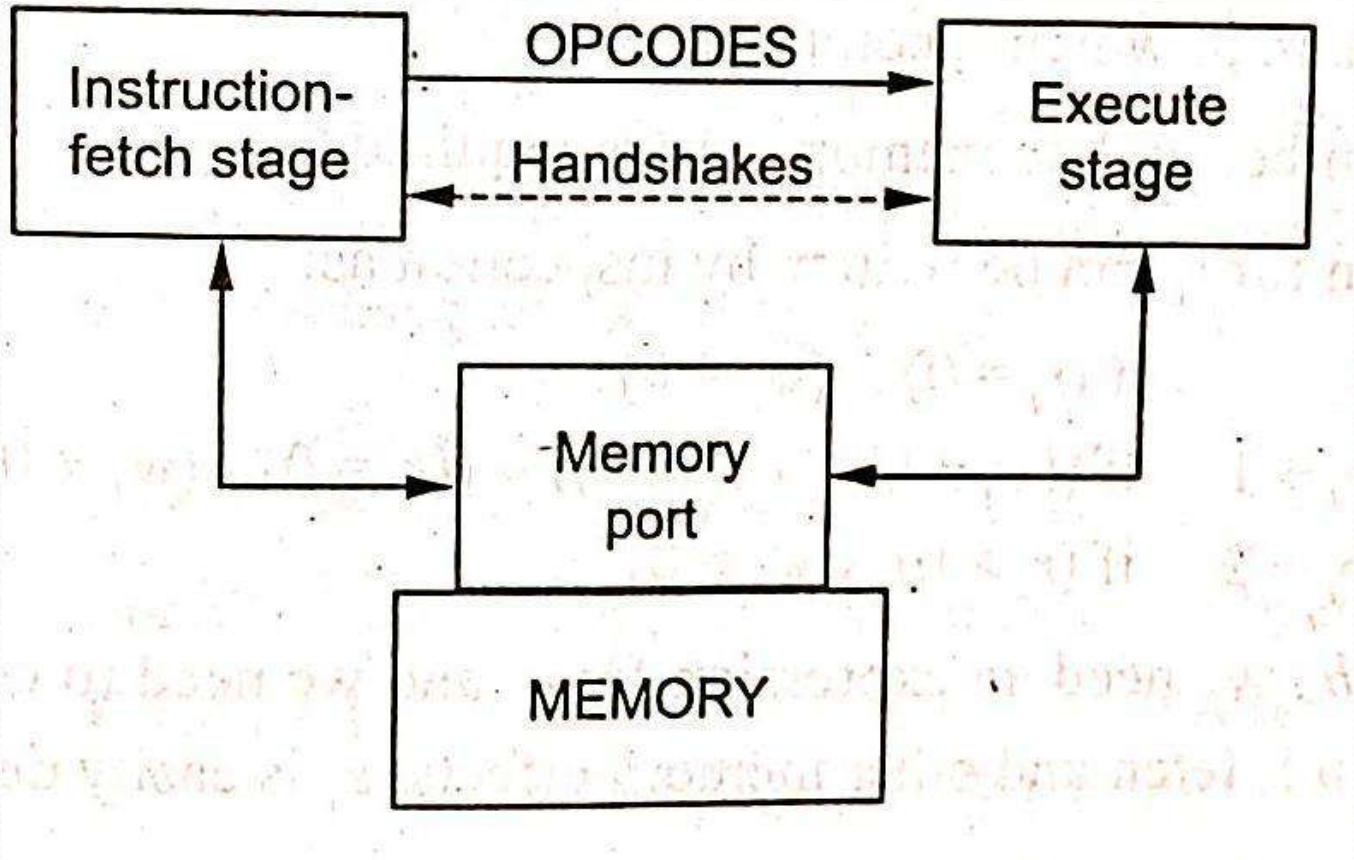
In the case where B1 is false but B2 is true, the execution time is

$$T(B1) + T(B2) + T(S2) + T(JMP)$$



Schematic of a timing estimation system

- Accounting of Pipeline



Two Stage pipeline

$$t_i = \begin{cases} b_i & \text{if } (r_i = 0) \wedge (w_i = 0) \\ b_i + 1 & \text{if } ((r_i \neq 0) \wedge (w_i = 0)) \vee ((r_i = 0) \wedge (w_i \neq 0)) \\ b_i + 2 & \text{if } (r_i > 0) \wedge (w_i > 0) \end{cases}$$

We can

$$b_i = \begin{cases} e_i + m\,(v_i - f_{i-1}) - h_{i-1} & \text{if case b1 applies} \\ e_i & \text{if Case b2.1 applies} \\ e_i + m - h_{i-1} & \text{if case b2.2 applies} \end{cases}$$

- Cache Memory
- Virtual Memory

4.3 Task Assignment and Scheduling

- A Task requires some execution time on a processor
- Also a task may required certain amount of memory or access to a bus
- Sometimes a resource must be exclusively held by a task
- In other cases resource may be exclusive or non exclusive depending on the operation to be performed on it

- Release Time
  - A task is a time at which all the data that are required to begin executing the Task are available
- Deadline
  - The deadline is the time by which the task must complete its execution
  - The deadline must be hard or soft
  - Task are classified as
    - Periodic
    - Sporadic
    - Aperiodic

- **Periodic**

  A task ti is periodic if it is released periodically. say every pi seconds pi is called the period of task Ti

- **Sporadic Task**
  - Sporadic task is a not periodic task, but may be invoked at irregular interval
  - Sporadic tasks are characterized by an upper bound on the rate at which they may be invoked

- **APeriodic Task**
- Tasks to be those tasks which are not periodic and which also have no upper bound on their invocation rate

- **Task Assignment / Schedule**
  - All task starts after the release time and complete before their deadline

  - A schedule may be
    - Precomputed(Offline scheduling)
    - Dynamically(Online Scheduling)

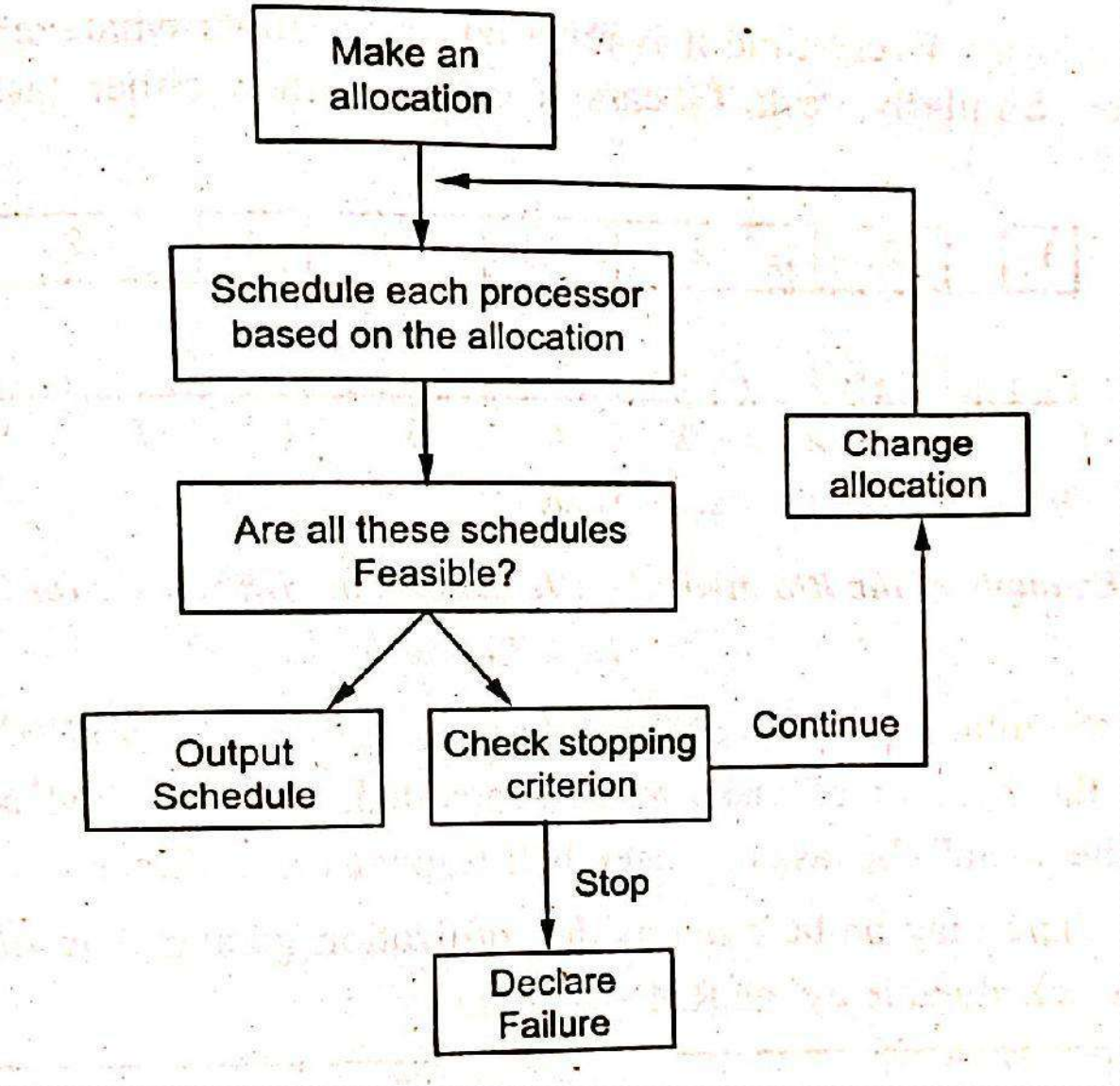
- Precomputed
  - Advance the operation with specification of periodic tasks will be run and slots for the sporadic / aperiodic tasks in the event that they are involved.

- Dynamically
  - Tasks are scheduled as they arrive in the system
  - The algorithm used in online scheduling must be fast and it takes to meet their deadlines is clearly useless
  - Two types priority algorithms are used
    - Static priority algorithm
    - Dynamic priority algorithm

- Static priority algorithm
  - Static priority algorithm assume that the task priority does not change within a mode
    - Example Rate monotonic algorithm

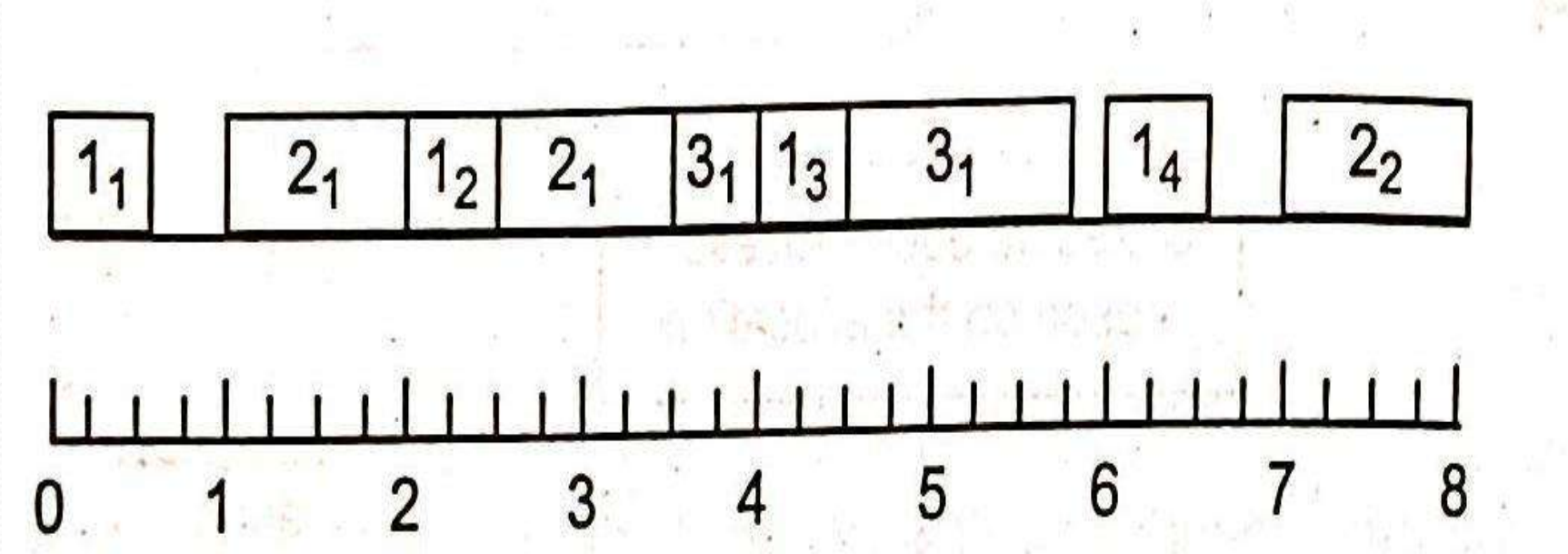
- Dynamic priority algorithm
  - algorithm assume that the task priority can change within a time
  - Example Earliest Deadline First (EDF)algorthim

- Classical uni-processor Scheduling Algorthim

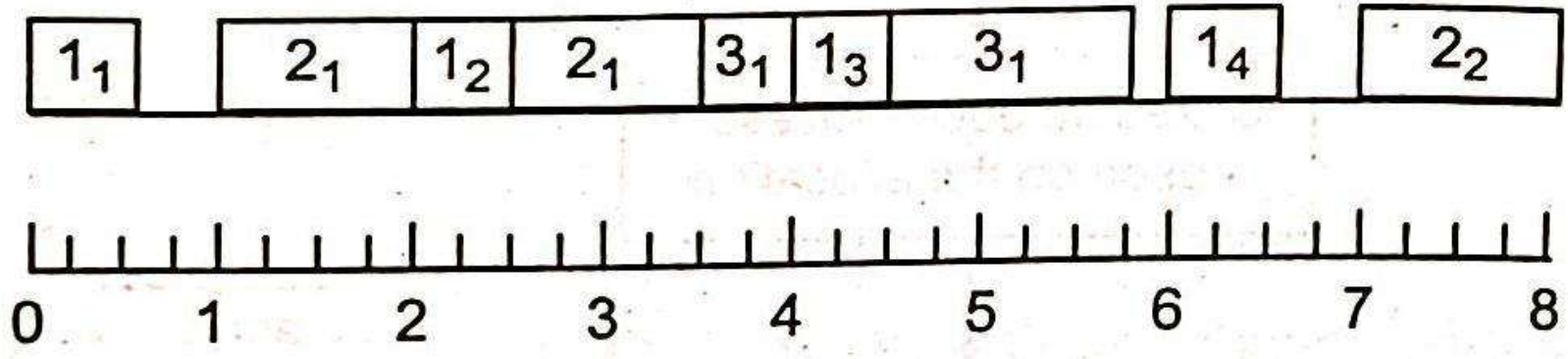


Example for Rate monotonic scheduling

There are three tasks, with $P_1 = 2$, $P_2 = 6$, $P_3 = 10$. The execution times are $e_1 = 0.5$, $e_2 = 2.0$, $e_3 = 1.75$ and $I_1 = 0$, $I_2 = 1$, $I_3 = 3$. Since $P_1 < P_2 < P_3$, task T_1 has highest priority. Every time it is released, it preempts whatever is running on the processor. Similarly, task T_3 cannot execute when either task T_1 or T_2 is unfinished.


1_1
2_1
1_2
2_1
3_1
1_3
3_1
1_4
2_2
0
1
2
3
4
5
6
7
8

There are three tasks, with $P_1 = 2$, $P_2 = 6$, $P_3 = 10$. The execution times are $e_1 = 0.5$, $e_2 = 2.0$, $e_3 = 1.75$ and $I_1 = 0$, $I_2 = 1$, $I_3 = 3$. Since $P_1 < P_2 < P_3$, task T_1 has highest priority. Every time it is released, it preempts whatever is running on the processor. Similarly, task T_3 cannot execute when either task T_1 or T_2 is unfinished.



UNIT V

PROCESSES AND OPERATING SYSTEMS

Introduction – Multiple tasks and multiple processes – Multirate systems- Preemptive real-time operating systems- Priority based scheduling- Interprocess communication mechanisms – Evaluating operating system performance- power optimization strategies for processes – Example Real time operating systems-POSIX-Windows-CE. Distributed embedded systems – MPSoCs and shared memory multiprocessors. – Design Example - Audio player, Engine control unit – Video accelerator.

INTRODUCTION

- Simple applications can be programmed on a microprocessor by writing a single piece of code.
- But for a complex application, multiple operations must be performed at widely varying times.
- Two fundamental abstractions that allow us to build complex applications on microprocessors.

1. ***Process***→ defines the state of an executing program
2. ***operating system (OS)***→provides the mechanism for switching execution between the processes.

MULTIPLE TASKS AND MULTIPLE PROCESSES

- Systems which are capable of performing multiprocessing known as multiple processor system.
- Multiprocessor system can execute multiple processes simultaneously with the help of multiple CPU.
- Multi-tasking→ The ability of an operating system to hold multiple processes in memory and switch the processor for executing one process.

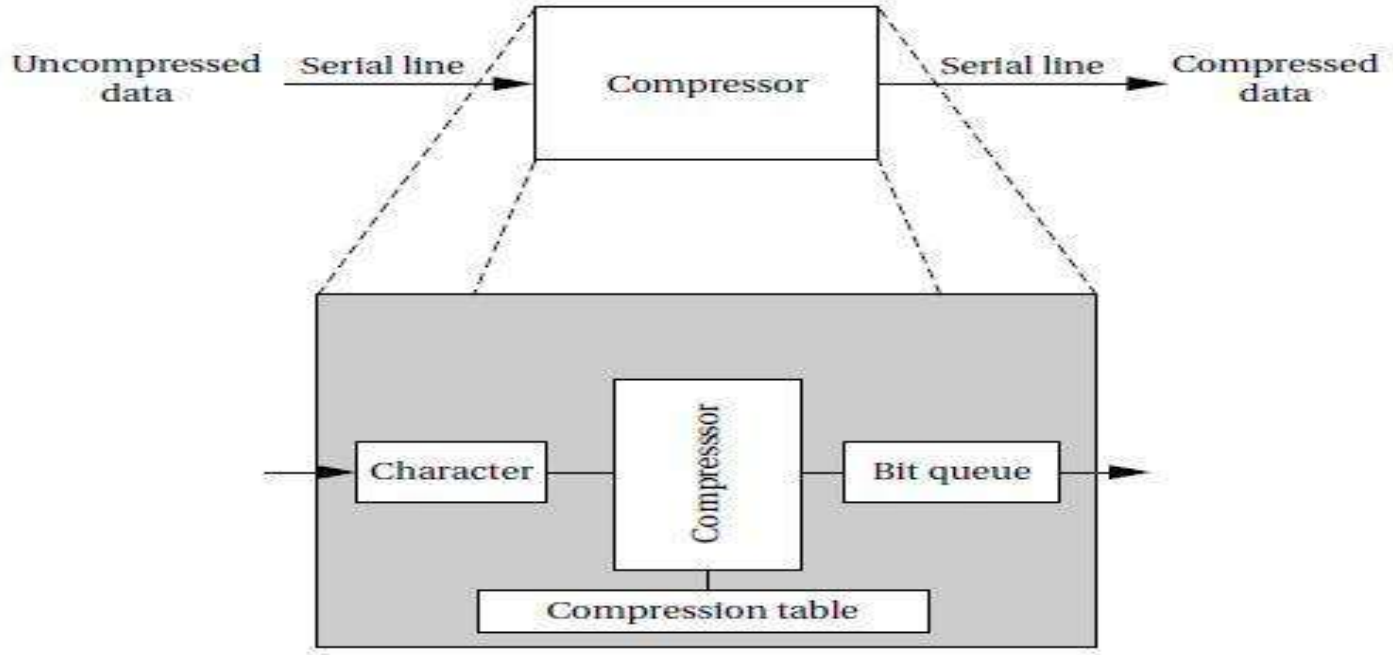
Tasks and Processes

- Task is nothing but different parts of functionality in a single system.
- Eg-Mobile Phones
- When designing a telephone answering machine, we can define recording a phone call ,answering a call and operating the user's control panel as distinct tasks, at different rates.
- Each application in a system is called a task.

Process

- A process is a single execution of a program.
- If we run the same program two different times, we have created two different processes.
- Each process has its own state that includes not only its registers but all of its memory.
- In some OSs, the memory management unit is used to keep each process in a separate address space.
- In others, particularly lightweight RTOSs, the processes run in the same address space.
- Processes that share the same address space are often called threads.

- This device is connected to serial ports on both ends.
- The input to the box is an uncompressed stream of bytes.
- The box emits a compressed string of bits, based on a compression table.



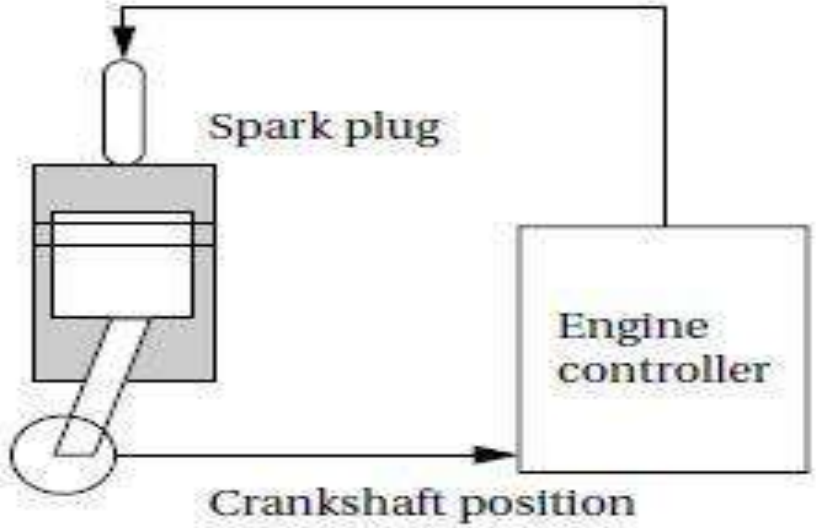
- Ex: compress data being sent to a modem.
- The program's need to receive and send data at different rates
- Eg→The program may emit 2 bits for the first byte and then 7 bits for the second byte— will obviously find itself reflected in the structure of the code.
- if we spend too much time in packaging and emitting output characters,we may drop an input character.

Asynchronous input

- Ex:A control panel on a machine provides a different type of rate.
- The control panel of the compression box include a compression mode button that disables or enables compression, so that the input text is passed through unchanged when compression is disabled.
- Sampling the button's state too slowly→ machine will miss a button depression entirely.
- Sampling it too frequently→ the machine will do incorrectly compress data.
- To solve this problem→ every n times the compression loop is executed.

Multi-rate Systems

- In operating system implementing code for satisfies timing requirements is more complex when multiple rates of computation must be handled.
- Multirate embedded computing systems→Ex: automobile engines, printers, and cell phones.
- In all these systems, certain operations must be executed periodically with its own rate.
- Eg→Automotive engine control



- The simplest automotive engine controllers, such as the ignition controller for a basic motorcycle engine, perform only one task—timing the firing of the spark plug, which takes the place of a mechanical distributor.

Spark Plug

- The spark plug must be fired at a certain point in the combustion cycle.

Microcontroller

- Using a microcontroller that senses the engine crankshaft position allows the spark timing to vary with engine speed.
- Firing the spark plug is a periodic process.

Engine controller

- Automobile engine controllers use additional sensors, including the gas pedal position and an oxygen sensor used to control emissions.
- They also use a multimode control scheme. one mode may be used for engine warm-up, another for cruise, and yet another for climbing steep hills.
- The engine controller takes a variety of inputs that determine the state of the engine.
- It then controls two basic engine parameters: the spark plug firings and the fuel/air mixture.

Task performed by engine controller unit

Variable	Time to move full range (ms)	Update period (ms)
Engine spark timing	300	2
Throttle	40	2
Airflow	30	4
Battery voltage	80	4
Fuel flow	250	10
Recycled exhaust gas	500	25
Set of status switches	100	50
Air temperature	seconds	500
Barometric pressure	seconds	1000
Spark/dwell	10	1
Fuel adjustments	80	4
Carburetor adjustments	500	25
Mode actuators	100	100

Timing Requirements on Processes

- Processes can have several different types of timing requirements based on the application.
- The timing requirements on a set of processes strongly depends on the type of scheduling.
- A scheduling policy must define the timing requirements that it uses to determine whether a schedule is valid.

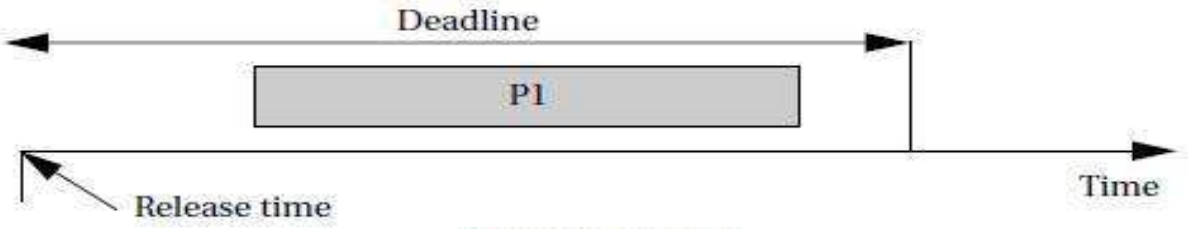
1. Release time→

- The time at which the process becomes ready to execute.
- simpler systems→ the process may become ready at the beginning of the period.
- sophisticated systems→ set the release time at the arrival time of certain data, at a time after the start of the period.

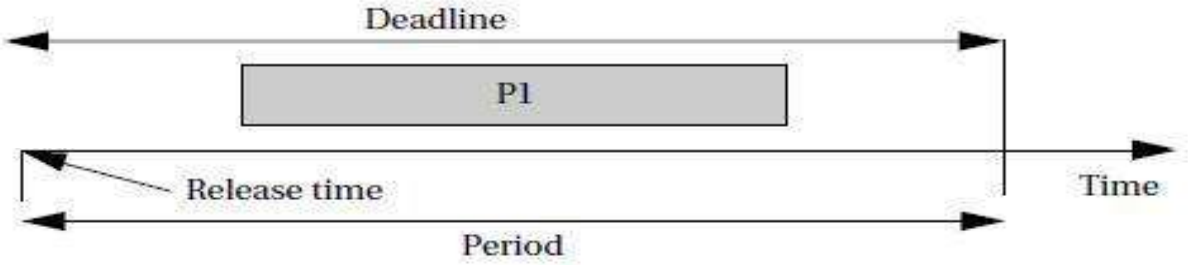
2. Deadline

- specifies when a computation must be finished.
- The deadline for an a periodic process is generally measured from the release time or initiation time.
- The deadline for a periodic process may occur at the end of the period.
- The period of a process is the time between successive executions.
- The process's rate is the inverse of its period.
- In a Multi rate system, each process executes at its own distinct rate.

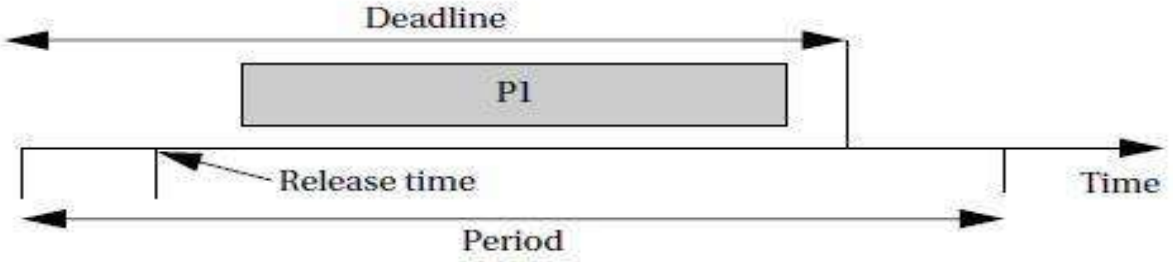
Example definitions of release times and deadlines



Aperiodic process

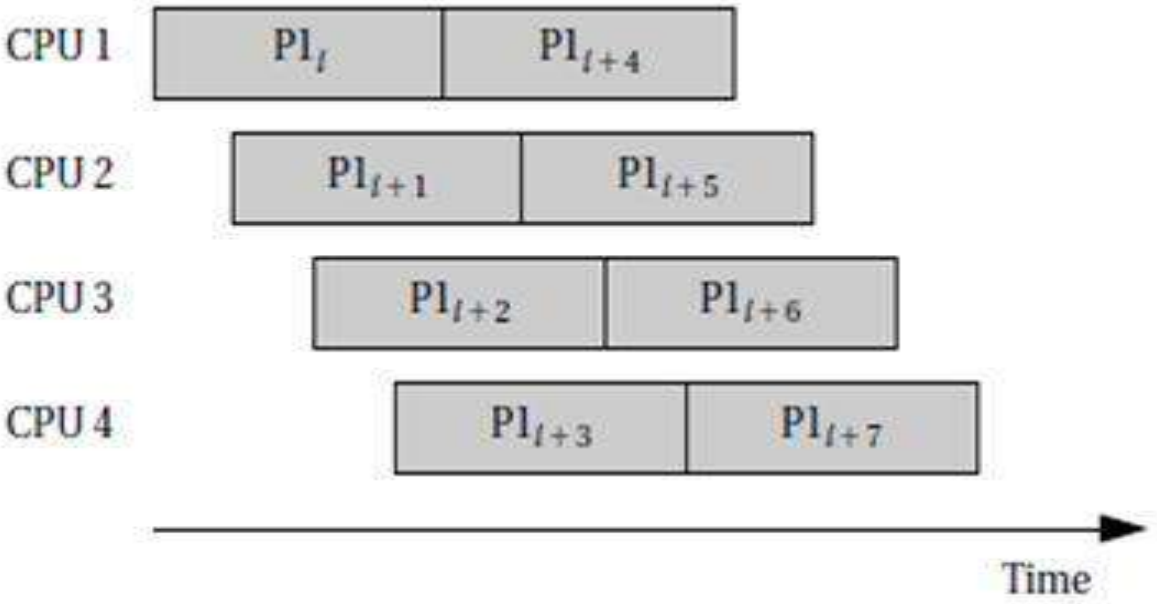


Periodic process initiated at start of period



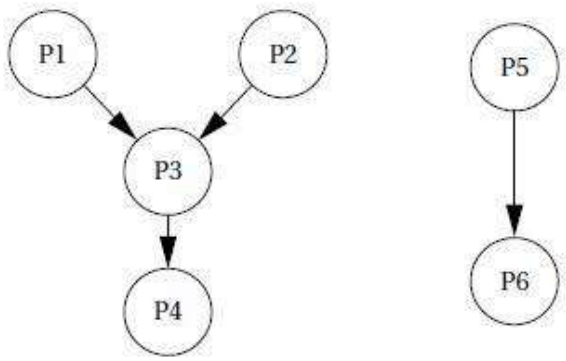
Periodic process released by event

A sequence of processes with a high initiation rate



- In this case, the initiation interval is equal to one fourth of the period.
- It is possible for a process to have an initiation rate less than the period even in single-CPU systems.
- If the process execution time is less than the period, it may be possible to initiate multiple copies of a program at slightly offset times.

Data dependencies among processes

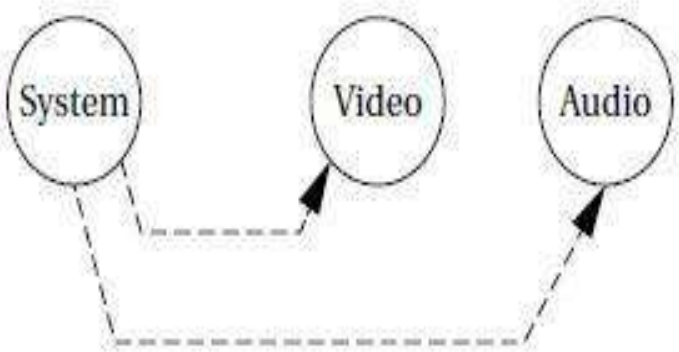


- The data dependencies define a partial ordering on process execution.
- P1 and P2 can execute in any order but must both complete before P3, and P3 must complete before P4.
- All processes must finish before the end of the period.

Directed Acyclic Graph (DAG)

- It is a directed graph that contains no cycles.
- The data dependencies must form a directed acyclic graph.
- A set of processes with data dependencies is known as a task graph.

Communication among processes at different rates (MPEG audio/Video)



- The system decoder process demultiplexes the audio and video data and distributes it to the appropriate processes.
- Missing Deadline
- Missing deadline in a multimedia system may cause an audio or video glitch.
- The system can be designed to take a variety of actions when a deadline is missed.

CPU Metrics

- CPU metrics are described by initiation time and completion time.
- Initiation time→It is the time at which a process actually starts executing on the CPU.
- Completion time→It is the time at which the process finishes its work.
- The CPU time of process i is called Ci .
- The CPU time is not equal to the completion time minus initiation time.
- The total CPU time consumed by a set of processes is

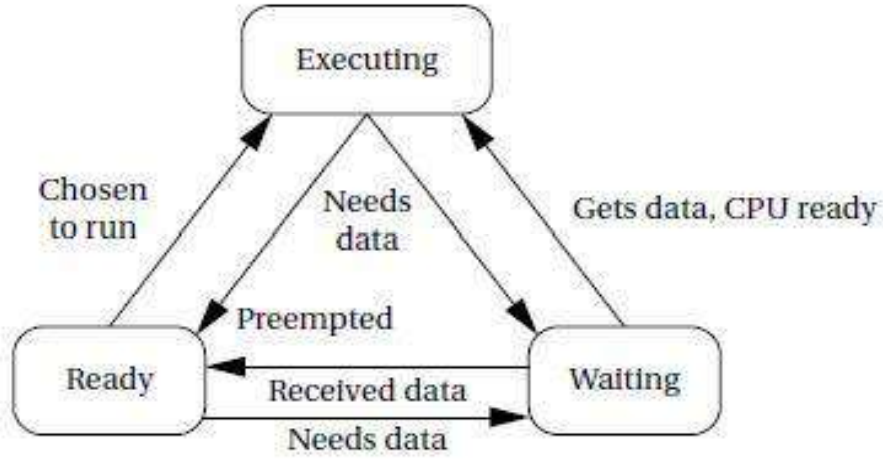
$$T = \sum_{1 \le i \le n} T_i.$$

- The simplest and most direct measure is utilization.

$$U = \frac{\text{CPU time for useful work}}{\text{total available CPU time}}.$$

Process State and Scheduling

- The first job of the OS is to determine that process runs next.
- The work of choosing the order of running processes is known as scheduling.
- There three basic scheduling ,such as waiting, ready and executing.



- A process goes into the waiting state when it needs data that it has finished all its work for the current period.
- A process goes into the ready state when it receives its required data, when it enters a new period.
- Finally a process can go into the executing state only when it has all its data, is ready to run, and the scheduler selects the process as the next process to run.

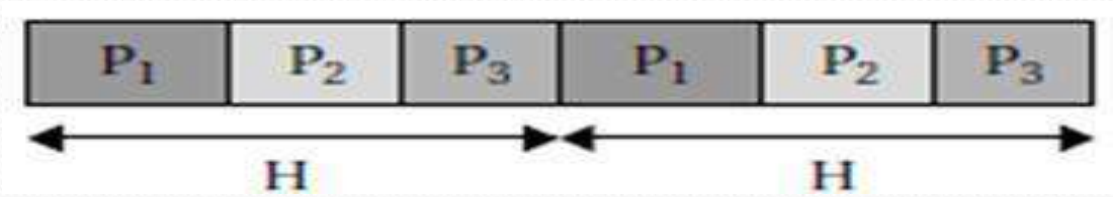
Scheduling Policies

- A scheduling policy defines how processes are selected for promotion from the ready state to the running state.
- Scheduling→Allocate time for execution of the processes in a system .
- For periodic processes, the length of time that must be considered is the hyper period, which is the least-common multiple of the periods of all the processes.
- Unrolled schedule →The complete schedule for the least-common multiple of the periods.

Types of scheduling

1. Cyclostatic scheduling or Time Division Multiple Access scheduling

- Schedule is divided into equal-sized time slots over an interval equal to the length of the hyperperiod *H*. (run in the same time slot)

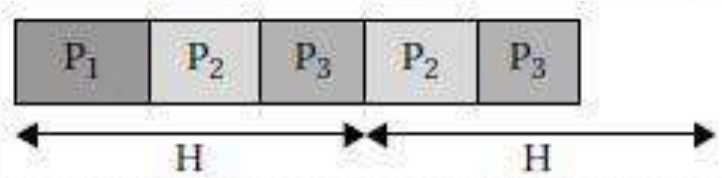


Two factors affect this scheduling

- The number of time slots used
- The fraction of each time slot that is used for useful work.

2)Round Robin-scheduling

- Uses the same hyper period as does cyclostatic.
- It also evaluates the processes in order.
- If a process does not have any useful work to do, the scheduler moves on to the next process in order to fill the time slot with useful work.



- All three processes execute during the first hyperperiod.
- During the second one, P1 has no useful work and is skipped so P3 is directly move on to the next process.

Scheduling overhead

- The execution time required to choose the next execution process, which is incurred in addition to any context switching overhead.

To calculate the utilization of CPU

Utilization of a set of processes

We are given three processes, their execution times, and their periods:

Process	Period	Execution time
P1	1.0×10^{-3}	1.0×10^{-4}
P2	1.0×10^{-3}	2.0×10^{-4}
P3	5.0×10^{-3}	3.0×10^{-4}

The least common multiple of these periods is 5×10^{-3} s.

We can now determine the utilization over the hyperperiod:

$$U = \frac{5.1 \times 10^{-4} + 5.2 \times 10^{-4} + 1.3 \times 10^{-4}}{5 \times 10^{-3}} = 0.36$$

This is well below our maximum utilization of 1.0.

Preemptive Real-Time Operating Systems(RTOS)

- A pre emptive OS →solves the fundamental problem in multitasking system.
- It executes processes based upon timing requirements provided by the system designer.
- To meet timing constraints accurately is to build a preemptive OS and to use priorities to control what process runs at any given time.

Preemption

- Preemption is an alternative to the C function call to control execution.
- To be able to take full advantage of the timer, change the process as something more than a function call.
- Break the assumptions of our high-level programming language.
- Create new routines that allow us to jump from one subroutine to another at any point in the program.
- The timer, will allow us to move between functions whenever necessary based upon the system's timing constraints.

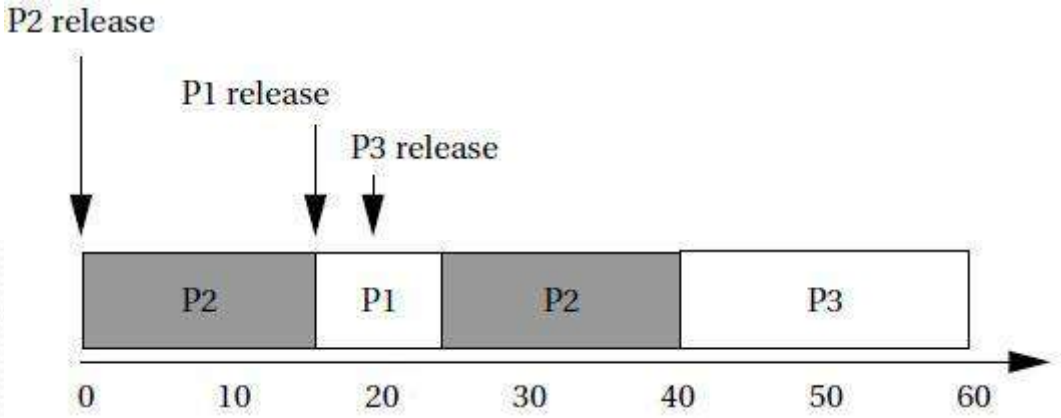
Kernel

- It is the part of the OS that determines what process is running.
- The kernel is activated periodically by the timer.
- It determines what process will run next and causes that process to run.

Priorities

- Based on the priorities →kernel can do the processes sequentially.
- which ones actually want to execute and select the highest priority process that is ready to run.
- This mechanism is both flexible and fast.
- The priority is a non-negative integer value.

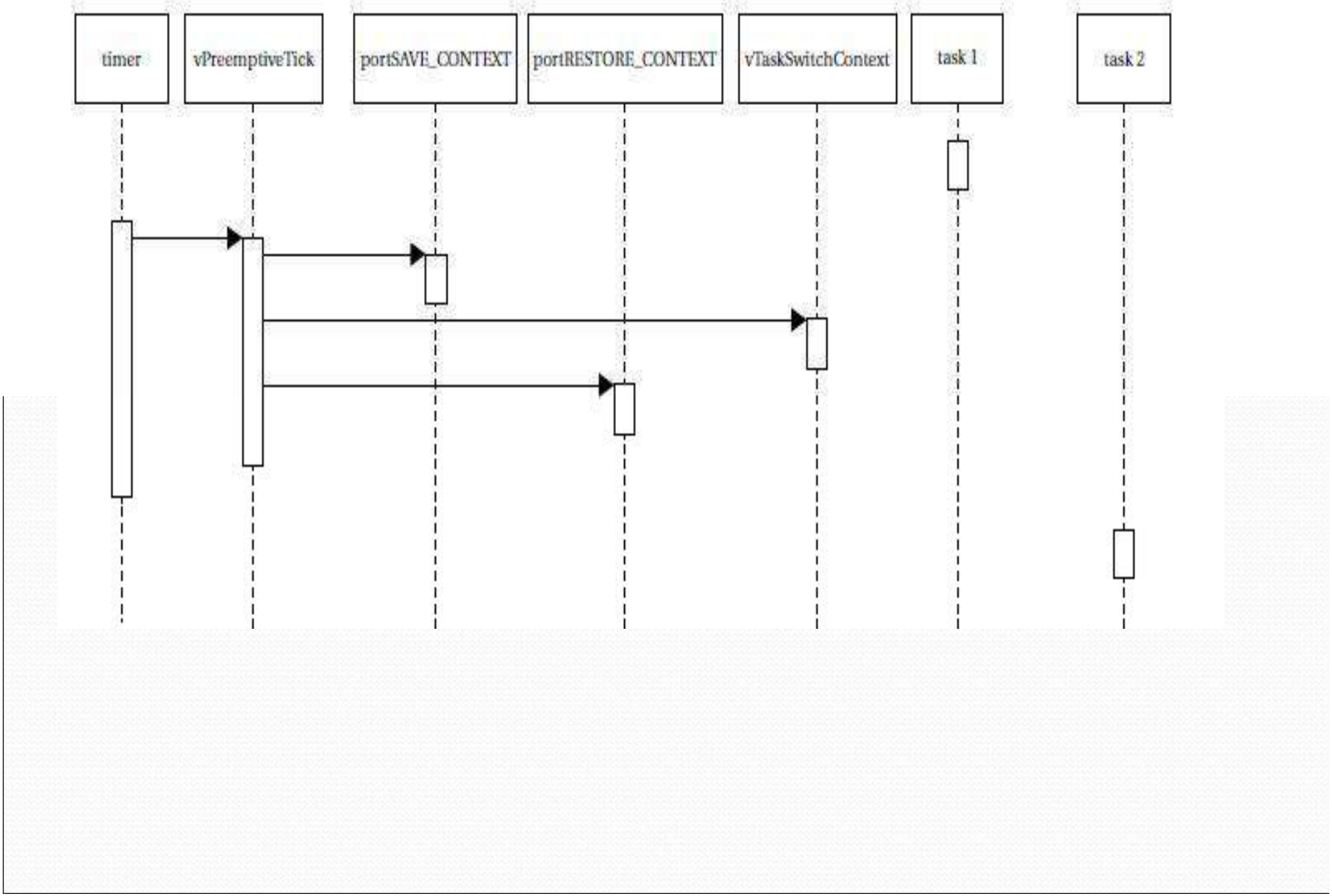
Process	Priority	Execution time
P1	1	10
P2	2	30
P3	3	20



•When the system begins execution,P2 is the only ready process, so it is selected for execution.
•At T=15, P1 becomes ready; it preempts P2 because p1 has a higher priority, so it execute immediately
•P3's data arrive at time 18, it has lowest priority.
•P2 is still ready and has higher priority than P3.
•Only after both P1 and P2 finish can P3 execute

- 5.4.4) Context Switching

- To understand the basics of a context switch, let's assume that the set of tasks is in steady state.
- Everything has been initialized, the OS is running, and we are ready for a timer interrupt.
- This diagram shows the application tasks, the hardware timer, and all the functions in the kernel that are involved in the context switch.
- vPreemptiveTick() → it is called when the timer ticks.
- portSAVE_CONTEXT()→ swaps out the current task context.
- vTaskSwitchContext ( ) →chooses a new task.
- portRESTORE_CONTEXT()→ swaps in the new context


timer
vPreemptiveTick
portSAVE_CONTEXT
portRESTORE_CONTEXT
vTaskSwitchContext
task 1
task 2

PRIORITY-BASED SCHEDULING

- Operating system is to allocate resources in the computing system based on the priority.
- After assigning priorities, the OS takes care of the rest by choosing the highest-priority ready process.
- There are two major ways to assign priorities.
- **Static priorities→** that do not change during execution
- **Dynamic priorities→** that do change during execution
- Types of scheduling process

1. Rate-Monotonic Scheduling
2. Earliest-Deadline-First Scheduling

Rate-Monotonic Scheduling(RMS)

- Rate-monotonic scheduling (RMS)→ is one of the first scheduling policies developed for real-time systems.
- RMS is a static scheduling policy.
- It assigns fixed priorities are sufficient to efficiently schedule the processes in many situations.

RMS is known as rate-monotonic analysis (RMA), as summarized below.

- All processes run periodically on a single CPU.
- Context switching time is ignored.
- There are no data dependencies between processes.
- The execution time for a process is constant.
- All deadlines are at the ends of their periods.
- The highest-priority ready process is always selected for execution.
- Priorities are assigned by rank order of period, with the process with the shortest period being assigned the highest priority.

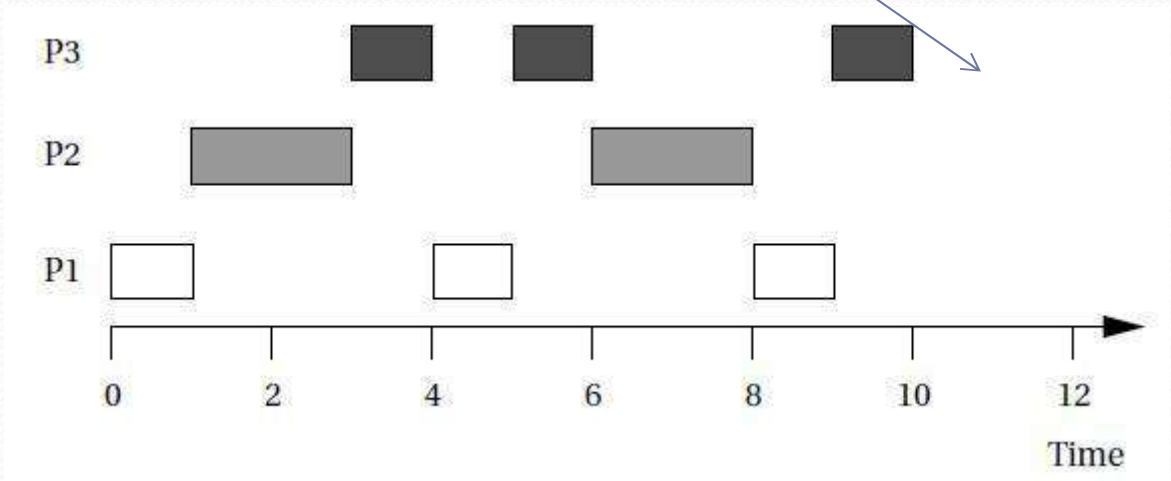
Example-Rate-monotonic scheduling

- set of processes and their characteristics

Process	Execution time	Period
P1	1	4
P2	2	6
P3	3	12

- According to RMA → Assign highest priority for least execution period.
- Hence P1 the highest priority, P2 the middle priority,and P3 the lowest priority.
- First execute P1 then P2 and finally P3.(T1>T2>T3)
- After assigning priorities, construct a time line equal in length to hyper period, which is 12 in this case.

- Every 4 time intervals P1 executes 1 units.(Execution time intervals for P1 0-4,4-8,8-12)
- Every 6 time intervals P2 executes 2 units. .(Execution time intervals for P2 0-6,6-12)
- Every 12 intervals P3 executes 3 units. .(Execution time intervals for P3 0-12)
- Time interval from 10-12 no scheduling available because no process will be available for execution. All process are executed already.



- P1 is the highest-priority process, it can start to execute immediately.
- After one time unit, P1 finishes and goes out of the ready state until the start of its next period.
- At time 1, P2 starts executing as the highest-priority ready process.
- At time 3, P2 finishes and P3 starts executing.
- P1's next iteration starts at time 4, at which point it interrupts P3.
- P3 gets one more time unit of execution between the second iterations of P1 and P2, but P3 does not get to finish until after the third iteration of P1.
- Consider the following different set of execution times.

Process	Execution time	Period
P1	2	4
P2	3	6
P3	3	12

- In this case, Even though each process alone has an execution time significantly less than its period, combinations of processes can require more than 100% of the available CPU cycles.
- During one 12 time-unit interval, we must execute P1 -3 times, requiring 6 units of CPU time; P2 twice, costing 6 units and P3 one time, costing 3 units.
- The total of 6 + 6 + 3 = 15 units of CPU time is more than the 12 time units available, clearly exceeding the available CPU capacity(12units).

RMA priority assignment analysis

- Response time→ The time at which the process finishes.
- Critical instant→The instant during execution at which the task has the largest response time.
- Let the periods and computation times of two processes P1 and P2 be τ1, τ2 and *T1, T2, with* τ *1* < τ *2.*
- let P1 have the higher priority. In the worst case we then execute P2 once during its period and as many iterations of P1 as fit in the same interval.
- Since there are τ2/ τ1 *iterations of P1 during a single* period of P2.
- The required constraint on CPU time, ignoring context switching overhead, is

$$\left\lceil \frac{\tau_2}{\tau_1} \right\rceil T_1 + T_2 \leq \tau_2.$$

- we give higher priority to P2, then execute all of P2 and all of P1 in one of P1's periods in the worst case.

$$T_1 + T_2 \leq \tau_1.$$

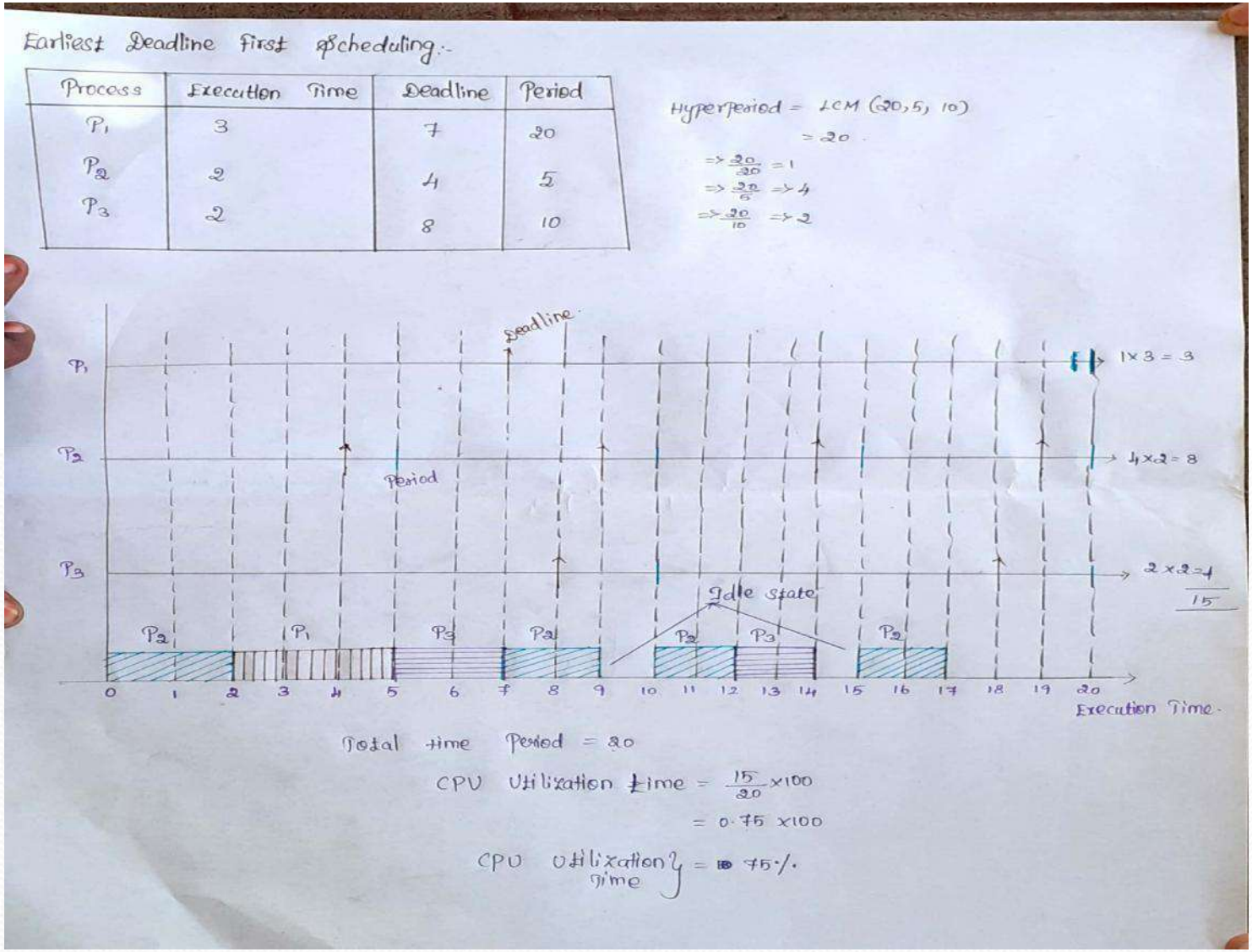
- Total CPU utilization for a set of *n tasks is* $U = \sum_{i=1}^{n} \frac{T_i}{\tau_i}.$

Earliest-Deadline-First Scheduling(EDF)

- Earliest deadline first (EDF)→ is a dynamic priority scheme.
- It changes process priorities during execution based on initiation times.
- As a result, it can achieve higher CPU utilizations than RMS.
- The EDF policy is also very simple.
- It assigns priorities in order of deadline.
- Assign highest priority to a process who has Earliest deadline.
- Assign lowest priority to a process who has farthest deadline.
- After assigning scheduling procedure, the highest-priority process is chosen for execution.
- Consider the following Example

- Hyper-period is 60

Dead line Table



- There is one time slot left at *t= 30, giving a CPU utilization of 59/60.*
- EDF can achieve 100% utilization
- **RMS vs. EDF**

	RMS	EDF
(*i*)	Less utilization of CPU	More utilization of CPU
(*ii*)	It is static priority	It is dynamic priority
(*iii*)	It can diagnose the overload	It cannot diagnose the overload
(*iv*)	It very easier to ensure all the deadline	Much difficult compare to RMS
(*v*)	While CPU utilization is less, this gives less problem to complete the deadline.	It is more Problematic

Ex:Priority inversion

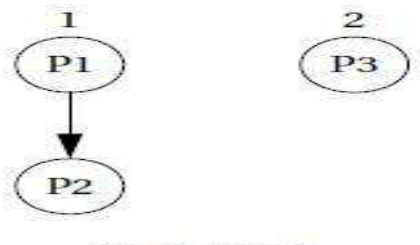
- Low-priority process blocks execution of a higher priority process by keeping hold of its resource.

Consider a system with two processes

- Higher-priority P1 and the lower-priority P2.
- Each uses the microprocessor bus to communicate to peripherals.
- When P2 executes, it requests the bus from the operating system and receives it.
- If P1 becomes ready while P2 is using the bus, the OS will preempt P2 for P1, leaving P2 with control of the bus.
- When P1 requests the bus, it will be denied the bus, since P2 already owns it.
- Unless P1 has a way to take the bus from P2, the two processes may deadlock.

Eg:Data dependencies and scheduling

- Data dependencies imply that certain combinations of processes can never occur. Consider the simple example.



Task graph

Task	Deadline
1	10
2	8

Task rates

Process	CPU time
P1	2
P2	1
P3	4

Execution times

- We know that P1 and P2 cannot execute at the same time, since P1 must finish before P2 can begin.
- P3 has a higher priority, it will not preempt both P1 and P2 in a single iteration.
- If P3 preempts P1, then P3 will complete before P2 begins.
- if P3 preempts P2, then it will not interfere with P1 in that iteration.
- Because we know that some combinations of processes cannot be ready at the same time, worst-case CPU requirements are less than would be required if all processes could be ready simultaneously.

Inter-process communication mechanisms

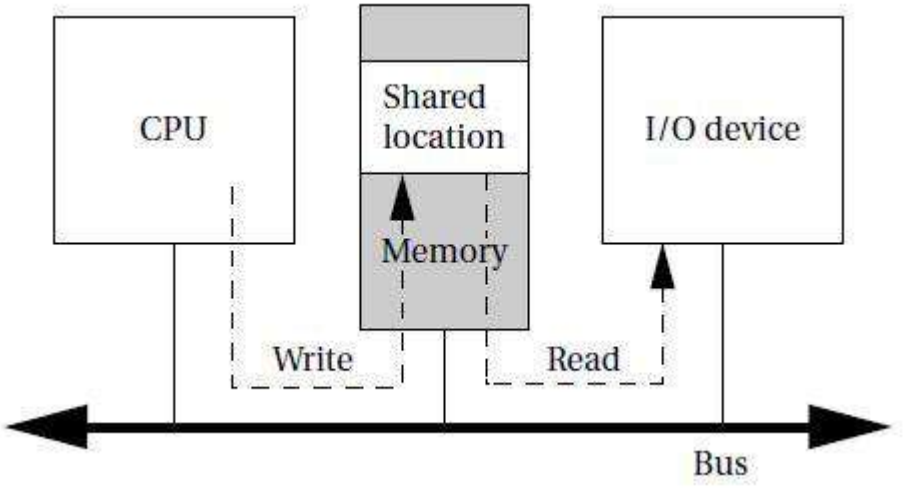
- It is provided by the operating system as part of the process abstraction.
- **Blocking Communication→** The process goes into the waiting state until it receives a response
- **Non-blocking Communication→**It allows a process to continue execution after sending the communication.

Types of inter-process communication

1. Shared Memory Communication
2. Message Passing
3. Signals

Shared Memory Communication

- The communication between inter-process is used by bus-based system.
- CPU and an I/O device, communicate through a shared memory location.
- The software on the CPU has been designed to know the address of the shared location.
- The shared location has also been loaded into the proper register of the I/O device.
- If CPU wants to send data to the device, it writes to the shared location.
- The I/O device then reads the data from that location.
- The read and write operations are standard and can be encapsulated in a procedural interface.



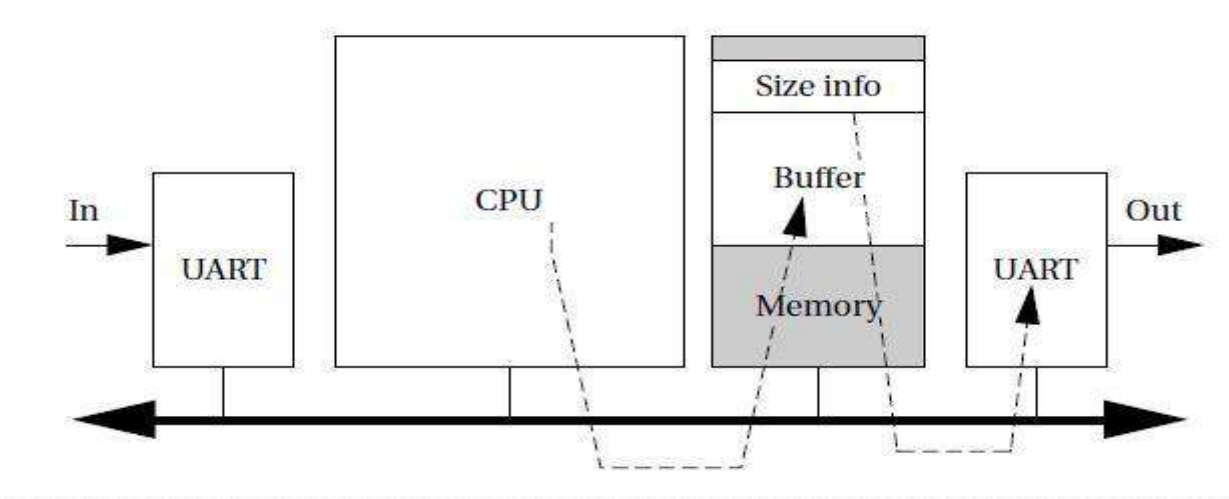
- CPU and the I/O device want to communicate through a shared memory block.
- There must be a flag that tells the CPU when the data from the I/O device is ready.
- The flag value of 0 when the data are not ready and 1 when the data are ready.
- If the flag is used only by the CPU, then the flag can be implemented using a standard memory write operation.
- If the same flag is used for bidirectional signaling between the CPU and the I/O device, care must be taken.

Consider the following scenario to call flag

1. CPU reads the flag location and sees that it is 0.
2. I/O device reads the flag location and sees that it is 0.
3. CPU sets the flag location to 1 and writes data to the shared location.
4. I/O device erroneously sets the flag to 1 and overwrites the data left by the CPU.

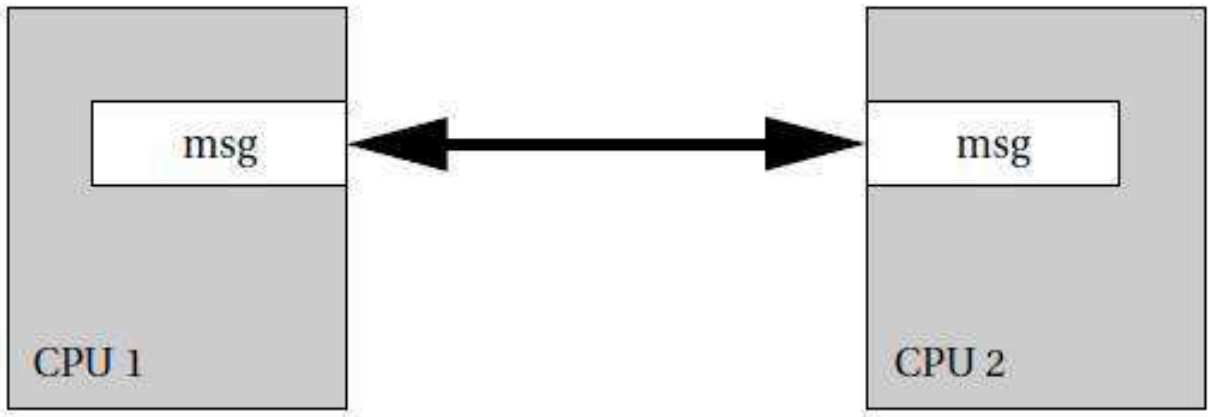
Ex: Elastic buffers as shared memory

- The text compressor is a good example of a shared memory.
- The text compressor uses the CPU to compress incoming text, which is then sent on a serial line by a UART.
- The input data arrive at a constant rate and are easy to manage.
- But the output data are consumed at a variable rate, these data require an elastic buffer.
- The CPU and output UART share a memory area—the CPU writes compressed characters into the buffer and the UART removes them as necessary to fill the serial line.
- Because the number of bits in the buffer changes constantly, the compression and transmission processes need additional size information.
- CPU writes at one end of the buffer and the UART reads at the other end.
- The only challenge is to make sure that the UART does not overrun the buffer.



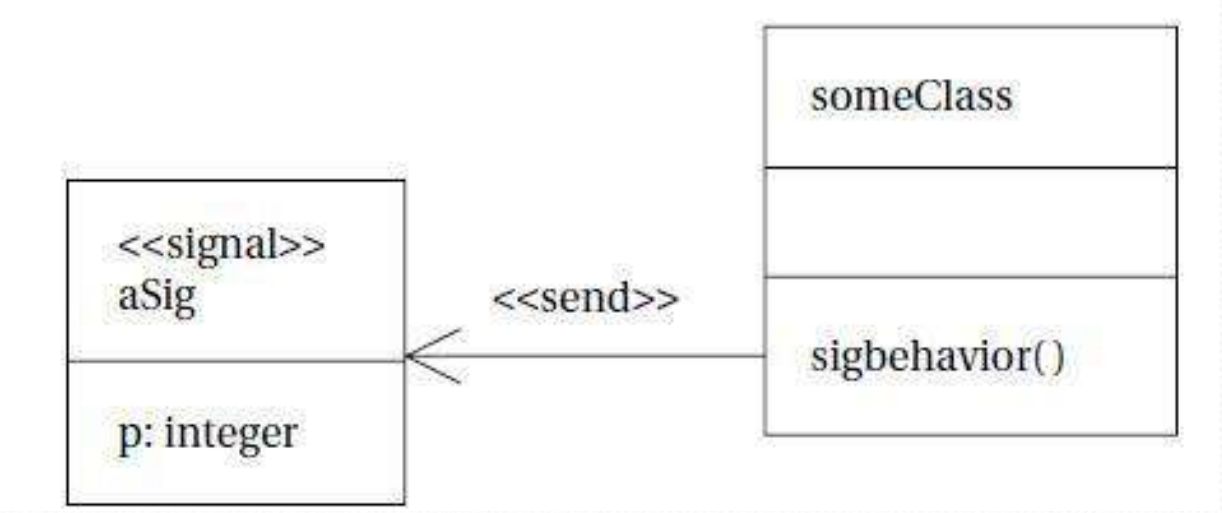
Message Passing

- Here each communicating entity has its own message send/receive unit.
- The message is not stored on the communications link, but rather at the senders/ receivers at the end points.
- Ex:Home control system
- It has one microcontroller per household device—lamp, thermostat, faucet, appliance.
- The devices must communicate relatively infrequently.
- Their physical separation is large enough that we would not naturally think of them as sharing a central pool of memory.
- Passing communication packets among the devices is a natural way to describe coordination between these devices.



Signals

- Generally signal communication used in Unix .
- A signal is analogous to an interrupt, but it is entirely a software creation.
- A signal is generated by a process and transmitted to another process by the OS.
- A UML signal is actually a generalization of the Unix signal.
- Unix signal carries no parameters other than a condition code.
- UML signal is an object, carry parameters as object attributes.
- The sigbehavior( ) →behavior of the class is responsible for throwing the signal, as indicated by<<send>>.
- The signal object is indicated by the <<signal>>



Evaluating operating system performance

- Analysis of scheduling policies is made by the following 4 assumptions
- Assumed that context switches require zero time. Although it is often reasonable to neglect context switch time when it is much smaller than the process execution time, context switching can add significant delay in some cases.
- We have largely ignored interrupts. The latency from when an interrupt is requested to when the device's service is complete is a critical parameter of real time performance.
- We have assumed that we know the execution time of the processes.
- We probably determined worst-case or best-case times for the processes in isolation.

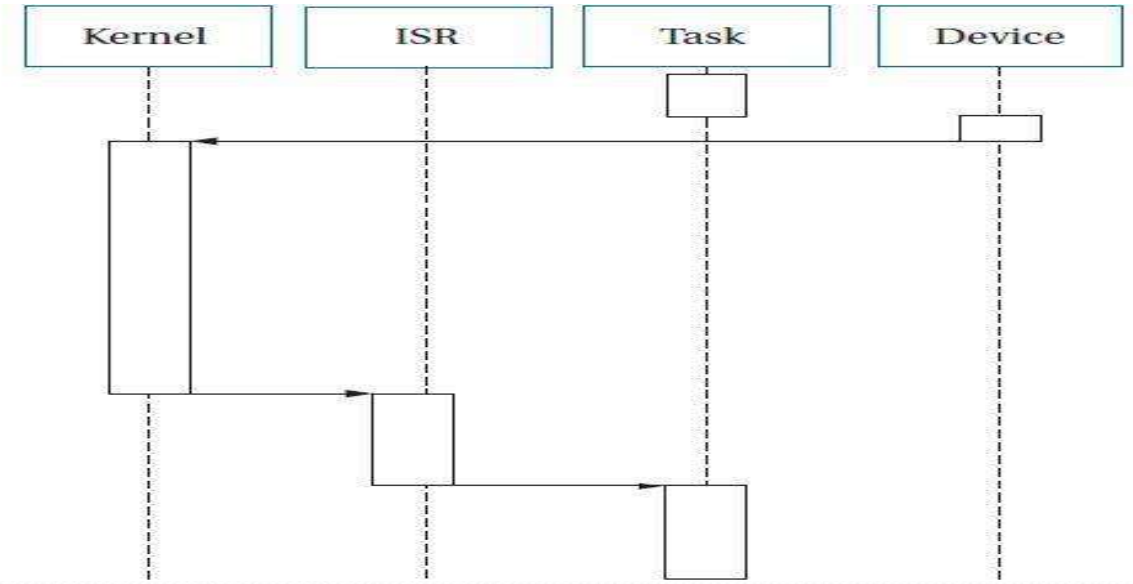
Context switching time

It depends on following factors

- The amount of CPU context that must be saved.
- Scheduler execution time.

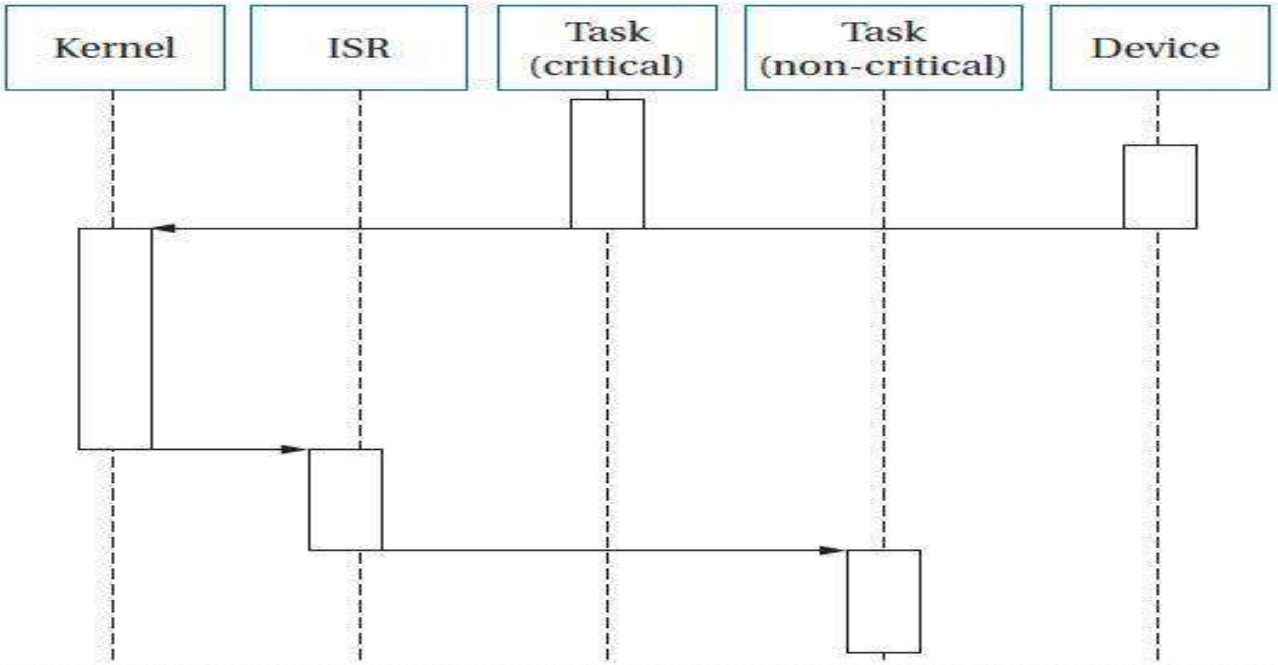
Interrupt latency

- Interrupt latency →It is the duration of time from the assertion of a device interrupt to the completion of the device's requested operation.
- Interrupt latency is critical because data may be lost when an interrupt is not serviced in a timely fashion.



- A task is interrupted by a device.
- The interrupt goes to the kernel, which may need to finish a protected operation.
- Once the kernel can process the interrupt, it calls the interrupt service routine (ISR), which performs the required operations on the device.
- Once the ISR is done, the task can resume execution.

- Several factors in both hardware and software affect interrupt latency:
- The processor interrupt latency
- The execution time of the interrupt handler
- Delays due to RTOS scheduling
- RTOS delay the execution of an interrupt handler in two ways.
- Critical sections and interrupt latency
- Critical sections in the kernel will prevent the RTOS from taking interrupts.
- Some operating systems have very long critical sections that disable interrupt handling for very long periods.



- If a device interrupts during a critical section, that critical section must finish before the kernel can handle the interrupt.
- The longer the critical section, the greater the potential delay.
- Critical sections are one important source of scheduling jitter because a device may interrupt at different points in the execution of processes and hit critical sections at different points.

Interrupt priorities and interrupt latency

- A higher-priority interrupt may delay a lower-priority interrupt.
- A hardware interrupt handler runs as part of the kernel, not as a user thread.
- The priorities for interrupts are determined by hardware.
- Any interrupt handler preempts all user threads because interrupts are part of the CPU's fundamental operation.
- We can reduce the effects of hardware preemption by dividing interrupt handling into two different pieces of code.
- **Interrupt service handler (ISH)→** performs the minimal operations required to respond to the device.
- **Interrupt service routine (ISR)→** Performs updating user buffers or other more complex operation.

- RTOS performance evaluation tools
- Some RTOSs provide simulators or other tools that allow us to view the operation of the processes,context switching time, interrupt response time, and other overheads.

Windows CE provides several performance analysis toolsAn instrumentation routine in the kernel that measures both interrupt serviceroutine and interrupt service thread latency.

- OS Bench measures the timing of operating system tasks such as critical section access, signals, and so on
- Kernel Tracker provides a graphical user interface for RTOS events.

Power optimization strategies for processes

- A power management policy is a strategy for determining when to perform certain power management operations.
- The system can be designed based on the static and dynamic power management mechanisms.

Power saving straegies

- Avoiding a power-down mode can cost unnecessary power.
- Powering down too soon can cause severe performance penalties.
- Re-entering run mode typically costs a considerable amount of time.
- A straightforward method is to power up the system when a request is received.

Predictive shutdown

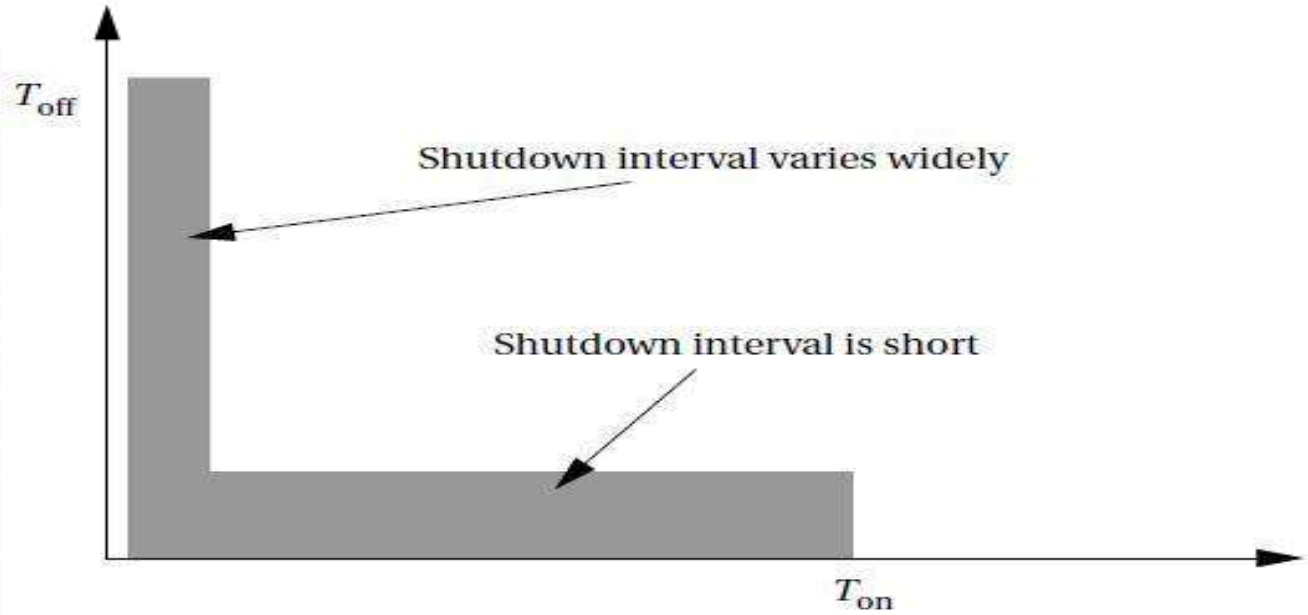
- The goal is to predict when the next request will be made and to start the system just before that time, saving the requestor the start-up time.
- Make guesses about activity patterns based on a probabilistic model of expected behavior.

This can cause two types of problems

- The requestor may have to wait for an activity period.
- In the worst case,the requestor may not make a deadline due to the delay incurred by system

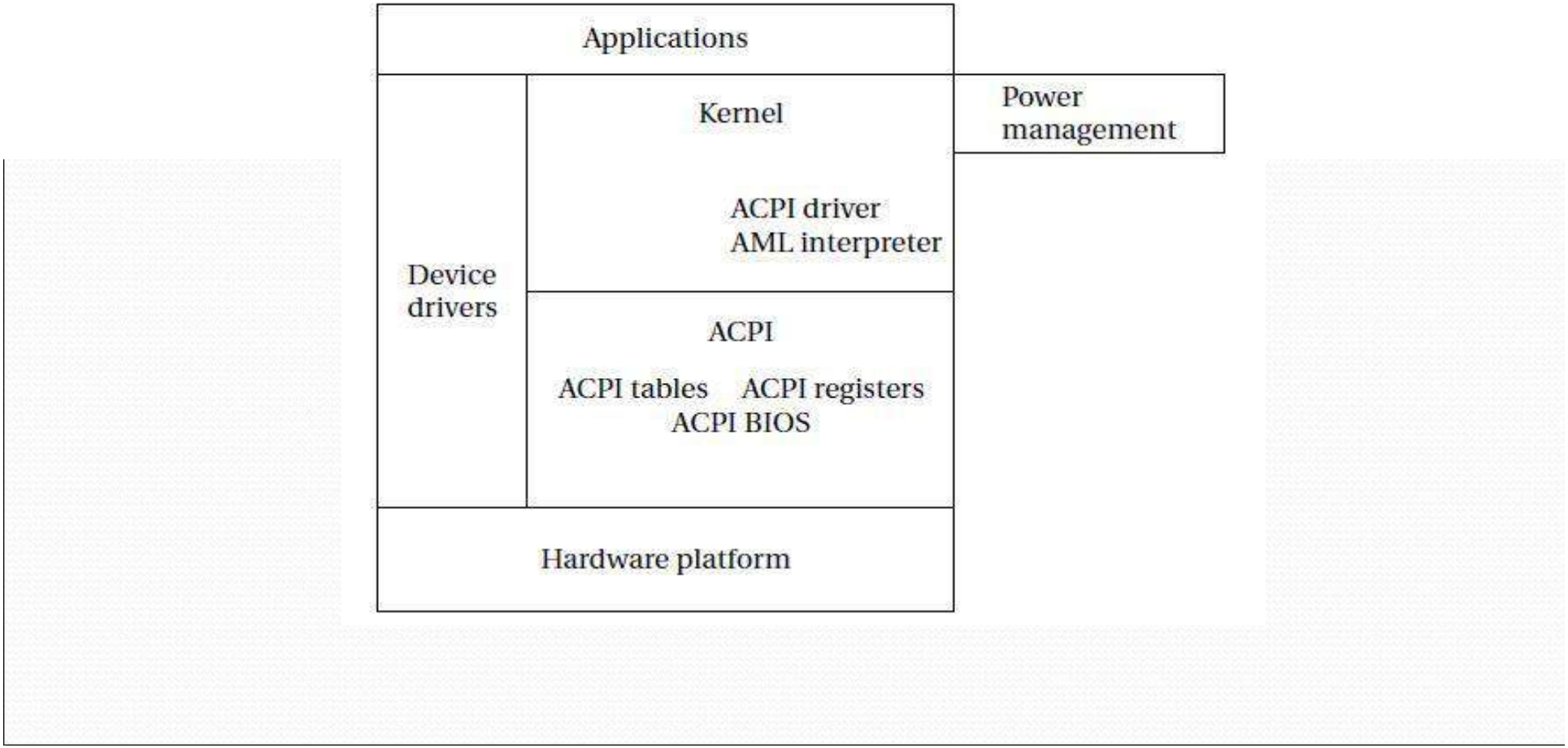
An L-shaped usage distribution

- A very simple technique is to use fixed times.
- If the system does not receive inputs during an interval of length *Ton, it shuts down.*
- Powered-down system waits for a period *Toff before* returning to the power-on mode.
- In this distribution, the idle period after a long active period is usually very short, and the length of the idle period after a short active period is uniformly distributed.
- Based on this distribution, shutdown when the active period length was below a threshold, putting the system in the vertical portion of the *L distribution.*



Advanced Configuration and Power Interface (ACPI)

- It is an open industry standard for power management services.
- It is designed to be compatible with a wide variety of OSs.
- A decision module →determines power management actions.



ACPI supports the following five basic global power states.

1. G3, the mechanical off state, in which the system consumes no power.
2. G2, the soft off state, which requires a full OS reboot to restore the machine to working condition. This state has four sub-states:

- S1, a low wake-up latency state with no loss of system context
- S2, a low wake-up latency state with a loss of CPU and system cache state
- S3, a low wake-up latency state in which all system state except for main
- memory is lost.

S4, the lowest-power sleeping state, in which all devices are turned off.

3. G1, the sleeping state, in which the system appears to be off.
4. G0, the working state, in which the system is fully usable.
5. The legacy state, in which the system does not comply with ACPI.

Example Real time operating systems

POSIX

- POSIX is a Unix operating system created by a standards organization.
- POSIX-compliant operating systems are source-code compatible.
- Application can be compiled and run without modification on a new POSIX platform.
- It has been extended to support real time requirements.
- Many RTOSs are POSIX-compliant and it serves as a good model for basic RTOS techniques.
- The Linux operating system has a platform for embedded computing.
- Linux is a POSIX-compliant operating system that is available as open source.
- Linux was not originally designed for real-time operation .
- Some versions of Linux may exhibit long interrupt latencies,
- To improve interrupt latency,A dual-kernel approach uses a specialized kernel, the co-kernel, for real-time processes **and** the standard kernel for non-real-time processes.

Process in POSIX

- A new process is created by making a copy of an existing process.
- The copying process creates two different processes both running the same code.
- The complex task is to ensuring that one process runs the code intended for the new process while the other process continues the work of the old process.

Scheduling in POSIX

- A process makes a copy of itself by calling the fork() function.
- That function causes the operating system to create a new process (the child process) which is a nearly exact copy of the process that called fork() (the parent process).
- They both share the same code and the same data values with one exception, the return value
- of fork().
- The parent process is returned the process ID number of the child process, while the child process gets a return value of 0.
- We can therefore test the return value of fork() to determine which process is the child

```
childid = fork();
if (childid == 0) { /* must be the child */
/* do child process here */
}
```

- execv() function takes as argument the name of the file that holds the child's code and the array of arguments.
- It overlays the process with the new code and starts executing it from the main() function.
- In the absence of an error, execv() should never return.
- The code that follows the call to perror() and exit(), take care of the case where execv() fails and returns to the parent process.
- The exit() function is a C function that is used to leave a process

```
childid = fork();
if (childid == 0) { /* must be the child */
execv("mychild",childargs);
perror("execv");
exit(1);
}
```

- The wait functions not only return the child process's status, in many implementations of POSIX they make sure that the child's resources .
- The parent stuff() function performs the work of the parent function.

```
childid = fork();
if (childid == 0) { /* must be the child */
execv("mychild",childargs);
perror("execl");
exit(1);
}
else { /* is the parent */
parent_stuff(); /* execute parent functionality */
wait(&cstatus);
exit(0);
}
```

The POSIX process model

- Each POSIX process runs in its own address space and cannot directly access the data or code.

Real-time scheduling in POSIX

- POSIX supports real-time scheduling in the POSIX_PRIORITY_SCHEDULING resource.
- POSIX supports Rate-monotonic scheduling in the SCHED_FIFO scheduling policy.
- It is a strict priority-based scheduling scheme in which a process runs until it is preempted or terminates.
- The term FIFO simply refers→ processes run in first-come first-served order.

POSIX semaphores

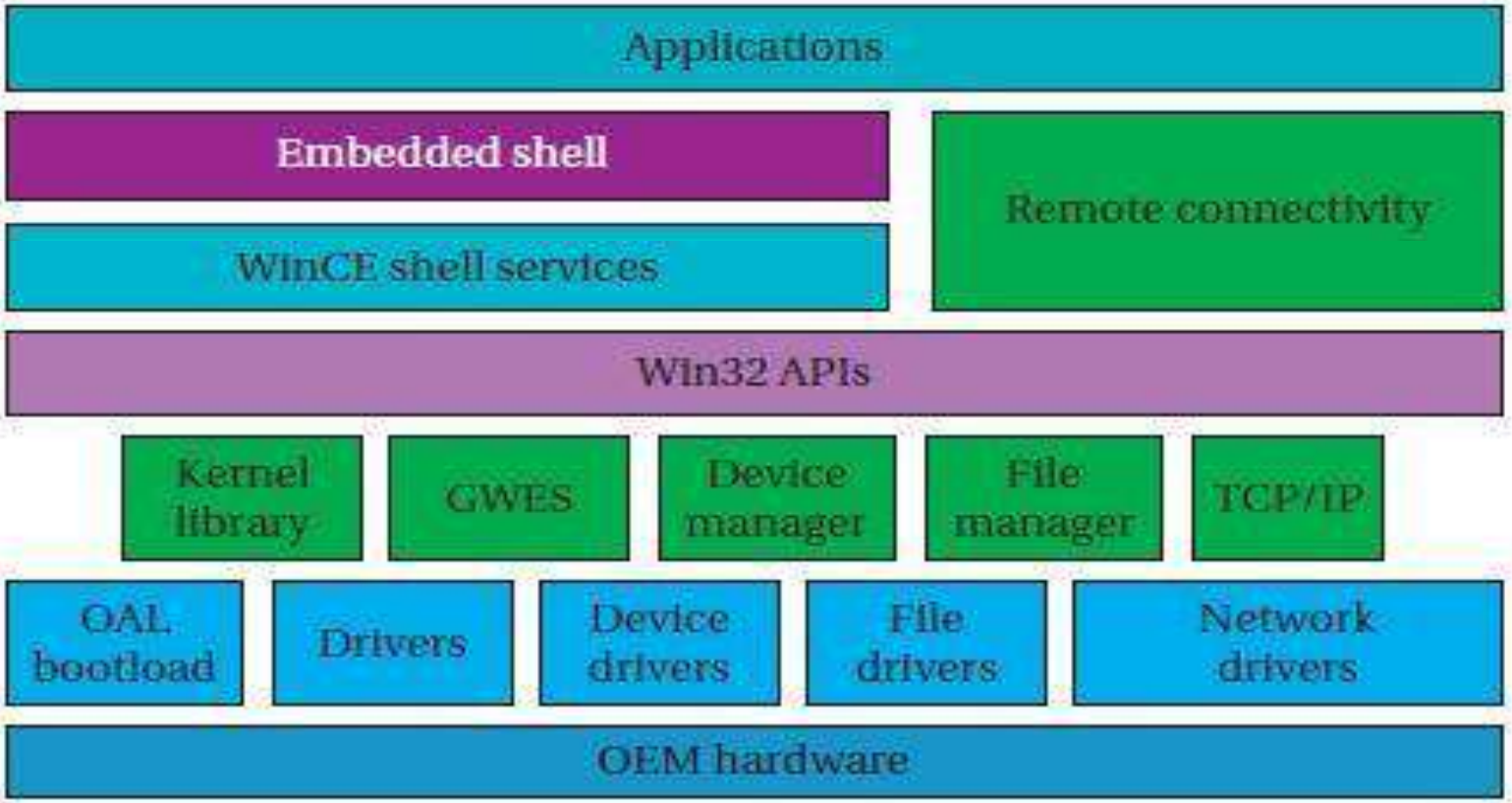
- POSIX supports semaphores and also supports a direct shared memory mechanism.
- POSIX supports counting semaphores in the _POSIX_SEMAPHORES option.
- A counting semaphore allows more than one process access to a resource at a time.
- If the semaphore allows up to N resources, then it will not block until N processes have
- simultaneously passed the semaphore;
- The blocked process can resume only after one of the processes has given up its semaphore.
- When the semaphore value is 0, the process must wait until another process gives up the semaphore and increments the count.

POSIX pipes

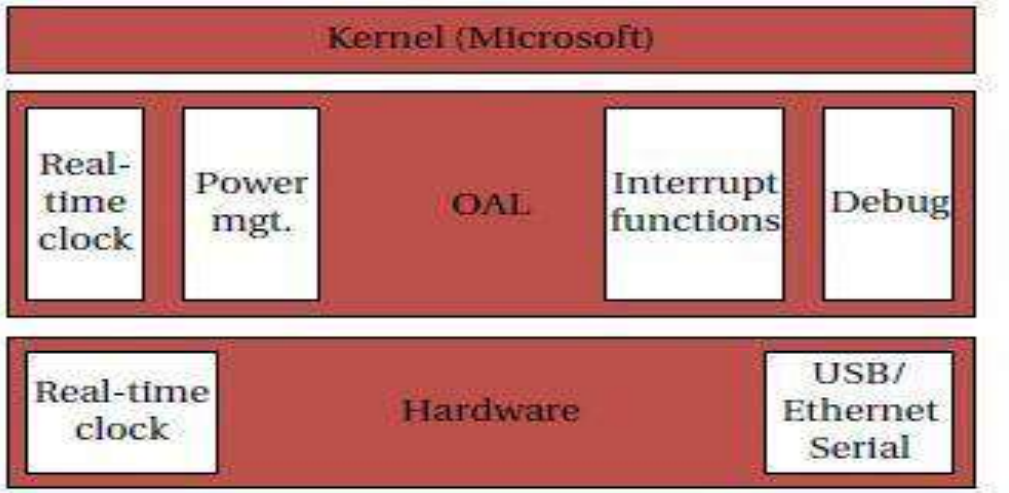
- Parent process uses the pipe() function to create a pipe to talk to a child.
- Each end of a pipe appears to the programs as a file.
- The pipe() function returns an array of file descriptors, the first for the write end and the second for the read end.
- POSIX also supports message queues under the _POSIX_MESSAGE_PASSING facility..

Windows CE

- Windows CE is designed to run on multiple hardware platforms and instruction set architectures.
- It supports devices such as smart phones, electronic instruments etc..,



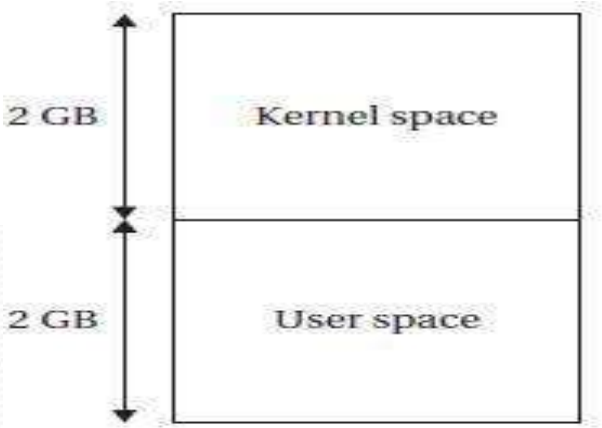
- Applications run under the shell and its user interface.
- The Win32 APIs manage access to the operating system.
- OEM Adaption Layer (OAL)→ provides an interface to the hardware and software architecture.



- OAL → provides services such as a real-time clock, power management, interrupts, and a debugging interface.
- A Board Support Package (BSP) for a particular hardware platform includes the OAL and drivers.

Memory Space

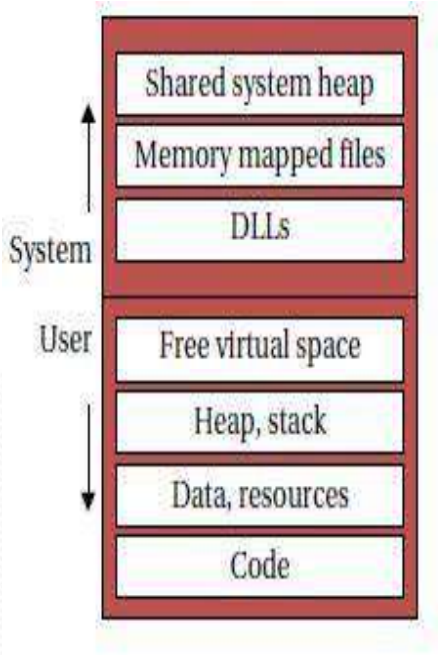
- It support for virtual memory with a flat 32-bit virtual address space.
- A virtual address can be statically mapped into main memory for key kernel-mode code.
- An address can also be dynamically mapped, which is used for all user-mode and some kernel-mode code.
- Flash as well as magnetic disk can be used as a backing store



- The top 1 GB is reserved for system elements such as DLLs, memory mapped files, and shared system heap.
- The bottom 1 GB holds user elements such as code, data, stack, and heap.

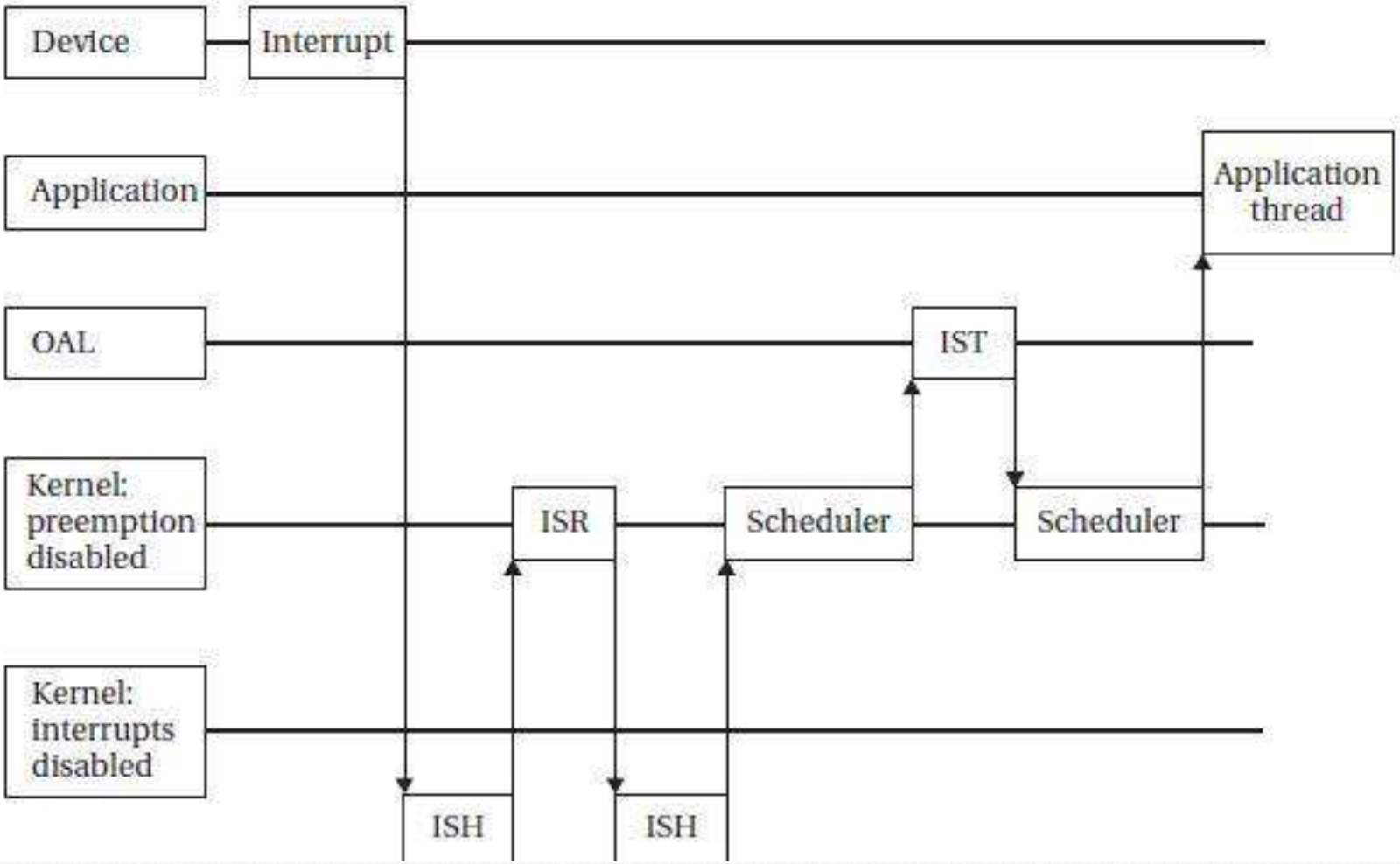
User address space in windows CE

- Threads are defined by executable files while drivers are defined by dynamically-linked libraries (DLLs).
- A process can run multiple threads.
- Threads in different processes run in different execution environments.
- Threads are scheduled directly by the operating system.
- Threads may be launched by a process or a device driver.
- A driver may be loaded into the operating system or a process.
- Drivers can create threads to handle interrupts
- Each thread is assigned an integer priority.
- 0 is the highest priority and 255 is the lowest priority.
- Priorities 248 through 255 are used for non-real-time threads .
- The operating system maintains a queue of ready processes at each priority level.



- Execution of a thread can also be blocked by a higher-priority thread.
- Tasks may be scheduled using either of two policies: a thread runs until the end of its quantum; or a thread runs until a higher-priority thread is ready to run.
- Within each priority level, round-robin scheduling is used.
- WinCE supports priority inheritance.
- When priorities become inverted, the kernel temporarily boosts the priority of the lower-priority thread to ensure that it can complete and release its resources.
- Kernel will apply priority inheritance to only one level.
- If a thread that suffers from priority inversion in turn causes priority inversion for another thread, the kernel will not apply priority inheritance to solve the nested priority inversion.

Sequence diagram for an interrupt



- Interrupt handling is divided among three entities
- The interrupt service handler (ISH)→ is a kernel service that provides the first response to the interrupt.
- The ISH selects an interrupt service routine (ISR) to handle the interrupt.
- The ISR in turn calls an interrupt service thread (IST) which performs most of the work required to handle the interrupt.
- The IST runs in the OAL and so can be interrupted by a higher-priority interrupt.
- ISR→determines which IST to use to handle the interrupt and requests the kernel to schedule that thread.
- The ISH then performs its work and signals the application about the updated device status as appropriate.
- kernel-mode and user-mode drivers use the same API.

Distributed Embedded Systems (DES)

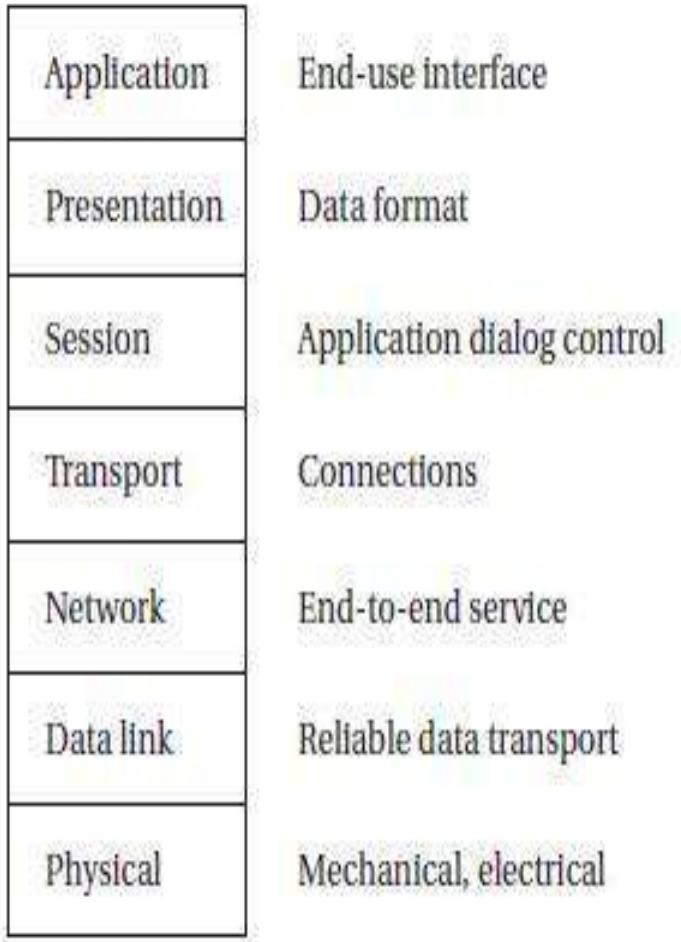
- It is a collection of hardware and software and its communication.
- It also has many control system performance.
- Processing Element (PE)is a basic unit of DES.
- It allows the network to communicate.
- PE is an instruction set processor such as DSP,CPU and Microcontroller.

Network abstractions

- Networks are complex systems.
- It provide high-level services such as data transmission from the other components in the system.
- ISO has developed a seven-layer model for networks known as Open Systems Interconnection (OSI) models.

OSI model layers

- Physical layer→ defines the basic properties of the interface between systems, including the physical connections, electrical properties & basic procedures for exchanging bits.
- Data link layer→ used for error detection and control across a single link.
- Network layer→ defines the basic end-to-end data transmission service.
- Transport layer→ defines connection-oriented services that ensure that data are delivered in the proper order .
- Session layer→ provides mechanisms for controlling the interaction of end-user services across a network, such as data grouping and checkpointing.
- Presentation layer→ layer defines data exchange formats
- Application layer→ provides the application interface between the network and end-user programs.

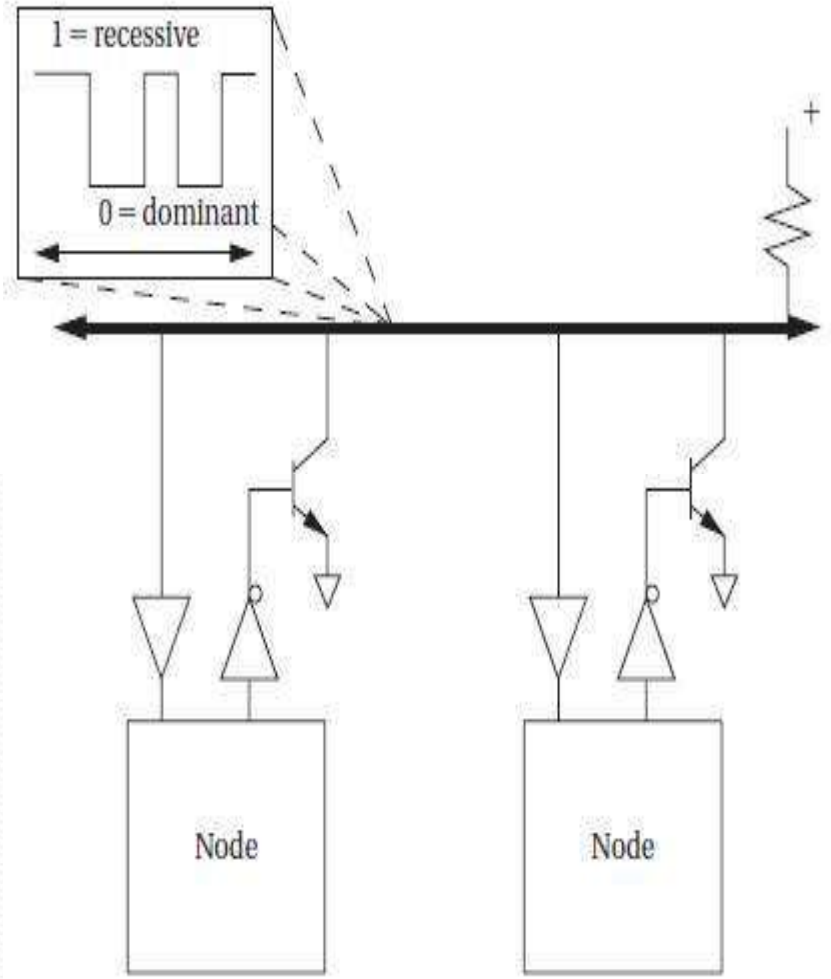


Controller Area Network(CAN)Bus

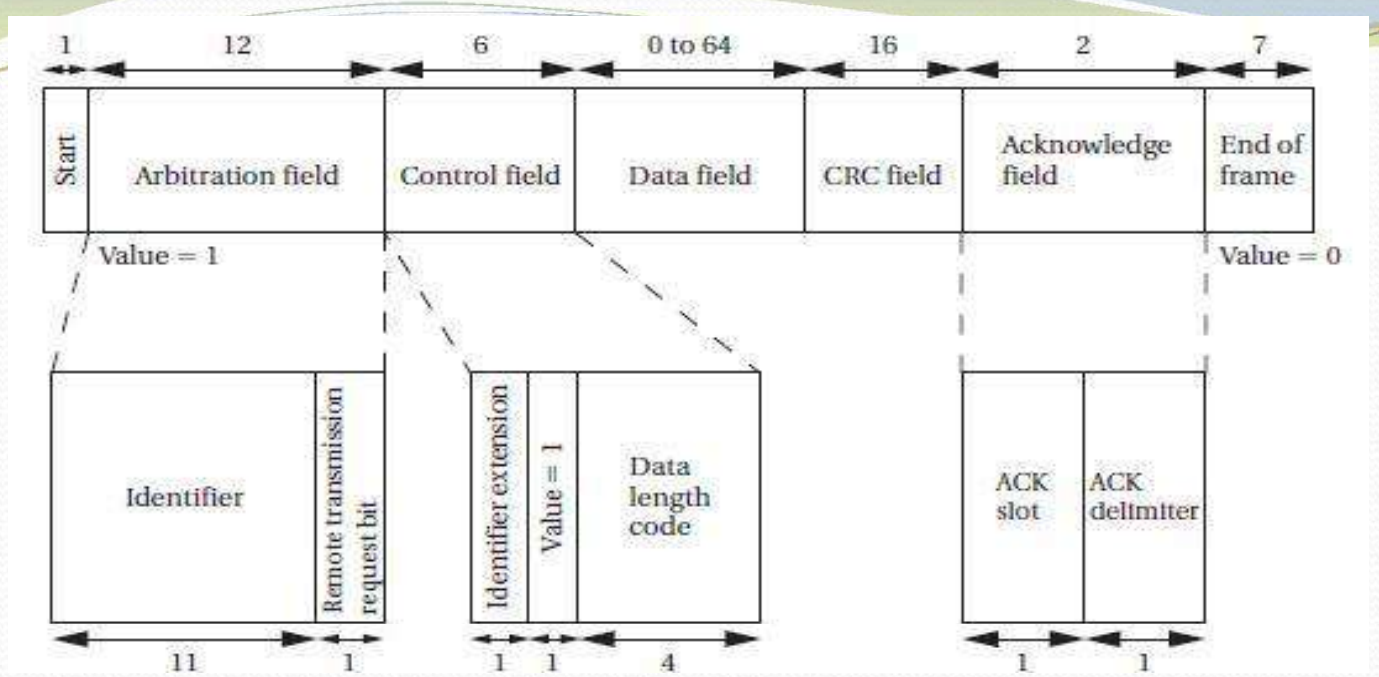
- It was designed for automotive electronics and was first used in production cars in 1991.
- It uses bit-serial transmission.
- CAN can run at rates of 1 Mbps over a twisted pair connection of 40 meters.
- An optical link can also be used.

4.7.2.1)Physical-electrical organization of a CAN bus

- Each node in the CAN bus has its own electrical drivers and receivers that connect the node to the bus in wired-AND fashion.
- When all nodes are transmitting 1s, the bus is said to be in the recessive state.
- when a node transmits a 0s, the bus is in the dominant state.



Data Frame



- Arbitration field→ The first field in the packet contains the packet's destination address 11 bits
- Remote Transmission Request (RTR) bit is set to 0 if the data frame is used to request data from the destination identifier.
- When RTR = 1, the packet is used to write data to the destination identifier.
- Control field→ 4-bit length for the data field with a 1 in between.
- Data field→0 to 64 bytes, depending on the value given in the control field.
- CRC→ It is sent after the data field for error detection.
- Acknowledge field →identifier signal whether the frame was correctly received.(sender puts a bit (1) in the ACK slot , if the receiver detected an error, it put (0) value)

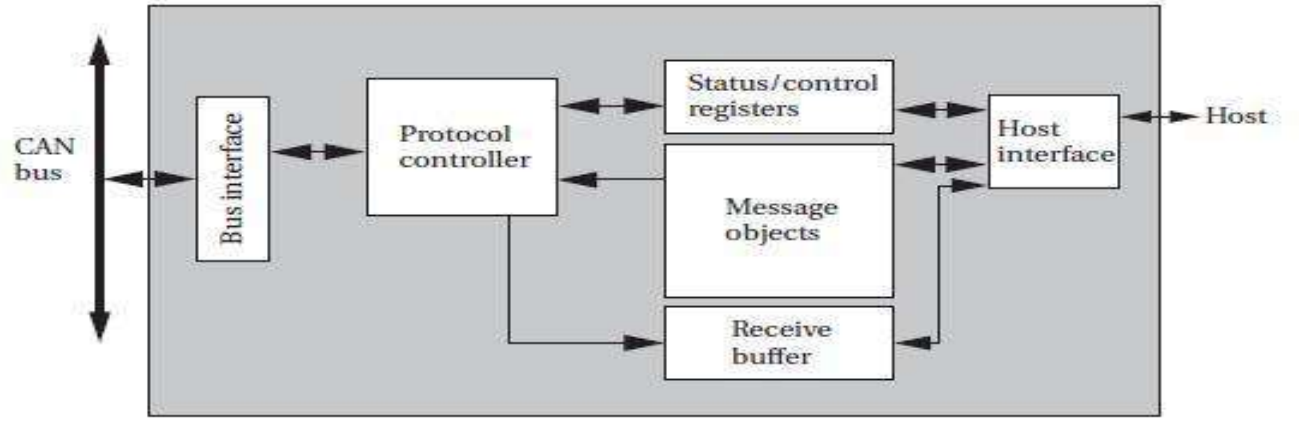
Arbitration

- It uses a technique known as Carrier Sense Multiple Access with Arbitration on Message Priority (CSMA/AMP).
- When a node hears a dominant bit in the identifier when it tries to send a recessive bit, it stops transmitting.
- By the end of the arbitration field, only one transmitter will be left.
- The identifier field acts as a priority identifier, with the all-o having the highest priority

Error handling

- An error frame can be generated by any node that detects an error on the bus.
- Upon detecting an error, a node interrupts the current transmission.
- Error flag field followed by an error delimiter field of 8 recessive bits.
- Error delimiter field allows the bus to return to the quiescent state so that data frame transmission can resume.
- Overload frame signals that a node is overloaded and will not be able to handle the next message. Hence the node can delay the transmission of the next frame .

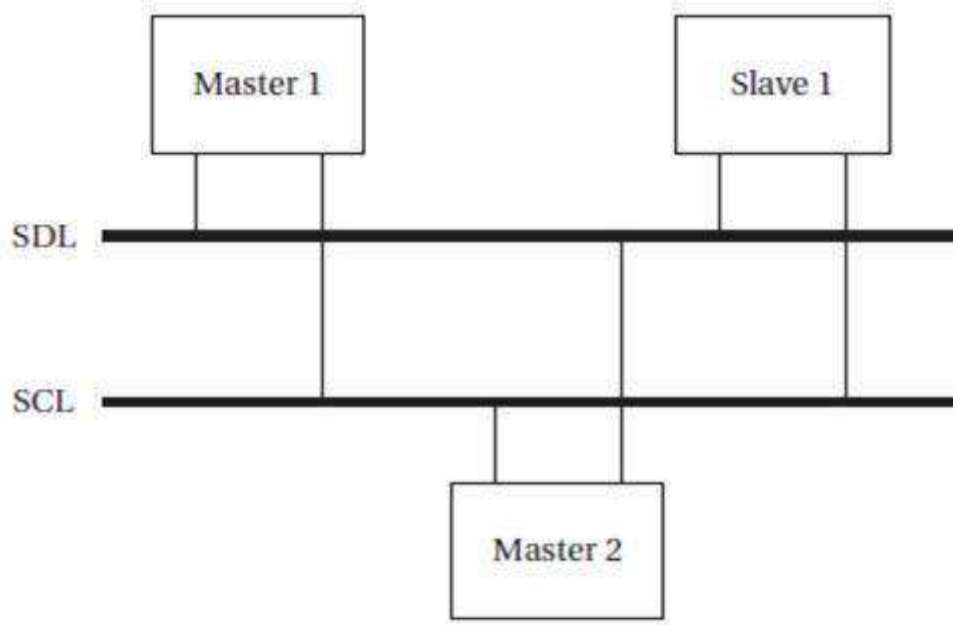
Architecture of a CAN controller



- The controller implements the physical and data link layers.
- CAN does not need network layer services to establish end-to-end connections.
- The protocol control block is responsible for determining when to send messages, when a message must be resent and when a message should be received.

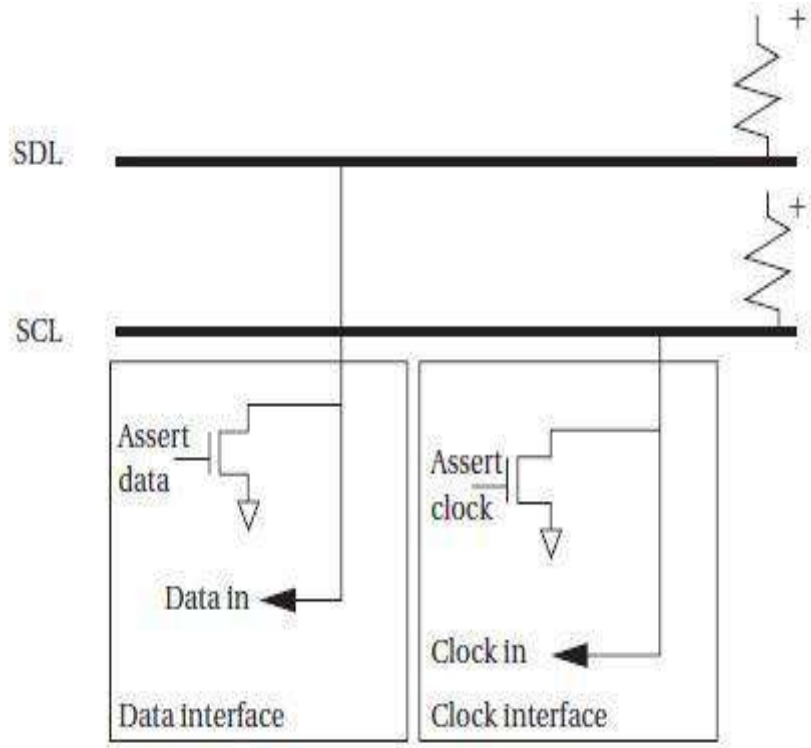
I²C bus

- I²C bus→ used to link microcontrollers into systems.
- I²C is designed to be low cost, easy to implement, and of moderate speed (up to 100kbps for the standard bus and up to 400 kbps for the extended bus).
- Serial data line (SDL) for data transmission.
- Serial clock line (SCL)→ indicates when valid data are on the data line.
- Every node in the network is connected to both SCL and SDL.
- Some nodes may act as bus masters .
- Other nodes may act as slaves that only respond to requests from masters.

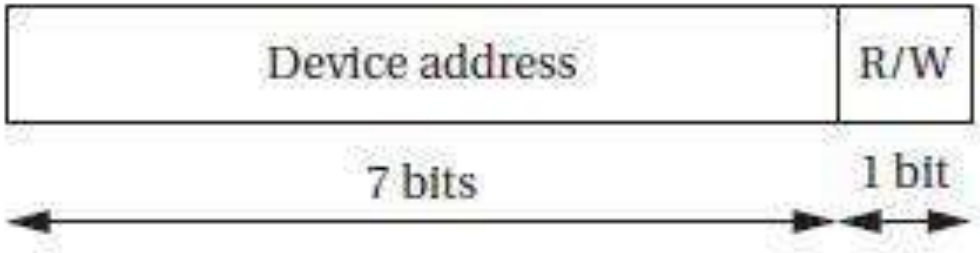


Electrical interface to the I2C bus

- Both bus lines are defined by an electrical signal.
- Both bus signals use open collector/open drain circuits.
- The open collector/open drain circuitry allows a slave device to stretch a clock signal during a read.
- The master is responsible for generating the SCL clock.
- The slave can stretch the low period of the clock.
- It is a multi master bus so different devices may act as the master at various times.
- Master drives both SCL and SDL when it is sending data.
- When the bus is idle, both SCL and SDL remain high.
- When two devices try to drive either SCL or SDL , the open collector/open drain circuitry prevents errors.
- Each master device make sure that it is not interfering with another message.

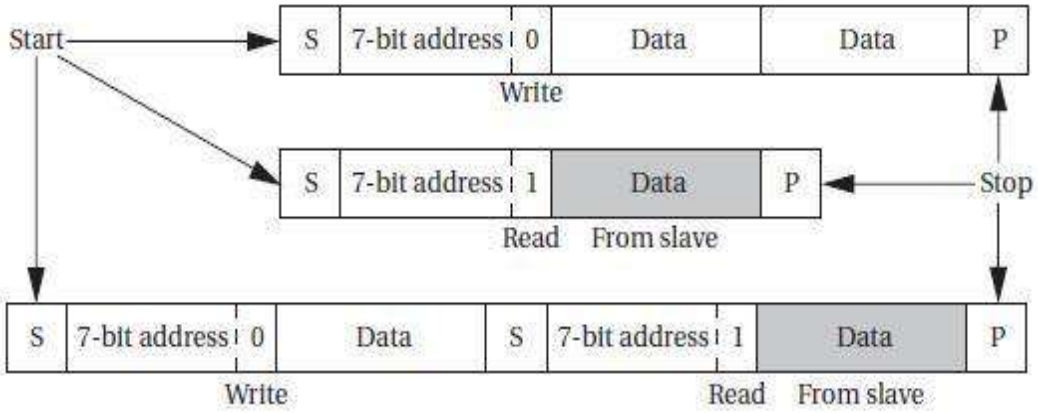


Format of an I2C address transmission



- Every I²C device has an separate address.
- A device address is 7 bits and 1 bit for read/write data.
- The address 0000000 ,which can be used to signal all devices simultaneously.
- The address 1110XX is reserved for the extended 10-bit addressing scheme.

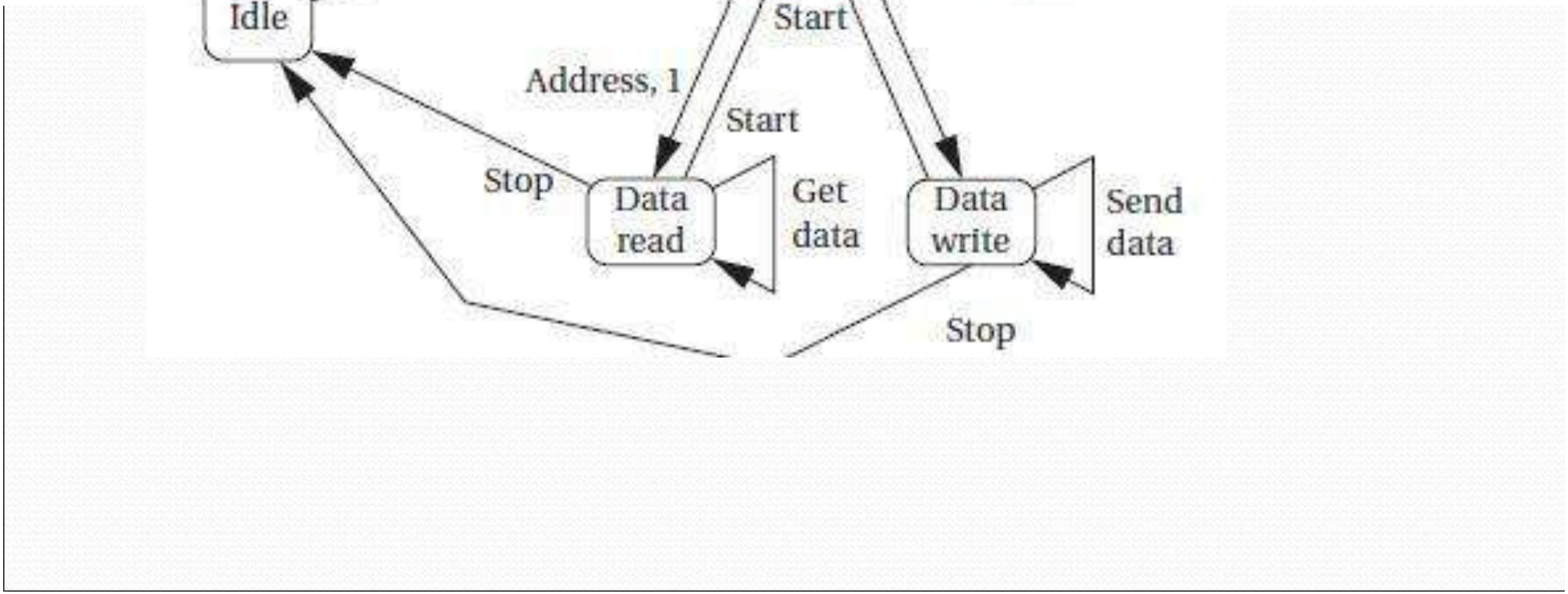
Bus transactions on the I2C bus



- When a master wants to write a slave, it transmits the slave's address followed by the data.
- When a master send a read request with the slave's address and the slave transmit the data.
- Transmission address has 7-bit and 1 bit for data direction.(0 for writing from the master to the slave and 1 for reading from the slave to the master)
- A bus transaction is initiated by a start signal and completed with an end signal.
- A start is signaled by leaving the SCL high and sending a 1 to 0 transition on SDL.
- A stop is signaled by setting the SCL high and sending a 0 to 1 transition on SDL.

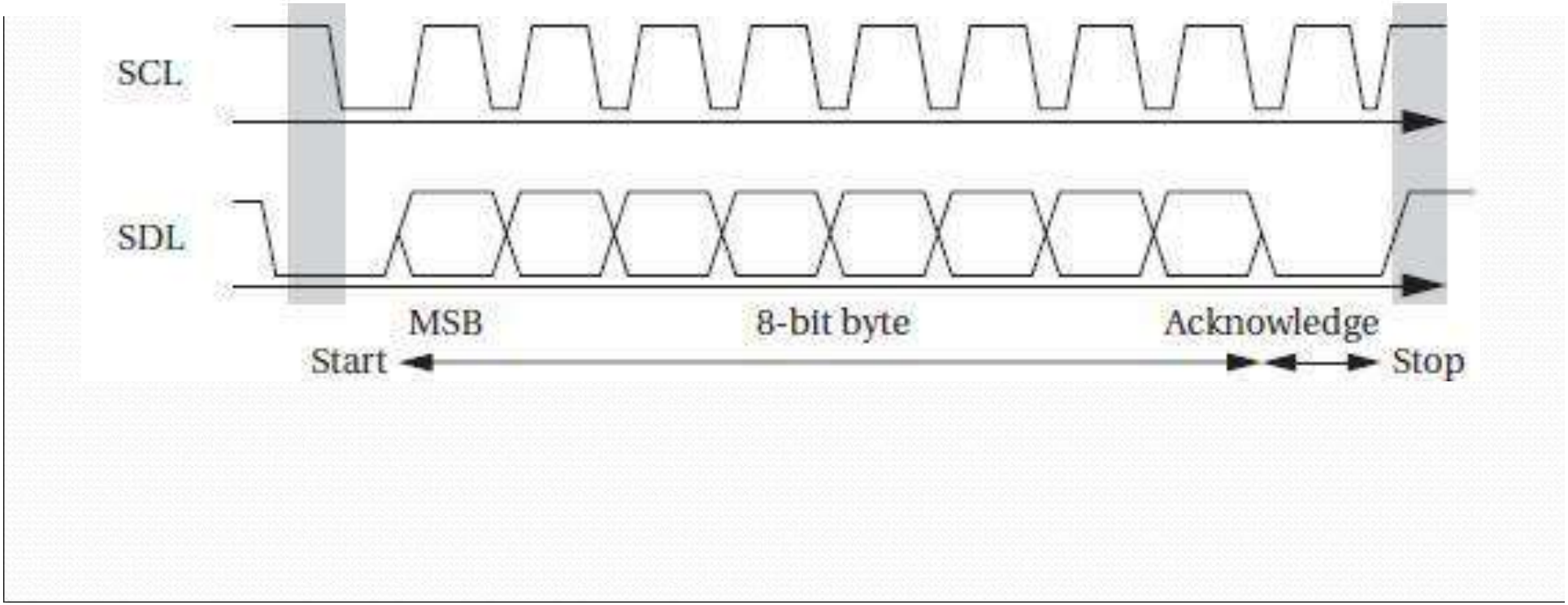
State transition graph for an I2C bus master

- Starts and stops must be paired.
- A master can write and then read by sending a start after the data transmission, followed by another address transmission and then more data.



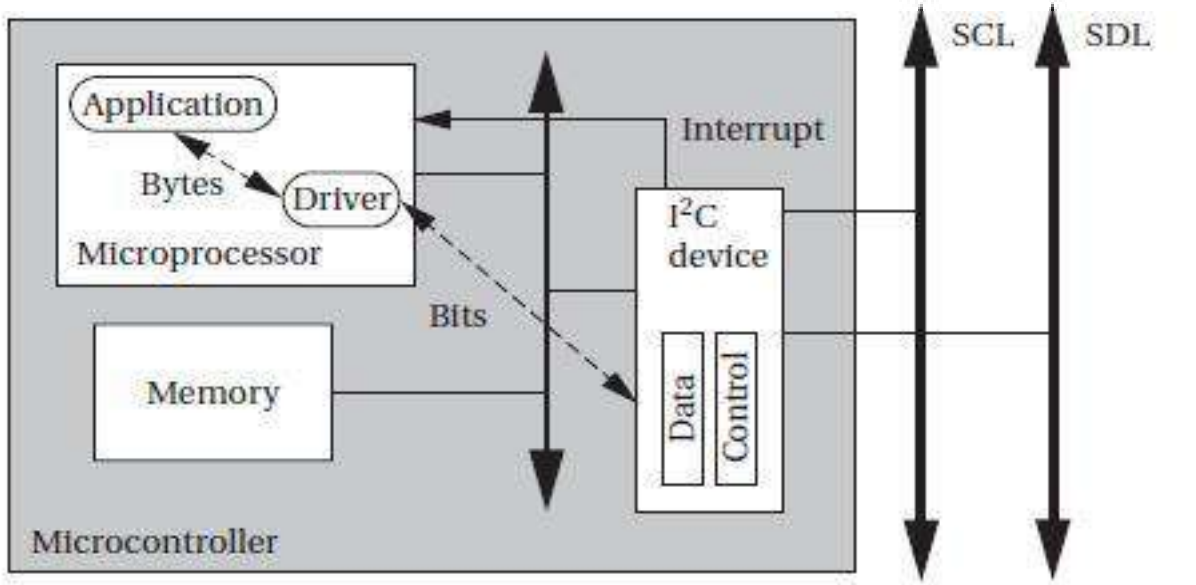
Transmitting a byte on the I2C bus

- The transmission starts when SDL is pulled low while SCL remains high.
- The clock is pulled low to initiate the data transfer.
- At each bit, the clock goes high while the data line assumes its proper value of 0 or 1.
- An acknowledgment is sent at the end of every 8-bit transmission, whether it is an address or data.
- After acknowledgment, the SDL goes from low to high while the SCL is high, signaling the stop condition.



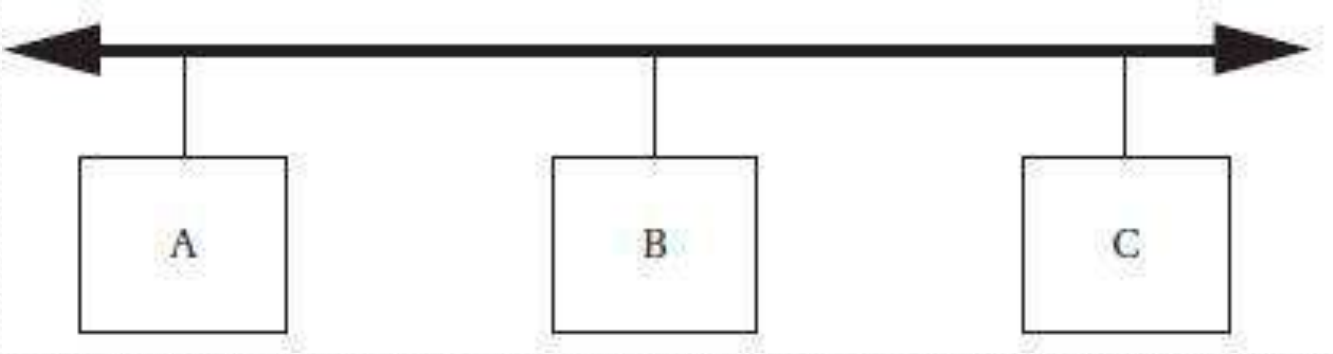
I2C interface in a microcontroller

- System has a 1-bit hardware interface with routines for byte-level functions.
- I²C device used to generates the clock and data.
- Application code calls routines to send an address, data byte, and also generates the SCL ,SDL and acknowledges.
- Timers is used to control the length of bits on the bus.
- When Interrupts used in master mode, polled I/O may be acceptable.
- If no other pending tasks can be performed, because masters initiate their own transfers.



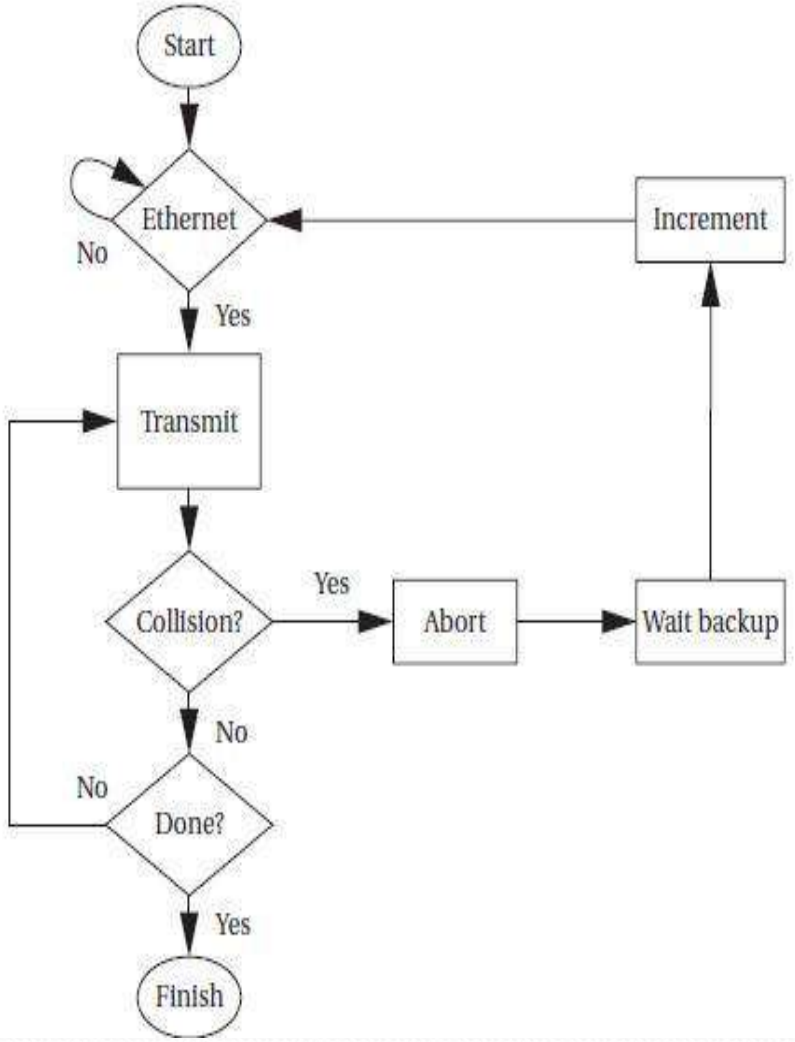
ETHERNET

- It is widely used as a local area network for general-purpose computing.
- It is also used as a network for embedded computing.
- It is particularly useful when PCs are used as platforms, making it possible to use standard components, and when the network does not have to meet real-time requirements.
- It is a bus with a single signal path.
- It supports both twisted pair and coaxial cable.
- Ethernet nodes are not synchronized, if two nodes decide to transmit at the same time,the message will be ruined.

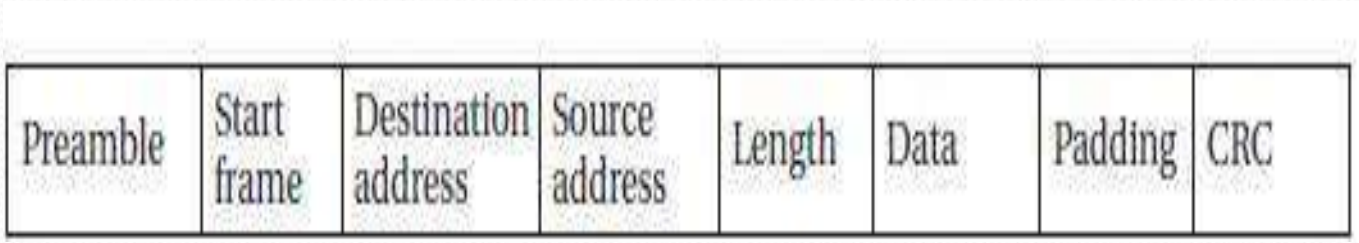


Ethernet CSMA/CD algorithm

- A node that has a message waits for the bus to become silent and then starts transmitting.
- It simultaneously listens, and if it hears another transmission that interferes with its transmission, it stops transmitting and waits to retransmit.
- The waiting time is random, but weighted by an exponential function of the number of times the message has been aborted



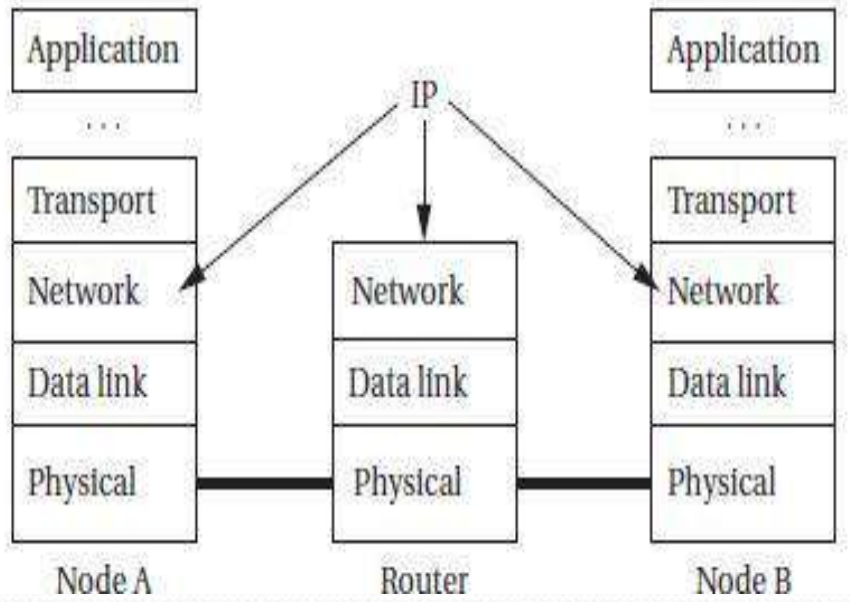
Ethernet-Packet format



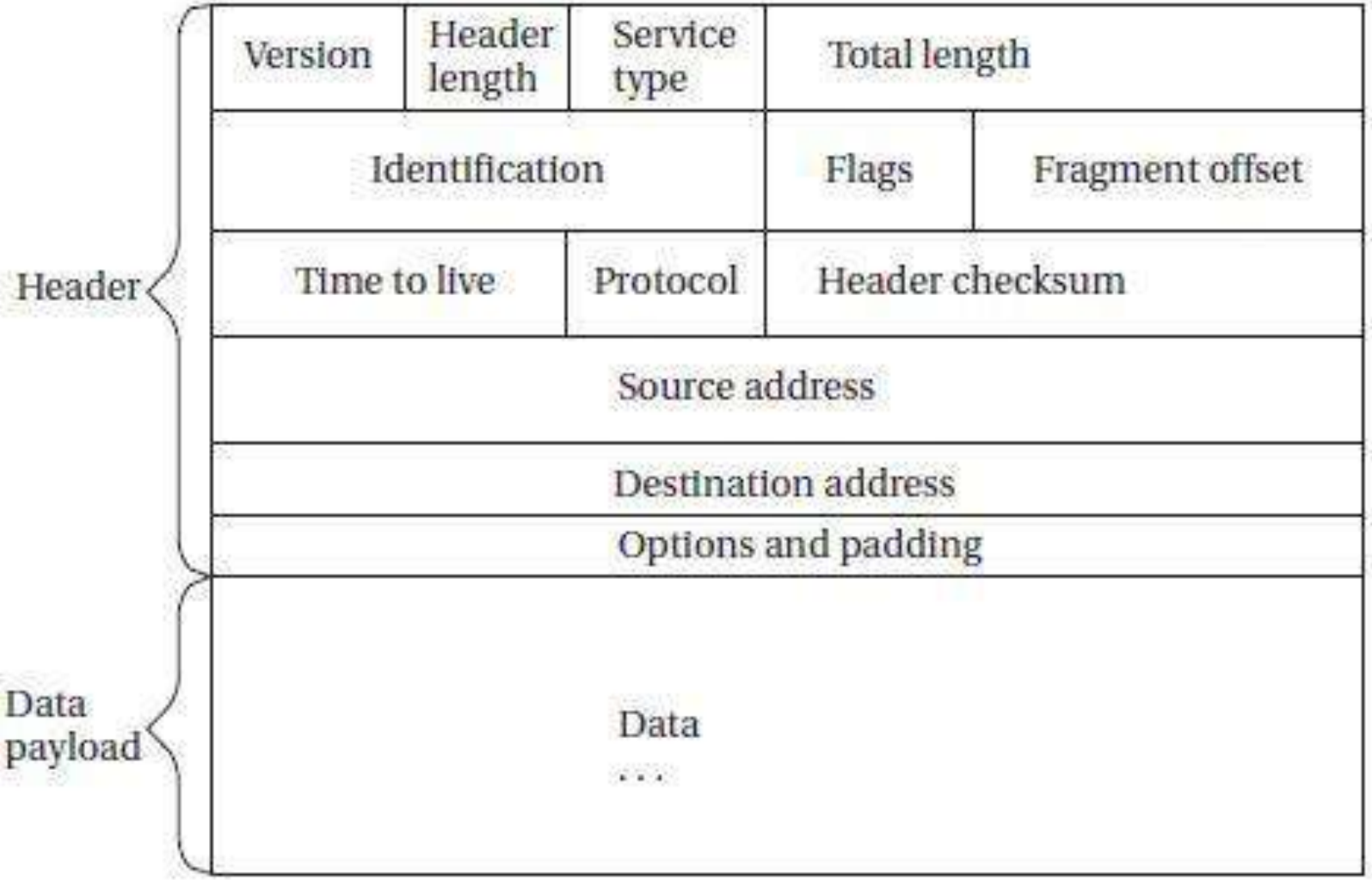
- Preamble→ 56-bit of alternating 1 and 0 bits, allowing devices on the network to easily synchronize their receiver clocks.
- SFD→8-bit ,indicates the beginning of the Ethernet frame
- Physical or MAC addresses → destination and the source(48-bit length)
- Length data payload→The minimum payload is 42 octets

INTERNET PROTOCOL(IP)

- It is the fundamental protocol on the Internet.
- It provides connection orientded, packet-based communication.
- It transmits packet over different networks from source to destination.
- It allows data to flow seamlessly from one end user t another.
- When node A wants to send data to node B, the data pass through several layers of the protocol stack to get to the Internet Protocol.
- IP creates packets for routing to the destination, which are then sent to the data link and physical layers.
- A packet may go through many routers to get to its destination.
- IP works at the network layer → does not guarantee that a packet is delivered to its destination.
- It supports best-effort routing packets→ packets that do arrive may come out of order.



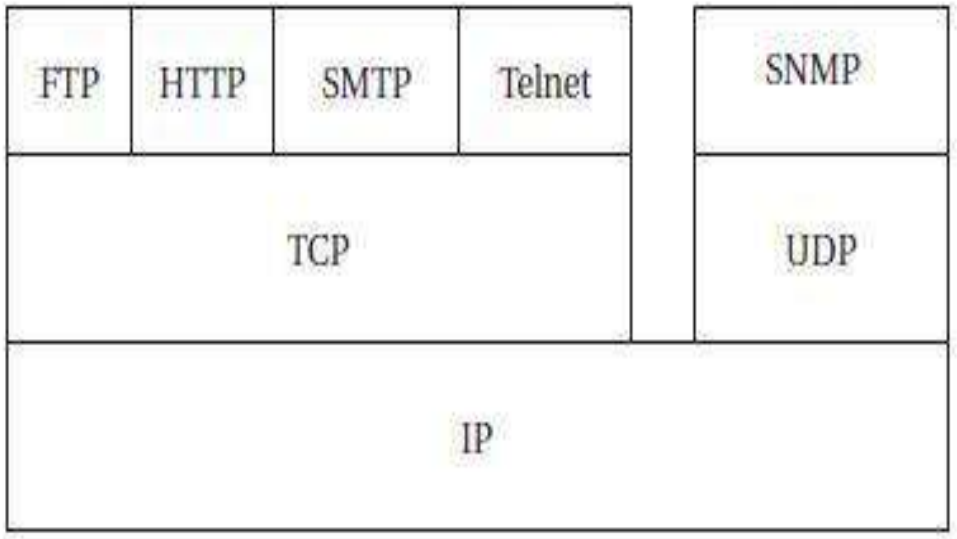
IP packet structure



- Version→ it ia s 4-bit field.used to identify v4 or v6.
- Header Length (HL)→It is a 4 bits, field.Indicates the length of the header.
- Service Type→it is a 8 bit field ,used to specify the type of service.
- Total length→Including header and data payload is 65,535 bytes.
- Identification→ identifying the group of fragments of a single IP datagram.
- Flags→ bit 0 Reserved.
  - bit 1: Don't Fragment (DF)
  - bit 2: More Fragments (MF)
- Fragment Offset→ It is 13 bits long , specifies the offset of a particular fragment relative to the beginning of the original unfragmented IP datagram
- Time To Live (TTL)→It is a 8 bit wide, indicates th datagram's lifetime
- Protocol→ protocol used in the data portion of the IP datagram
- Header Checksum→(16 bit) used for error-checking of the header
- Source address→ Sender packet address(32-bits size)
- Destination address→ Receiver packet address(32-bits size)

Transmission Control Protocol(TCP)

- It provides a connection-oriented service.
- It ensures that data arrive in the appropriate order.
- It uses an acknowledgment protocol to ensure that packets arrive.
- TCP is used to provide File Transport Protocol (FTP) for batch file transfers.
- Hypertext Transport Protocol (HTTP) for World Wide Web service.
- Simple Mail Transfer Protocol (SMTP) for email.
- Telnet for virtual terminals.
- User Datagram Protocol (UDP), is used to provide connection-less services.
- Simple Network Management Protocol (SNMP) provides the network management services.



MPSoCs and shared memory multiprocessors

- Shared memory processors are well-suited to applications that require a large amount of data to be processed(Signal processing systems)
- Most MPSoCs are shared memory systems.
- Shared memory allows for processors to communicate with varying patterns.
- If the pattern of communication is very fixed and if the processing of different steps is performed in different units, then a networked multiprocessor may be most appropriate.
- If one processing element is used for several different steps, then shared memory also allows the required flexibility in communication.

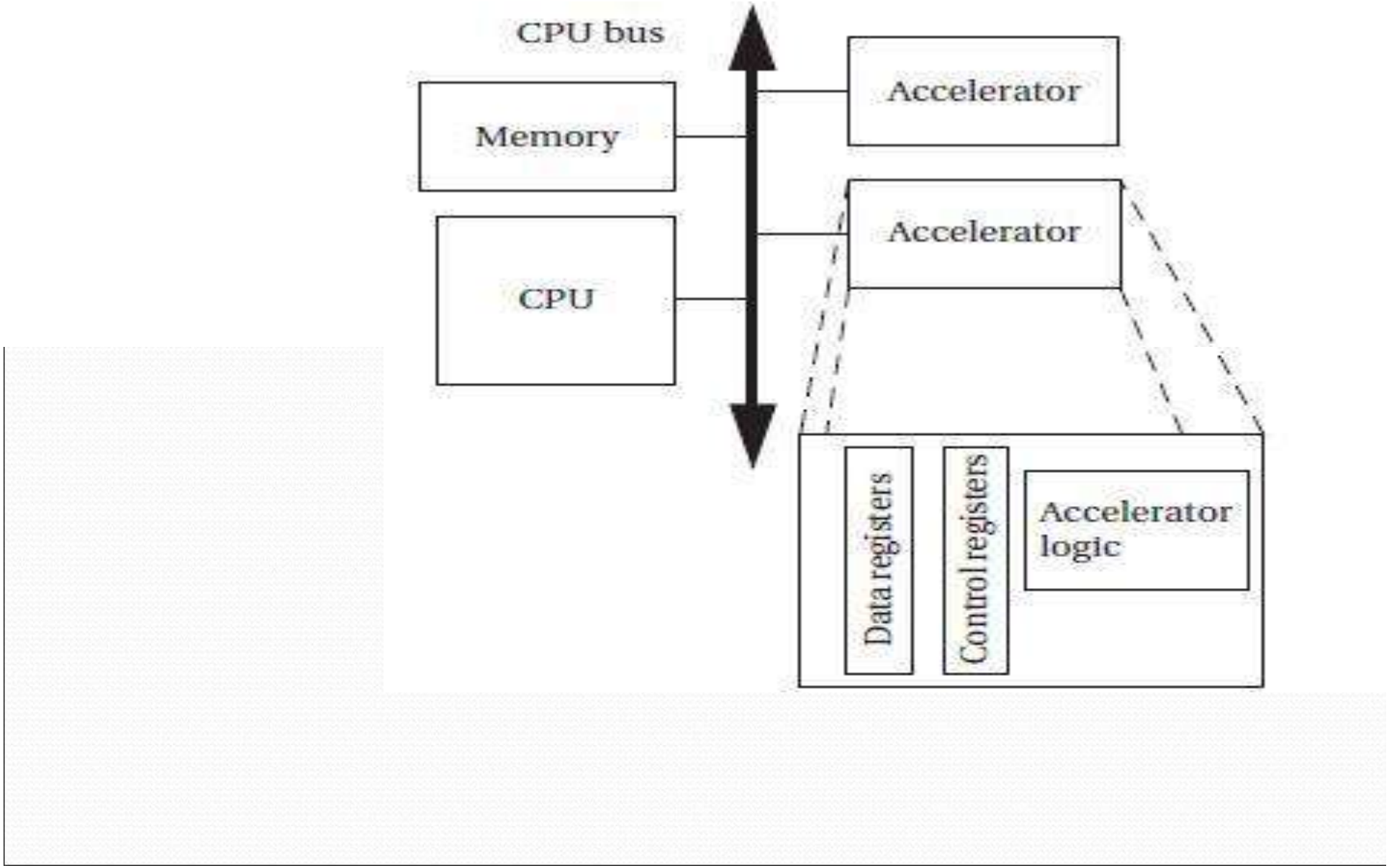
Heterogeneous shared memory multiprocessors

- Many high-performance embedded platforms are heterogeneous multiprocessors.
- Different processing elements (PE)perform different functions.
- PEs may be programmable processors with different instruction sets or specialized accelerators.
- Processors with different instruction sets can perform different tasks faster and using less energy.
- Accelerators provide even faster and lower-power operation for a narrow range of functions.

Accelerators

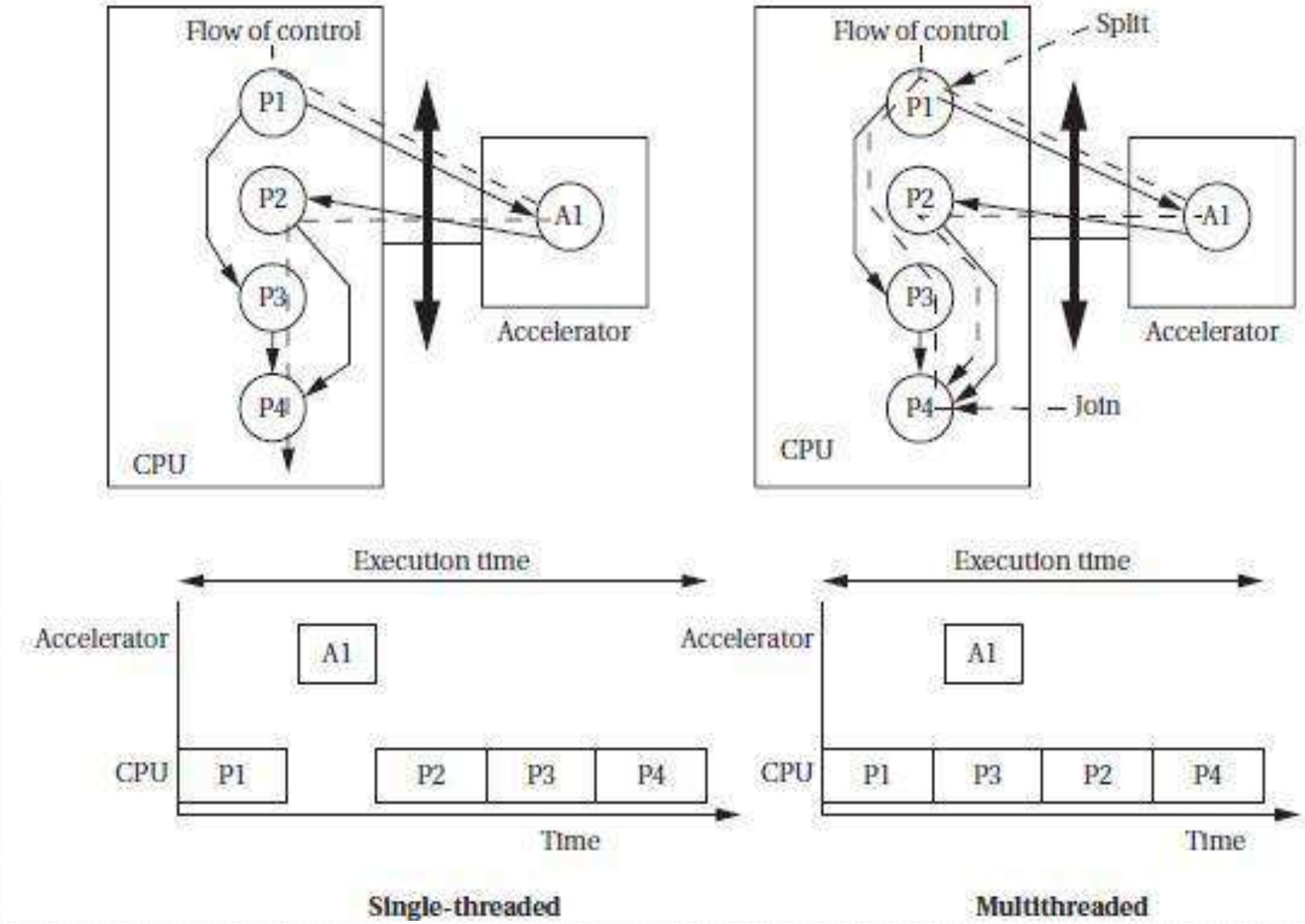
- It is the important processing element for embedded multiprocessors.
- It can provide large performance increases for applications with computational kernels .
- It can also provide critical speedups for low-latency I/O functions.
- CPU(host) accelerator is attached to the CPU bus.
- CPU talks to the accelerator through data and control registers in the accelerator.
- Control registers allow the CPU to monitor the accelerator's operation and to give the accelerator commands.
- The CPU and accelerator may also communicate via shared memory.
- The accelerator operate on a large volume of data with efficient data in memory.
- Accelerator read and write memory directly .
- The CPU and accelerator use synchronization mechanisms to ensure that they do not
- destroy each other's data.
- An accelerator is not a co-processor.
- A co-processor is connected to the internals of the CPU and processes instructions.
- An accelerator interacts with the CPU through the programming model interface.
- It does not execute instructions.
- CPU and accelerators performs computations for specification.

CPU accelerators in a system



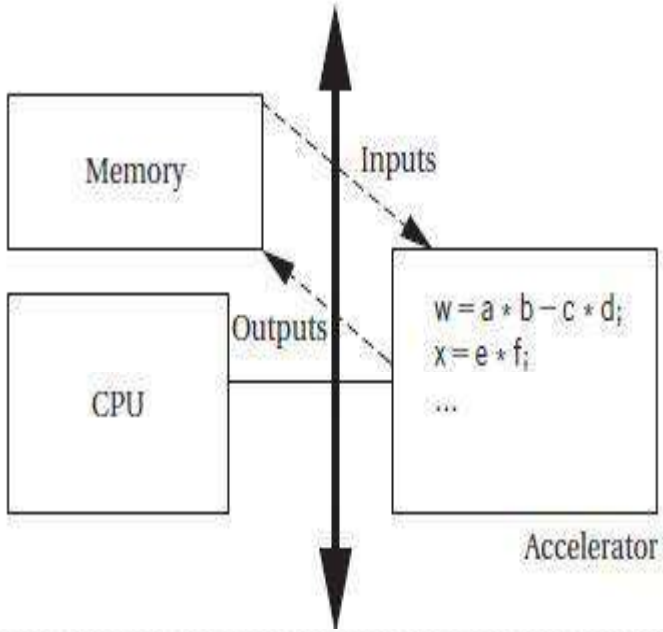
Accelerator Performance Analysis

- The speed factor of accelerator will depend on the following factors.
- Single threaded→CPU is in idle state while the accelerator runs.
- Multithreaded→CPU do some useful work in parallel with accelerator.
- Blocking→ CPU's scheduler block other operations wait for the accelerator call to complete.
- Non-blocking→ CPU's run some other work parallel with accelerator.
- Data dependencies allow P2 and P3 to run independently on the CPU.
- P2 relies on the results of the A1 process that is implemented by the accelerator.
- Single-threaded→ CPU blocks to wait for the accelerator to return the results of its computation.t, it doesn't matter whether P2 or P3 runs next on the CPU.
- Multithreaded → CPU continues to do useful work while the accelerator runs, so the CPU can start P3 just after starting the accelerator and finish the task earlier.


Flow of control
P1
P2
P3
P4
A1
Accelerator
CPU
Flow of control
Split
P1
P2
P3
P4
A1
Accelerator
Join
CPU
Execution time
Accelerator
A1
CPU
P1
P2
P3
P4
Time
Single-threaded
Execution time
Accelerator
A1
CPU
P1
P3
P2
P4
Time
Multithreaded

Components of execution time for an accelerator

- Execution time of a accelerator depends on the time required to execute the accelerator's function.
- It also depends on the time required to get the data into the accelerator and back out of it.
- Accelerator will read all its input data, perform the required computation,and write all its results.
- Total execution time given as
- $t_{acccel}=t_x+t_{in}+t_{out}$
- t_x →execution time of the accelerator
- T_{in} →times required for reading the required variables
- t_{out}- →times required for writing the required variables

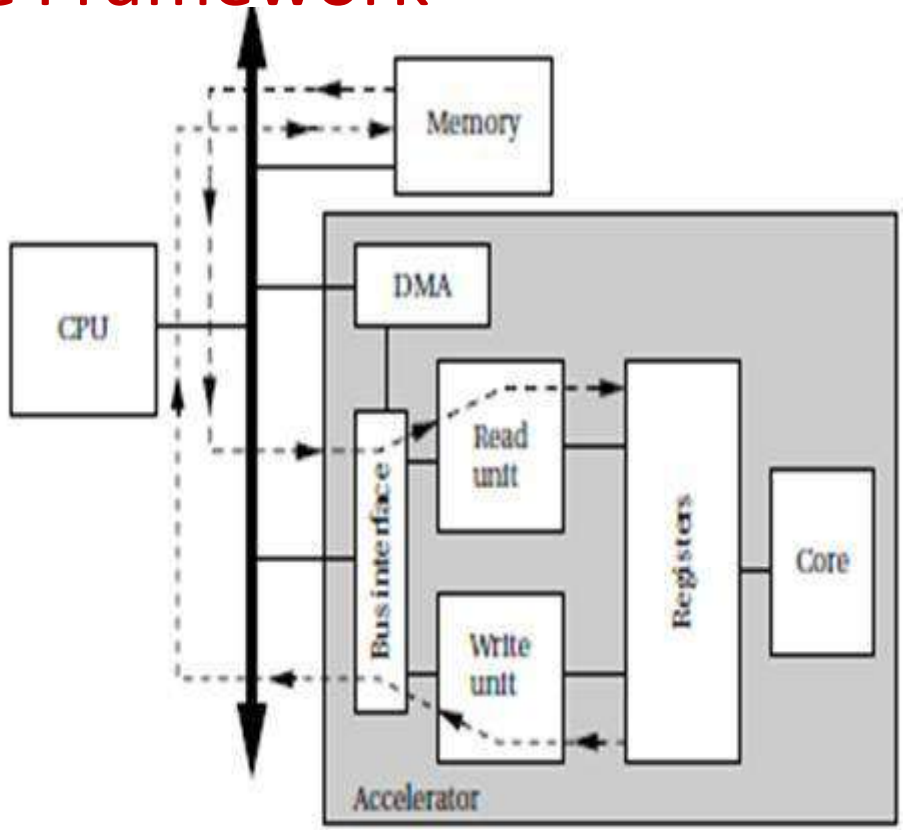


System Architecture Framework

- Architectural design depends on the application.

An accelerator can be considered from two angles.

- Accelerator core functionality
- Accelerator interface to the CPU bus.
- The accelerator core typically operates off internal registers.
- Requirement of number of registers is an important design decision.
- Main memory accesses will probably take multiple clock cycles.
- Status registers used to test the accelerator's state an to perform basic operations(starting, stopping, and resetting the accelerator)
- A register file in the accelerator acts as a buffer betwe main memory and the accelerator core.
- Read unit can read the accelerator's requirements and load the registers with the next required data.
- Write unit can send recently completed values to main memory.

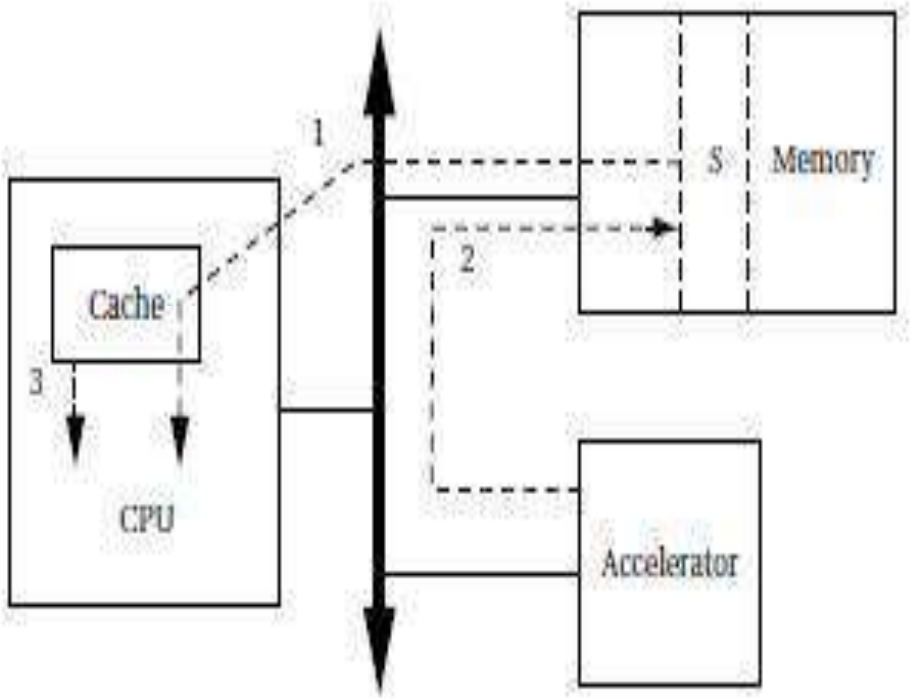


cache problem in an accelerated system

- CPU cache can cause problems for accelerators.

1. The CPU reads location S.
2. The accelerator writes S.
3. The CPU again reads S.

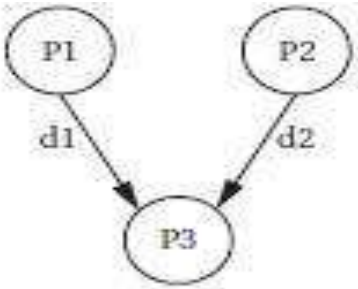
- If the CPU has cached location S ,the program will not see the value of S written by the accelerator. It will instead get the old value of S stored in the cache
- To avoid this problem, the CPU's cache must update the cache by setting cache entry is invalid.



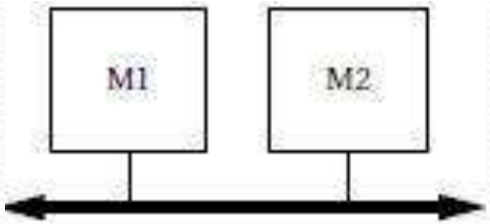
Scheduling and allocation

- Designing a distributed embedded system, depends upon the scheduling and allocation of resources.
- We must schedule operations in time, including communication on the network and computations on the processing elements.
- The scheduling of operations on the PEs and the communications between the PEs are linked.
- If one PE finishes its computations too late, it may interfere with another communication on the network as it tries to send its result to the PE that needs it.
- This is bad for both the PE that needs the result and the other PEs whose communication is interfered with.
- We must allocate computations to the processing elements.
- The allocation of computations to the PEs determines what communications are required—if a value computed on one PE is needed on another PE, it must be transmitted over the network.

- We can specify the system as a task graph. However, different processes may end up on different processing elements. Here is a task graph



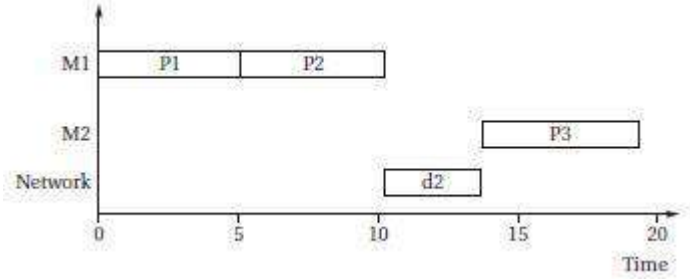
- We have labeled the data transmissions on each arc ,We want to execute the task on the platform below.



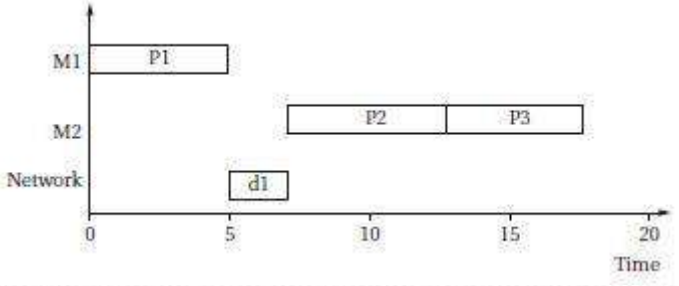
- The platform has two processing elements and a single bus connecting both PEs. Here are the process speeds:

	M1	M2
P1	5	5
P2	5	6
P3	—	5

- As an initial design, let us allocate P1 and P2 to M1 and P3 to M2This schedule shows what happens on all the processing elements and the network.



- The schedule has length 19. The d1 message is sent between the processes internal to
- P1 and does not appear on the bus.
- Let's try a different allocation. P1 on M1 and P2 and P3 on M2. This makes P2 run more slowly. Here is the new schedule:.
- The length of this schedule is 18, or one time unit less than the other schedule. The
- increased computation time of P2 is more than made up for by being able to transmit a
- shorter message on the bus. If we had not taken communication into account when analyzing total execution time, we could have made the wrong choice of which processes to put on the same processing element



Audio player/MP3 Player

Operation and requirements

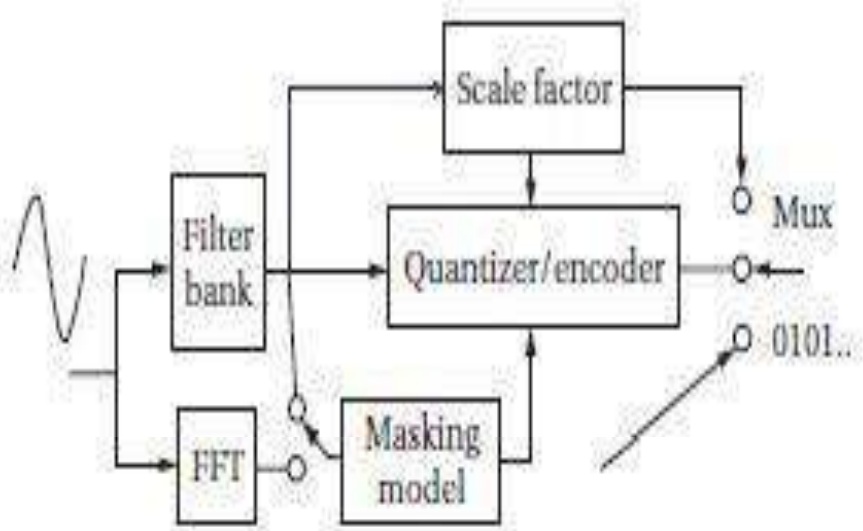
- MP3 players use either flash memory or disk drives to store music.
- It performs the following functions such as audio storage, audio decompression, and user interface.
- Audio compression→ It is a lossy process. The coder eliminates certain features of the audio stream so that the result can be encoded in fewer bits.
- Audio decompression→ The incoming bit stream has been encoded using a Huffman style code, which must be decoded.
- Masking→ One tone can be masked by another if the tones are sufficiently close in frequency.

Audio compression standards

- Layer 1 (MP1) →uses a lossless compression of sub bands and simple masking model.
- Layer 2 (MP2) →uses a more advanced masking model.
- Layer 3 (MP3)→ performs additional processing to provide lower bit rates.

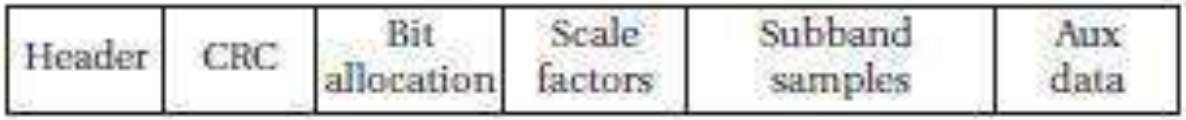
MPEG Layer 1 encoder

- Filter bank→ splits the signal into a set of 32 sub-bands that are equally spaced in the frequency domain and together cover the entire frequency range of the audio.
- Encoder→It reduce the bit rate for the audio signals.
- Quantizer →scales each sub-band(fits within 6 bits), then quantizes based upon the current scale factor for that sub-band.
- Masking model → It is driven by a separate Fast Fourier transform (FFT), the filter bank could be used for masking, a separate FFT provides better results.
- The masking model chooses the scale factors for the sub-bands, which can change along with the audio stream.
- Multiplexer→ output of the encoder passes along all the required data.



MPEG Layer 1 data frame format

- A frame carries the basic MPEG data, error correction codes, and additional information.
- After disassembling the data frame, the data are un-scaled and inverse quantized to produce sample streams for the sub-band.



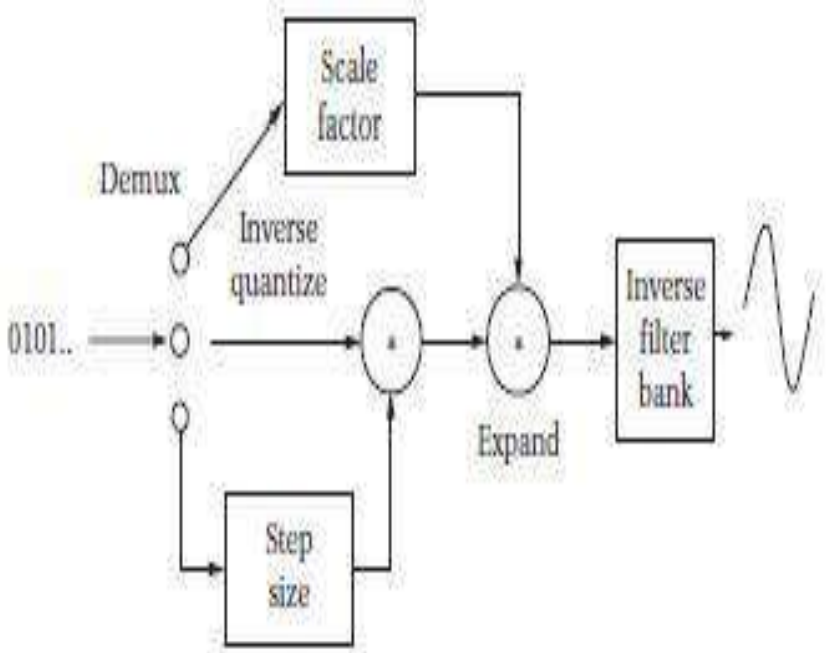
MPEG Layer 1 decoder

•After disassembling the data frame, the data are un-scaled and inverse quantized to produce sample streams for the sub-band.

•An inverse filter bank then reassembles the sub-bands into the uncompressed signal.

User interface→ MP3 player is simple both the physical size and power consumption of the device. Many players provide only a simple display and a few buttons.

File system→ player generally must be compatible with PCs. CD/MP3 players used compact discs that had been created on PCs.

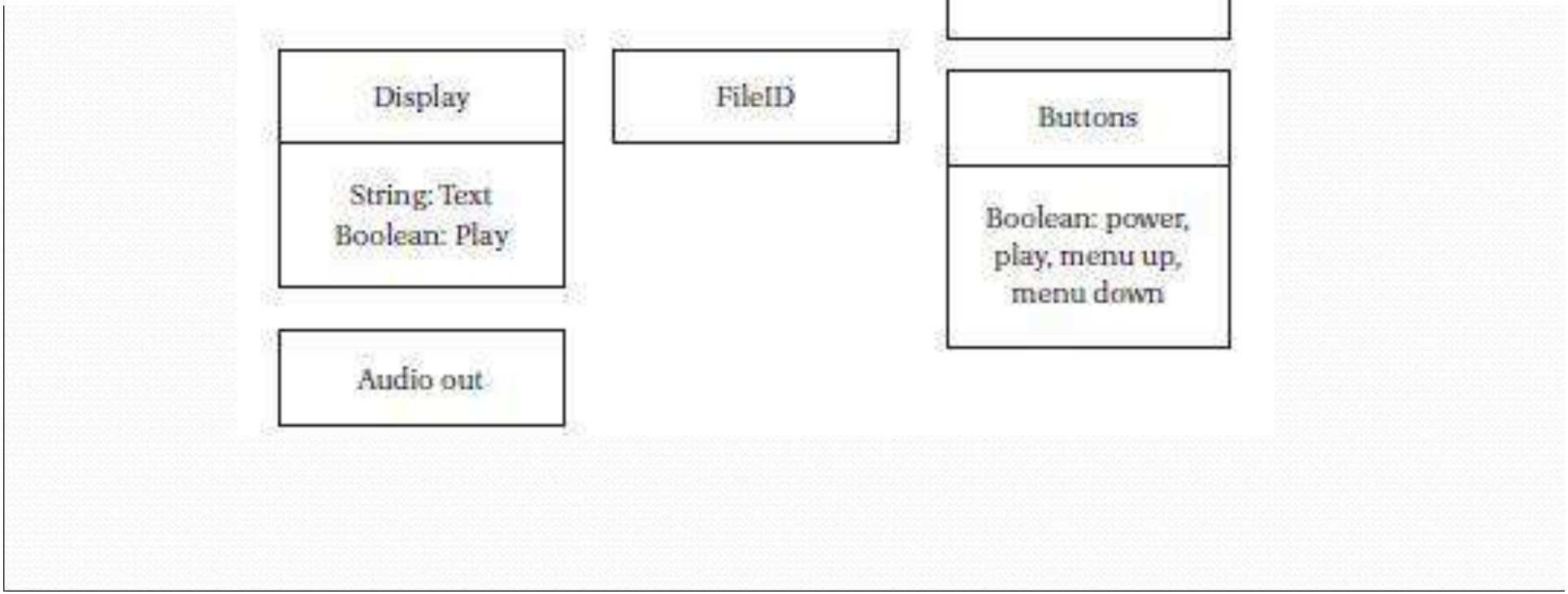


Requirements

Name	Audio player
Purpose	Play audio from files.
Inputs	Flash memory socket, on/off, play/stop, menu up/down.
Outputs	Speaker
Functions	Display list of files in flash memory, select file to play, play file.
Performance	Sufficient to play audio files at required rate.
Manufacturing cost	Approximately $25
Power	1 AAA battery
Physical size and weight	Approx. 1 in x 2 in, less than 2 oz

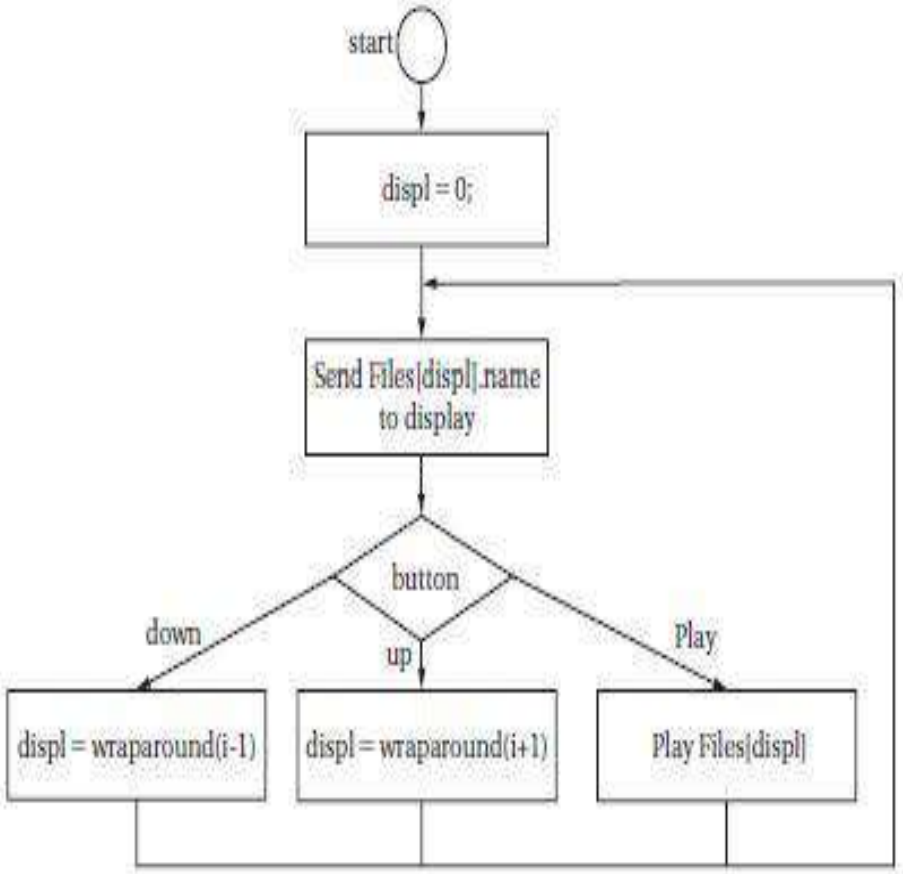
Specification

- The File ID class is an abstraction of a file in the flash file system.
- The controller class provides the method that operates the player.



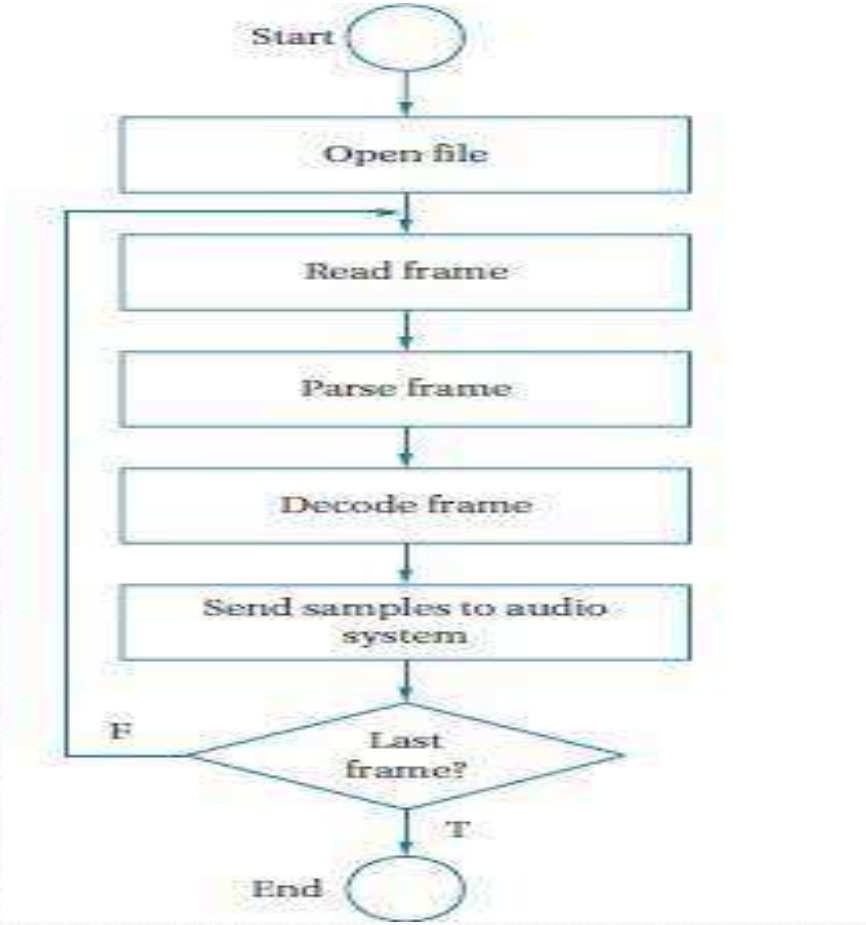
State diagram for file display and selection

- This specification assumes that all files are in the root directory and that all files are playable audio.



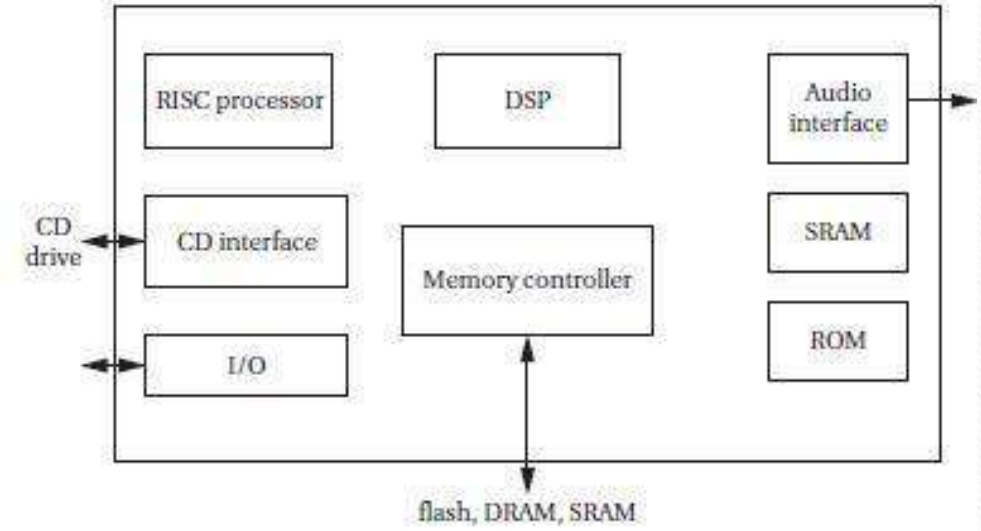
State diagram for Audio Playback

- It refers to sending the samples to the audio system.
- Playback and reading the next data frame must be overlapped to ensure continuous operation.
- The details of playback depend on the hardware platform selected, but will probably involve a DMA transfer.



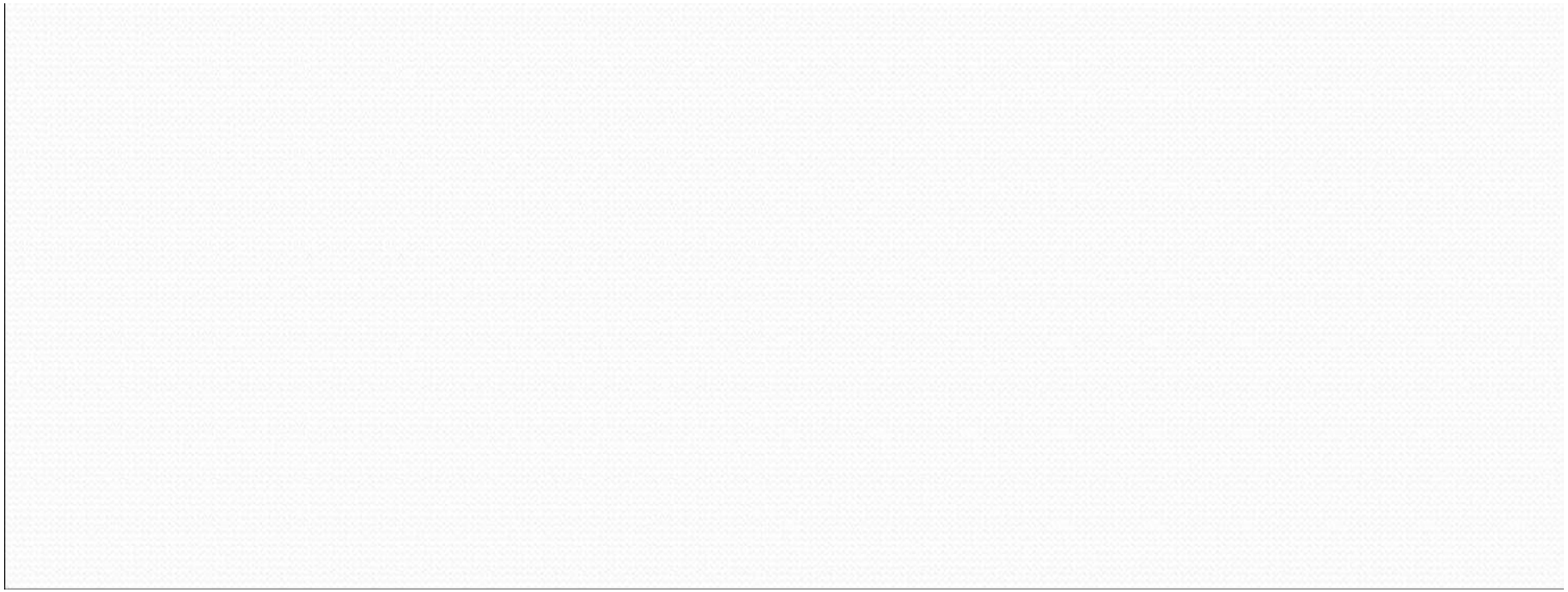
System architecture

- The audio controller includes two processors.
- The 32-bit RISC processor is used to perform system control and audio decoding.
- The 16-bit DSP is used to perform audio effects such as equalization.
- The memory controller can be interfaced to several different types of memory.
- Flash memory can be used for data or code storage.
- DRAM can be used to handle temporary disruptions of the CD data stream.
- The audio interface unit puts out audio in formats that can be used by A/D converters.
- General- purpose I/O pins can be used to decode buttons, run displays.



Component design and testing

- The audio output system should be tested separately from the compression system.
- Testing of audio decompression requires sample audio files.
- The standard file system can either implement in a DOS FAT or a new file system.
- While a non-standard file system may be easier to implement on the device, it also requires software to create the file system.
- The file system and user interface can be tested independently .



System integration and debugging

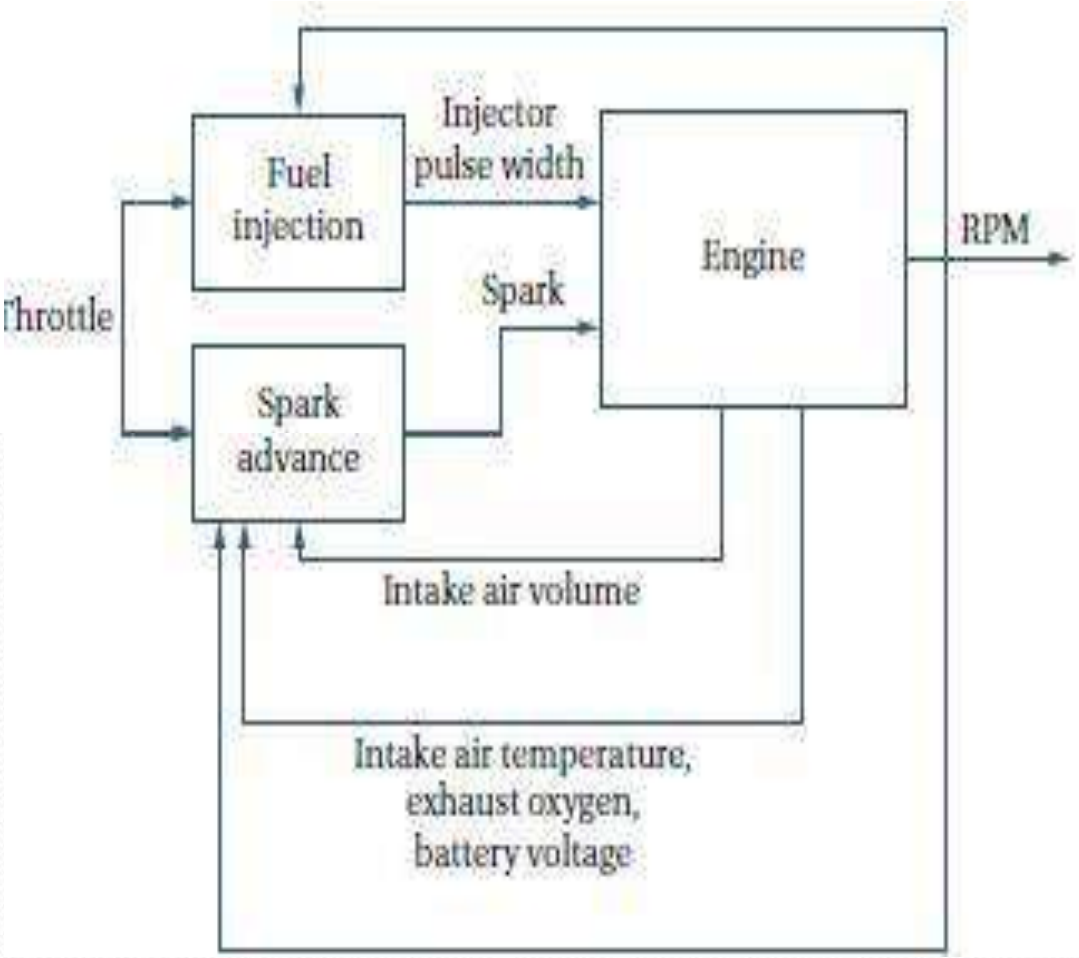
- It ensure that audio plays smoothly and without interruption.
- Any file access and audio output that operate concurrently should be separately tested, ideally using an easily recognizable test signal.

Engine Control Unit

- This unit controls the operation of a fuel-injected engine based on several measurements taken from the running engine.

Operation and Requirements

- The throttle is the command input.
- The engine measures throttle, RPM, intake air volume, and other variables.
- The engine controller computes injector pulse width and spark.



Requirements

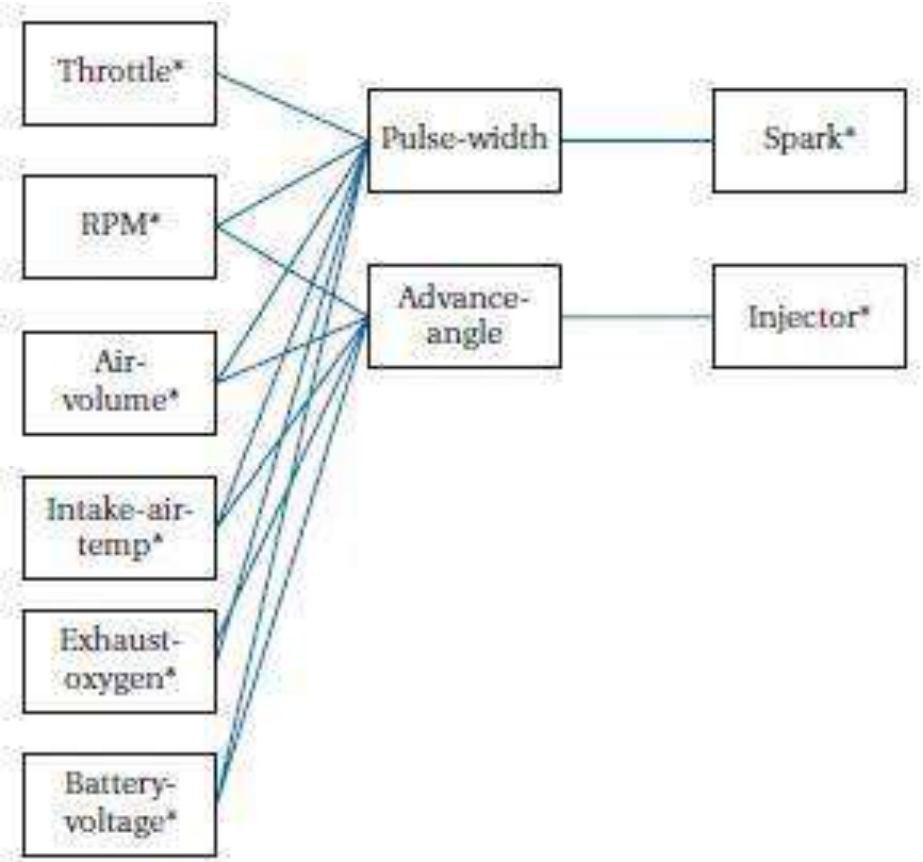
Name	ECU
Purpose	Engine controller for fuel-injected engine
Inputs	Throttle, RPM, intake air volume, intake manifold pressure
Outputs	Injector pulse width, spark advance angle
Functions	Compute injector pulse width and spark advance angle as a function of throttle, RPM, intake air volume, intake manifold pressure
Performance	Injector pulse updated at 2-ms period, spark advance angle updated at 1-ms period
Manufacturing cost	Approximately $50
Power	Powered by engine generator
Physical size and weight	Approx 4 in × 4 in, less than 1 pound.

Specification

- The engine controller must deal with processes at different rates
- ΔNE and ΔT to represent the change in RPM and throttle position.
- Controller computes two output signals, injector pulse width PW and spark advance angle S.
  - $S=k_2X\ \Delta NE-k_3VS$
- The controller then applies corrections to these initial values
- If intake air temperature (THA) increases during engine warm-up, the controller reduces the injection duration.
- If the throttle opens, the controller temporarily increases the injection frequency.
- Controller adjusts duration up or down based upon readings from the exhaust oxygen sensor (OX).

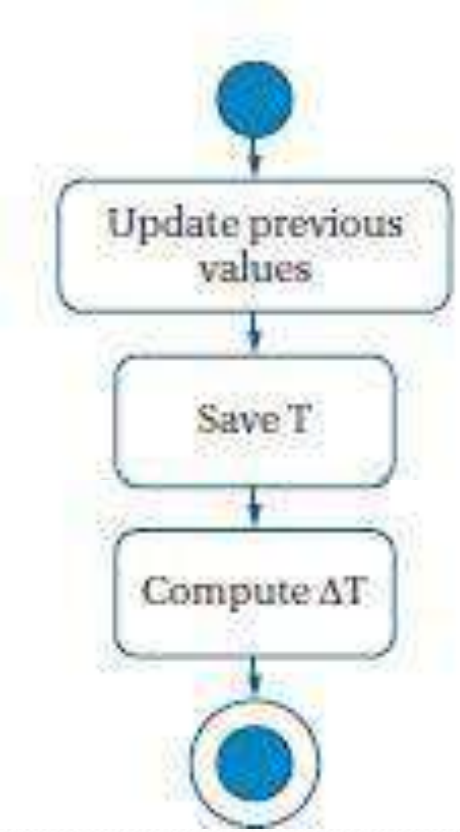
System architecture

- The two major processes, pulse-width and advance-angle, compute the control parameters for the spark plugs and injectors.
- Control parameters rely on changes in some of the input signals.
- Physical sensor classes used to compute these values.
- Each change must be updated at the variable's sampling rate.



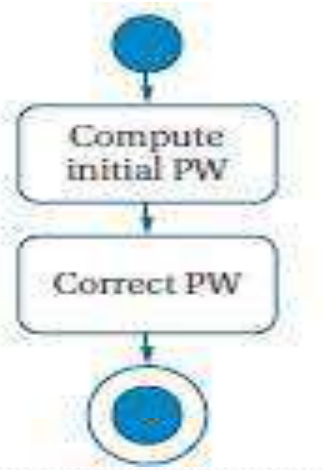
State diagram for throttle position sensing

- Throttle sensing, which saves both the current value and change in value of the throttle.

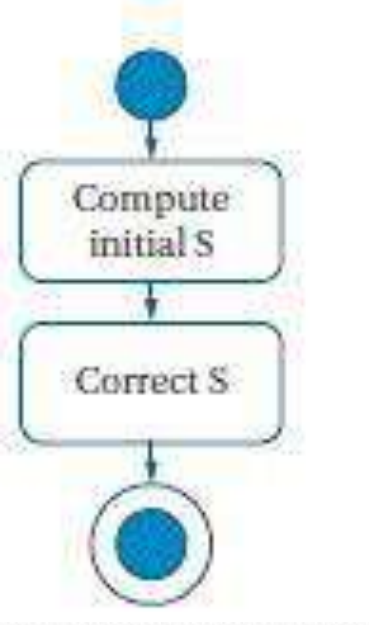


State diagram for injector pulsewidth

- In each case, the value is computed in two stages, first an initial value followed by a correction.

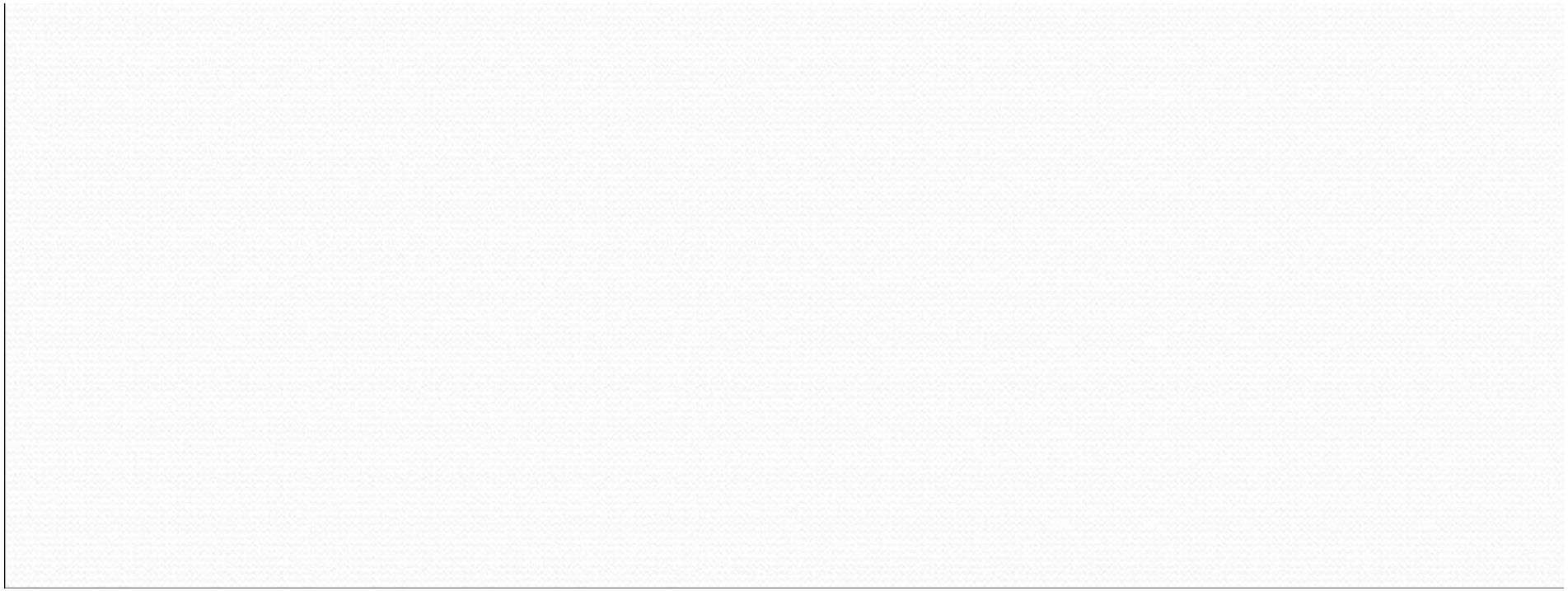


State diagram for spark advance angle



Component design and testing

- Various tasks must be coded to satisfy the requirements of RTOS processes.
- Variables that are maintained across task execution, such as the change-of-state variables, must be allocated and saved in appropriate memory locations.
- Some of the output variables depend on changes in state, these tasks should be tested with multiple input variable sequences to ensure that both the basic and adjustment calculations are performed correctly.

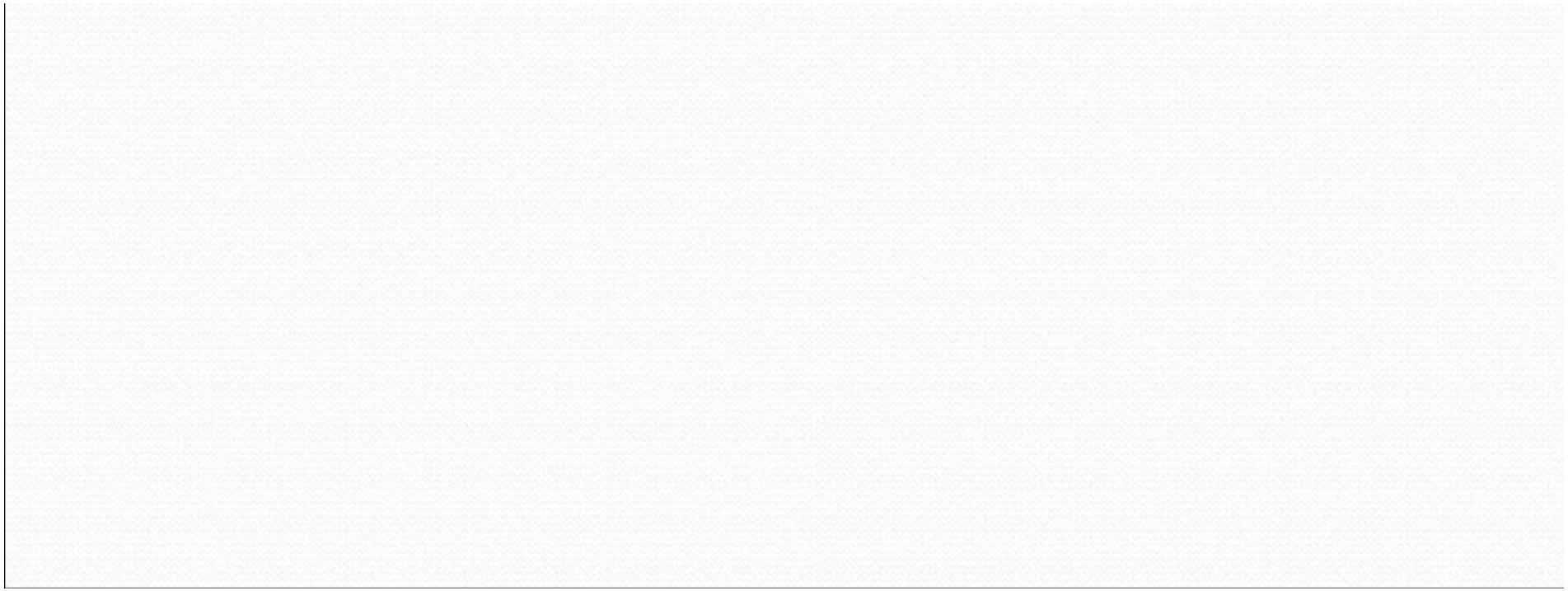


System integration and testing

- Engines generate huge amounts of electrical noise that can cripple digital electronics.
- They also operate over very wide temperature ranges.

1. hot during engine operation,
2. potentially very cold before the engine is started.

- Any testing performed on an actual engine must be conducted using an engine controller that has been designed to withstand the harsh environment of the engine compartment.



Video Accelerator

- It is a hardware circuits on a display adapter that speed up fill motion video.
- Primary video accelerator functions are color space conversion, which converts YUV to RGB.
- Hardware scaling is used to enlarge the image to full screen and double buffering which moves the frames into the frame buffer faster.

Video compression

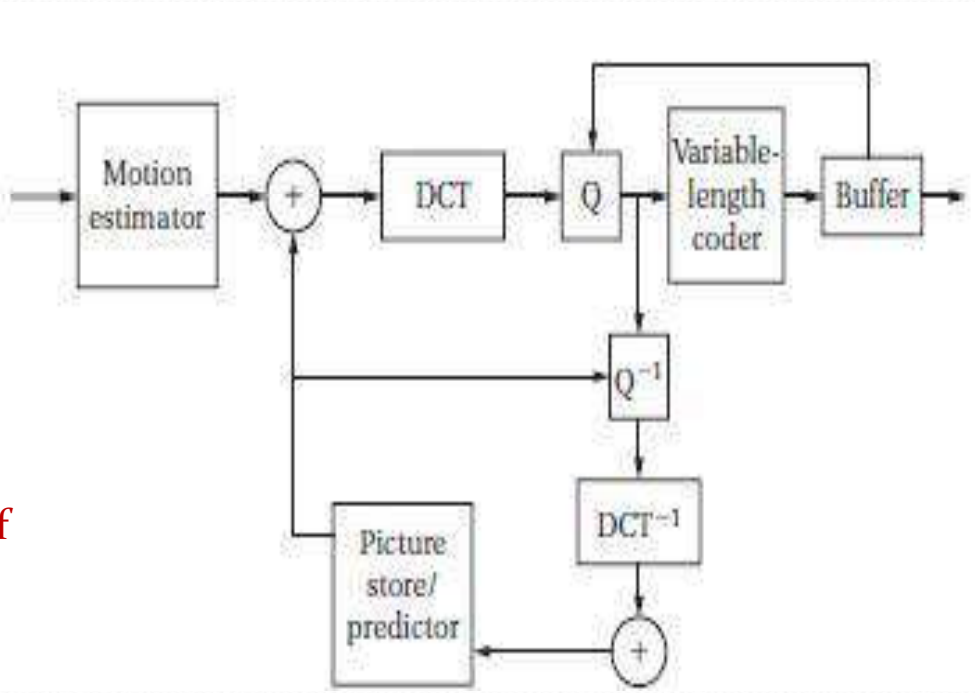
•MPEG-2 forms the basis for U.S. HDTV broadcasting.

• This compression uses several component algorithms together in a feedback loop.

•Discrete cosine transform (DCT) used in JPEG and MPEG-2.

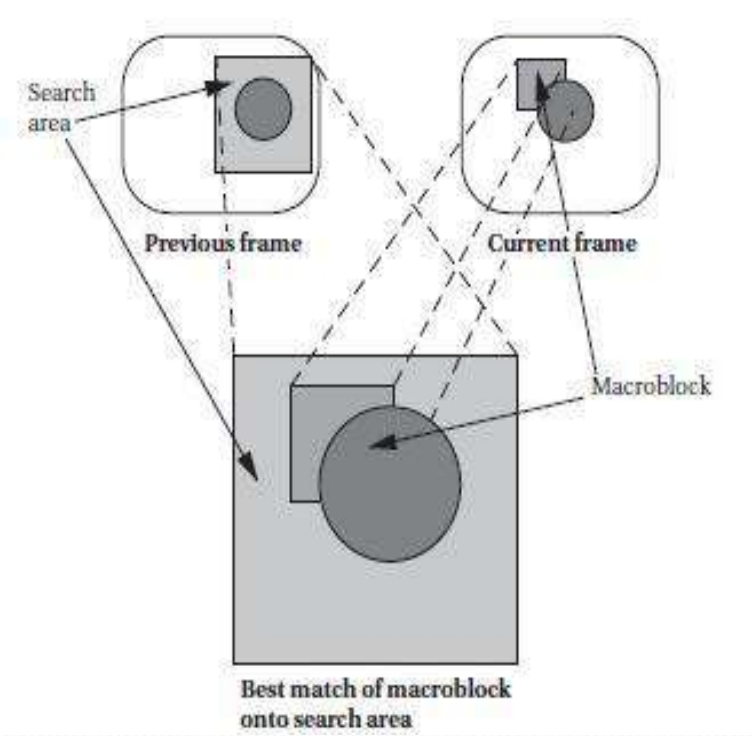
•DCT used a block of pixels which is quantized for lossy compression.

•Variable-length coder→assign number of bits required to represent the block.



Block motion Estimation

- MPEG uses motion to encode one frame in terms of another.
- Block motion estimation→some frames are sent as modified forms of other frames
- During encoding, the frame is divided into macro blocks.
- Encoder uses the encoding information to recreate the lossily-encoded picture, compares it to the original frame, and generates an error signal.
- Decoder keep recently decoded frames in memory so that it can retrieve the pixel values of macro-blocks.



5.13.2).Concept of Block motion estimation

- To find the best match between regions in the two frames.
- Divide the current frame into 16 x 16 macro blocks.
- For every macro block in the frame, to find the region in the previous frame that most closely matches the macro block.
- Measure similarity using the following sum-of-differences measure

$$\sum_{1 \le i,j \le n} |M(i,j) - S(i - o_x, j - o_y)|$$

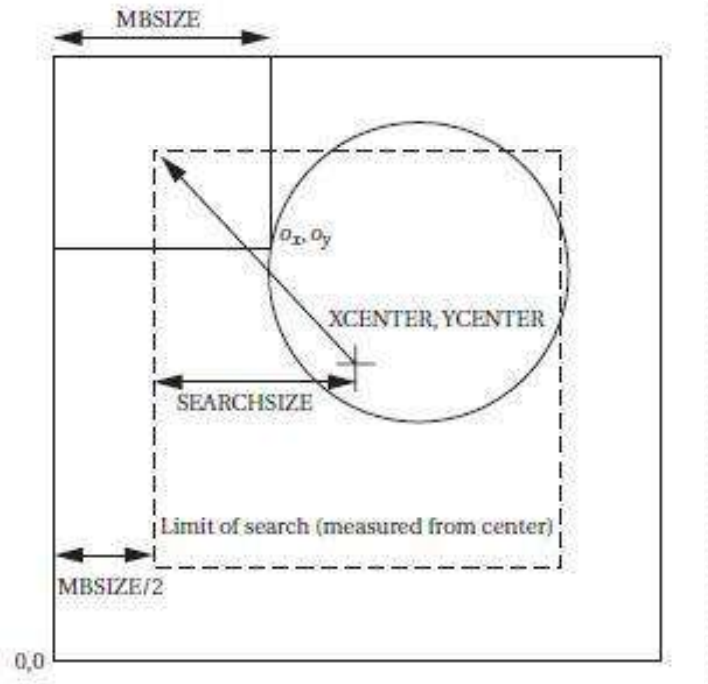
- M(i,j) → intensity of the macro block at pixel i,j,
- S(i,j) → intensity of the search region
- N→ size of the macro block in one dimension
- <ox, oy>→offset between the macro block and search region
- We choose the macro block position relative to the search area that gives us the smallest value for this metric.
- The offset at this chosen position describes a vector from the search area center to the macro block's center that is called the motion vector.

Algorithm and requirements

- C code for a single search, which assumes that the search region does not extend past the boundary of the frame.
- The arithmetic on each pixel is simple, but we have to process a lot of pixels.
- If MBSIZE is 16 and SEARCHSIZE is 8, and remembering that the search distance in each
- dimension is 8 + 1 + 8, then we must perform

$$n_{ops} = (16 \times 16) \times (17 \times 17) = 73{,}984$$

```
bestx = 0; besty = 0; /* initialize best location--none yet */
bestsad = MAXSAD; /* best sum-of--difference thus far */
for (ox = -SEARCHSIZE; ox < SEARCHSIZE; ox++) {
    /* x search ordinate */
    for (oy = -SEARCHSIZE; oy < SEARCHSIZE; oy++) {
        /* y search ordinate */
        int result = 0;
        for (i = 0; i < MBSIZE; i++) {
            for (j = 0; j < MBSIZE; j++) {
                result = result + iabs(mb[i][j] -
                    search[i -ox + XCENTER][j - oy + YCENTER]);
            }
        }
        if (result <= bestsad) { /* found better match */
            bestsad = result;
            bestx = ox; besty = oy;
        }
    }
}
```

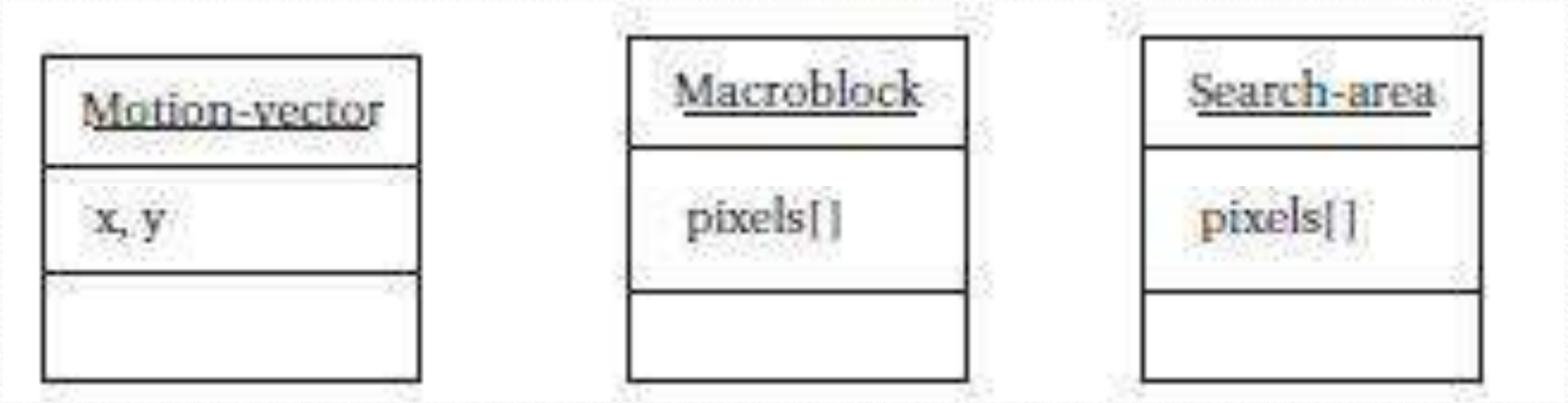


Requirements

Name	Block motion estimator
Purpose	Perform block motion estimation within a PC system
Inputs	Macroblocks and search areas
Outputs	Motion vectors
Functions	Compute motion vectors using full search algorithms
Performance	As fast as we can get
Manufacturing cost	Hundreds of dollars
Power	Powered by PC power supply
Physical size and weight	Packaged as PCI card for PC

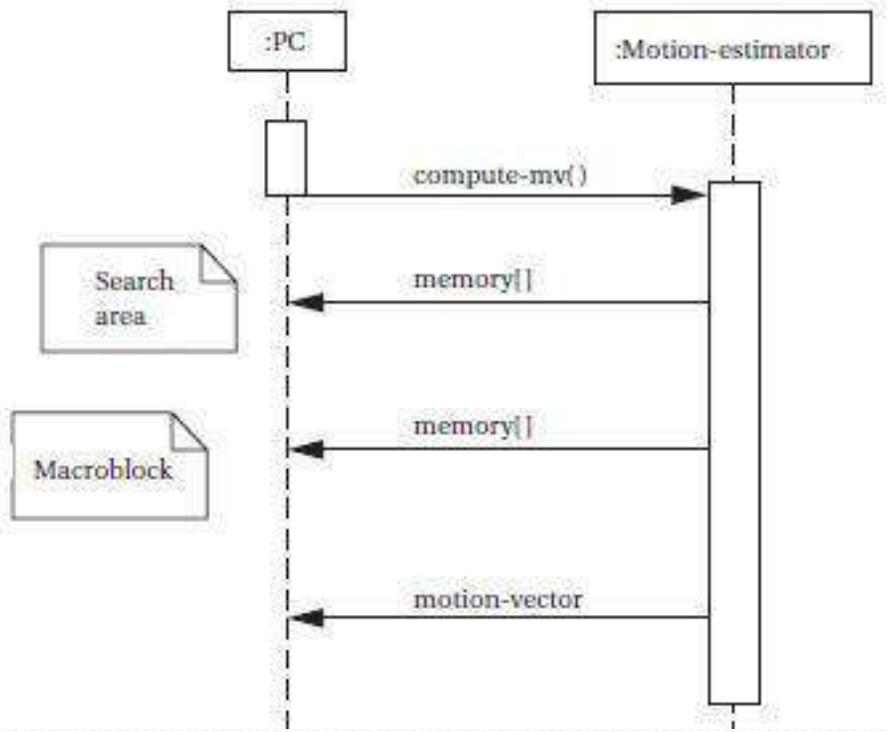
Specification

- Specification for the system is relatively straightforward because the algorithm is simple.
- The following classes used to describe basic data types in the system motion vector, macro block, search area.



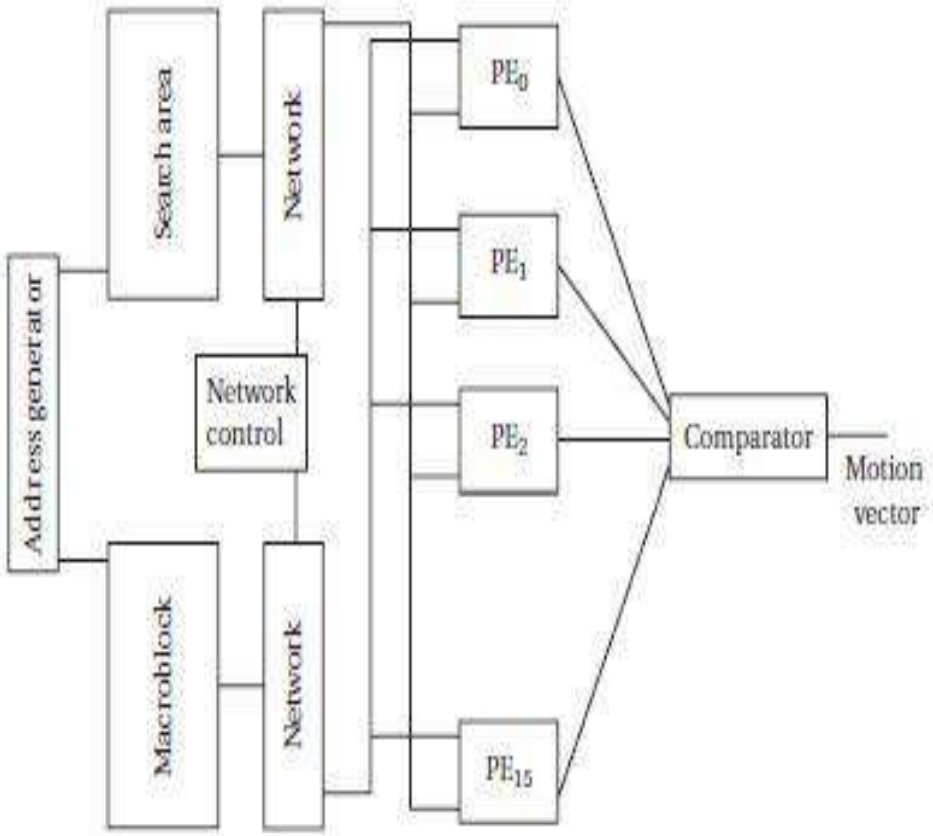
Sequence Diagram

- The accelerator provides a behavior compute-mv() that performs the block motion estimation algorithm.
- After initiating the behavior, the accelerator reads the search area and macro block from the PC, after computing the motion vector, it returns it to the PC.



Architecture

- The macro block has 16 x16 = 256.
- The search area has $(8 + 8 + 1 + 8 + 8)^2 =$ 1,089 pixels.
- FPGA probably will not have enough memory to hold 1,089 (8-bit)values.
- The machine has two memories, one for the macro block and another for the search memories.
- It has 16 processing elements that perform the difference calculation on a pair of pixels.
- Comparator sums them up and selects the best value to find the motion vector.



System testing

- Testing video algorithms requires a large amount of data.
- we are designing only a motion estimation accelerator and not a complete video compressor, it is probably easiest to use images, not video, for test data.
- use standard video tools to extract a few frames from a digitized video and store them in JPEG format.
- Open source for JPEG encoders and decoders is available.
- These programs can be modified to read JPEG images and put out pixels in the format required by your accelerator.

MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

EC8094 SATELLITE COMMUNICATION

Semester - 08

Notes

MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

Vision

To excel in providing value based education in the field of Electronics and Communication Engineering, keeping in pace with the latest technical developments through commendable research, to raise the intellectual competence to match global standards and to make significant contributions to the society upholding the ethical standards.

Mission

- ✓ To deliver Quality Technical Education, with an equal emphasis on theoretical and practical aspects.
- ✓ To provide state of the art infrastructure for the students and faculty to upgrade their skills and knowledge.
- ✓ To create an open and conducive environment for faculty and students to carry out research and excel in their field of specialization.
- ✓ To focus especially on innovation and development of technologies that is sustainable and inclusive, and thus benefits all sections of the society.
- ✓ To establish a strong Industry Academic Collaboration for teaching and research, that could foster entrepreneurship and innovation in knowledge exchange.
- ✓ To produce quality Engineers who uphold and advance the integrity, honour and dignity of the engineering.

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

1. To provide the students with a strong foundation in the required sciences in order to pursue studies in Electronics and Communication Engineering.
2. To gain adequate knowledge to become good professional in electronic and communication engineering associated industries, higher education and research.
3. To develop attitude in lifelong learning, applying and adapting new ideas and technologies as their field evolves.
4. To prepare students to critically analyze existing literature in an area of specialization and ethically develop innovative and research oriented methodologies to solve the problems identified.
5. To inculcate in the students a professional and ethical attitude and an ability to visualize the engineering issues in a broader social context.

PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO1: Design, develop and analyze electronic systems through application of relevant electronics, mathematics and engineering principles.

PSO2: Design, develop and analyze communication systems through application of fundamentals from communication principles, signal processing, and RF System Design & Electromagnetics.

PSO3: Adapt to emerging electronics and communication technologies and develop innovative solutions for existing and newer problems.

SATELLITE COMMUNICATION

UNIT I SATELLITE ORBITS

THEORY

1.1 Introduction to satellite communication

Satellites are specifically made for telecommunication purpose. They are used for mobile applications such as communication to ships, vehicles, planes, hand -held terminals and for TV and radio broadcasting.

They are responsible for providing these services to an assigned region (area) on the earth. The power and bandwidth of these satellites depend upon the preferred size of the footprint, complexity of the traffic control protocol schemes and the cost of ground stations.

A satellite works most efficiently when the transmissions are focused with a desired area.

When the area is focused, then the emissions don ‟ t go outside that designated area and thus minimizing the interference to the other systems. This leads more efficient spectrum usage.

Satellites antenna patterns play an important role and must be designed to best cover the designated geographical area (which is generally irregular in shape).

Satellites should be designed by keeping in mind its usability for short and long term effects throughout its life time.

The earth station should be in a position to control the satellite if it drifts from its orbit it is subjected to any kind of drag from the external forces.

Applications of Satellites:

- Weather Forecasting
- Radio and TV Broadcast
- Military Satellites
- Navigation Satellites
- Global Telephone
- Connecting Remote Area
- Global Mobile Communication

1.2 Kepler's laws

1.2.1 Kepler's law Introduction

Satellites (spacecraft) orbiting the earth follow the same laws that govern the motion of the planets around the sun.

Kepler's laws apply quite generally to any two bodies in space which interact through gravitation. The more massive of the two bodies is referred to as the *primary,* the other, the *secondary* or *satellite*.

1.2.2 Kepler's First Law

Kepler's first law states that the path followed by a satellite around the primary will be an ellipse. An ellipse hast Two focal points shown as $F1$ and $F2$in Fig. 2.1. The center of mass of the two-body system, termed the *bary center,* is always center of the foci.

The semi major axis of the ellipse is denoted by a, and the semi minor axis, by b. The eccentricity e is given by

$$e = \frac{\sqrt{a^2 - b^2}}{a}$$

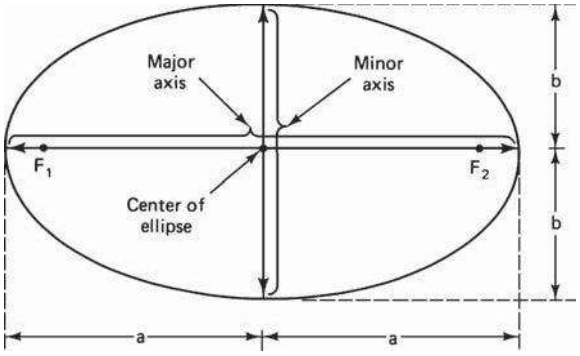


Figure 1.1 The foci $F1$ and $F2$, the semi major axis a, and the semi minor axis b of an ellipse.

1.2.3 Kepler's Second Law

Kepler's second law states that, for equal time intervals, a satellite will sweep out equal areas in its orbital plane, focused at the barycenter. Referring to Fig. 2.2, assuming the satellite travels distances $S1$ and $S2$ meters in 1 s, then the areas $A1$ and $A2$ will be equal. The average velocity in each case is $S1$ and $S2$ m/s, and because of the equal area law, it follows that the velocity at $S2$ is less than that at $S1$.

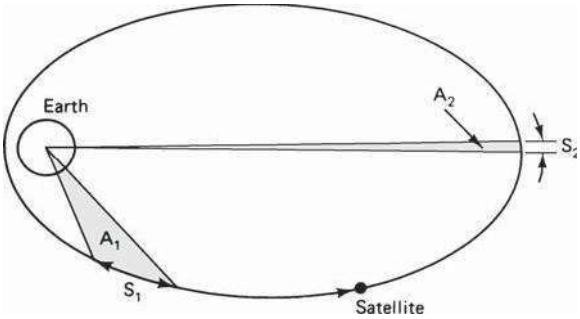


Figure 1.2 Kepler's second law. The areas $A1$and $A2$swept out in unit time are equal.

1.2.4 Kepler's Third Law

Kepler's third law states that the square of the periodic time of orbit is proportional to the cube of the mean distance between the two bodies. The mean distance is equal to the semi major axis *a*.

For the artificial satellites orbiting the earth, Kepler's third law can be written in the form

$$a^3 = \mu/n^2$$

Where n is the mean motion of the satellite in radians per second and is the earth's geocentric gravitational constant μ=3.986005 X 1014m3/s2

1.3. Newton's law:

1.3.1 Newton's first law

An object at rest will remain at rest unless acted on by an unbalanced force. An object in motion continues in motion with the same speed and in the same direction unless acted upon by an unbalanced force. This law is often called "the law of inertia".

1.3.2 Newton's second law

Acceleration is produced when a force acts on a mass. The greater the mass (of the object being accelerated) the greater the amount of force needed (to accelerate the object).

1.3.3 Newton's first law

For every action there is an equal and opposite re-action. This means that for every force there is a reaction force that is equal in size, but opposite in direction. That is to say that whenever an object pushes another object it gets pushed back in the opposite direction equally hard.

1.4. orbital parameters

Apogee: A point for a satellite farthest from the Earth. It is denoted as **ha.**

Perigee**:** A point for a satellite closest from the Earth. It is denoted as **hp.**

Line of Apsides: Line joining perigee and apogee through centre of the Earth. It is the major axis of the orbit. One-half of this line‟s length is the semi-major axis equivalents to satellite‟s mean distance from the Earth.

Ascending Node: The point where the orbit crosses the equatorial plane going from north to south.

Descending Node: The point where the orbit crosses the equatorial plane going from south to north.

Inclination: the angle between the orbital plane and the Earth‟s equatorial plane. Its measured at the ascending node from the equator to the orbit, going from East to North. Also, this angle is commonly denoted as **i.**

Line of Nodes: the line joining the ascending and descending nodes through the centre of Earth.

Prograde Orbit: an orbit in which satellite moves in the same direction as the Earth‟s rotation. Its inclination is always between 00 to 900. Many satellites follow this path as Earth‟s velocity makes it easier to lunch these satellites.

Retrograde Orbit: an orbit in which satellite moves in the same direction counter to the Earth‟s rotation.

Argument of Perigee: An angle from the point of perigee measure in the orbital plane at the Earth‟s centre, in the direction of the satellite motion.

Right ascension of ascending node: The definition of an orbit in space, the position of ascending node is specified. But as the Earth spins, the longitude of ascending node changes and cannot be used for reference. Thus for practical determination of an orbit, the longitude and time of crossing the ascending node is used.For absolute measurement, a fixed reference point in space is required.

It could also be defined as "*right ascension of the ascending node; right ascension is the angular position measured eastward along the celestial equator from the vernal equinox vector to the hour circle of the object*".

Mean anamoly: It gives the average value to the angular position of the satellite with reference to the perigee.

True anamoly: It is the angle from point of perigee to the satellite‟s position, measure at the Earth‟s centre.

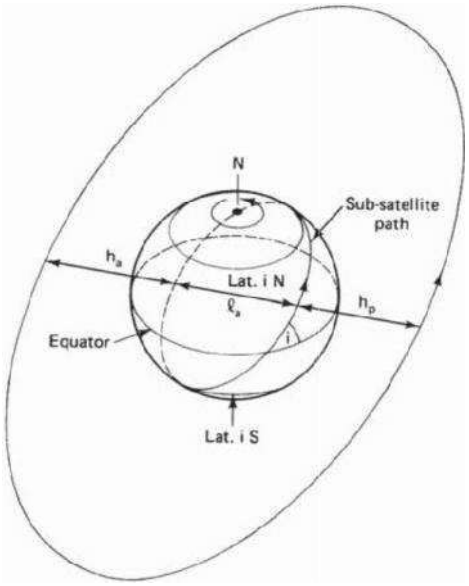


Figure1.2 Apogee height h_a, perigee height h_p, and inclination i. $L_a$is the line of apsides.

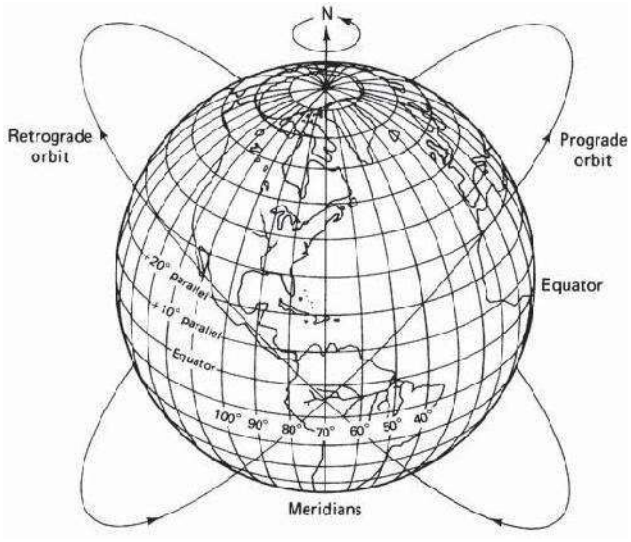


Figure 1.3(a) Prograde and retrograde orbits.

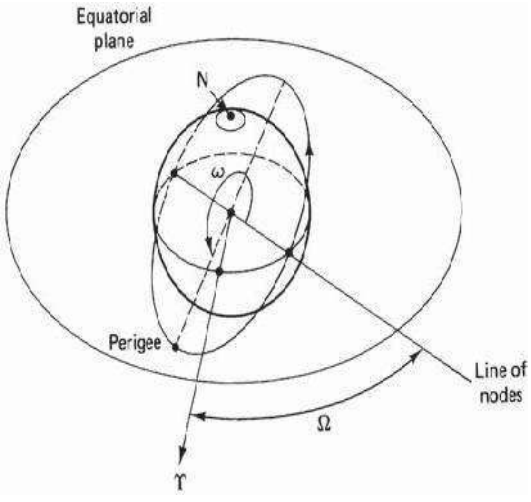


Figure.1.4 The argument of perigee w & right ascension of the ascending node Ω.

1.5. Orbital Perturbations

Theoretically, an orbit described by Kepler is ideal as Earth is considered to be a perfect sphere and the force acting around the Earth is the centrifugal force. This force is supposed to balance the gravitational pull of the earth.

In reality, other forces also play an important role and affect the motion of the satellite. These forces are the gravitational forces of Sun and Moon along with the atmospheric drag.

Effect of Sun and Moon is more pronounced on geostationary earth satellites where as the atmospheric drag effect is more pronounced for low earth orbit satellites.

1.5.1 Effects of non-Spherical Earth

As the shape of Earth is not a perfect sphere, it causes some variations in the path followed by the satellites around the primary. As the Earth is bulging from the equatorial belt, and keeping in mind that an orbit is not a physical entity, and it is the forces resulting from an oblate Earth which act on the satellite produce a change in the orbital parameters.

This causes the satellite to drift as a result of regression of the nodes and the latitude of the point of perigee (point closest to the Earth). This leads to rotation of the line of apsides. As the orbit itself is moving with respect to the Earth, the resultant changes are seen in the values of argument of perigee and right ascension of ascending node.

Due to the non-spherical shape of Earth, one more effect called as the "Satellite Graveyard" is seen. The non-spherical shape leads to the small value of eccentricity (10-5) at the equatorial plane. This causes a gravity gradient on GEO satellite and makes them drift to one of the two stable points which coincide with minor axis of the equatorial ellipse.

1.5.2 Atmospheric Drag

For Low Earth orbiting satellites, the effect of atmospheric drag is more pronounces. The impact of this drag is maximum at the point of perigee. Drag (pull towards the Earth) has an effect on velocity of Satellite (velocity reduces).

This causes the satellite to not reach the apogee height successive revolutions. This leads to a change in value of semi-major axis and eccentricity. Satellites in service are maneuvered by the earth station back to their original orbital position.

1.6 Station Keeping

In addition to having its attitude controlled, it is important that a geo-stationary satellite be kept in its correct orbital slot. The equatorial ellipticity of the earth causes geostationary satel- lites to drift slowly along the orbit, to one of two stable points, at 75°E and 105°W.

To counter this drift, an oppositely directed velocity com-ponent is imparted to the satellite by means of jets, which are pulsed once every 2 or 3 weeks.

These maneuvers are termed *east-west station-keeping maneuvers*. Satellites in the 6/4-GHz band must be kept within 0.1° of the desig- nated longitude, and in the 14/12-GHz band, within 0.05°.

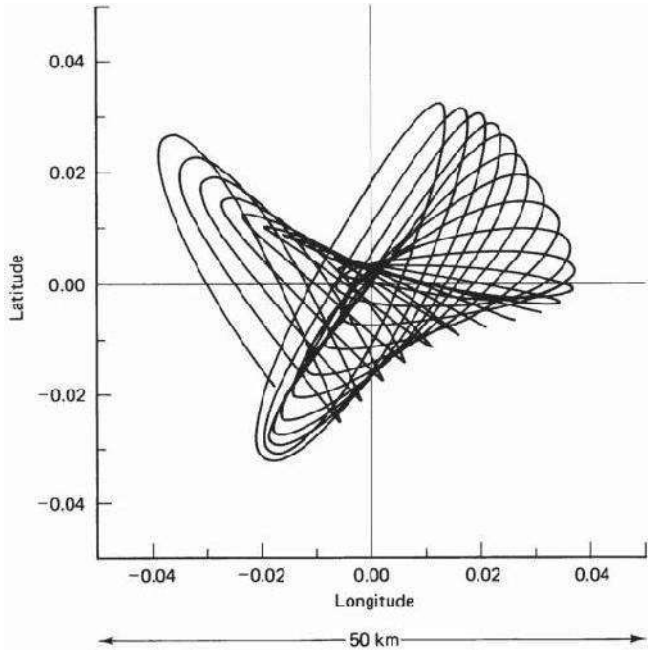


Figure 1.5 Typical satellite motion.(*CourtesyofTelesat,Canada,1983.*)

1.7. Geo stationary and Non Geo-stationary orbits

1.7.1 Geo stationary

A **geostationary** orbit is one in which a satellite orbits the earth at exactly the same speed as the earth turns and at the same latitude, specifically zero, the latitude of the equator. A satellite orbiting in a geostationary orbit appears to be hovering in the same spot in the sky, and is directly over the same patch of ground at all times.

A **geosynchronous** orbit is one in which the satellite is synchronized with the earth's rotation, but the orbit is tilted with respect to the plane of the equator. A satellite in a geosynchronous orbit will wander up and down in latitude, although it will stay over the same line of longitude. Although the terms 'geostationary' and 'geosynchronous' are sometimes used interchangeably, they are not the same technically; geostationary orbit is a subset of all possible geosynchronous orbits.

The person most widely credited with developing the concept of geostationary orbits is noted science fiction author Arthur C. Clarke (Islands in the Sky, Childhood's End, Rendezvous with Rama, and the movie 2001: a Space Odyssey). Others had earlier pointed out that bodies traveling a certain distance above the earth on the equatorial plane would remain motionless with respect to the earth's surface. But Clarke published an article in 1945's Wireless World that made the leap from the Germans' rocket research to suggest permanent manmade satellites that could serve as communication relays.

Geostationary objects in orbit must be at a certain distance above the earth; any closer and the orbit would decay, and farther out they would escape the earth's gravity altogether. This distance is 35,786 kilometers (22,236 miles) from the surface.

The first geosynchrous satellite was orbited in 1963, and the first geostationary one the following year. Since the only geostationary orbit is in a plane with the equator at 35,786 kilometers, there is only one circle around the world where these conditions obtain.

This means that geostationary 'real estate' is finite. While satellites are in no danger of bumping in to one another yet, they must be spaced around the circle so that their frequencies do not interfere with the functioning of their nearest neighbors.

Geostationary Satellites

There are 2 kinds of manmade satellites in the heavens above: One kind of satellite ORBITS the earth once or twice a day, and the other kind is called a communications satellite and it is PARKED in a STATIONARY position 22,300 miles (35,900 km) above the equator of the STATIONARY earth.

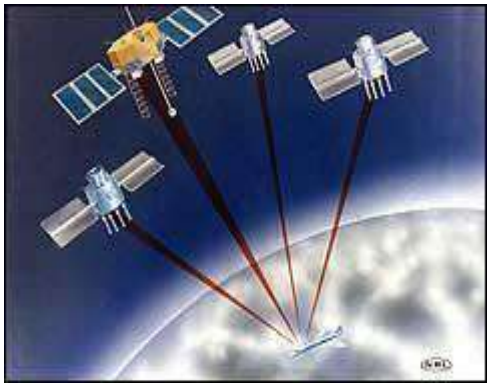
A type of the orbiting satellite includes the space shuttle and the international space station which keep a low earth orbit (LEO) to avoid the deadly Van Allen radiation belts.

The most prominent satellites in medium earth orbit (MEO) are the satellites which comprise the GLOBAL POSITIONING SYSTEM or GPS as it is called.

The Global Positioning System

The global positioning system was developed by the U.S. military and then opened to civilian use. It is used today to track planes, ships, trains, cars or literally anything that moves. Anyone can buy a receiver and track their exact location by using a GPS receiver.

GPS satellites orbit at a height of about 12,000 miles (19,300 km) and orbit the earth once every 12 hours.

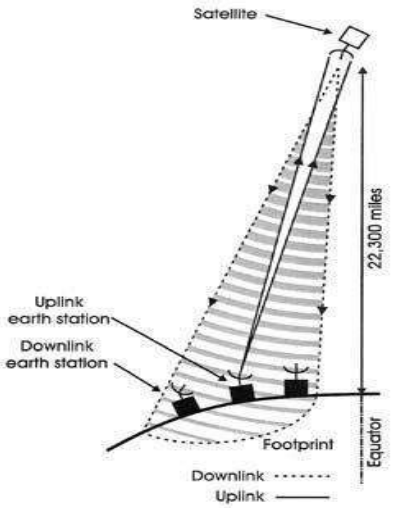


About 24 GPS satellites orbit the earth every 12 hours.

These satellites are traveling around the earth at speeds of about 7,000 mph (11,200 kph). GPS satellites are powered by solar energy. They have backup batteries onboard to keep them running in the event of a solar eclipse, when there's no solar power.

Small rocket boosters on each satellite keep them flying in the correct path. The satellites have a lifetime of about 10 years until all their fuel runs out.

At exactly 22,300 miles above the equator, the force of gravity is cancelled by the centrifugal force of the rotating universe. This is the ideal spot to park a stationary satellite.



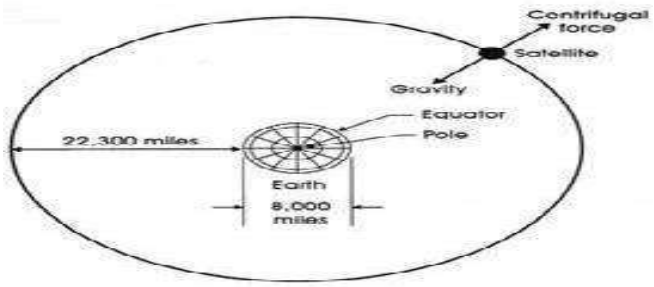


Figure. 1.6 & 1.7 At exactly 22,000 miles (35,900 km) above the equator, the earth's force of gravity is canceled by the centrifugal force of the rotating universe. .

1.7.2 Non Geo-Stationary Orbit

For the geo- stationary case, the most important of these are the gravitational fields of the moon and the sun, and the nonspherical shape of the earth.

Other significant forces are solar radiation pressure and reaction of the satellite itself to motor movement within the satellite. As a result, station-keeping maneuvers must be carried out to maintain the satel- lite within set limits of its nominal geostationary position.

An exact geostationary orbit therefore is not attainable in practice, and the orbital parameters vary with time. The two-line orbital elements are published at regular intervals.

The period for a geostationary satellite is 23 h, 56 min, 4 s, or 86,164 s. The reciprocal of this is 1.00273896 rev/day, which is about the value tabulated for most of the satellites in Fig.

Thus these satellites are *geo- synchronous,* in that they rotate in synchronism with the rotation of the earth. However, they are not geostationary. The term *geosynchronous satellite* is used in many cases instead of *geostationary* to describe these near-geostationary satellites.

It should be noted, however, that in gen- eral a geosynchronous satellite does not have to be near-geostationary, and there are a number of geosynchronous satellites that are in highly elliptical orbits with comparatively large inclinations (e.g., the Tundra satellites).

The small inclination makes it difficult to locate the position of the ascending node, and the small eccentricity makes it difficult to locate the position of the perigee.

However, because of the small inclination, the angles w and Ω can be assumed to be in the same plane.The longitude of the subsatellite point (thesatellitelongitude) is the east early rotation from the Greenwich meridian.

$$\phi_{SS} = \omega + \Omega + v - \text{GST}$$

The *Greenwich sidereal time* (GST) gives the eastward position of the Greenwich meridian relative to the line of Aries, and hence the subsatellite point is at longitudeand the mean longitude of the satellite is given by

$$\phi_{SSmean} = \omega + \Omega + M - \text{GST}$$

Equation(2.31)can be used to calculate the trueanomaly, and because of the small eccentricity, this can be approximated as v= M + 2$e$sin$M$.

1.8 Look Angle Determination

The look angles for the ground station antenna are Azimuth and Elevation angles. They are required at the antenna so that it points directly at the satellite. Look angles are calculated by considering the elliptical orbit. These angles change in order to track the satellite.

For geostationary orbit, these angels values does not change as the satellites are stationary with respect to earth. Thus large earth stations are used for commercial communications.

For home antennas, antenna beamwidth is quite broad and hence no tracking is essential. This leads to a fixed position for these antennas.

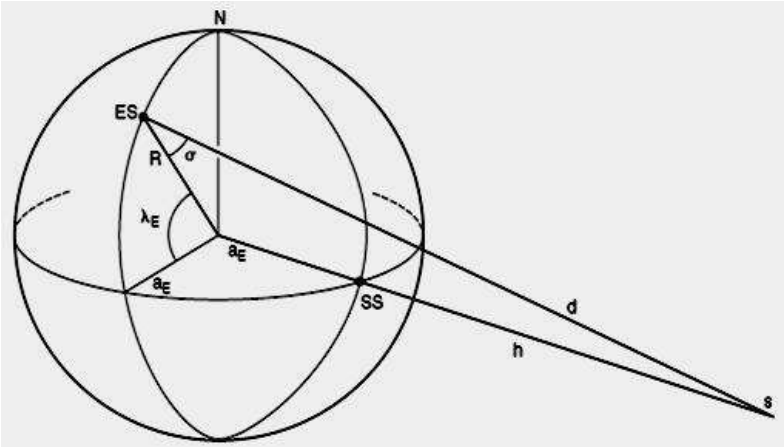


Figure 1.8: The geometry used in determining the look angles for Geostationary Satellites.

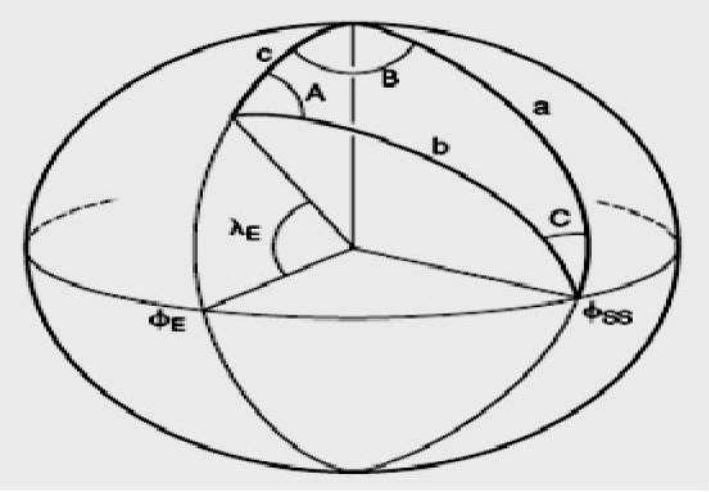


Figure 1.9: The spherical geometry related to figure 1.8

With respect to the figure 1.8 and 1.9, the following information is needed to determine the look angles of geostationary orbit.

1. Earth Station Latitude: λE
2. Earth Station Longitude: ΦE

3. Sub-Satellite Point‟s Longitude: ΦSS
4. ES: Position of Earth Station
5. SS: Sub-Satellite Point
6. S: Satellite
7. d: Range from ES to S
8. ζ: angle to be determined

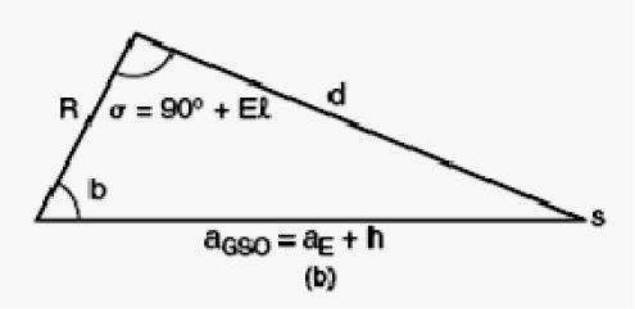


Figure 1.10: A plane triangle obtained from figure 1.8

Considering figure 3.3, it‟s a spherical triangle. All sides are the arcs of a great circle. Three sides of this triangle are defined by the angles subtended by the centre of the earth.

o Side a: angle between North Pole and radius of the sub-satellite point.

o Side b: angle between radius of Earth and radius of the sub-satellite point.

o Side c: angle between radius of Earth and the North Pole.

a =90^0 and such a spherical triangle is called quadrantal triangle. c = 90^0 – λ

Angle B is the angle between the plane containing c and the plane containing a.

Thus, B = ΦE-ΦSS

Angle A is the angle between the plane containing b and the plane containing c.

Angle C is the angle between the plane containing a and the plane containing b.

Thus, $a = 90^0$

$c = 90^0 - \lambda E$

$B = \Phi E - \Phi SS$

Thus, $b = \arccos(\cos B \cos \lambda E)$

And $A = \arcsin(\sin |B| / \sin b)$

Applying the cosine rule for plane triangle to the triangle of figure

$$d = \sqrt{R^2 + a_{GSO}{}^2 - 2Ra_{GSO} \cos b}$$

Applying the sine rule for plane triangles to the triangle of figure 3.3 allows the angle of elevation to be found:

$$El = \arccos\left(\frac{a_{GSO}}{d} \sin b\right)$$

1.9. Limits of visibility

The east and west limits of geostationary are visible from any given Earth station. These limits are set by the geographic coordinates of the Earth station and antenna elevation.

The lowest elevation is zero (in theory) but in practice, to avoid reception of excess noise from Earth. Some finite minimum value of elevation is issued. The earth station can see a satellite over a geostationary arc bounded by **+- (81.30)** about the earth station‟s longitude.

1.10. Eclipse

It occurs when Earth‟s equatorial plane coincides with the plane f he Earth‟s orbit around the sun.

Near the time of spring and autumnal equinoxes, when the sun is crossing the equator, the satellite passes into sun‟s shadow. This happens for some duration of time every day.

These eclipses begin 23 days before the equinox and end 23 days after the equinox. They last for almost 10 minutes at the beginning and end of equinox and increase for a maximum period of 72 minutes at a full eclipse.

The solar cells of the satellite become non-functional during the eclipse period and the satellite is made to operate with the help of power supplied from the batteries.

A satellite will have the eclipse duration symmetric around the time t=Satellite Longitude/15 + 12 hours. A satellite at Greenwich longitude 0 will have the eclipse duration symmetric around 0/15

UTC +12hours = 00:00 UTC.

The eclipse will happen at night but for satellites in the east it will happen late evening local time.

For satellites in the west eclipse will happen in the early morning hour's local time.

An earth caused eclipse will normally not happen during peak viewing hours if the satellite is located near the longitude of the coverage area. Modern satellites are well equipped with batteries for operation during eclipse.

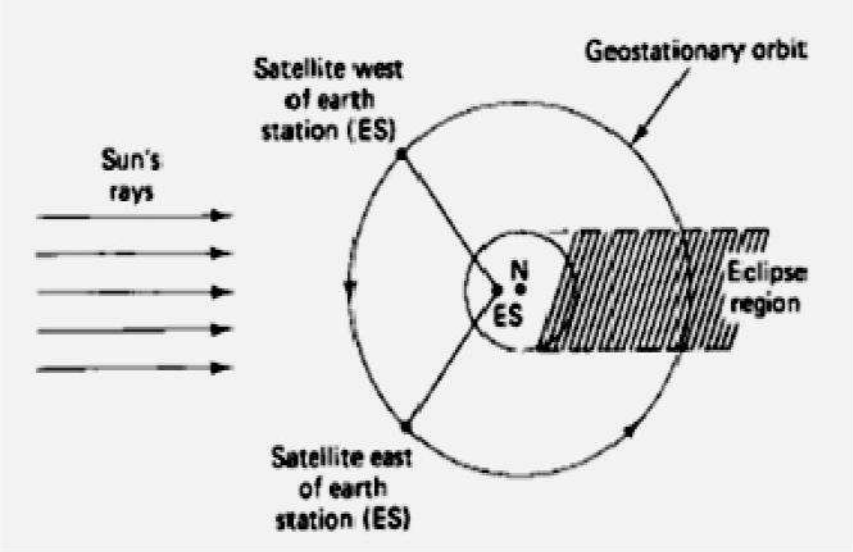


Figure 1.11(i): A satellite east of the earth station enters eclipse during daylight busy) hours at the earth station. A Satellite west of earth station enters eclipse during night and early morning hours (non busy time).

1.11. Sub satellite Point

Point at which a line between the satellite and the center of the Earth intersects the Earth's surface

Location of the point expressed in terms of latitude and longitude

If one is in the US it is common to use

- Latitude – degrees north from equator
- Longitude – degrees west of the Greenwich meridian

Location of the sub satellite point may be calculated from coordinates of the rotating system as:

$$L_s = \cos^{-1}\left(\frac{z_r}{\sqrt{x_r^2 + y_r^2 + z_r^2}}\right)$$

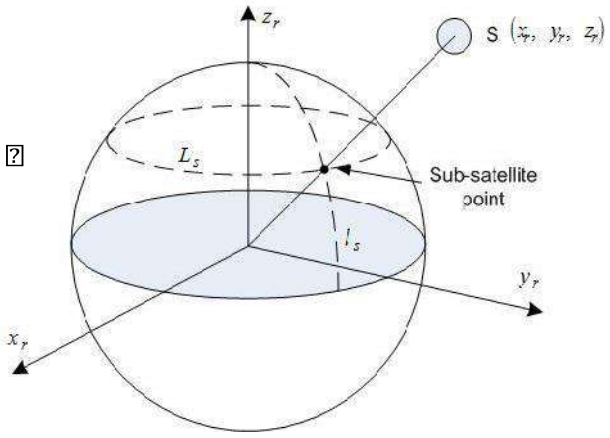


Figure 1.11(ii) **Sub satellite Point**

1.12. Sun Transit Outage

Sun transit outage is an interruption in or distortion of geostationary satellite signals caused by interference from solar radiation.

Sun appears to be an extremely noisy source which completely blanks out the signal from satellite. This effect lasts for 6 days around the equinoxes. They occur for a maximum period of 10 minutes.

Generally, sun outages occur in February, March, September and October, that is, around the time of the equinoxes.

At these times, the apparent path of the sun across the sky takes it directly behind the line of sight between an earth station and a satellite.

As the sun radiates strongly at the microwave frequencies used to communicate with satellites (C-band, Ka band and Ku band) the sun swamps the signal from the satellite.

The effects of a sun outage can include partial degradation, that is, an increase in the error rate, or total destruction of the signal.

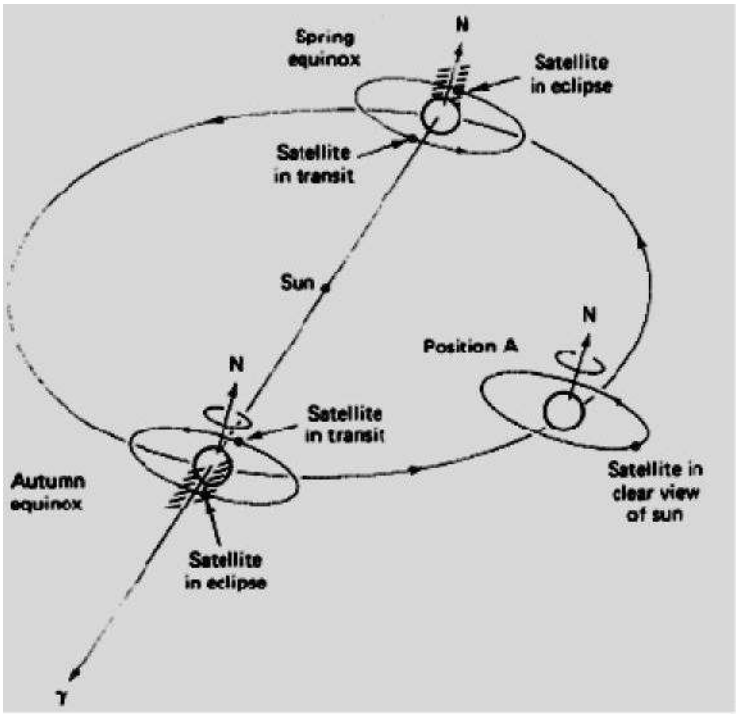


Figure 1.12 : Earth Eclipse of a Satellite and Sun transit Outage

1.13. Launching Procedures

1.13.1 Intoduction

Low Earth Orbiting satellites are directly injected into their orbits. This cannot be done incase of GEOs as they have to be positioned 36,000kms above the Earth‟s surface.

Launch vehicles are hence used to set these satellites in their orbits. These vehicles are reusable. They are also known as „Space Transportation System‟(STS).

When the orbital altitude is greater than 1,200 km it becomes expensive to directly inject the satellite in its orbit.

For this purpose, a satellite must be placed in to a transfer orbit between the initial lower orbit and destination orbit. The transfer orbit is commonly known as *Hohmann-Transfer Orbit.

1.13.2 Orbit Transfer

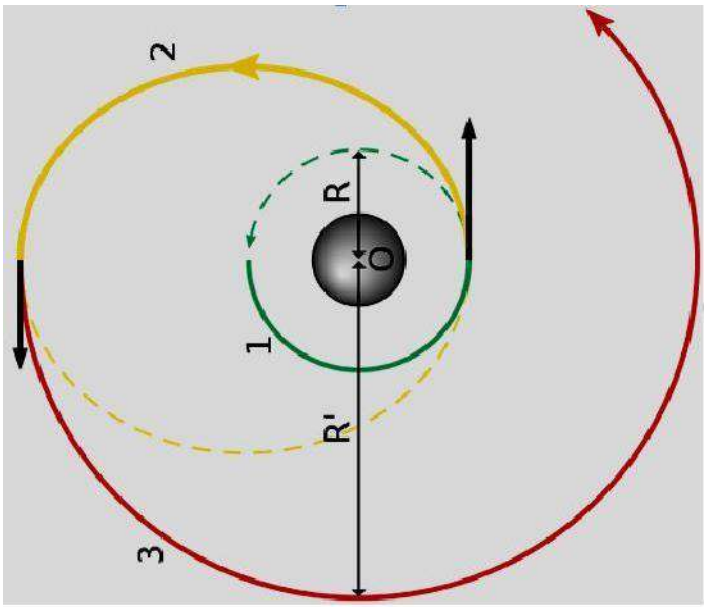


***Figure 1.13**: Orbit Transfer positions*

(*About Hohmann Transfer Orbit: This manoeuvre is named for the German civil engineer who first proposed it, Walter Hohmann, who was born in 1880. He didn't work in rocketry professionally (and wasn't associated with military rocketry), but was a key member of Germany's pioneering Society for Space

Travel that included people such as Willy Ley, Hermann, and Werner von Braun. He published his concept of how to transfer between orbits in his 1925 book, The Attainability of Celestial Bodies.)

The transfer orbit is selected to minimize the energy required for the transfer. This orbit forms a tangent to the low attitude orbit at the point of its perigee and tangent to high altitude orbit at the point of its apogee.

1.14 Launch vehicles and propulsion

The rocket injects the satellite with the required thrust** into the transfer orbit. With the STS, the satellite carries a perigee kick motor*** which imparts the required thrust to inject the satellite in its transfer orbit. Similarly, an apogee kick motor (AKM) is used to inject the satellite in its destination orbit.

Generally it takes 1-2 months for the satellite to become fully functional. The Earth Station performs the Telemetry Tracking and Command**** function to control the satellite transits and functionalities.

(**Thrust: It is a reaction force described quantitatively by Newton's second and third laws. When a system expels or accelerates mass in one direction the accelerated mass will cause a force of equal magnitude but opposite direction on that system.)

Kick Motor refers to a rocket motor that is regularly employed on artificial satellites destined for a geostationary orbit. As the vast majority of geostationary satellite launches are carried out from spaceports at a significant distance away from Earth's equator.

The carrier rocket would only be able to launch the satellite into an elliptical orbit of maximum apogee 35,784-kilometres and with a non-zero inclination approximately equal to the latitude of the launch site.

TT&C: it‟s a sub-system where the functions performed by the satellite control network to maintain health and status, measure specific mission parameters and processing over time a sequence of these measurement to refine parameter knowledge, and transmit mission commands to the satellite. Detailed study of TT&C in the upcoming units.

1.14.1 Transfer Orbit

It is better to launch rockets closer to the equator because the Earth rotates at a greater speed here than that at either pole. This extra speed at the equator means a rocket needs less thrust (and therefore less fuel) to launch into orbit.

In addition, launching at the equator provides an additional 1,036 mph (1,667 km/h) of speed once the vehicle reaches orbit. This speed bonus means the vehicle needs less fuel, and that freed space can be used to carry more pay load.

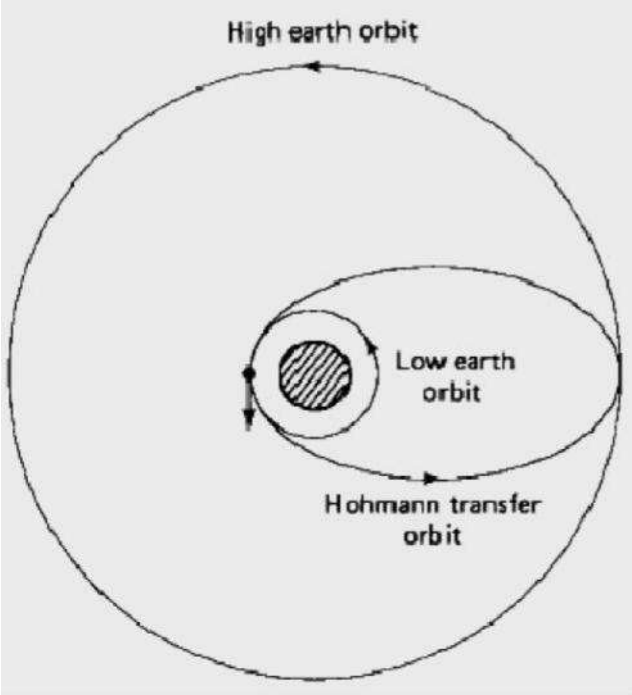


***Figure 1.14**: Hohmann Transfer Orbit*

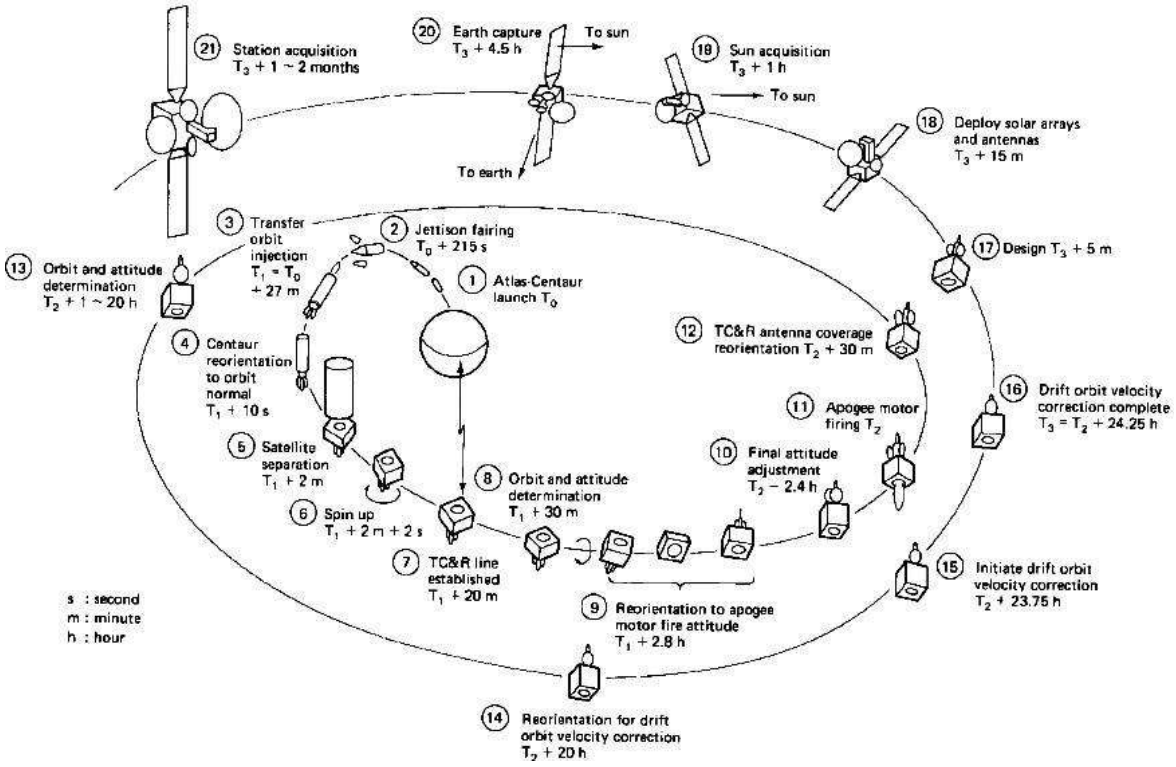


Figure 1.15*: Launching stages of a GEO (example INTELSAT)*

Rocket launch

A **rocket launch** is the takeoff phase of the flight of a rocket. Launches for orbital spaceflights, or launches into interplanetary space, are usually from a fixed location on the ground, but may also be from a floating platform (such as the Sea Launch vessel) or, potentially, from a super heavy An-225-class airplane

Launches of suborbital flights (including missile launches), can also be from:

- a missile silo
- a mobile launcher vehicle
- a submarine
- air launch:
  - from a plane (e.g. Scaled Composites Space Ship One, Pegasus Rocket, X-15)
  - from a balloon (Rockoon, da Vinci Project (under development))
  - a surface ship (Aegis Ballistic Missile Defense System)
  - an inclined rail (e.g. rocket sled launch)

"Rocket launch technologies" generally refers to the entire set of systems needed to successfully launch a vehicle, not just the vehicle itself, but also the firing control systems, ground control station, launch pad, and tracking stations needed for a successful launch and/or recovery.

Orbital launch vehicles commonly take off vertically, and then begin to progressively lean over, usually following a gravity turn trajectory.

Once above the majority of the atmosphere, the vehicle then angles the rocket jet, pointing it largely horizontally but somewhat downwards, which permits the vehicle to gain and then maintain altitude while increasing horizontal speed. As the speed grows, the vehicle will become more and more horizontal until at orbital speed, the engine will cut off.

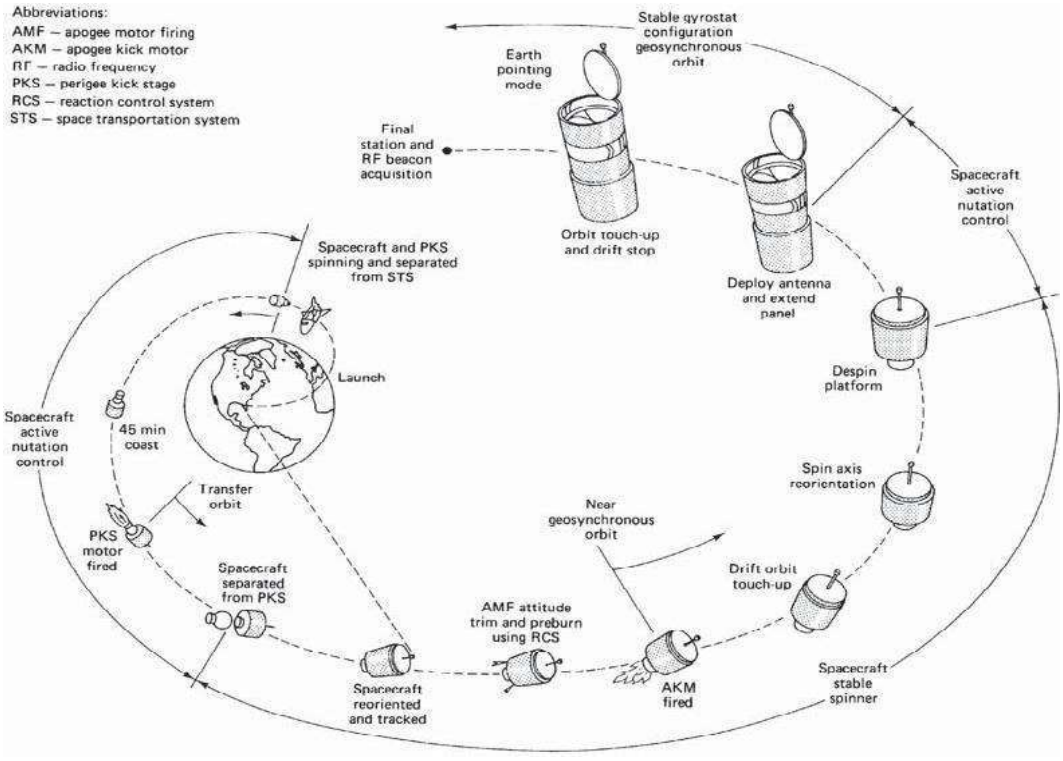


Figure 1.16 STS-7/Anik C2 mission scenario. (*From Anik C2 Launch Handbook; courtesy of Telesat, Canada.*)

APPLICATIONS

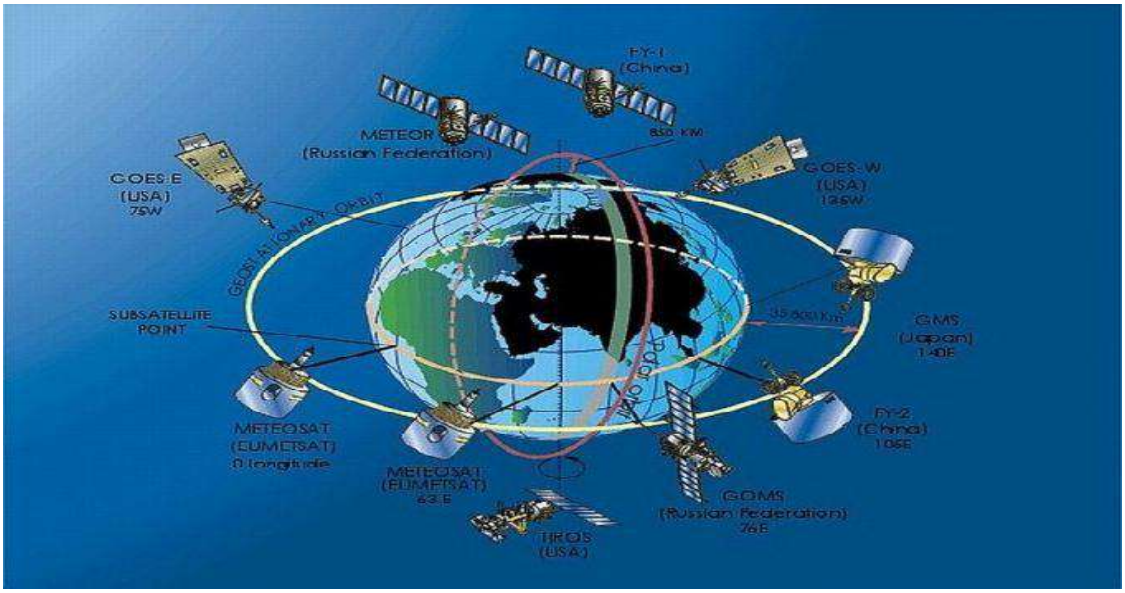


Figure example of geostationary satellites

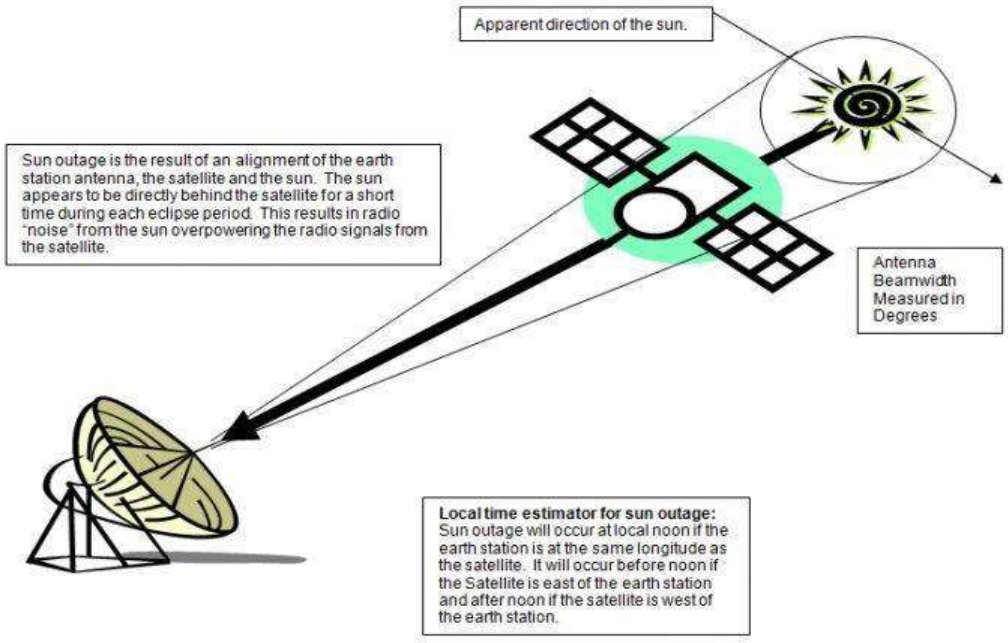


Figure example of sun transit outage

SATELLITE COMMUNICATION

UNIT II SPACE SEGMENT AND SATELLITE LINK DESIGN

2.1 Spacecraft Technology- Structure

A satellite communications system can be broadly divided into two segments—a ground segment and a space segment.

The space segment will obviously include the satellites, but it also includes the ground facilities needed to keep the satellites operational, these being referred to as the *tracking, telemetry, and command* (TT&C) facilities. In many networks it is common practice to employ a ground station solely for the purpose of TT&C.

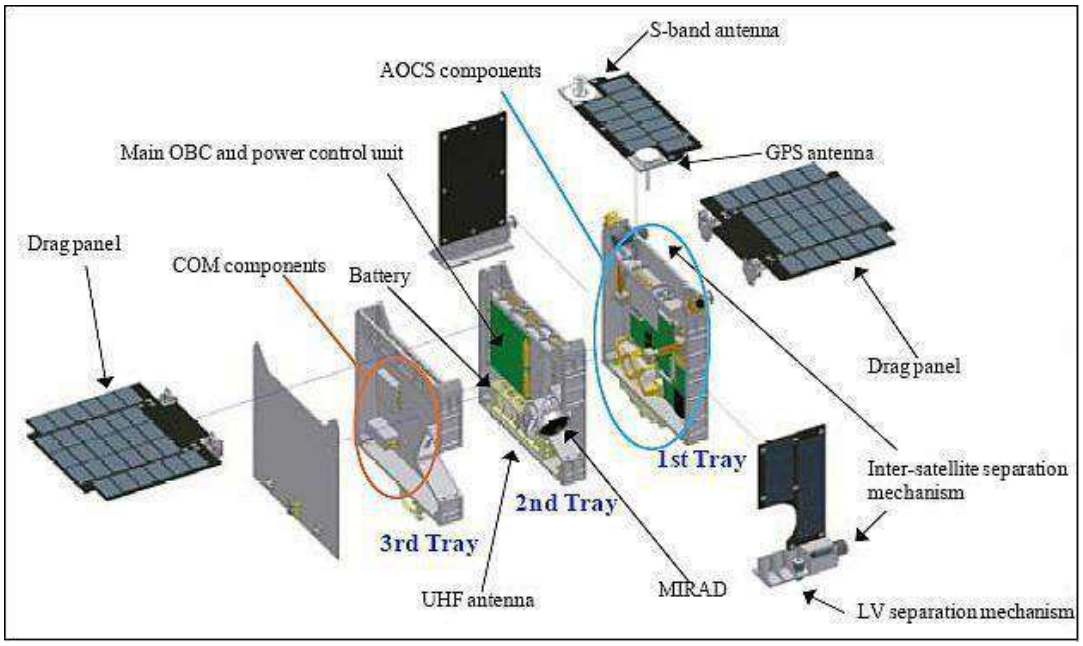


Figure 2.1 (a) Satellite Structure

The equipment carried aboard the satellite also can be classified according to function. The *payload* refers to the equipment used to pro- vide the service for which the satellite has been launched.

In a communications satellite, the equipment which provides the con- necting link between the satellite's transmit and receive antennas is referred to as the *transponder*. The transponder forms one of the main sections of the payload, the other being the antenna subsystems.In this chapter the main characteristics of certain bus systems and payloads are described.

2.2 The Power Supply

The primary electrical power for operating the electronic equipment is obtained from solar cells. Individual cells can generate only small amounts of power, and therefore, arrays of cells in series-parallel connection are required.

Figure shows the solar cell panels for the HS 376 satellite manufactured by Hughes Space and Communications Company.

In geostationary orbit the telescoped panel is fully extended so that both are exposed to sun- light. At the beginning of life, the panels produce 940 W dc power, which may drop to 760 W at the end of 10 years.

During eclipse, power is provided by two nickel-cadmium (Ni-Cd) long-life batteries, which will deliver 830 W. At the end of life, battery recharge time is less than 16 h.

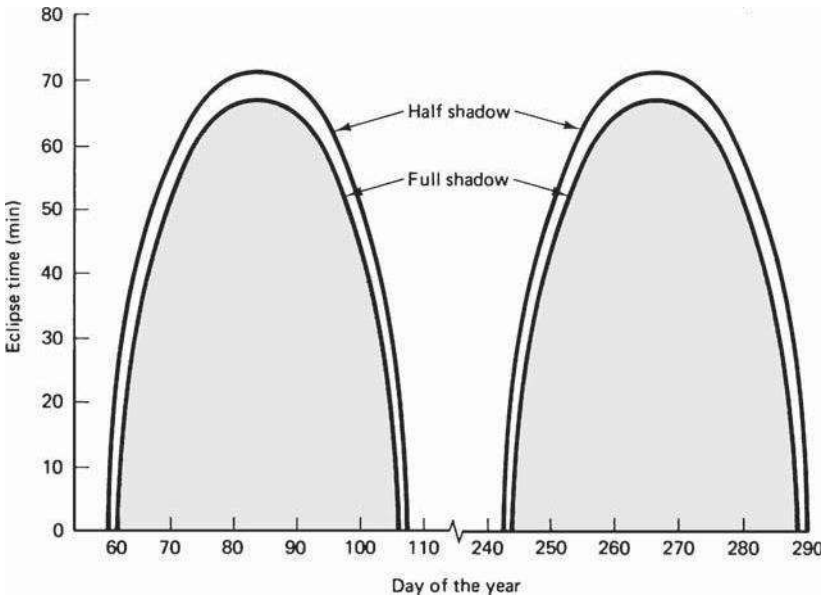


Figure 2.1.(b) Satellite eclipse time as a function of the current day of the year. (*Courtesy of*

Spilker, 1977. Reprinted by permission of Prentice-Hall, Englewood Cliffs, NJ.)

capacity of cylindrical and solar-sail satellites, the cross-over point is esti- mated to be about 2 kW, where the solar-sail type is more economical than the cylindrical type (Hyndman, 1991).

2.3 Attitude Control & Orbit Control

The *attitude* of a satellite refers to its orientation in space. Much of the equipment carried aboard a satellite is there for the purpose of control- ling its attitude. Attitude control is necessary, for example, to ensure that directional antennas point in the proper directions.

In the case of earth environmental satellites, the earth-sensing instruments must cover the required regions of the earth, which also requires attitude control. A number of forces, referred to as *disturbance torques*, can alter the attitude, some examples being the gravitational fields of the earth and the moon, solar radiation, and meteorite impacts.

Attitude control must not be con- fused with station keeping, which is the term used for maintaining a satellite in its correct orbital position, although the two are closely related.

To exercise attitude control, there must be available some measure of a satellite's orientation in space and of any tendency for this to shift. In one method, infrared sensors, referred to as *horizon detectors*, are used to detect the rim of the earth against the background of space.

With the use of four such sensors, one for each quadrant, the center of the earth can be readily established as a reference point.

Usually, the attitude-control process takes place aboard the satellite, but it is also possible for control signals to be transmitted from earth, based on attitude data obtained from the satellite.

Also, where a shift in attitude is desired, an *attitude maneuver* is executed. The control signals needed to achieve this maneuver may be transmitted from an earth station.

Controlling torques may be generated in a number of ways. *Passive attitude control* refers to the use of mechanisms which stabilize the satellite without putting a drain on the satellite's energy supplies; at most, infrequent use is made of these supplies, for example, when thruster jets are impulsed to provide corrective torque. Examples of passive attitude control are *spin stabilization* and *gravity gradient sta- bilization*.

The other form of attitude control is *active control.* With active attitude control, there is no overall stabilizing torque present to resist the disturbance torques. Instead, corrective torques are applied as required in response to disturbance torques. Methods used to generate active control torques include momentum wheels, electromagnetic coils, and mass expulsion devices, such as gas jets and ion thrusters.

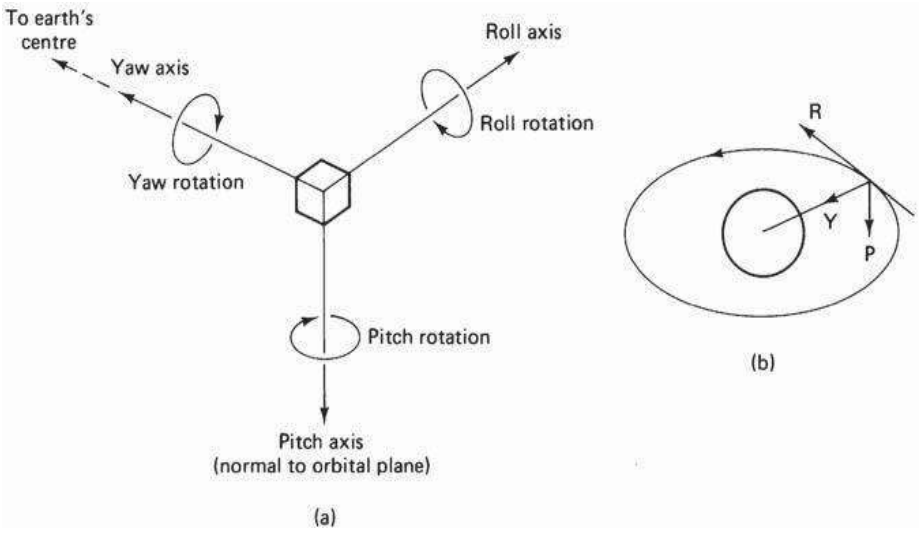


Figure 2.2 (*a*) Roll, pitch, and yaw axes. The yaw axis is directed toward the earth's center, the pitch axis is normal to the orbital plane, and the roll axis is perpendicular to the other two. (*b*) RPY axes for the geostationary orbit. Here, the roll axis is tangential to the orbit and lies along the satellite velocity vector.

The three axes which define a satellite's attitude are its *roll, pitch*, and *yaw* (RPY) axes. These are shown relative to the earth in Fig. 7.4. All three axes pass through the center of gravity of the satellite. For an equatorial orbit, movement of the satellite about the roll axis moves the antenna footprint north and south; movement about the pitch axis moves the footprint east and west; and movement about the yaw axis rotates the antenna footprint.

2.3.1 Spinning satellite stabilization

Spin stabilization may be achieved with cylindrical satellites. The satellite is constructed so that it is mechanically balanced about one partic- ular axis and is then set spinning around this axis. For geostationary satellites, the spin axis is adjusted to be parallel to the N-S axis of the earth, as illustrated in Fig. 7.5. Spin rate is typically in the range of 50 to 100 rev/min. Spin is initiated during the launch phase by means of small gas jets.

In the absence of disturbance torques, the spinning satellite would maintain its correct attitude relative to the earth. Disturbance torques are generated in a number of ways, both external and internal to the satellite.

Solar radiation, gravitational gradients, and meteorite impacts are all examples of external forces which can give rise to disturbance torques. Motor-bearing friction and the movement of satellite elements such as the antennas also can give rise to disturbance torques. The

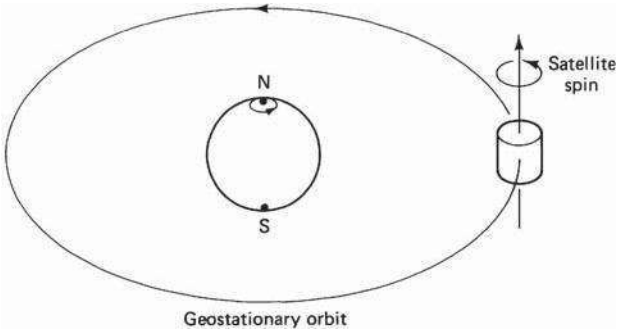


Figure 2.3 Spin stabilization in the geostationary orbit. The spin axis lies along the pitch axis, parallel to the earth's N-S axis.

overall effect is that the spin rate will decrease, and the direction of the angular spin axis will change. Impulse-type thrusters, or jets, can be used to increase the spin rate again and to shift the axis back to its cor- rect N-S orientation.

Nutation, which is a form of wobbling, can occur as a result of the disturbance torques and/or from misalignment or unbalance of the control jets. This nutation must be damped out by means of energy absorbers known as *nutation dampers*.

The antenna feeds can therefore be connected directly to the transponders without the need for radiofrequency (rf) rotary joints, while the complete platform is despun. Of course, control signals and power must be transferred to the despun section, and a mechanical bearing must be provided.

The complete assembly for this is known as the *bearing and power transfer assembly* (BAPTA). Figure 2.4 shows a photograph of the internal structure of the HS 376.

Certain dual-spin spacecraft obtain spin stabilization from a spinning fly-wheel rather than by spinning the satellite itself. These flywheels are termed *momentum wheels*, and their average momentum is referred to as *momentum bias*

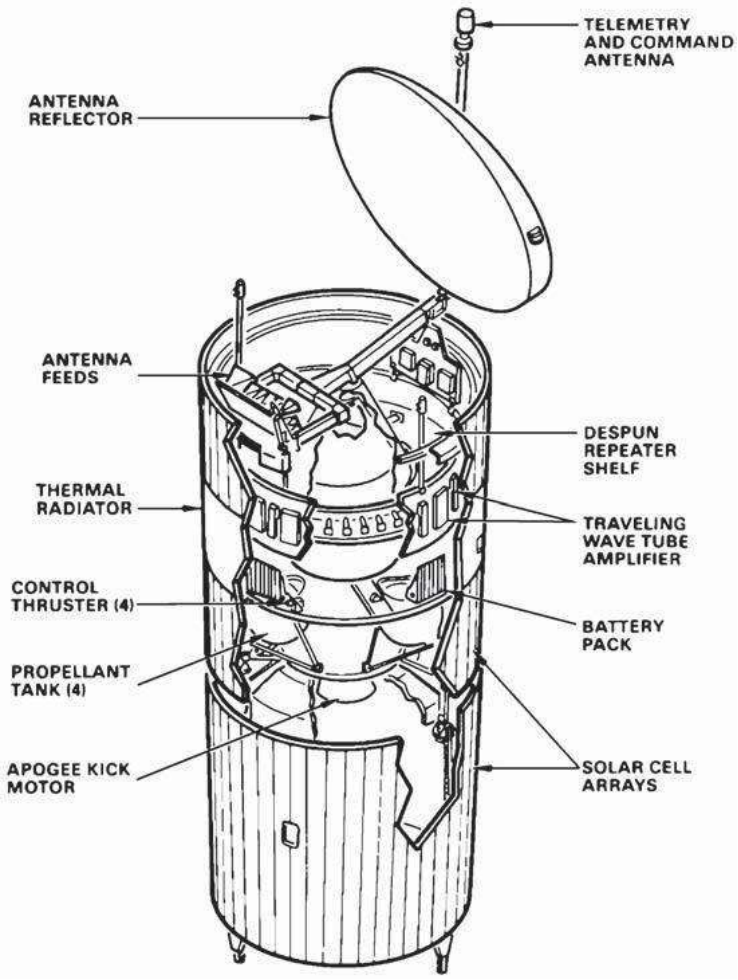


Figure 2.4 HS 376 spacecraft. (*Courtesy of Hughes Aircraft Company Space and Communication Group.*)

2.3.2 Momentum wheel stabilization

In the previous section the gyroscopic effect of a spinning satellite was shown to provide stability for the satellite attitude.

Stability also can be achieved by utilizing the gyroscopic effect of a spinning flywheel, and this approach is used in satellites with cube-like bodies (such as shown in Fig. and the INTELSAT V type satellites shown in Fig. These are known as *body-stabilized* satellites.

The complete unit, termed a momentum wheel, consists of a flywheel, the bearing assembly, the casing, and an electric drive motor with associated electronic con- trol circuitry.

The flywheel is attached to the rotor, which consists of a permanent magnet providing the magnetic field for motor action. The stator of the motor is attached to the body of the satellite.

Thus the motor provides the coupling between the flywheel and the satellite structure. Speed and torque control of the motor is exercised through the currents fed to the stator.

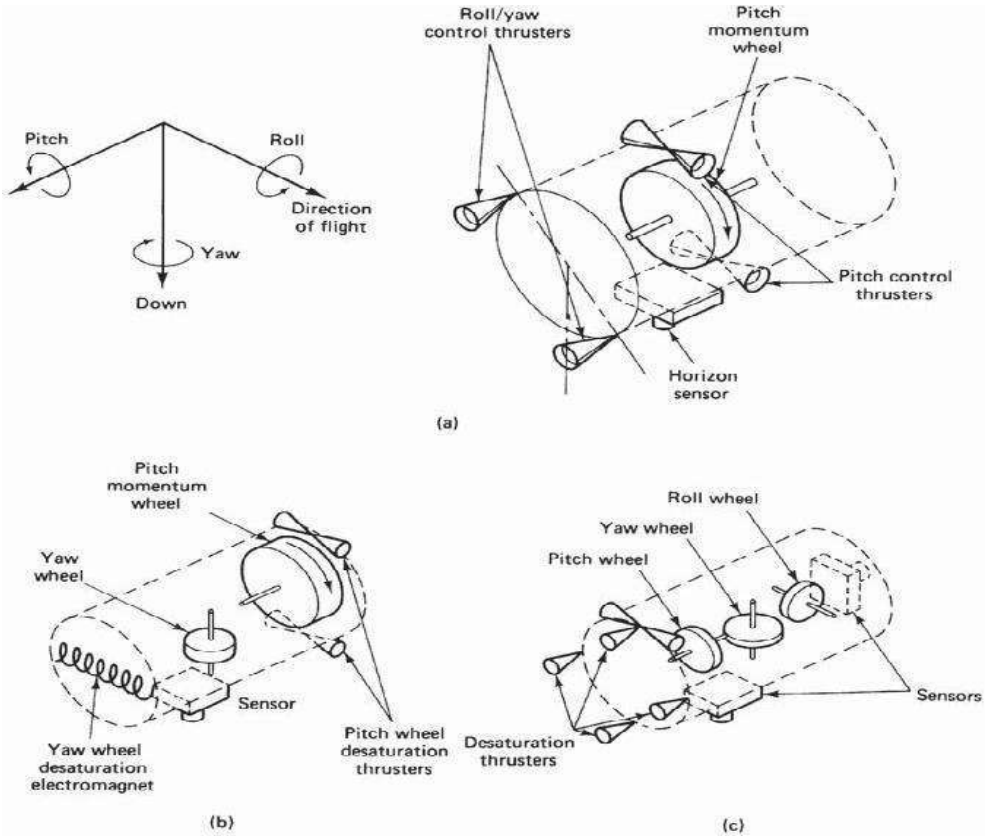


Figure 2.5 Alternative momentum wheel stabilization systems: (*a*) one-wheel, (*b*) two- wheel, (*c*) three-wheel.

When a momentum wheel is operated with zero momentum bias, it is generally referred to as a *reaction wheel*. Reaction wheels are used in three-axis stabilized systems. Here, as the name suggests, each axis is stabilized by a reaction wheel, as shown in Fig. 7.8*c*. Reaction wheels can also be combined with a momentum wheel to provide the control needed (Chetty, 1991).

Random and cyclic disturbance torques tends to produce zero momentum on average. However, there will always be some disturbance torques that causes a cumulative increase in wheel momentum, and eventually at some point the wheel *saturates*.

In effect, it reaches its maximum allowable angular velocity and can no longer take in any more momentum. Mass expulsion devices are then used to unload the wheel, that is, remove momentum from it (in the same way a brake removes energy from a moving vehicle). Of course, operation of the mass expulsion devices consumes part of the satellite's fuel supply.

2.4 Thermal Control and Propulsion

Satellites are subject to large thermal gradients, receiving the sun's radiation on one side while the other side faces into space. In addition, thermal radiation from the earth and the earth's *albedo*, which is the fraction of the radiation falling on earth which is reflected, can be sig- nificant for low-altitude earth-orbiting satellites, although it is negligi- ble for geostationary satellites.

Equipment in the satellite also generates heat which has to be removed. The most important consideration is that the satellite's equipment should operate as nearly as possible in a stable temperature environment. Various steps are taken to achieve this. Thermal blankets and shields may be used to provide insulation. Radiation mirrors are often used to remove heat from the communications payload.

The mirrored thermal radiator for the Hughes HS 376 satellite can be seen in Fig. These mirrored drums surround the communications equipment shelves in each case and pro- vide good radiation paths for the generated heat to escape into the surrounding space.

One advantage of spinning satellites compared with body-stabilized is that the spinning body provides an averaging of the temperature extremes experienced from solar flux and the cold back- ground of deep space.

In order to maintain constant temperature conditions, heaters may be switched on (usually on command from ground) to make up for the heat reduction which occurs when transponders are switched off. The INTELSAT VI satellite used heaters to maintain propulsion thrusters and line temperatures (Pilcher, 1982).

2.5 Communication Payload & Supporting Subsystems

The physical principle of establishing communication connections between remote communication devices dates back to the late 1800s when scientists were beginning to understand electromagnetism and discovered that electromagnetic (EM) radiation (also called EM waves) generated by one device can be detected by another located at some distance away.

By controlling certain aspect s of the radiation (through a process called modulation , explained in Section 4.4), useful information can be embedded in the EM waves and transmitted from one device to another.

The second major module is the communication payload, which is made up of transponders. A transponder is capable of :

- Receiving uplinked radio signals from earth satellite transmission stations (antennas).
- Amplifying received radio signals
- Sorting the input signals and directing the output signals through input/output signal multiplexers to the proper downlink antennas for retransmission to earth satellite receiving stations (antennas).

2.6 TT&C Subsystem

The TT&C subsystem performs several routine functions aboard the spacecraft. The telemetry, or telemetering, function could be interpreted as *measurement at a distance*. Specifically, it refers to the overall oper- ation of generating an electrical signal proportional to the quantity being measured and encoding and transmitting this to a distant station, which for the satellite is one of the earth stations.

Data which are trans- mitted as telemetry signals include attitude information such as that obtained from sun and earth sensors; environmental information such as the magnetic field intensity and direction, the frequency of meteorite impact, and so on; and spacecraft information such as temperatures, power supply voltages, and stored-fuel pressure.

Telemetry and command may be thought of as complementary functions. The telemetry subsystem transmits information about the satellite to the earth station, while the command subsystem receives command signals from the earth station, often in response to telemetered information. The command subsystem

demodulates and, if necessary, decodes the command signals and routes these to the appropriate equipment needed to exe- cute the necessary action.

Thus attitude changes may be made, communication transponders switched in and out of circuits, antennas redirected, and station-keeping maneuvers carried out on command. It is clearly important to prevent unauthorized commands from being received and decoded, and for this reason, the command signals are often encrypted.

Encrypt is derived from a Greek word *kryptein*, meaning *to hide*, and represents the process of concealing the command signals in a secure code. This differs from the normal process of encoding which converts characters in the command signal into a code suitable for transmission.

Tracking of the satellite is accomplished by having the satellite transmit beacon signals which are received at the TT&C earth stations.

Tracking is obviously important during the transfer and drift orbital phases of the satellite launch. Once it is on station, the position of a geostationary satellite will tend to be shifted as a result of the various disturbing forces, as described previously.

Therefore, it is necessary to be able to track the satellite's movement and send correction signals as required.

2.6.1 Transponders

A transponder is the series of interconnected units which forms a single communications channel between the receive and transmit antennas in a communications satellite.

Some of the units utilized by a transponder in a given channel may be common to a number of transponders. Thus, although reference may be made to a specific transponder, this must be thought of as an equipment *channel* rather than a single item of equipment.

Before describing in detail the various units of a transponder, the overall frequency arrangement of a typical C-band communications satellite will be examined briefly. The bandwidth allocated for C-band service is 500 MHz, and this is divided into subbands, one transponder.

A typical transponder bandwidth is 36 MHz, and allowing for a 4-MHz guardband between transponders, 12 such transponders can be accommodated in the 500-MHz bandwidth.

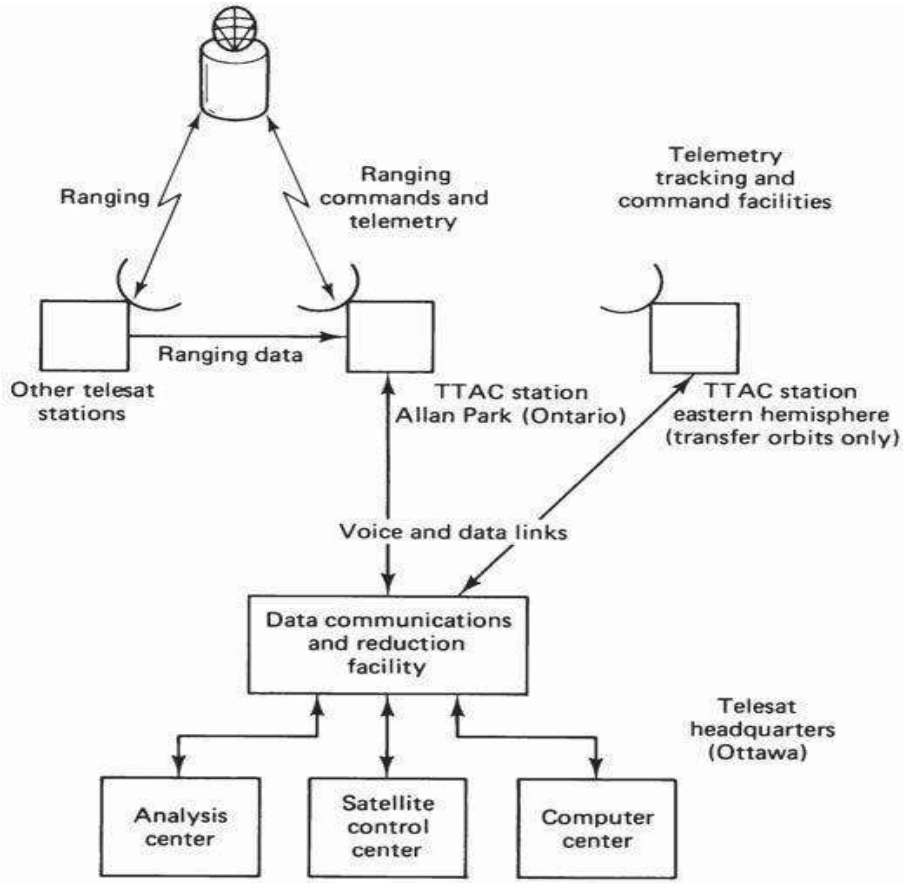
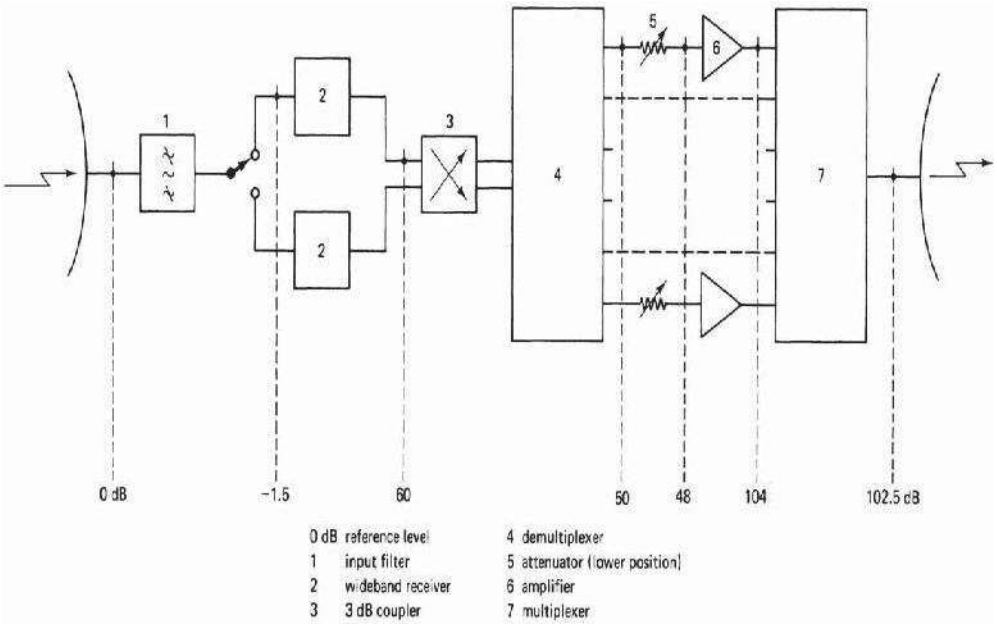


Figure 2.8 Satellite control system. (*Courtesy of Telesat Canada, 1983.*)

By making use of *polar- ization isolation*, this number can be doubled. Polarization isolation refers to the fact that carriers, which may be on the same frequency but with opposite senses of polarization, can be isolated from one another by receiving antennas matched to the incoming polarization.

With linear polarization, vertically and horizontally polarized carriers can be separated in this way, and with circular polarization, left-hand circular and right-hand circular polarizations can be separated.

Because the carriers with opposite senses of polarization may overlap in frequency, this technique is referred to as *frequency reuse*. Figure 2.9 shows part of the frequency and polarization plan for a C-band communications satellite.



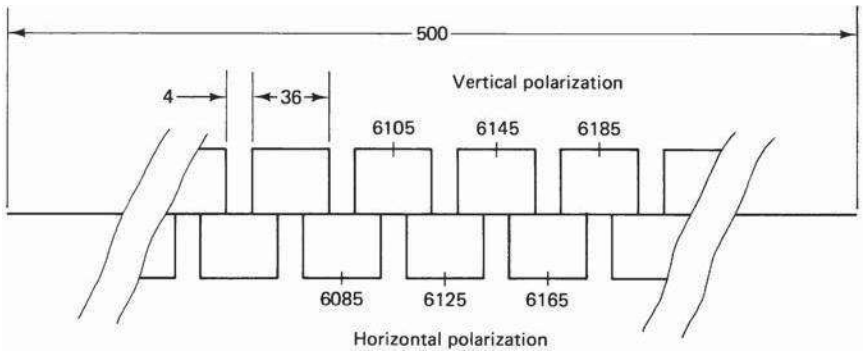


Figure 2.9 Section of an uplink frequency and polarization plan. Numbers refer to frequency in megahertz.

Frequency reuse also may be achieved with spot-beam antennas, and these may be combined with polarization reuse to provide an effective bandwidth of 2000 MHz from the actual bandwidth of 500 MHz.

For one of the polarization groups, Fig. 2.9 shows the channeling scheme for the 12 transponders in more detail. The incoming, or uplink, frequency range is 5.925 to 6.425 GHz.

The frequency conversion shifts the carriers to the downlink frequency band, which is also 500 MHz wide, extending from 3.7 to 4.2 GHz. At this point the signals are channelized into frequency bands which represent the individual transponder bandwidths.

2.6.2 The wideband receiver

The wideband receiver is shown in more detail in Fig. 2.10. A duplicate receiver is provided so that if one fails, the other is automatically switched in. The combination is referred to as a *redundant receiver*, meaning that although two are provided, only one is in use at a given time.

The first stage in the receiver is a *low-noise amplifier* (LNA). This amplifier adds little noise to the carrier being amplified, and at the same time it provides sufficient amplification for the carrier to override the higher noise level present in the following mixer stage.

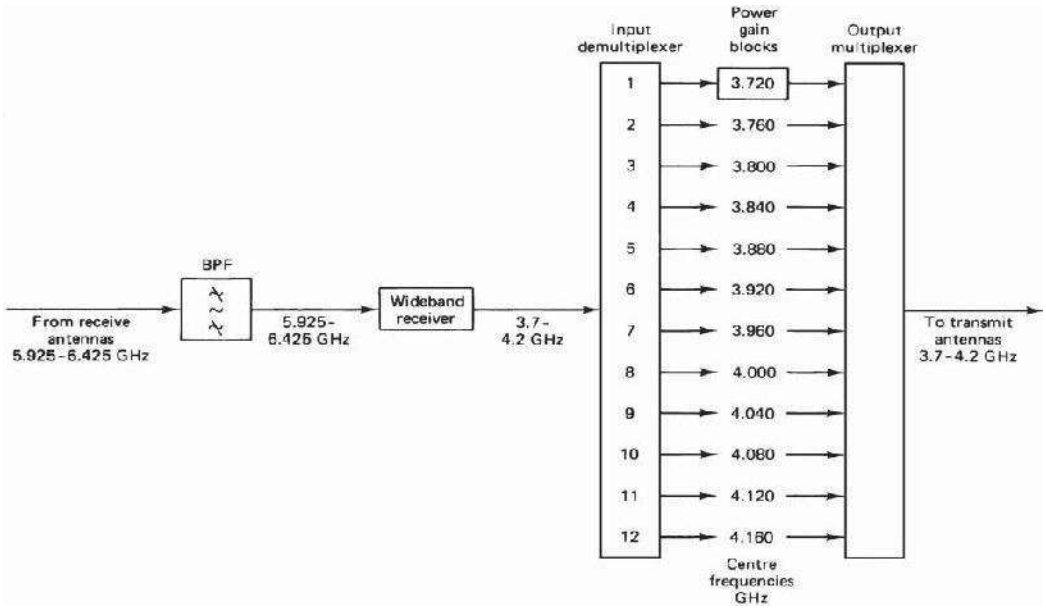


Figure 2.10 Satellite transponder channels

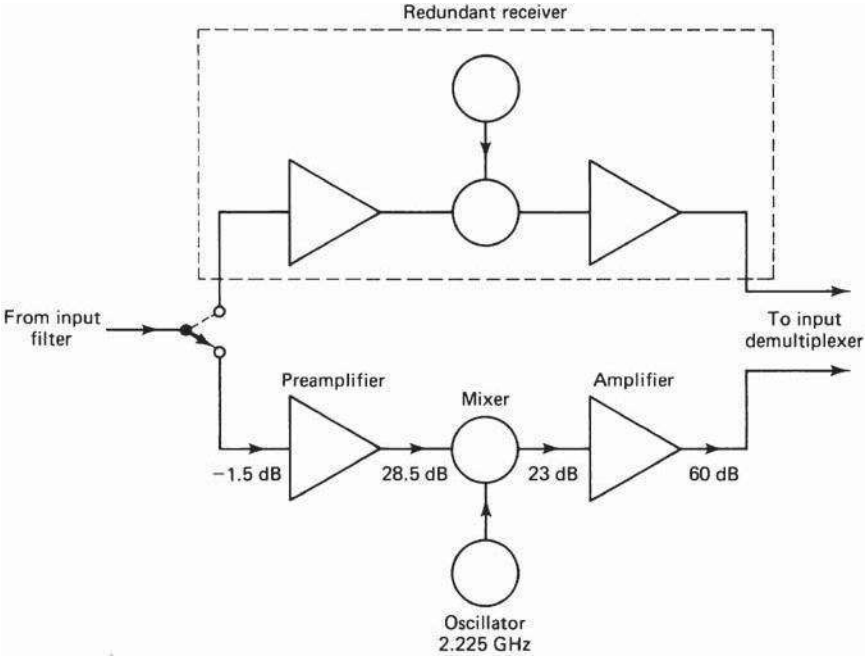


Figure 2.11 Satellite wideband receiver. (*Courtesy of CCIR, CCIR Fixed Satellite Services Handbook, final draft 1984.*)

involving noise, it is usually more convenient to refer all noise levels to the LNA input, where the total receiver noise may be expressed in terms of an equivalent noise temperature.

In a well-designed receiver, the equivalent noise temperature referred to the LNA input is basically that of the LNA alone. The overall noise temperature must take into account the noise added from the antenna, and these calculations are presented in detail in Chap. 12. The equivalent noise temperature of a satellite receiver may be on the order of a few hundred kelvins.

The LNA feeds into a mixer stage, which also requires a *local oscillator* (LO) signal for the frequency-conversion process.

With advances in *field-effect transistor* (FET) technology, FET amplifiers, which offer equal or better performance, are now available for both bands. Diode mixer stages are used.

The amplifier following the mixer may utilize *bipolar junction transistors* (BJTs) at 4 GHz and FETs at 12 GHz, or FETs may in fact be used in both bands.

2.6.3 The input demultiplexer

The input demultiplexer separates the broadband input, covering the frequency range 3.7 to 4.2 GHz, into the transponder frequency channels.

This provides greater frequency separation between adjacent channels in a group, which reduces adjacent channel interference.

The output from the receiver is fed to a power splitter, which in turn feeds the two separate chains of circulators.

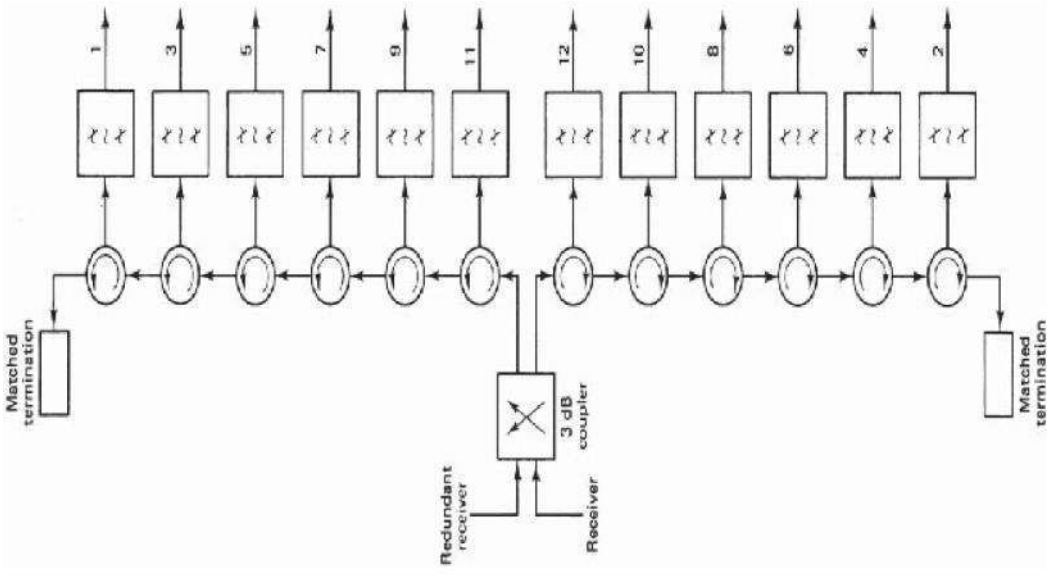


Figure 2.12 Satellite input multiplexer

The full broadband signal is transmitted along each chain, and the channelizing is achieved by means of channel filters con- nected to each circulator,

Each filter has a bandwidth of 36 MHz and is tuned to the appropriate center frequency, as shown in Fig. 2.11.

Although there are considerable losses in the demultiplexer, these are easily made up in the overall gain for the transponder channels.

2.6.4 The power amplifier

The fixed attenuation is needed to balance out variations in the input attenuation so that each transpon- der channel has the same nominal attenuation, the necessary adjust- ments being made during assembly.

The variable attenuation is needed to set the level as required for different types of service (an example being the requirement for input power backoff discussed later). Because this variable attenuator adjustment is an operational requirement, it must be under the control of the ground TT&C station.

Traveling-wave tube amplifiers (TWTAs) are widely used in transpon- ders to provide the final output power required to the transmit antenna. Figure 2.13 shows the schematic of a *traveling wave tube* (TWT) and its power supplies.

In the TWT, an electron-beam gun assembly consisting of a heater, a cathode, and focusing electrodes is used to form an elec- tron beam. A magnetic field is required to confine the beam to travel along the inside of a wire helix.

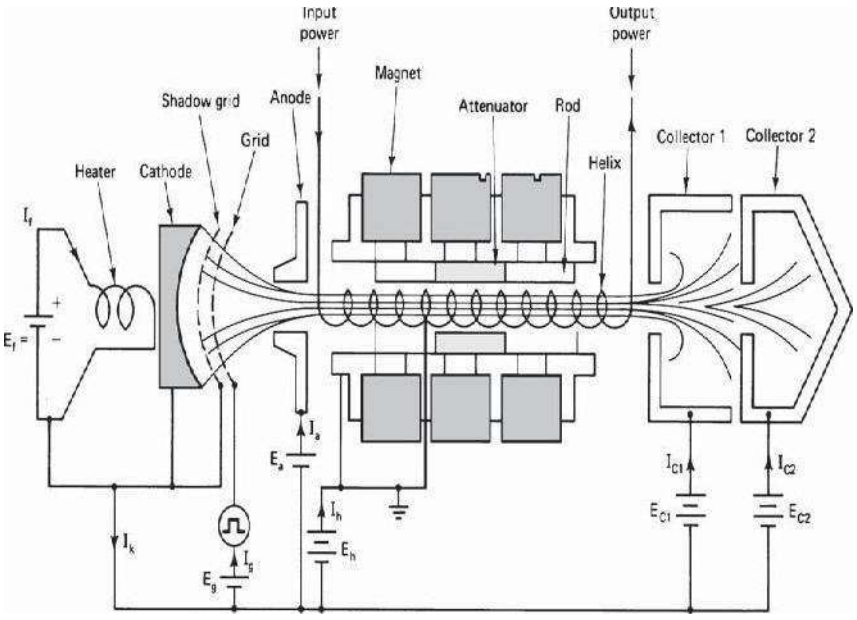


Figure 2.13 Satellite TWTA

used in ground stations, the magnetic field can be provided by means of a solenoid and dc power supply. The comparatively large size and high power consumption of solenoids make them unsuitable for use aboard satellites, and lower-power TWTs are used which employ permanent- magnet focusing.

The wave actually will travel around the helical path at close to the speed of light, but it is the axial component of wave velocity which interacts with the electron beam.

This component is less than the velocity of light approximately in the ratio of helix pitch to circumference. Because of this effective reduction in phase velocity, the helix is referred to as a *slowwave structure*.

The advantage of the TWT over other types of tube amplifiers is that it can provide amplification over a very wide bandwidth. Input levels to the TWT must be carefully controlled, however, to minimize the effects of certain forms of distortion.

The worst of these result from the nonlinear transfer characteristic of the TWT, illustrated in Fig. 2.14.

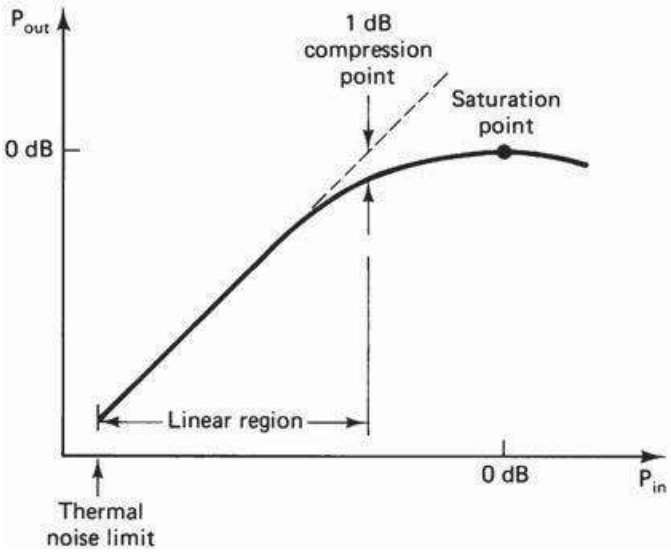


Figure 2.14 Power transfer characteristics of a TWT. The saturation point is used as 0-dB reference for both input and output.

At low-input powers, the output-input power relationship is linear; that is, a given decibel change in input power will produce the same decibel change in output power. At higher power inputs, the output power sat- urates, the point of maximum power output being known as the *satu- ration point*.

The saturation point is a very convenient reference point, and input and output quantities are usually referred to it. The linear region of the TWT is defined as the region bound by the thermal noise limit at the low end and by what is termed the *1-dB compression point* at the upper end. This is the point where the actual transfer curve drops

2.7. Satellite uplink and downlink Analysis and Design

2.7.1 Introduction

This chapter describes how the link-power budget calculations are made. These calculations basically relate two quantities, the transmit power and the receive power, and show in detail how the difference between these two powers is accounted for.

Link-budget calculations are usually made using decibel or decilog quantities. These are explained in App. G. In this text [square] brackets are used to denote decibel quantities using the basic power definition.

Where no ambiguity arises regarding the units, the abbreviation dB is used. For example, Boltzmann's constant is given as 228.6 dB, although, strictly speaking, this should be given as 228.6 deci logs relative to 1 J/K.

2.7.2 Equivalent Isotropic Radiated Power

A key parameter in link-budget calculations is the *equivalent isotropic radiated power*, conventionally denoted as EIRP. From Eqs, the maximum power flux density at some distance *r* from a transmitting antenna of gain *G* i

$$\text{Pr} = \frac{GP}{4\pi^2}$$

An isotropic radiator with an input power equal to GPS would produce the same flux density. Hence, this product is referred to as the EIRP, or EIRP is often expressed in decibels relative to 1 W, or dBW. Let PS be in watts; then [EIRP] = [PS] x [G] dB ,where [PS] is also in dBW and [G] is in dB.

2.7.3 Transmission Losses

The [EIRP] may be thought of as the power input to one end of the transmission link, and the problem is to find the power received at the other end. Losses will occur along the way, some of which are constant.

Other losses can only be estimated from statistical data, and some of these are dependent on weather conditions, especially on rainfall.

The first step in the calculations is to determine the losses for *clear- weather* or *clear-sky conditions*. These calculations take into account the losses, including those calculated on a statistical basis, which do not vary significantly with time. Losses which are weather-related, and other losses which fluctuate with time, are then allowed for by introducing appropriate *fade margins* into the transmission equation.

Free-space transmission:

As a first step in the loss calculations, the power loss resulting from the spreading of the signal in space must be determined.

Feeder losses:

Losses will occur in the connection between the receive antenna and the receiver proper. Such losses will occur in the connecting waveguides, filters, and couplers. These will be denoted by RFL, or [RFL] dB, for *receiver feeder losses*.

Antenna misalignment losses

When a satellite link is established, the ideal situation is to have the earth station and satellite antennas aligned for maximum gain, as shown in Fig. There are two possible sources of off-axis loss, one at the satellite and one at the earth station, as shown in Fig.

The off-axis loss at the satellite is taken into account by designing the link for operation on the actual satellite antenna contour; this is described in more detail in later sections. The off-axis loss at the earth station is referred to as the *antenna pointing loss*. Antenna pointing losses are usually only a few tenths of a decibel;

In addition to pointing losses, losses may result at the antenna from misalignment of the polarization direction (these are in addition to the polarization losses described in Chap. 5). The polarization misalign- ment losses

are usually small, and it will be assumed that the antenna misalignment losses, denoted by [AML], include both pointing and polar- ization losses resulting from antenna misalignment. It should be noted

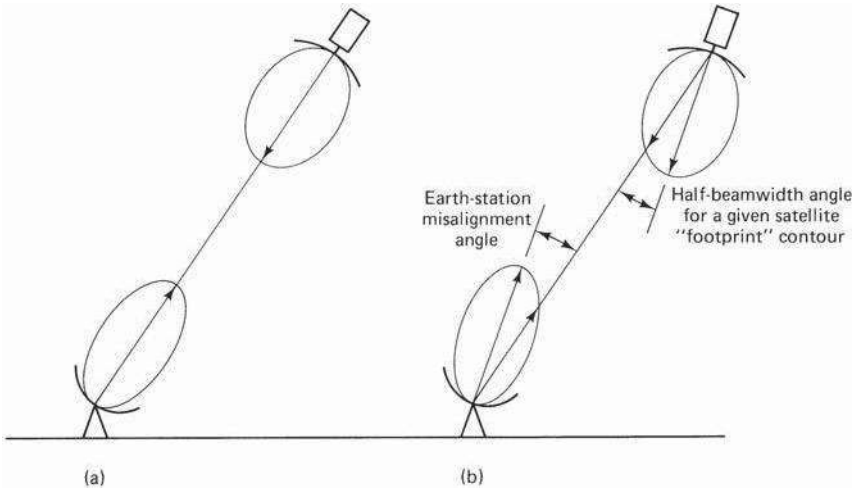


Figure 2.15 (*a*) Satellite and earth-station antennas aligned for maximum gain; (*b*) earth station situated on a given satellite "footprint," and earth-station antenna misaligned.

2.8 The Link-Power Budget Equation

Now that the losses for the link have been identified, the power at the receiver, which is the power output of the link, may be calculated simply as [EIRP] [LOSSES] [G_R], where the last quantity is the receiver antenna gain. Note carefully that decibel addition must be used.

The major source of loss in any ground-satellite link is the free-space spreading loss [FSL], the basic link-power budget equation taking into account this loss only. However, the other losses also must be taken into account, and these are simply added to [FSL]. The losses for clear-sky conditions are

[LOSSES] = [FSL] + [RFL] + [AML] + [AA] - [PL] equation for the received power is then

$$[P_R] = [\text{EIRP}] \times [G_R] - [\text{LOSSES}]$$

where [PR] received power, dBW

[EIRP] - equivalent isotropic radiated power, dBW [FSL] free-space spreading loss, dB

[RFL] - receiver feeder loss, dB

[AML] - antenna misalignment loss, dB

[AA] - atmospheric absorption loss, dB [PL] polarization mismatch loss, dB

2.9 Amplifier noise temperature

Consider first the noise representation of the antenna and the *low noise amplifier* (LNA) shown in Fig. 2.15.

The available power gain of the amplifier is denoted as G, and the noise power output, as P_{no}.

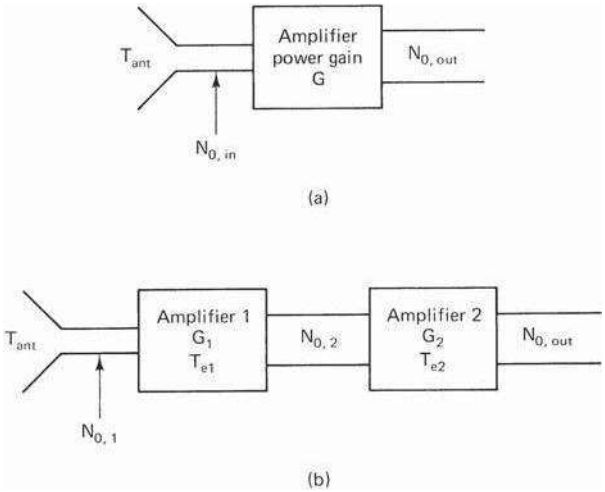


Figure 2.15 LNA Amplifier gain

For the moment we will work with the noise power per unit bandwidth, which is simply noise energy in joules as shown by Eq.

The input noise energy coming from the antenna is

$$N_{0,\text{ant}} = kT_{\text{ant}}$$

2.10 The Uplink

The uplink of a satellite circuit is the one in which the earth station is transmitting the signal and the satellite is receiving it specifically that the uplink is being considered.

$$\frac{C}{N} = [EIRP]] - [LOSSES] \;] + k]$$

In this Eq the values to be used are the earth station EIRP, the satellite receiver feeder losses, and satellite receiver G/T. The free-space loss and other losses which are frequency-dependent are calculated for the uplink frequency.

2.10.1 Input backoff

Number of carriers are present simultaneously in a TWTA, the operating point must be backed off to a linear portion of the transfer characteristic to reduce the effects of inter modulation distortion. Such multiple carrier operation occurs with *frequency- division multiple access* (FDMA), which is described in Chap. 14. The point to be made here is that *backoff* (BO) must be allowed for in the link- budget calculations.

Suppose that the saturation flux density for single-carrier operation is known. Input BO will be specified for multiple-carrier operation, referred to the single-carrier saturation level. The earth-station EIRP will have to be reduced by the specified BO, resulting in an uplink value of

$$[\text{EIRP}]_U = [\text{EIRP}_S]_U + [\text{BO}]_i$$

2.10.2 The earth station HPA

The earth station HPA has to supply the radiated power plus the transmit feeder losses, denoted here by TFL, or [TFL] dB. These include waveguide, filter, and coupler losses between the HPA output and the transmit antenna. Referring back to Eq. (12.3), the power output of

The earth station itself may have to transmit multiple carriers, and its output also will require back off, denoted by $[\text{BO}]_{\text{HPA}}$. The earth station HPA must be rated for a saturation power output given by

$$[P_{\text{HPA,sat}}] = [P_{\text{HPA}}] + [\text{BO}]_{\text{HPA}}$$

2.11 Downlink

The downlink of a satellite circuit is the one in which the satellite is transmitting the signal and the earth station is receiving it. Equation can be applied to the downlink, but subscript D will be used to denote specifically that the downlink is being considered. Thus Eq. becomes

$$\frac{C}{N} = [EIRP]] - [LOSSES] \quad] + [\quad]$$

In Eq. the values to be used are the satellite EIRP, the earth- station receiver feeder losses, and the earth-station receiver G/T. The free space and other losses are calculated for the downlink frequency. The resulting carrier-to-noise density ratio given by Eq. is that which appears at the detector of the earth station receiver.

2.11.1 Output back-off

Where input BO is employed as described in a corresponding output BO must be allowed for in the satellite EIRP. As the curve of Fig. 2.16 shows, output BO is not linearly related to input BO. A rule of thumb, frequently used, is to take the output BO as the point on the curve which is 5 dB below the extrapolated linear portion, as shown in Fig. 12.7. Since the linear portion gives a 1:1 change in decibels, the relationship between input and output BO is $[BO]_0$ $[BO]_i$ 5 dB. For example, with an input BO of $[BO]_i$ 11 dB, the corresponding output BO is $[BO]_0$

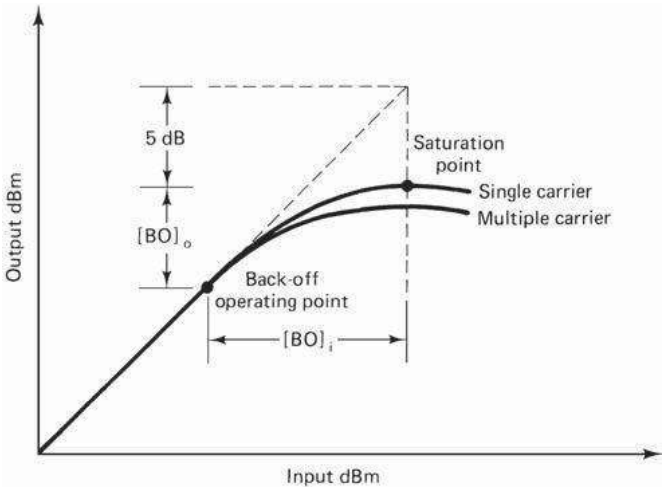


Figure 2.16 Input and output back- off relationship for the satellite traveling-wave-tube amplifier; $[BO]_i$ $[BO]_0$ 5 dB.

2.11.2 Effects of Rain

In the C band and, more especially, the Ku band, rainfall is the most significant cause of signal fading. Rainfall results in attenuation of radio waves by scattering and by absorption of energy from the wave.

Rain attenuation increases with increasing frequency and is worse in the Ku band compared with the C band.

This produces a depolarization of the wave; in effect, the wave becomes elliptically polarized. This is true for both linear and circular polar- izations, and the effect seems to be much worse for circular polarization (Freeman, 1981).

The C/N_0 ratio for the downlink alone, not counting the P_{NU} contri- bution, is P_R/P_{ND}, and the combined C/N_0 ratio at the ground receiver is

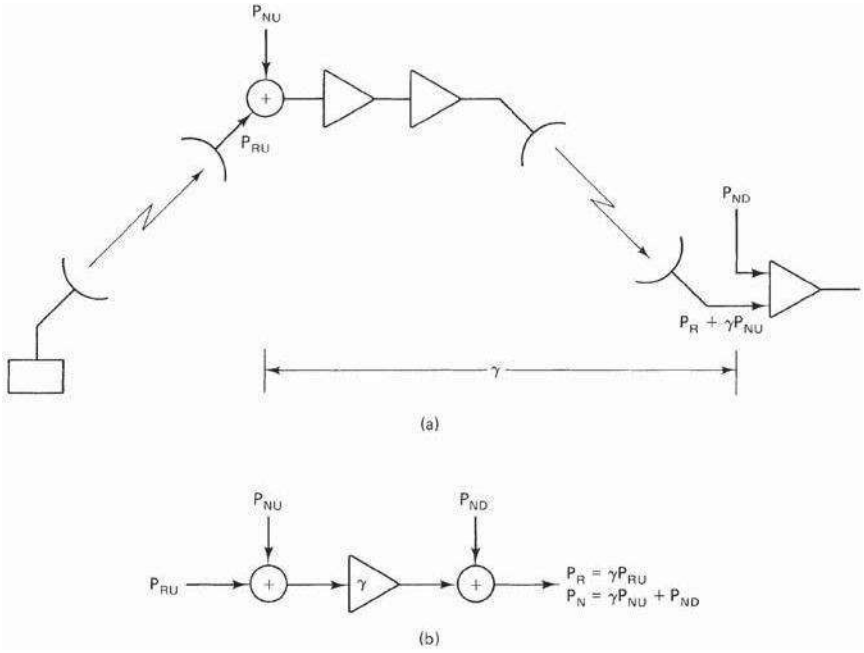


Figure 2.17 (a) Combined uplink and downlink; (b) power flow diagram

The reason for this reciprocal of the sum of the reciprocals method is that a single signal power is being transferred through the system, while the various noise powers, which are present are additive. Similar reasoning applies to the carrier-to-noise ratio, C/N.

2.12. inter modulation and interference

Intermodulation interference is the undesired combining of several signals in a nonlinear device, producing new, unwanted frequencies, which can cause interference in adjacent receivers located at repeater sites.

Not all interference is a result of intermodulation distortion. It can come from co-channel interference, atmospheric conditions as well as man-made noise generated by medical, welding and heating equipment.

Most intermodulation occurs in a transmitter's nonlinear power amplifier (PA). The next most common mixing point is in the front end of a receiver. Usually it occurs in the unprotected first mixer of older model radios or in some cases an overdriven RF front-end amp.

Intermodulation can also be produced in rusty or corroded tower joints, guy wires, turnbuckles and anchor rods or any nearby metallic object, which can act as a nonlinear "mixer/rectifier" device.

2.13. Propagation Characteristics and Frequency considerations

2.13.1 Introduction

A number of factors resulting from changes in the atmosphere have to be taken into account when designing a satellite communications system in order to avoid impairment of the wanted signal.

Generally, a margin in the required carrier-to-noise ratio is incorporated to accommodate such effects.

2.13.2 Radio Noise

Radio noise emitted by matter is used as a source of information in radio astronomy and in remote sensing. Noise of a thermal origin has a continuous spectrum, but several other radiation mechanisms cause the emission to have a spectral-line structure. Atoms and molecules are distinguished by their different spectral lines.

For other services such as satellite communications noise is a limiting factor for the receiving system; generally, it is inappropriate to use receiving systems with noise temperatures which are much less than those specified by the minimum external noise.

From about 30 MHz to about 1 GHz cosmic noise predominates over atmospheric noise except during local thunderstorms, but will generally be exceeded by man-made noise in populated areas.

In the bands of strong gaseous absorption, the noise temperature reaches maximum values of some 290 K. At times, precipitation will also increase the noise temperature at frequencies above 5 GHz.

Figure 6.1 gives an indication of sky noise at various elevation angles and frequencies.

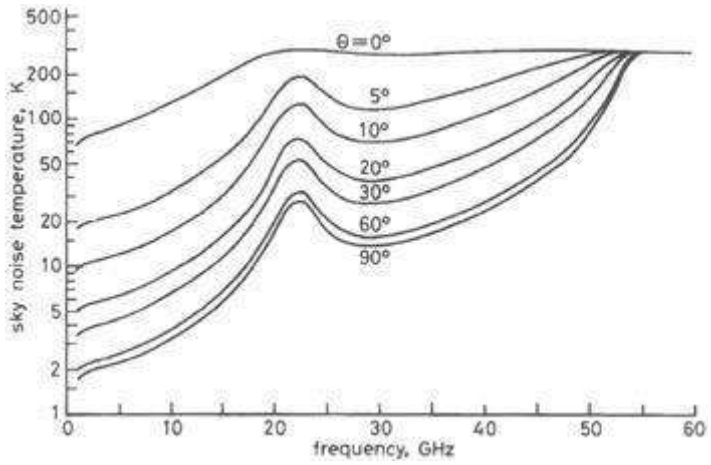


Figure 2.18 Sky-Noise Temperature for Clear Air

2.14. System reliability and design lifetime

2.14.1 System reliability

Satellites are designed to operate dependably throughout their operational life, usually a number of years.

This is achieved through stringent quality control and testing of parts and subsystems before they are used in the construction of the satellite.

Redundancy of key components is often built in so that if a particular part or subassembly fails, another can perform its functions.

In addition, hardware and software on the satellite are often designed so that ground controllers can reconfigure the satellite to work around a part that has failed.

2.14.2. Design lifetime

The Milstar constellation has demonstrated exceptional reliability and capability, providing vital protected communications to the warfighter," said Kevin Bilger, vice president and general manager, Global Communications Systems, Lockheed Martin Space Systems in Sunnyvale.

"Milstar's robust system offers our nation worldwide connectivity with flexible, dependable and highly secure satellite communications."

The five-satellite Milstar constellation has surpassed 63 years of combined successful operations, and provides a protected, global communication network for the joint forces of the U.S. military. In addition, it can transmit voice, data, and imagery, and offers video teleconferencing capabilities.

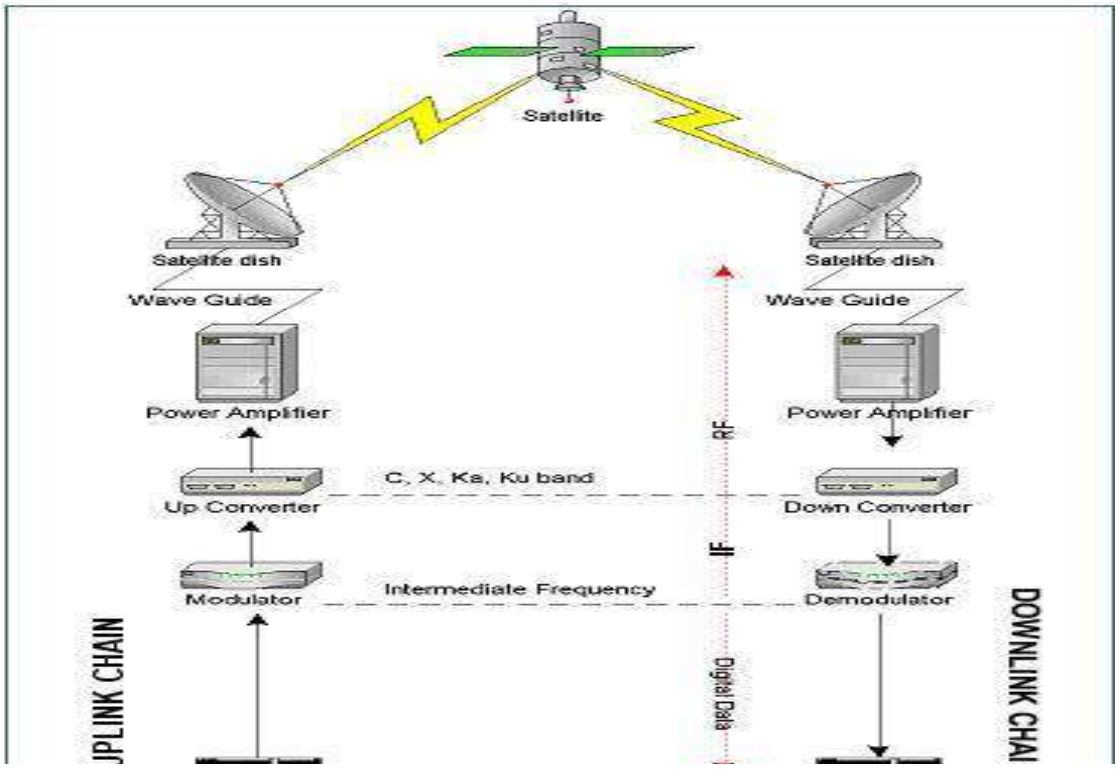
The system is the principal survivable, endurable communications structure that the President, the Secretary of Defense and the Commander, U.S. Strategic Command use to maintain positive command and control of the nation's strategic forces.

In addition to this 10-year milestone for Flight-5, each of the first two Milstar satellites have been on orbit for over 16 years – far exceeding their 10-year design life.

The next-generation Lockheed Martin-built Advanced EHF satellites, joining the Milstar constellation, provide five times faster data rates and twice as many connections, permitting transmission of strategic and tactical military communications, such as real-time video, battlefield maps and targeting data. Advanced EHF satellites are designed to be fully interoperable and backward compatible with Milstar.

Headquartered in Bethesda, Md., Lockheed Martin is a global security company that employs about 123,000 people worldwide and is principally engaged in the research, design, development, manufacture, integration and sustainment of advanced technology systems, products and services. The Corporation's net sales for 2011 were $46.5 billion.

APPLICATIONS



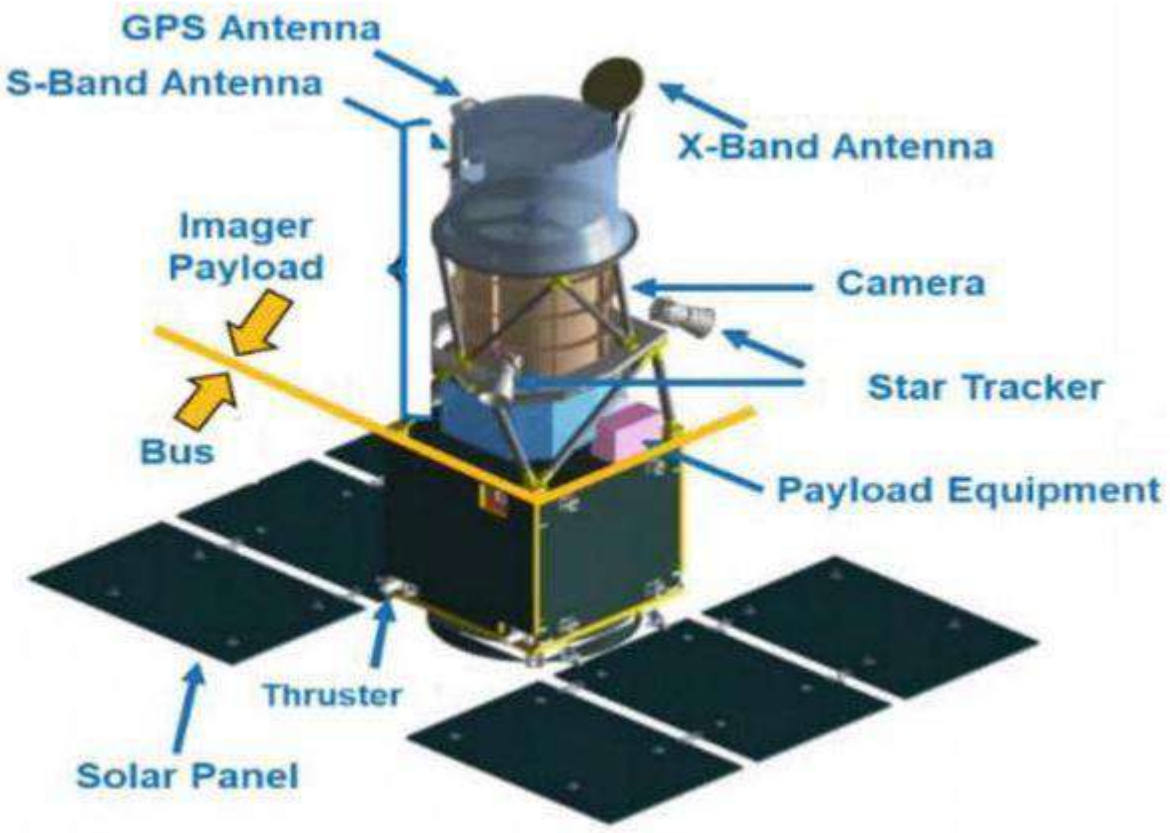


Figure typical satellite with bus and payload separation

SATELLITE COMMUNICATION

UNIT III EARTH SEGMENT

1.1 Earth Station Technology

The earth segment of a satellite communications system consists of the transmit and receive earth stations. The simplest of these are the home *TV receive-only* (TVRO) systems, and the most complex are the terminal stations used for international communications networks. Also included in the earth segment are those stations which are on ships at sea, and commercial and military land and aeronautical mobile stations.

As mentioned in earth stations that are used for logistic sup- port of satellites, such as providing the *telemetry, tracking, and command* (TT&C) functions, are considered as part of the space segment.

1.1.1 Terrestrial Interface

Earth station is a vital element in any satellite communication network. The function of an earth station is to receive information from or transmit information to, the satellite network in the most cost-effective and reliable manner while retaining the desired signal quality. The design of earth station configuration depends upon many factors and its location. But it is fundamentally governed by its

Location which are listed below,

- In land
- On a ship at sea
- Onboard aircraft

The factors are

- Type of services
- Frequency bands used
- Function of the transmitter
- Function of the receiver
- Antenna characteristics

1.1.2 Transmitter and Receiver

Any earth station consists of four major subsystems

- Transmitter
- Receiver
- Antenna • Tracking equipment

Two other important subsystems are

- Terrestrial interface equipment
- Power supply

The earth station depends on the following parameters

- Transmitter power
- Choice of frequency
- Gain of antenna
- Antenna efficiency
- Antenna pointing accuracy
- Noise temperature

The functional elements of a basic digital earth station are shown in the below figure

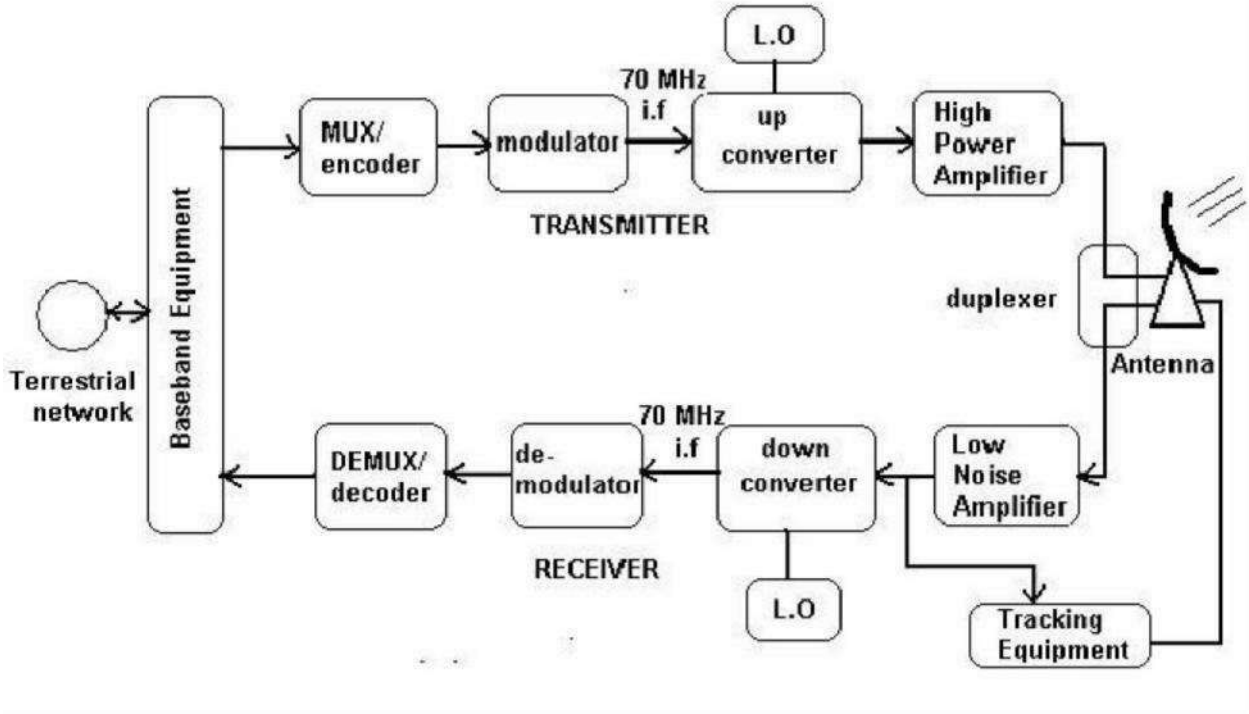


Figure 3.1 Transmitter- Receiver

Digital information in the form of binary digits from terrestrial networks enters earth station and is then processed (filtered, multiplexed, formatted etc.) by the base band equipment.

- The encoder performs error correction coding to reduce the error rate, by introducing extra digits into digital stream generated by the base band

equipment. The extra digits carry information.

- In satellite communication, I.F carrier frequency is chosen at 70 MHz for communication using a 36 MHz transponder bandwidth and at 140 MHz for a transponder bandwidth of 54 or 72 MHz.

- On the receive side, the earth station antenna receives the low -level modulated R.F carrier in the downlink frequency spectrum.

- The low noise amplifier (LNA) is used to amplify the weak rec eived signals and improve the signal to Noise ratio (SNR). The error rate requirements can be met more easily.
- R.F is to be reconverted to I.F at 70 or 140 MHz because it is easier design a demodulation to work at these frequencies than 4 or 12 GHz.

- The demodulator estimate which of the possible symbols was transmitted based on observation of the received if carrier.

- The decoder performs a function opposite that of the encoder. Because the sequence of symbols recovered by the demodulator may cont ain errors, the decoder must use the uniqueness of the redundant digits introduced by the encoder to correct the errors and recover information-bearing digits.

- The information stream is fed to the base-band equipment for processing for delivery to the terrestrial network.

- The tracking equipments track the satellite and align the beam towards it to facilitate communication.

3.1.3. Earth Station Tracking System

Tracking is essential when the satellite drift, as seen by an earth station antenna is a significant fraction of an earth station's antenna beam width.

An earth station's tracking system is required to perform some of the functions such as

i)Satellite acquisition
ii)Automatic tracking
iii) Manual tracking
iv) Program tracking.

1.2 Antenna Systems

The antenna system consist of

- Feed System
- Antenna Reflector
- Mount
- Antenna tracking System

1.2.1 FEED SYSTEM

The feed along with the reflector is the radiating/receiving element of electromagnetic waves. The reciprocity property of the feed element makes the earth station antenna system suitable for transmission and reception of electromagnetic waves.

The way the waves coming in and going out is called feed configuration
Earth Station feed systems most commonly used in satellite communication are:

i) Axi-Symmetric Configuration
ii)Asymmetric Configuration
i)Axi-Symmetric Configuration

In an axi-symmetric configuration the antenna axes are symmetrical with respect to the reflector ,which results in a relatively simple mechanical structure and antenna mount.

Primary Feed

In primary, feed is located at the focal point of the parabolic reflector. Many dishes use only a single bounce, with incoming waves reflecting off the dish surface to the focus in front of the dish, where the antenna is located. when the dish is used to transmit ,the transmitting antenna at the focus beams waves toward the dish, bouncing them off to space. This is the simplest arrangement.

Cassegrain

Many dishes have the waves make more than one bounce .This is generally called as folded systems. The advantage is that the whole dish and feed system is more compact. There are several folded configurations, but all have at least one secondary reflector also called a sub reflector, located out in front of the dish to redirect the waves.

A common dual reflector antenna called Cassegrain has a convex sub reflector positioned in front of the main dish, closer to the dish than the focus. This sub reflector bounces back the waves back toward a feed located on the main dish's center, sometimes behind a hole at the center of the main dish. Sometimes there are even more sub reflectors behind the dish to direct the waves to the fed for convenience or compactness.

Gregorian

This system has a concave secondary reflector located just beyond the primary focus. This also bounces the waves back toward the dish.

ii) Asymmetric Configuration

Offset or Off-axis feed

The performance of tan axi-symmetric configuration is affected by the blockage of the aperture by the feed and the sub reflector assembly. The result is a reduction in the antenna efficiency and an increase in the side lobe levels. The asymmetric configuration can remove this limitation..This is achieved by off - setting the mounting arrangement of the feed so that it does not obstruct the main beam. As a result ,the efficiency and side lobe level performance a re improved.

1.2.2 ANTENNA REFLECTOR

Mostly parabolic reflectors are used as the main antenna for the earth stations because of the high gain available from the reflector and the ability of focusing a parallel beam into a point at the focus where the feed,i.e., the receiving/radiating element is located .For large antenna system more than one reflector surfaces may be used in as in the cassegrain antenna system.

Earth stations are also classified on the basis of services for example:

1. Two way TV ,Telephony and data
2. Two way TV
3. TV receive only and two way telephony and data
4. Two way data

From the classifications it is obvious that the technology of earth station will vary considerably on the performance and the service require ments of earth station

For mechanical design of parabolic reflector the following parameters are required to be considered:

- Size of the reflector
- Focal Length /diameter ratio
- RMS error of main and sub reflector
- Pointing and tracking accuracies
- Speed and acceleration
- Type of mount
- Coverage Requirement

Wind Speed

The size of the reflector depends on transmit and receive gain requirement and beamwidth of the antenna. Gain is directly proportional to the antenna diameter whereas the beamwidth is inversely proportional to the antenna diameter .for high inclination angle of the satellite ,the tracking of the earth station becomes necessary when the beamwidth is too narrow.

The gain of the antenna is given by

Gain= $(\eta 4\Pi Aeff)/ \lambda 2$

Where Aeff is the aperture

Λ is wave length

H is efficiency of antenna system

For a parabolic antenna with circular aperture diameter D, the gain of the antenna is :

Gain= $(\eta 4\Pi/ \lambda 2)(\Pi D2/4)$

$= \eta (\Pi D/ \lambda)2$

The overall efficiency of the antenna is the net product of various factors such as

1. Cross Polarization
2. Spill over
3. Diffraction
4. Blockage
5. Surface accuracy
6. Phase error
7. Illumination

In the design of feed, the ratio of focal length F to the diameter of the

reflector D of the antenna system control the maximum angle subtended by the reflector surface on the focal point. Larger the F/D ratio larger is the aperture illumination efficiency and lower the cross polarization.

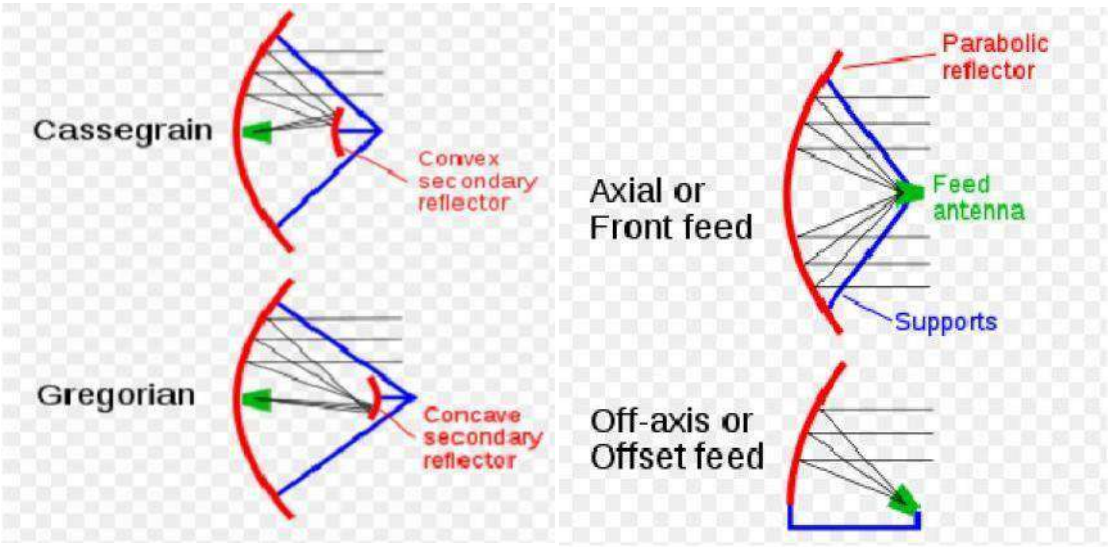


Figure 3.2 Antenna sub systems

1.2.3 ANTENNA MOUNT

Type of antenna mount is determined mainly by the coverage requirement and tracking requirements of the antenna systems. Different types of mounts used for earth station antenna are:

i) **The Azimuth –elevation mount**

This mount consists of a primary vertical axis. Rotation around this axis controls the azimuth angle. The horizontal axis is mounted over the primary axis, providing the elevation angle control.

ii) **The X-Y mount**

It consists of a horizontal primary axis (X-axis) and a secondary axis (Y-axis) and at right angles to it. Movement around these axes provides necessary steering.

1.2.4 ANTENNA TRACKING SYSTEM

Tracking is essential when the satellite drift, as seen by an earth station antenna is a significant fraction of an earth station's antenna beam width.

An earth station's tracking system is required to perform some of the functions such as

i)Satellite acquisition
ii)Automatic tracking
iii) Manual tracking
iv) Program tracking.

Recent Tracking Techniques

There have been some interesting recent developments in auto -track techniques which can potentially provide high accuracies at a low cost.

In one proposed technique the sequential lobing technique has been I implemented by using rapid electronic switching of a s single beam which effectively approximates simultaneous lobbing.

1.3 Receive-Only Home TV Systems

Planned broadcasting directly to home TV receivers takes place in the Ku (12-GHz) band. This service is known as *direct broadcast satellite* (DBS) service.

There is some variation in the frequency bands assigned to different geographic regions. In the Americas, for example, the down- link band is 12.2 to 12.7 GHz.

The comparatively large satellite receiving dishes [ranging in diame- ter from about 1.83 m (6 ft) to about 3-m (10 ft) in some locations], which may be seen in some "backyards" are used to receive downlink TV signals at C band (4 GHz).

Originally such downlink signals were never intended for home reception but for network relay to commercial TV outlets (VHF and UHF TV broadcast stations and cable TV "head-end" studios).

1.3.1 The Indoor unit

Equipment is now marketed for home reception of C-band signals, and some manufacturers provide dual C-band/Ku-band equipment. A single mesh type reflector may be used which focuses the signals into a dual feed- horn, which has two separate outputs, one for the C-band signals and onefor the Ku-band signals.

Much of television programming originates as *first generation signals*, also known as *master broadcast quality signals*.

These are transmitted via satellite in the C band to the network head- end stations, where they are retransmitted as compressed digital signals to cable and direct broadcast satellite providers.

- Another of the advantages, claimed for home C-band systems, is the larger number of satellites available for reception compared to what is available for direct broadcast satellite sys- terms.

- Although many of the C-band transmissions are scrambled, there are free channels that can be received, and what are termed "wild feeds."

- These are also free, but unannounced programs, of which details can be found in advance from various publications and Internet sources.

- C-band users can also subscribe to pay TV channels, and another advantage claimed is that subscription services are cheaper than DBS or cable because of the multiple-source programming available.

- The most widely advertised receiving system for C-band system appears to be 4DTV manufactured by Motorola.

This enables reception of

- Free, analog signals and "wild feeds"
- Video Cipher ll plus subscription services
- Free Digi Cipher 2 services
- Subscription DigiCipher 2 services

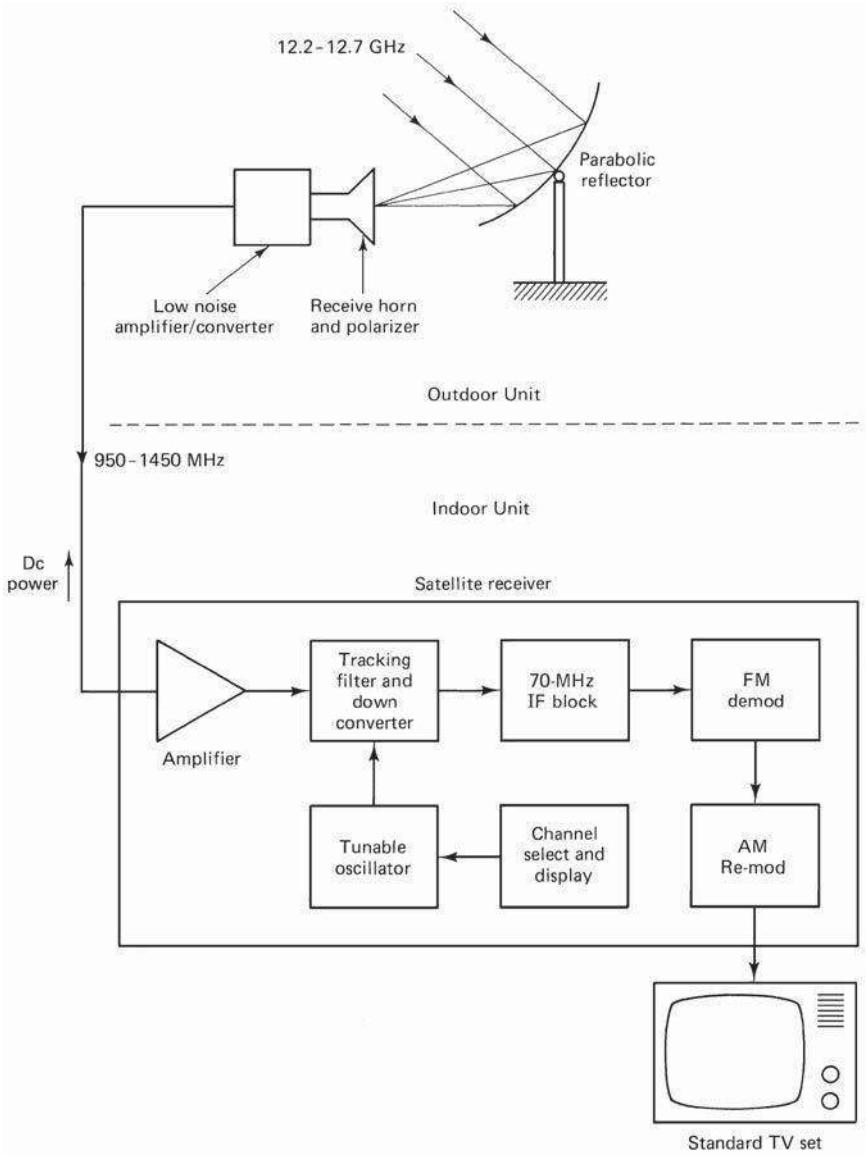


Figure 3.3 TVRO System block diagrams

.

1.3.2 The outdoor unit

This consists of a receiving antenna feeding directly into a low -noise amplifier/converter combination. A parabolic reflector is generally used, with the receiving horn mounted at the focus. A common design is to have the focus directly in front of the reflector, but for better interference rejection, an offset feed may be used as shown.

Comparing the gain of a 3-m dish at 4 GHz with a 1-m dish at 12 GHz, the ratio D/λ equals 40 in each case, so the gains will be about equal. Although the free-space losses are much higher at 12 GHz compared with 4 GHz.

The downlink frequency band of 12.2 to 12.7 GHz spans a range of 500 MHz, which accommodates 32 TV/FM channels, each of which is 24-MHz wide. Obviously, some overlap occurs between channels, but these are alternately polarized *left-hand circular* (LHC) and *right-hand circular* (RHC) or vertical/horizontal, to reduce interference to accept- able levels. This is referred to as *polarization interleaving*. A polarizer that may be switched to the desired polarization from the indoor con- trol unit is required at the receiving horn.

The receiving horn feeds into a *low-noise converter* (LNC) or possibly a combination unit consisting of a *low-noise amplifier* (LNA) followed by a converter.

The combination is referred to as an LNB, for *low-noise block*. The LNB provides gain for the broadband 12-GHz signal and then converts the signal to a lower frequency range so that a low-cost coaxial cable can be used as feeder to the indoor unit.

The signal fed to the indoor unit is normally a wideband signal cov- ering the range 950 to 1450 MHz. This is amplified and passed to a tracking filter which selects the desired channel, as shown in Fig.

As previously mentioned, polarization interleaving is used, and only half the 32 channels will be present at the input of the indoor unit for any one setting of the antenna polarizer. This eases the job of the tracking filter, since alternate channels are well separated in frequency.

The selected channel is again down converted, this time from the 950- to 1450-MHz range to a fixed intermediate frequency, usually 70 MHz although other values in the *very high frequency* (VHF) range are also used.

The 70-MHz amplifier amplifies the signal up to the levels required for demodulation. A major difference between DBS TV and conventional TV is that with DBS, frequency modulation is used, whereas with conventional TV, amplitude modulation in the form of *vestigial single side- band* (VSSB) is used.

The 70-MHz, FM *intermediate frequency* (IF) carrier therefore must be demodulated, and the baseband information used to generate a VSSB signal which is fed into one of the VHF/UHF channels of a standard TV set.

1.4 Master Antenna TV System

A *master antenna TV* (MATV) system is used to provide reception of DBS TV/FM channels to a small group of users, for example, to the tenants in an apartment building. It consists of a single outdoor unit (antenna and LNA/C) feeding a number of indoor units, as shown in Fig.

It is basically similar to the home system already described, but with each user having access to all the channels independently of the other users. The advantage is that only one outdoor unit is required, but as shown, separate LNA/Cs and feeder cables are required for each sense of polarization.

Compared with the single- user system, a larger antenna is also required (2- to 3-m diameter) in order to maintain a good signal-to-noise ratio at all the indoor units.

Where more than a few subscribers are involved, the distribution system used is similar to the *community antenna* (CATV) system described in the following section.

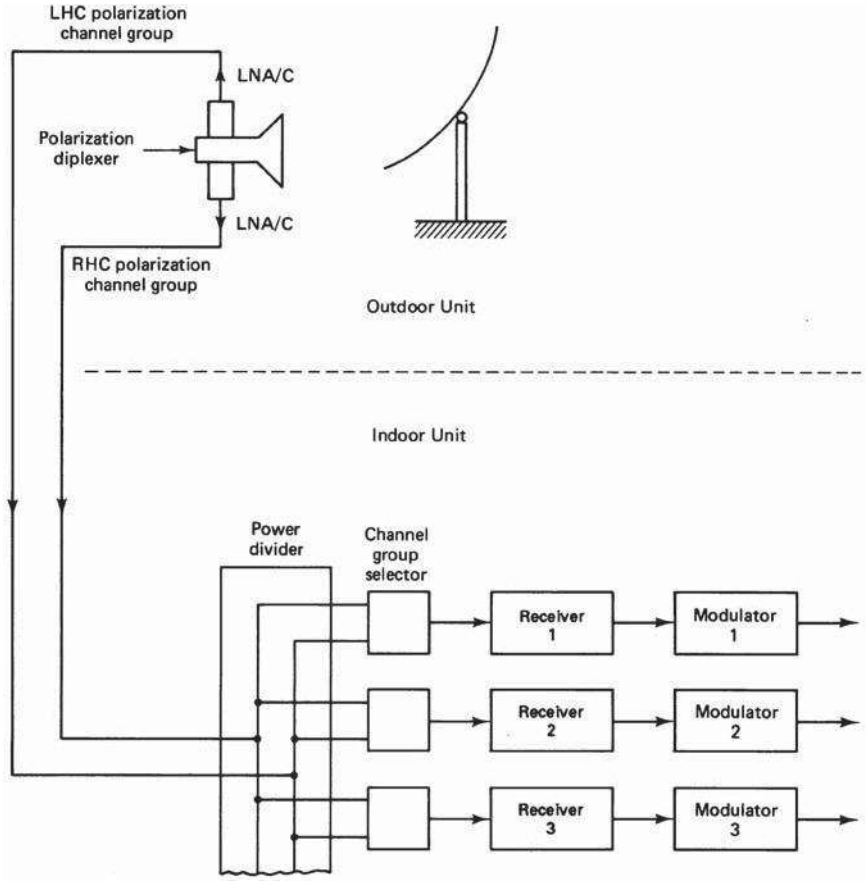


Figure 3.4 CATV System block diagrams

1.5 Community Antenna TV System

The CATV system employs a single outdoor unit, with separate feeds available for each sense of polarization, like the MATV system, so that all channels are made available simultaneously at the indoor receiver.

Instead of having a separate receiver for each user, all the carriers are demodulated in a common receiver-filter system, as shown in Fig. The channels are then combined into a standard multiplexed signal for transmission over cable to the subscribers.

In remote areas where a cable distribution system may not be installed, the signal can be rebroadcast from a low-power VHF TV transmitter.

Figure shows a remote TV station which employs an 8-m (26.2-ft) antenna for reception of the satellite TV signal in the C band.

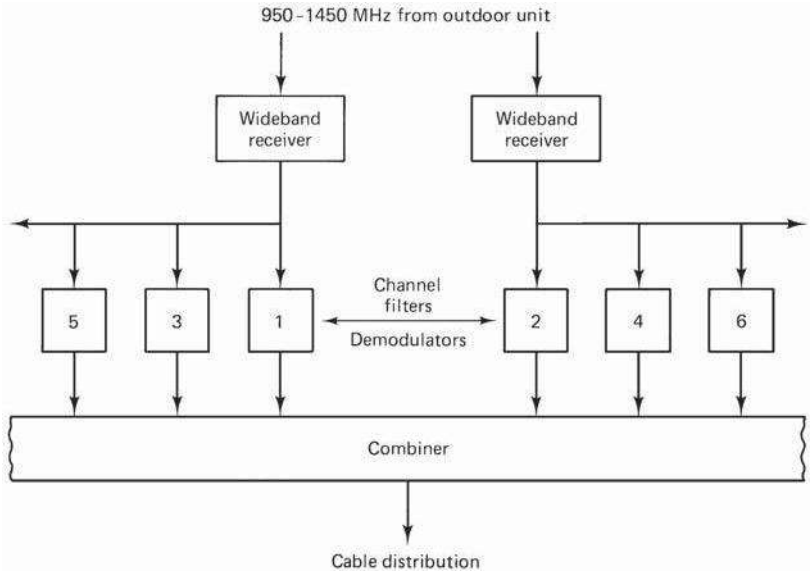


Figure 3.5 One possible arrangement for the indoor unit of a community antenna TV (CATV) system.

With the CATV system, local programming material also may be distributed to subscribers, an option which is not permitted in the MATV system.

1.6 Test Equipment Measurements on G/T, C/No, EIRP

Measurement of G/T of small antennas is easily and simply measured using the spectrum analyser method. For antennas with a diameter of less than 4.5 meters it is not normally necessary to point off from the satellite.

A step in frequency would be required into one of the satellite transponder guard bands.

However antennas with a G/T sufficiently large to enable the station to see the transponder noise floor either a step in frequency into one of the satellite transponder guard bands and/or in azimuth movement would be required.

The test signal can be provided from an SES WORLD SKIES beacon.

Procedure

(a) Set up the test equipment as shown below. Allow half an hour to warm up,
and then calibrate in accordance with the manufacturer's procedures.

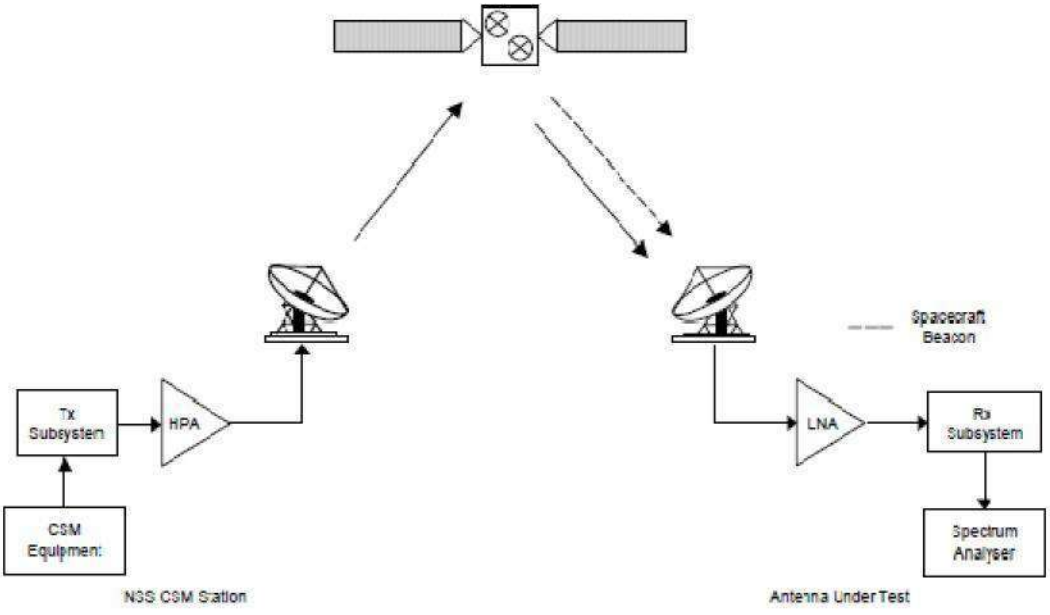


Figure 3.6 One possible arrangement for Measurement of G/T

(b) Adjust the centre frequency of your spectrum analyzer to receive the SES WORLD SKIES beacon (data to be provided on the satellite used for testing)

(c) Carefully peak the antenna pointing and adjust the polarizer by nulling the
cross polarized signal. You cannot adjust polarization when using the circularly polarized SES WORLD SKIES beacon.

(d) Configure the spectrum analyser as follows:

Centre Frequency: Adjust for beacon or test signal frequency (to be advised).

Use marker to peak and marker to centre functions.

- Frequency Span: 100 KHz
- Resolution Bandwidth: 1 KHz
- Video Bandwidth: 10 Hz (or sufficiently small to limit noise variance)
- Scale: 5 dB/div
- Sweep Time: Automatic
- Attenuator Adjust to ensure linear operation. Adjust to provide the "Noise floor delta" described in steps 7 and 8.

(e) To insure the best measurement accuracy during the following steps, adjust the spectrum analyser amplitude (reference level) so that the measured signal, carrier or noise, is approximately one division below the top line of the spectrum analyser display.

(f) Record the frequency and frequency offset of the test signal from the nominal frequency:

For example, assume the nominal test frequency is 11750 MHz but the spectrum analyser shows the peak at 11749 MHz. The frequency offset in this case is -1 MHz.

(g) Change the spectrum analyser centre frequency as specified by SES WORLD SKIES so that the measurement is performed in a transponder guard band so that only system noise power of the earth station and no satellite signals are received. Set the spectrum analyser frequency as follows:

Centre Frequency = Noise slot frequency provided by the PMOC

(h) Disconnect the input cable to the spectrum analyser and confirm that the noise floor drops by at least 15 dB but no more than 25dB. This confirms that the spectrum analyser's noise contribution has an insignificant effect on the measurement. An input attenuation value allowing a "Noise floor Delta" in excess of 25 dB may cause overloading of the spectrum analyser input. (i) Reconnect the input cable to the spectrum analyser.

(j) Activate the display line on the spectrum analyser.

(k) Carefully adjust the display line to the noise level shown on the spectrum analyser. Record the display line level.

(l) Adjust the spectrum analyser centre frequency to the test carrier frequency
recorded in step (e).

(m) Carefully adjust the display line to the peak level of the test carrier on the
spectrum analyser. Record the display line level.

(n) Determine the difference in reference levels between steps (l) and (j) which is the (C+N)/N.

(o) Change the (C+N)/N to C/N by the following conversion:

This step is not necessary if the (C+N)/N ratio is more than 20 dB because the resulting correction is less than 0.1 dB.

$$\left(\frac{C}{N}\right) = 10\log_{10}\left(10^{\frac{\left(\frac{C+N}{N}\right)}{10}} - 1\right) \quad \text{dB}$$

(p) Calculate the carrier to noise power density ratio (C/No) using:

$$\left(\frac{C}{No}\right) = \left(\frac{C}{N}\right) - 2.5 + 10\log_{10}(RBW \times SA_{corr}) \quad \text{dB}$$

The 2.5 dB figure corrects the noise power value measured by the log converters in the spectrum analyser to a true RMS power level, and the SA corr

factor takes into account the actual resolution filter bandwidth. (q) Calculate the G/T using the following:

$$\left(\frac{G}{T}\right) = \left(\frac{C}{No}\right) - (EIRP_{SC} - A_{corr}) + (FSL + L_a) - 228.6 \quad dB/K$$

where,
EIRPSC – Downlink EIRP measured by the PMOC
(dBW) Acorr – Aspect correction supplied by the PMOC
(dB)
FSL – Free Space Loss to the AUT supplied by the PMOC (dB)
La – Atmospheric attenuation supplied by the PMOC (dB)

(r) Repeat the measurement several times to check consistency of the result.

1.7 Antenna Gain

Antenna gain is usually **defined** as the ratio of the power produced by the **antenna** from a far-field source on the **antenna's** beam axis to the power produced by a hypothetical lossless isotropic **antenna**, which is equally sensitive to signals from all directions.

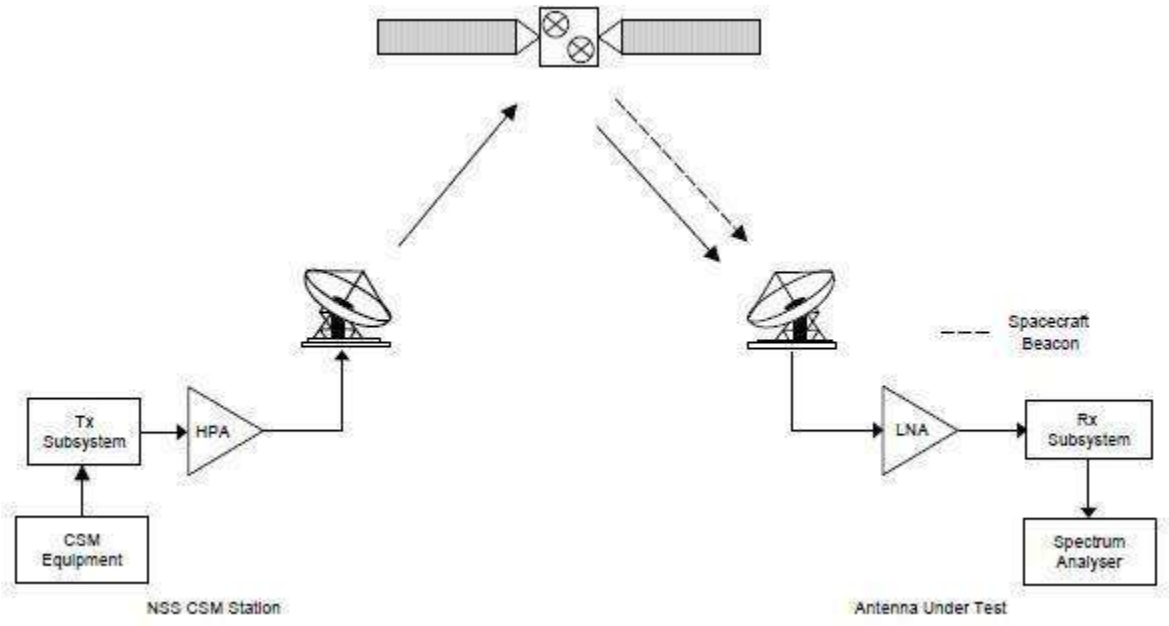


Figure 3.6 One possible arrangement for Measurement of Antenna Gain

Two direct methods of measuring the Rx gain can be used; integration of the Rx sidelobe pattern or by determination of the 3dB and 10dB beamwidths.

The use of pattern integration will produce the more accurate results but would require the AUT to have a tracking system. In both cases the test configurations for measuring Rx gain are identical, and are illustrated in Figure.

In order to measure the Rx gain using pattern integration the AUT measures the elevation and azimuth narrowband (±5° corrected) sidelobe patterns.

The AUT then calculates the directive gain of the antenna through integration of the sidelobe patterns. The Rx gain is then determine d by reducing the directive gain by the antenna inefficiencies.

In order to measure the Rx gain using the beamwidth method, the AUT measures the corrected azimuth and elevation 3dB/10dB beamwidths. From these results the Rx gain of the antenna can be directly calculated using the formula below.

$$G = 10\text{Log}_{10}\left[\frac{1}{2}\left(\frac{31000}{(Az_3)(El_3)} + \frac{91000}{(Az_{10})(El_{10})}\right)\right] - F_{Loss} - R_{Loss}$$

where:

G is the effective antenna gain (dBi)

Az3 is the corrected azimuth 3dB beamwidth

(°) El3 is the elevation 3dB beamwidth (°)

Az10 is the corrected azimuth 10dB beamwidth

(°) El10 is the elevation 10dB beamwidth (°)

FLoss is the insertion loss of the feed (dB)

RLoss is the reduction in antenna gain due to reflector inaccuracies, and is given by:

$$RLoss = 4.922998677(Sdev\ f)2\ dB$$

where: Sdev is the standard deviation of the actual reflector surface (inches)
f is the frequency (GHz)

APPLICATIONS

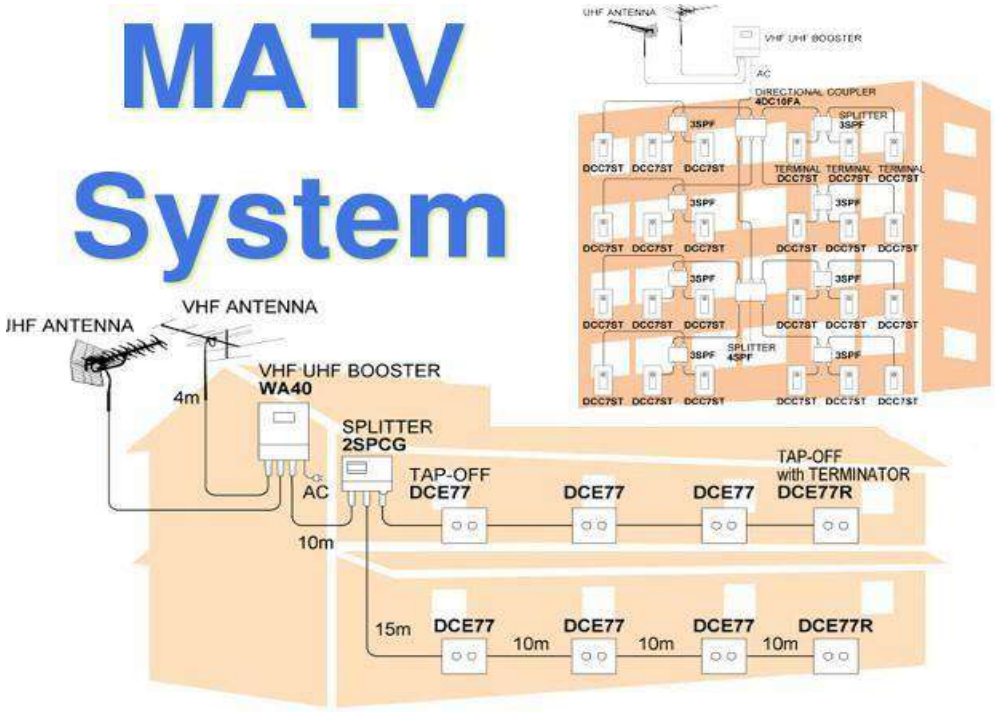


Figure an example of MATV system

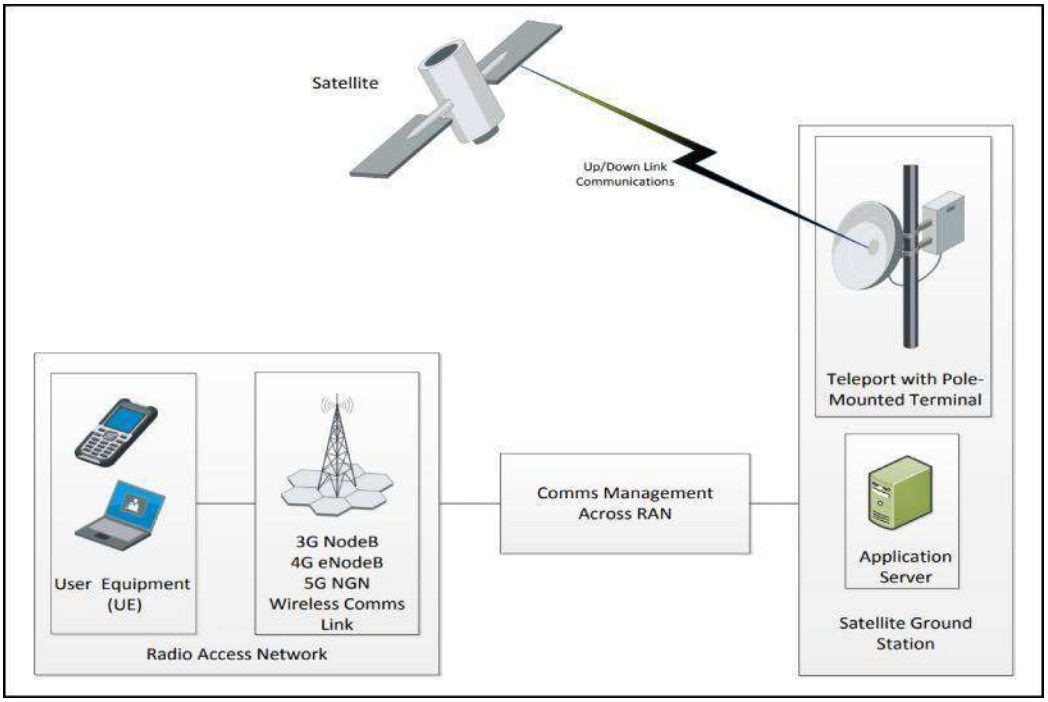


Figure an example of Satellite Earth Station

SATELLITE COMMUNICATION

UNIT IV SATELLITE ACCESS

4.1 Modulation and Multiplexing: Voice, Data, Video

Communications satellites are used to carry telephone, video, and data signals, and can use both analog and digital modulation techniques.

Modulation

Modification of a carrier's parameters (amplitude, frequency, phase, or a combination of them) in dependence on the symbol to be sent.

Multiplexing

Task of multiplexing is to assign space, time, frequency, and code to each communication channel with a minimum of interference and a maximum of medium utilization Communication channel refers to an association of sender(s) and receiver(s) that want to exchange data One of several constellations of a carrier's parameters defined by the used modulation scheme.

4.1.1 Voice, Data, Video

The modulation and multiplexing techniques that were used at this time were analog, adapted from the technology developed for The change to digital voice signals made it easier for long-distance.

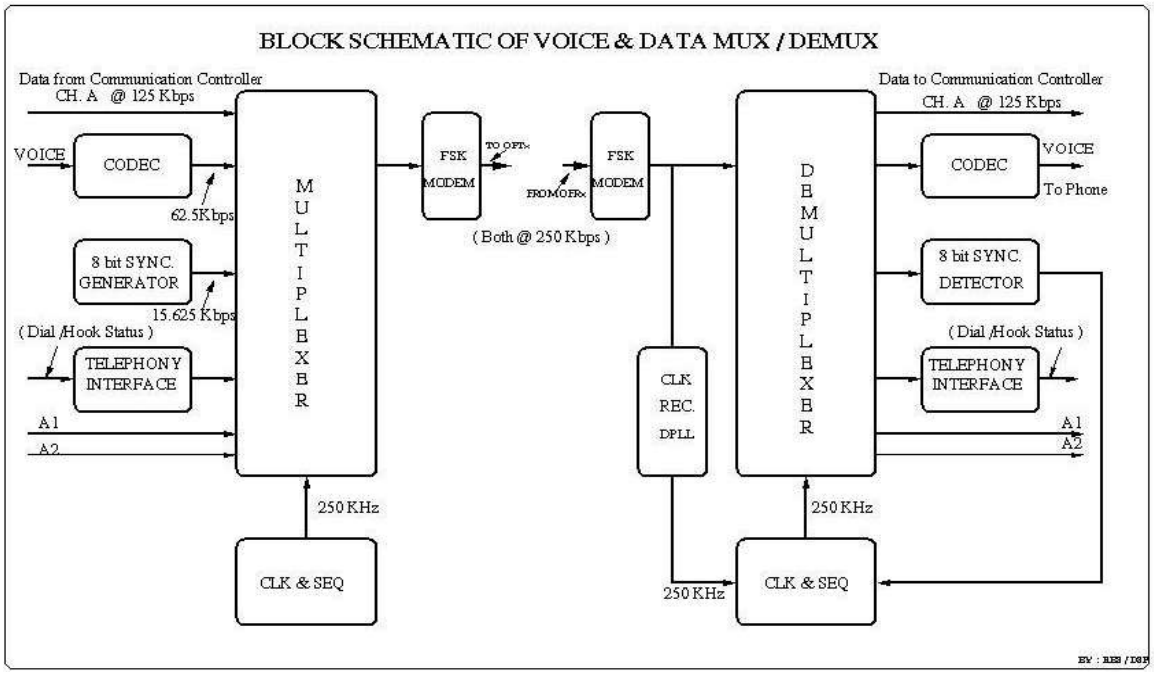


Figure 4.1 Modulation and Multiplexing: Voice/Data/Video

Communication carriers to mix digital data and telephone Fiber -optic Cable Transmission Standards System Bit rate (Mbps) 64 - kbps Voice channel capacity Stuffing bits and words are added to the satellite data stream as n eeded to fill empty bit and word spaces.

Primarily for video provided that a satellite link's overall carrier-to-noise but in to older receiving equipment at System and Satellite Specification Ku - band satellite parameters.

4.1.2 Modulation And Multiplexing

In analog television (TV) transmission by satellite, the baseband video signal and one or two audio subcarriers constitute a composite video signal.

Digital modulation is obviously the modulation of choice for transmitting digital data are digitized analog signals may conveniently share a channel with digital data, allowing a link to carry a varying mix of voice and data traffic.

Digital signals from different channels are interleaved for transmission through time division multiplexing TDM carry any type of traffic â€" the bent pipe transponder that can carry voice, video, or data as the marketplace demands.

Hybrid multiple access schemes can use time division multiplexing of baseband channels which are then modulate.

4.2 Analog – digital transmission system

4.2.1 Analog vs. Digital Transmission

Compare at two levels:

1. Data—continuous (audio) vs. discrete (text)
2. Signaling—continuously varying electromagnetic wave vs. sequence of voltage pulses.

Also Transmission—transmit without regard to signal content vs. being concerned with signal content. Difference in how attenuation is handled, but not focus on this. Seeing a shift towards digital transmission despite large analog base. Why?

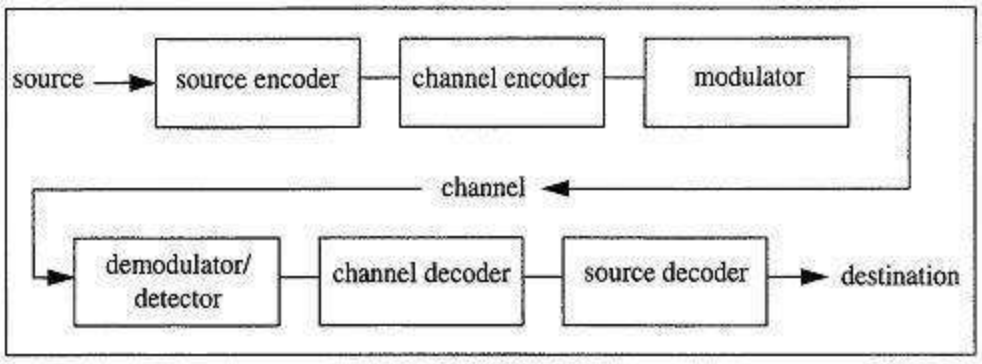


Figure 4.2 basic communication systems

- Improving digital technology
- Data integrity. Repeaters take out cumulative problems in transmission. Can thus transmit longer distances.
- Easier to multiplex large channel capacities with digital
- Easy to apply encryption to digital data
- Better integration if all signals are in one form. Can integrate voice, video and digital data.

4.2.2 Digital Data/Analog Signals

Must convert digital data to analog signal such device is a modem to translate between bit-serial and modulated carrier signals?

To send digital data using analog technology, the sender generates a carrier signal at some continuous tone (e.g. 1 -2 kHz in phone circuits) that looks like a sine wave. The following techniques are used to encode digital data into analog signals.

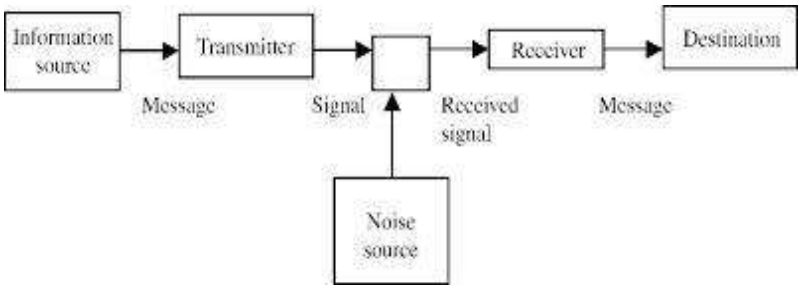


Figure 4.3 Digital /Analog Transmitter & receiver

Resulting bandwidth is centered on the carrier frequency.

- Amplitude-shift modulation (keying): vary the amplitude (e.g. voltage) of the signal. Used to transmit digital data over optical fiber.
- Frequency-shift modulation: two (or more tones) are used, which are near the carrier frequency. Used in a full-duplex modem (signals in both directions).
- Phase-shift modulation: systematically shift the carrier wave at uniformly spaced intervals.

For instance, the wave could be shifted by 45, 135, 225, 315 degree at each timing mark. In this case, each timing interval carries 2 bits of information.

Why not shift by 0, 90, 180, 270? Shifting zero degrees means no shift, and an extended set of no shifts leads to clock synchronization difficulties.

Frequency division multiplexing (FDM): Divide the frequency spectrum into smaller subchannels, giving each user exclusive use of a subchannel (e.g., radio and TV). One problem with FDM is that a user is given all of the frequency to use, and if the user has no data to send, bandwidth is wasted — it cannot be used by another user.

Time division multiplexing (TDM): Use time slicing to give each user the full bandwidth, but for only a fraction of a second at a time (analogous to time sharing in operating systems). Again, if the user doesn't have data to sent during his time slice, the bandwidth is not used (e.g., wasted).

Statistical multiplexing: Allocate bandwidth to arriving packets on demand. This leads to the most efficient use of channel bandwidth because it only carries useful data. That is, channel bandwidth is allocated to packets that are waiting for transmission, and a user generating no packets doesn't use any of the channel resources.

4.3. Digital Video Broadcasting (DVB)

- Digital Video Broadcasting (DVB) has become the synonym for digital television and for data broadcasting world-wide.

- DVB services have recently been introduced in Europe, in North- and
- South America, in Asia, Africa and Australia.

- This article aims at describing what DVB is all about and at introducing some of the technical background of a technology that makes possible the broadcasting.

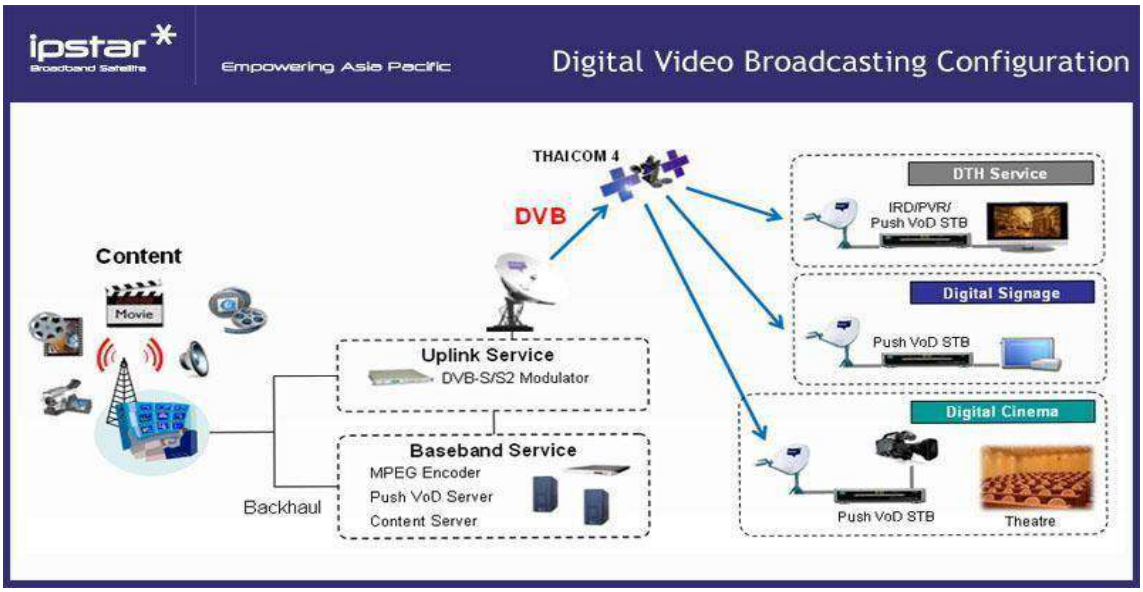


Figure 4.4 Digital Video Broadcasting systems

4.4 Multiple Access Techniques

- The transmission from the BS in the downlink can be heard by each and every mobile user in the cell, and is referred as *broadcasting*. Transmission from the mobile users in the uplink to the BS is many-to-one, and is referred to as multiple access.

- Multiple access schemes to allow many users to share simultaneously a finite amount of radio spectrum resources.

    - Should not result in severe degradation in the performance of the system as compared to a single user scenario.
    - Approaches can be broadly grouped into two categories: narrowband and wideband.

- Multiple Accessing Techniques : with possible conflict and conflict- free

- Random access
- Frequency division multiple access (FDMA)
- Time division multiple access (TDMA)
- Spread spectrum multiple access (SSMA) : an example is Code division multiple access (CDMA)
- Space division multiple access (SDMA)

Duplexing

For voice or data communications, must assure two way communication (duplexing, it is possible to talk and listen simultaneously). Duplexing may be done using frequency or time domain techniques.

- ❖ Forward (downlink) band provides traffic from the BS to the mobile
- ❖ Reverse (uplink) band provides traffic from the mobile to the BS.

4.4.1 Frequency division duplexing (FDD)

- □ Provides two distinct bands of frequencies for every user, one for downlink and one for uplink.

- □ A large interval between these frequency bands must be allowed so that interference is minimized.

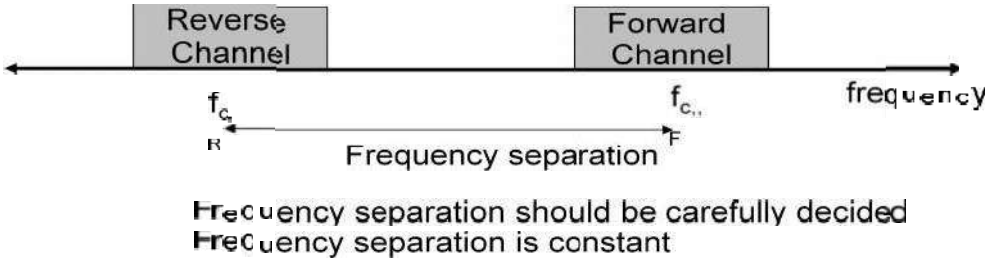


Figure 4.5 Frequency Separation

4.4.2. Time division duplexing (TDD)

- ❖ In TDD communications, both directions of transmission use one contiguous frequency allocation, but two separate time slots to provide both a forward and reverse link.

- Because transmission from mobile to BS and from BS to mobile alternates in time, this scheme is also known as "ping pong".
- As a consequence of the use of the same frequency band, the communication quality in both directions is the same. This is different from FDD.

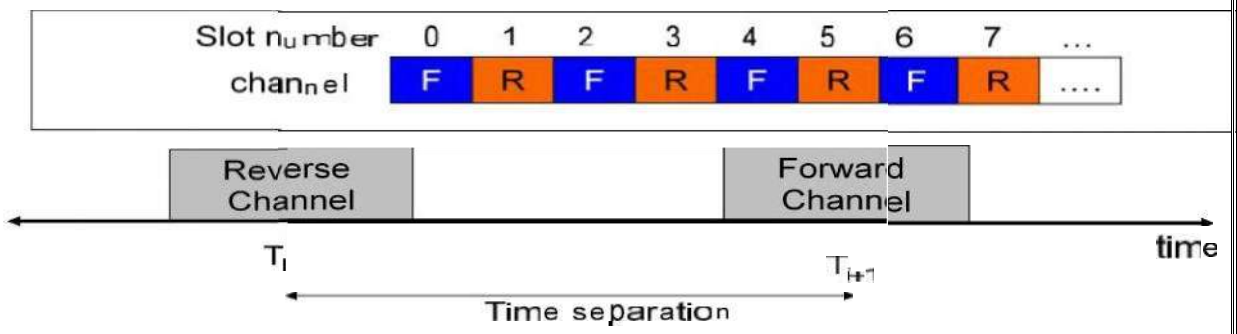


Figure 4.6 Time Slot

4.4.3 FDMA

- In FDMA, each user is allocated a unique frequency band or channel. During the period of the call, no other user can share the same frequency band.

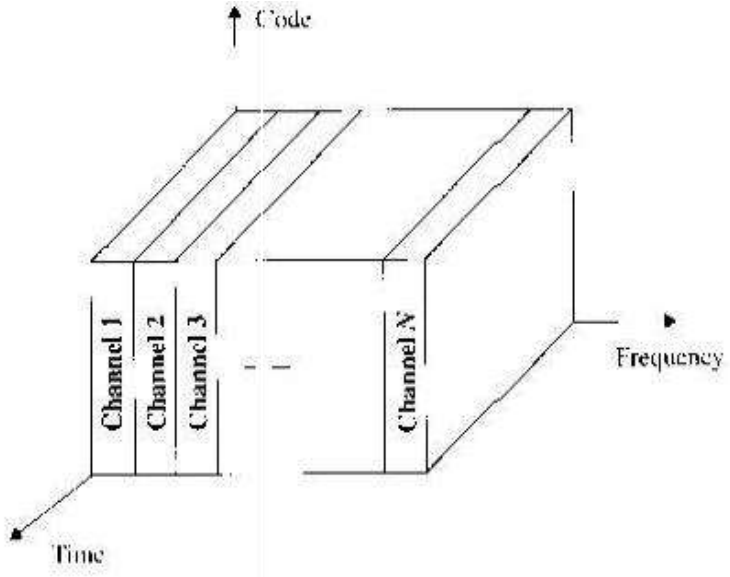


Figure 4.7 FDMA Channels

- All channels in a cell are available to all the mobiles. Channel assignment is carried out on a first-come first- served basis.

- The number of channels, given a frequency spectrum BT , depends on the modulation technique (hence Bw or Bc) and the guard bands between the channels $2Bguard$.

- These guard bands allow for imperfect filters and oscillators and can be used to minimize adjacent channel interference.
- FDMA is usually implemented in narrowband systems.

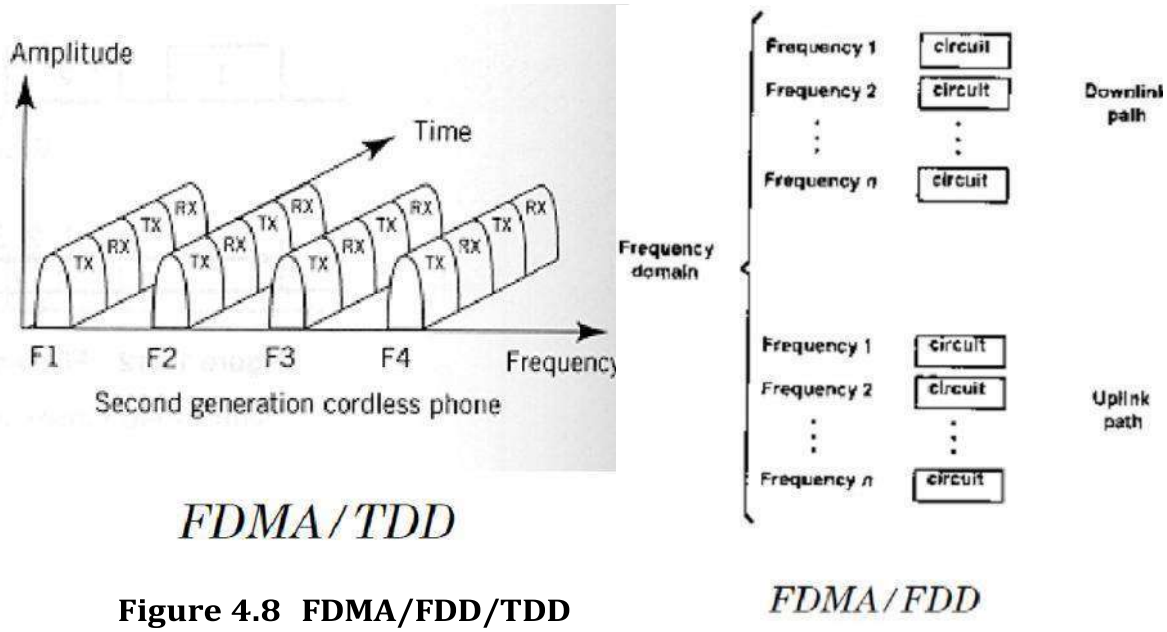


Figure 4.8 FDMA/FDD/TDD

Nonlinear effects in FDMA

- In a FD MA system, many channels share t he same antenna at the BS. The power amplifiers or the power combiners, when operated at or near saturation are non linear.
- The nonlinear ties generate inter-modulation frequencies.
- Undesirable harmonics generated outside the mobile radio band cause interference to adjacent services.
- Undesirable harmonics present inside the band ca use interference to other users in the mobile system.

4.4.4 TDMA

- TDMA systems divide the channel time into frames. Each frame is further partitioned into time slots. In each slot only one user is allowed to either transmit or receive.
- Unlike FDMA, only digital data and digital modulation must be used.
- Each user occupies a cyclically repeating time slot, so a channel may be thought of as a particular time slot of every frame, where *N* time slots comprise a frame.

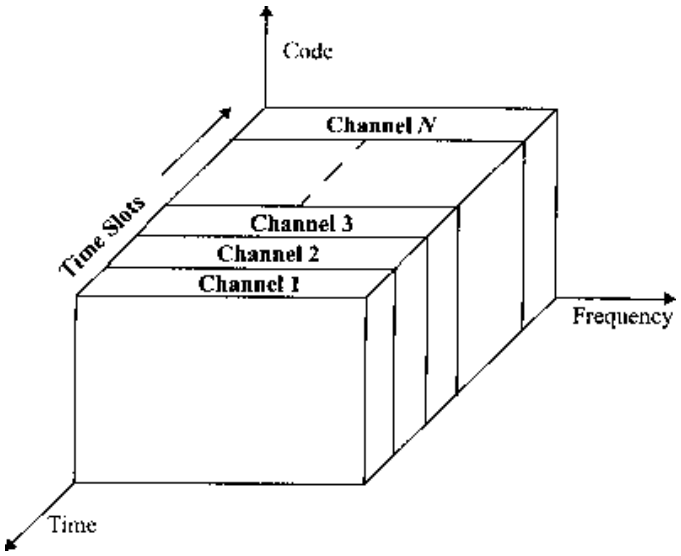


Figure 4.9 TDMA Channels

Features

- ❖ Multiple channels per carrier or RF channels.
- ❖ Burst transmission since channels are used on a timesharing basis.
  - Transmitter can be turned off during idle periods.
- ❖ Narrow or wide bandwidth – depends on factors such as modulation scheme, number of voice channels per carrier channel.
- ❖ High ISI – Higher transmission symbol rate, hence resulting in high ISI.
  - Adaptive equalizer required.

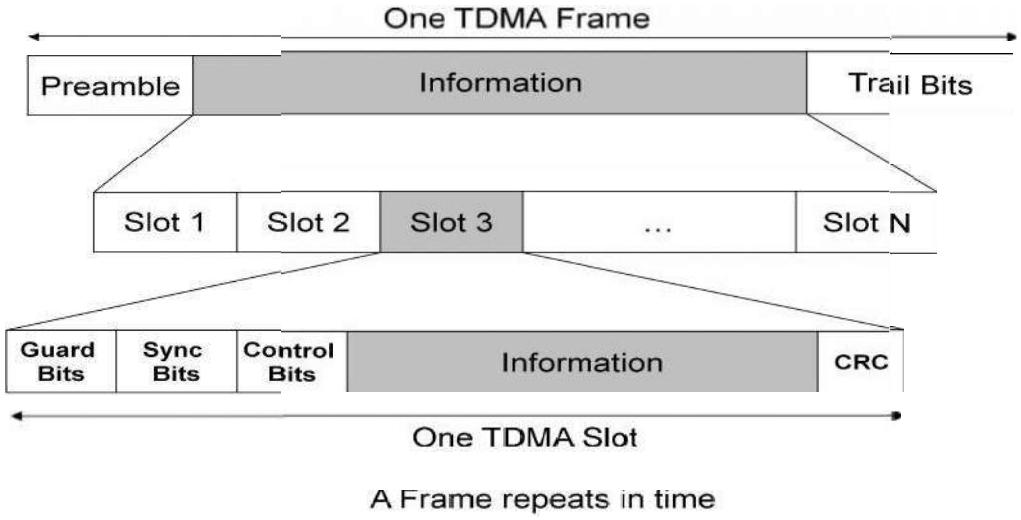


Figure 3.10 TDMA Channels time slot

- A guard time between the two time slots must be allowed in order to avoid
  - interference, especially in the uplink direction. All synchronize with BS to minimize interference.
- mobiles should

- Efficient power utilization : FDMA systems require a 3- to 6-dB power back
  - off in order to compensate for inter-modulation effects.

- Efficient handoff : TDMA systems can take advantage of the fact that the transmitter is switched off during idle time slots to improve the handoff procedure. An enhanced link control, such as that provided by mobile assisted handoff (MAHO) can be carried out by a subscriber by listening to
  - neighboring base station during the idle slot of the TDMA frame.

- Efficiency of TDMA

- Efficiency of TDMA is a measure of the percentage of bits per frame whi ch contain transmitted data. The transmitted data include source and channel coding bits.

$$\eta_f = \frac{b_T - b_{OH}}{b_T} \cdot 100\%$$

- bOH includes all overhead bits such as preamble, guard bits, etc.

4.4.5 Code Division Multiple Access (CDMA)

- Spreading signal (code) consists of chips
    - Has Chip period and and hence, chip rate
    - Spreading signal use a pseudo-noise (PN) sequence (a pseudo-random sequence)
    - PN sequence is called a codeword
    - Each user has its own cordword
    - Codewords are orthogonal. (low autocorrelation)
    - Chip rate is order of magnitude larger than the symbol rate.
- The receiver correlator distinguishes the senders signal by examining the wideband signal with the same time-synchronized spreading code
- The sent signal is recovered by despreading process at the receiver.

CDMA Advantages:

- Low power spectral density.
    - Signal is spread over a larger frequency band
    - Other systems suffer less from the transmitter
- Interference limited operation
    - All frequency spectrum is used
- Privacy
    - The codeword is known only between the sender and receiver. Hence other users can not decode the messages that are in transit
- Reduction of multipath affects by using a larger spectrum

CDMA data

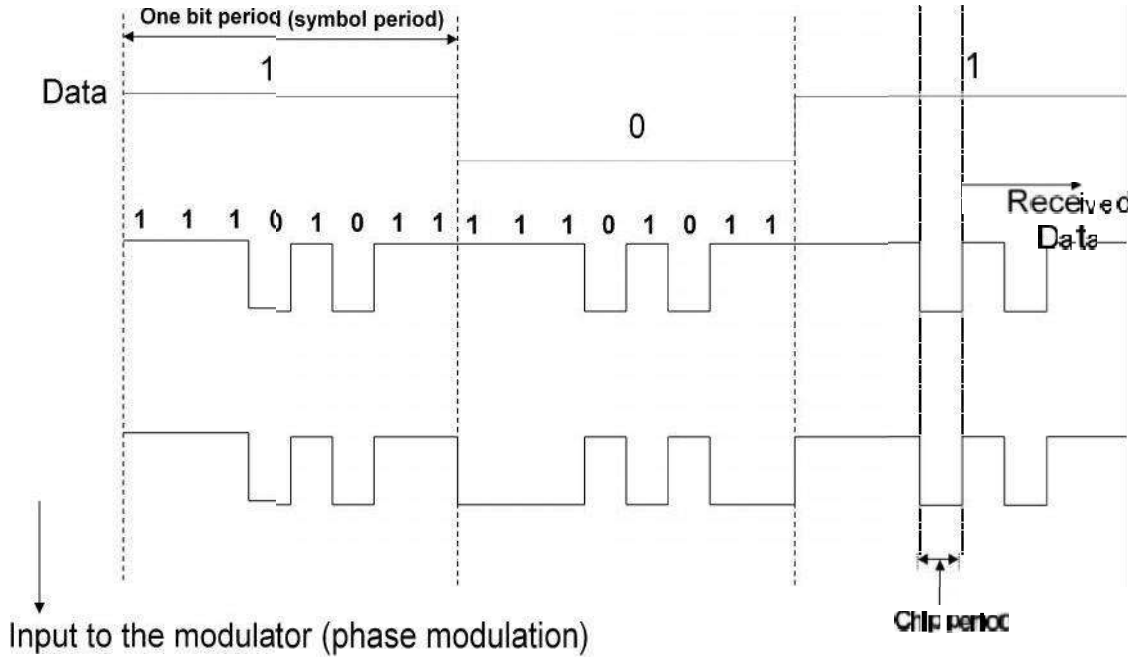


Figure 4.11 CDMA Channels transmission

DSSS Transmitter:

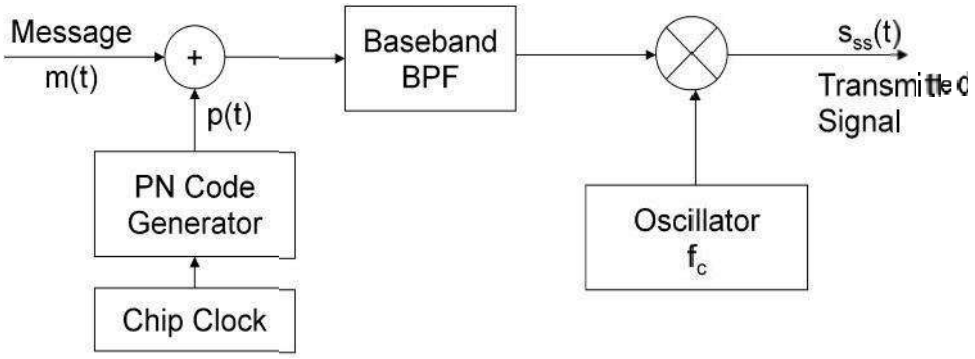
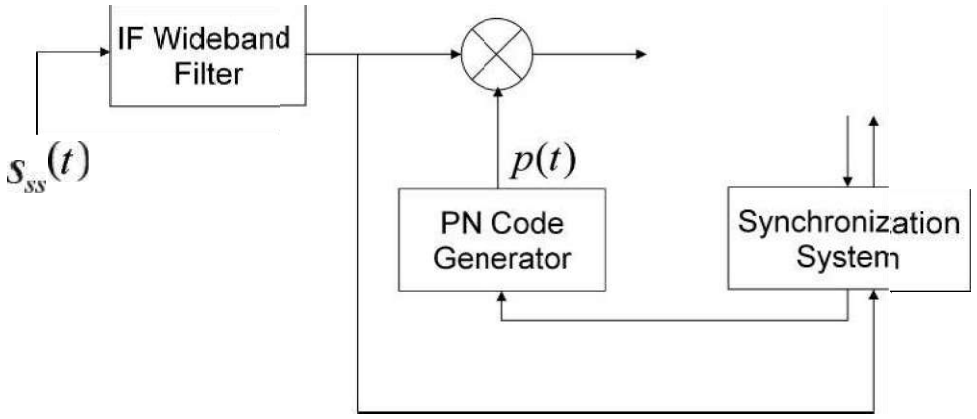


Figure 4.12 CDMA Transmitter

$$s_{ss}(t) = \sqrt{\frac{2E_s}{T_s}} m(t) p(t) \cos(2\pi f_c t + \theta)$$

DSSS Receiver



$$s_1(t) = \sqrt{\frac{2E_s}{T_s}} m(t) \cos(2\pi f_c t + \theta)$$

Figure 4.13 CDMA Receiver

- FDMA/CDMA
  - Available wideband spectrum is frequency divided into number narrowband radio channels. CDMA is employed inside each channel.
- DS/FHMA
  - The signals are spread using spreading codes (direct sequence signals
  - are obtained), but these signal are not transmitted over a constant
  - carrier frequency; they are transmitted over a frequency hopping carrier frequency.
- Time Division CDMA (TCDMA)
  - Each cell is using a different spreading code (CDMA employed between cells) that is conveyed to the mobiles in its range.
  - Inside each cell (inside a CDMA channel), TDMA is employed to multiplex multiple users.

- Time Division Frequency Hopping
  - At each time slot, the user is hopped to a new frequency according to a pseudo-random hopping sequence.
  - Employed in severe co-interference and multi-path environments.

Bluetooth and GSM are using this technique

- A large number of independently steered high-gain beams can be formed without any resulting degradation in SNR ratio.
- Beams can be assigned to individual users, thereby assuring that all links operate with maximum gain.
- Adaptive beam forming can be easily implemented to improve the system capacity by suppressing co channel interference.

Advantage of CDMA

- It is recognized that CDMA's capacity gains over TDMA
- FDMA are entirely due to Its tighter, dynamic control over the use of the power domain.
- Choosing a new non-orthogonal PN sequence a CDMA system does not encounter the difficulties of choosing a spare carrier frequency or time slot to carry a Traffic Channel
- Ensure that interference will not be too great if it begins to transmit -that there is still enough space left in the power domain.

Disadvantages of CDMA

- Satellite transponders are channelized too narrowly for roadband CDMA, which is the most attractive form of CDMA.
- Power control cannot be as tight as it is in a terrestrial system because of long round-trip delay.

4.5. Channel allocation schemes

In radio resource management for wireless and cellular network, channel allocation schemes are required to allocate bandwidth and communication channels to base stations, access points and terminal equipment.

The objective is to achieve maximum system spectral efficiency in bit/s/Hz/site by means of frequency reuse, but still assure a certain grade of service by avoiding co-channel interference and adjacent channel interference among nearby cells or networks that share the bandwidth. There are two types of strategies that are followed:-

- Fixed: FCA, fixed channel allocation: Manually assigned by the network operator
- Dynamic:
  - DCA, dynamic channel allocation,
  - DFS, dynamic frequency selection
  - Spread spectrum

4.5.1 FCA

In **Fixed Channel Allocation** or **Fixed Channel Assignment** (FCA) each cell is given a predetermined set of frequency channels.

FCA requires manual frequency planning, which is an arduous task in TDMA and FDMA based systems, since such systems are highly sensitive to co-channel interference from nearby cells that are reusing the same channel.

This results in traffic congestion and some calls being lost when traffic gets heavy in some cells, and idle capacity in other cells.

4.5.2. DCA and DFS

Dynamic Frequency Selection (DFS) may be applied in wireless networks with several adjacent non-centrally controlled access points.

A more efficient way of channel allocation would be **Dynamic Channel Allocation** or **Dynamic Channel Assignment** (DCA) in which voice channel are not allocated to cell permanently, instead for every call request base station request channel from MSC.

4.6 Spread spectrum

Spread spectrum can be considered as an alternative to complex DCA algorithms. Spread spectrum avoids cochannel interference between adjacent

cells, since the probability that users in nearby cells use the same spreading code is insignificant.

Thus the frequency channel allocation problem is relaxed in cellular networks based on a combination of Spread spectrum and FDMA, for example IS95 and 3G systems.

In packet based data communication services, the communication is bursty and the traffic load rapidly changing. For high system spectrum efficiency, DCA should be performed on a packet-by-packet basis.

Examples of algorithms for packet-by-packet DCA are **Dynamic Packet Assignment** (DPA), Dynamic Single Frequency Networks (DSFN) and **Packet and resource plan scheduling** (PARPS).

4.6.1 Spread spectrum Techniques

1 In telecommunication and radio communication, spread-spectrum techniques are methods by which a signal (e.g. an electrical, electromagnetic, or acoustic signal) generated with a particular bandwidth is deliberately spread in the frequency domain, resulting in a signal with a wider bandwidth.

2 These techniques are used for a variety of reasons, including the establishment of secure communications, increasing resistance to natural interference, noise and jamming, to prevent detection, and to limit power flux density (e.g. in satellite downlinks).

3 Spread-spectrum telecommunications this is a technique in which a telecommunication signal is transmitted on a bandwidth considerably larger than the frequency content of the original information.

4 Spread-spectrum telecommunications is a signal structuring technique that employs direct sequence, frequency hopping, or a hybrid of these, which can be used for multiple access and/or multiple functions.

5 Frequency-hopping spread spectrum (FHSS), direct-sequence spread spectrum (DSSS), time-hopping spread spectrum (THSS), chirp spread spectrum (CSS).

6 Techniques known since the 1940s and used in military communication systems since the 1950s "spread" a radio signal over a wide frequency range several magnitudes higher than minimum requirement.

7 Resistance to jamming (interference). DS (direct sequence) is good at resisting continuous-time narrowband jamming, while FH (frequency hopping) is better at resisting pulse jamming.

8 Resistance to fading. The high bandwidth occupied by spread- spectrum signals offer some frequency diversity, i.e. it is unlikely that the

signal will encounter severe multipath fading over its whole bandwidth, and in other cases the signal can be detected using e.g. a Rake receiver.

9 Multiple access capability, known as code-division multiple access (CDMA) or code-division multiplexing (CDM). Multiple users can transmit simultaneously in the same frequency band as long as they use different spreading codes.

4.7 Compression – Encryption

At the broadcast center, the high-quality digital stream of video goes through an MPEG encoder, which converts the programming to MPEG-4 video of the correct size and format for the satellite receiver in your house.

Encoding works in conjunction with compression to analyze each vi deo frame and eliminate redundant or irrelevant data and extrapolate information from other frames. This process reduces the overall size of the file. Each frame can be encoded in one of three ways:

- As an **intraframe**, which contains the complete image data for that frame. This method provides the least ompression.
- As a **predicted** frame, which contains just enough information to tell the satellite receiver how to display the frame based on the most recently displayed intraframe or predicted frame.
- As a **bidirectional** frame, which displays information from the surrounding intraframe or predicted frames. Using data from the closest surrounding frames, the receiver **interpolates** the position and color of each pixel.

This process occasionally produces **artifacts --** glitches in the video image. One artifact is **macroblocking,** in which the fluid picture temporarily dissolves into blocks. Macroblocking is often mistakenly called **pixilating,** a technically incorrect term which has been accepted as slang for this annoying artifact.

There really are pixels on your TV screen, but they're too small for your human eye to perceive them individually -- they're tiny squares of video data that make up the image you see. (For more information about pixels and perception, see How TV Works.)

The rate of compression depends on the nature of the programming. If the encoder is converting a newscast, it can use a lot more predicted frames because most of the scene stays the same from one frame to the next.

In more fast-paced programming, things change very quickly from one frame to the next, so the encoder has to create more intraframes. As a result, a newscast generally compresses to a smaller size than something like a car race.

4.7.1 Encryption and Transmission

After the video is compressed, the provider encrypts it to keep people from accessing it for free. Encryption scrambles the digital data in such a way that it can only be **decrypted** (converted back into usable data) if the receiver has the correct decryption algorithm and security keys.

Once the signal is compressed and encrypted, the broadcast center beams it directly to one of its satellites. The satellite picks up the signal with an onboard dish, amplifies the signal and uses another dish to beam the signal back to Earth, where viewers can pick it up.

In the next section, we'll see what happens when the signal reaches a viewer's house.

4.7.2 Video and Audio Compression

Video and Audio files are very large beasts. Unless we develop and maintain very high bandwidth networks (Gigabytes per second or more) we have to compress to data.

Relying on higher bandwidths is not a good option -- M25 Syndrome: Traffic needs ever increases and will adapt to swamp current limit whatever this is.

As we will compression becomes part of the representation or *coding* scheme which have become popular audio, image and video formats.

We will first study basic compression algorithms and then go on to study some actual coding formats.

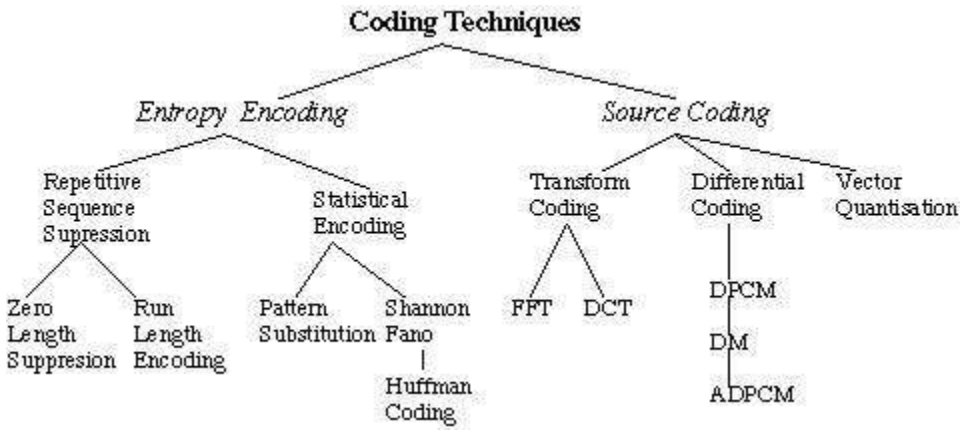


Figure 4.14 *coding* scheme

What is Compression?

Compression basically employs redundancy in the data:

- Temporal -- in 1D data, 1D signals, Audio etc.
- Spatial -- correlation between neighbouring pixels or data items
- Spectral -- correlation between colour or luminescence components. This uses the frequency domain to exploit relationships between frequency of change in data.
- psycho-visual -- exploit perceptual properties of the human visual system.

Compression can be categorized in two broad ways:

Lossless Compression

-- where data is compressed and can be reconstituted (uncompressed) without loss of detail or information. These are referred to as bit-preserving or reversible compression systems also.

Lossy Compression

-- where the aim is to obtain the best possible *fidelity* for a given bit-rate or minimizing the bit-rate to achieve a given fidelity measure. Video and audio compression techniques are most suited to this form of compression.

If an image is compressed it clearly needs to uncompressed (decoded) before it can viewed/listened to. Some processing of data may be possible in encoded form however. Lossless compression frequently involves some form of *entropy encoding* and are based in information theoretic techniques.

Lossy compression use source encoding techniques that may involve transform encoding, differential encoding or vector quantization.

4.7.3 MPEG Standards

All MPEG standards exist to promote system interoperability among your computer, television and handheld video and audio devices. They are:

- **MPEG-1:** the original standard for encoding and decoding streaming video and audio files.
- **MPEG-2:** the standard for digital television, this compresses files for transmission of high-quality video.
- **MPEG-4:** the standard for compressing high-definition video into smaller-scale files that stream to computers, cell phones and PDAs (personal digital assistants).
- **MPEG-21:** also referred to as the Multimedia Framework. The standard that interprets what digital content to provide to which individual user so that media plays flawlessly under any language, machine or user conditions.

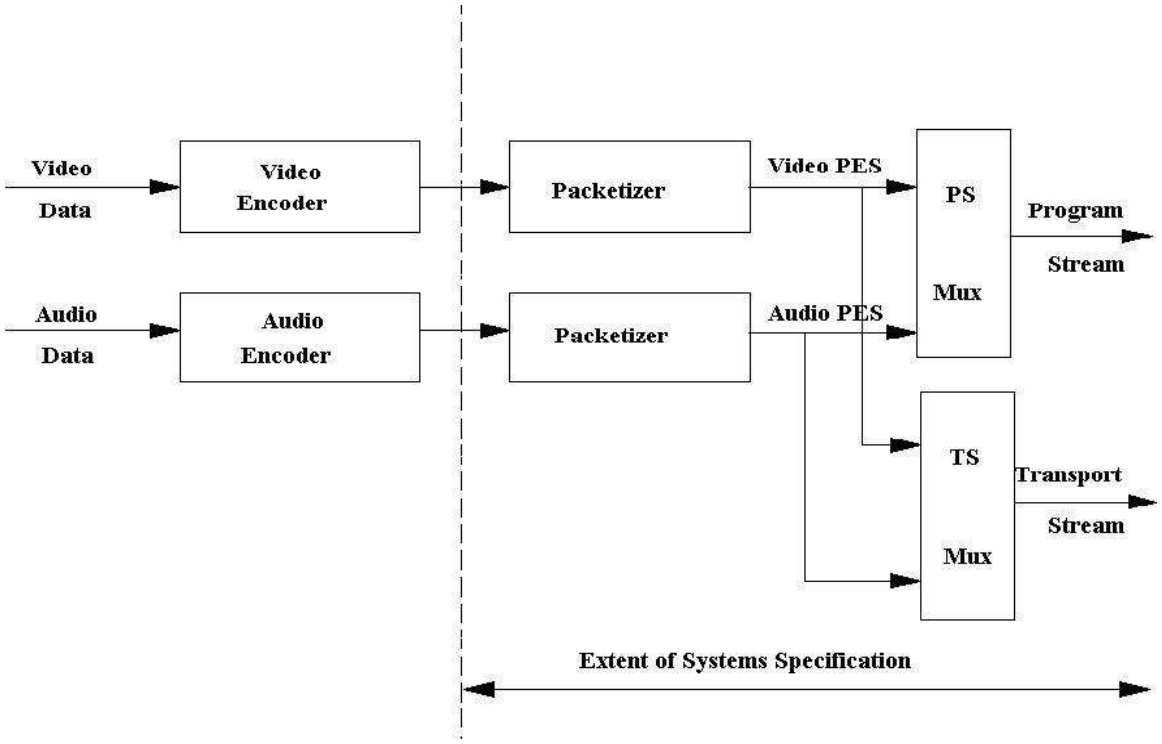


Figure 4.15 MPEG scheme

4.8 Encryption

It is the most effective way to achieve data security. To read an **encrypted** file, you must have access to a secret key or password that enables you to decrypt it. Unencrypted data is called **plain text** ; **encrypted** data is referred to as **cipher text.**

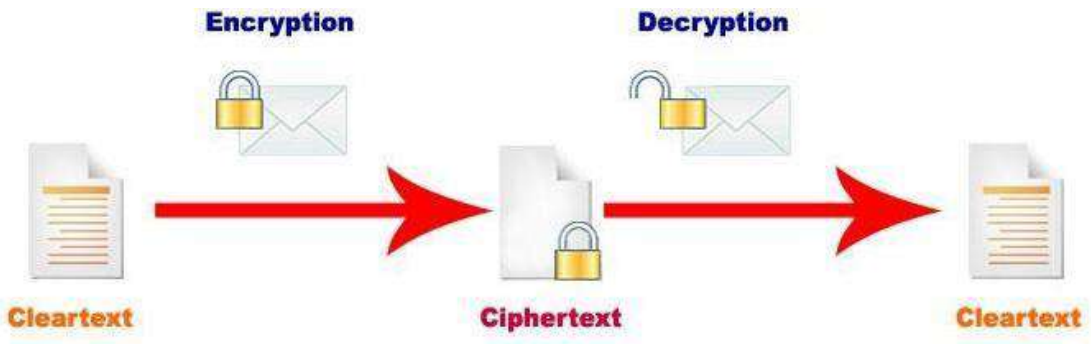


Figure 4.16 Encryption methods

4.8.1 Symmetric key encryption

In symmetric-key schemes, the encryption and decryption keys are the same. Thus communicating parties must have the same key before they can achieve secret communication.

In public-key encryption schemes, the encryption key is published for anyone to use and encrypt messages. However, only the receiving party has access to the decryption key that enables messages to be read.

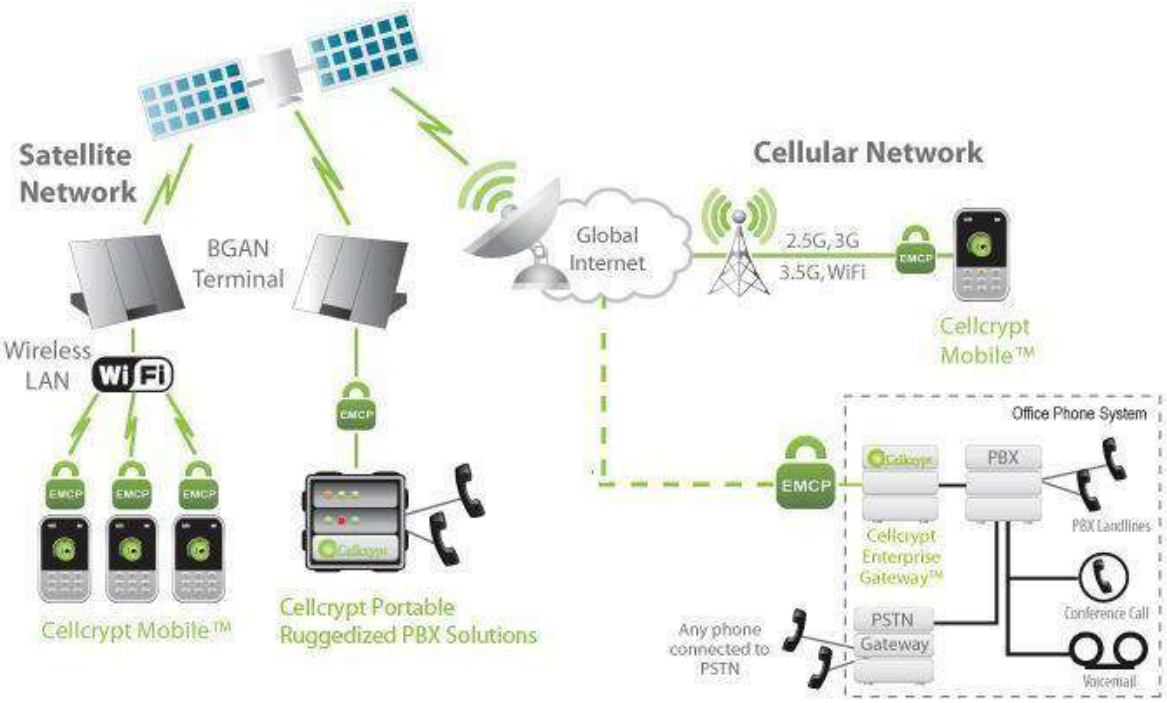


Figure 4.17 General block diagram Encryption methods

4.8.2 Decryption

It is the process of taking encoded or encrypted text or other data and converting it back into text that you or the computer are able to read and understand.

This term could be used to describe a method of un-encrypting the data manually or with un-encrypting the data using the proper codes or keys.

Data may be encrypted to make it difficult for someone to steal the information. Some companies also encrypt data for general protection of company data and trade secrets. If this data needs to be viewable, it may require decryption.

APPLICATIONS

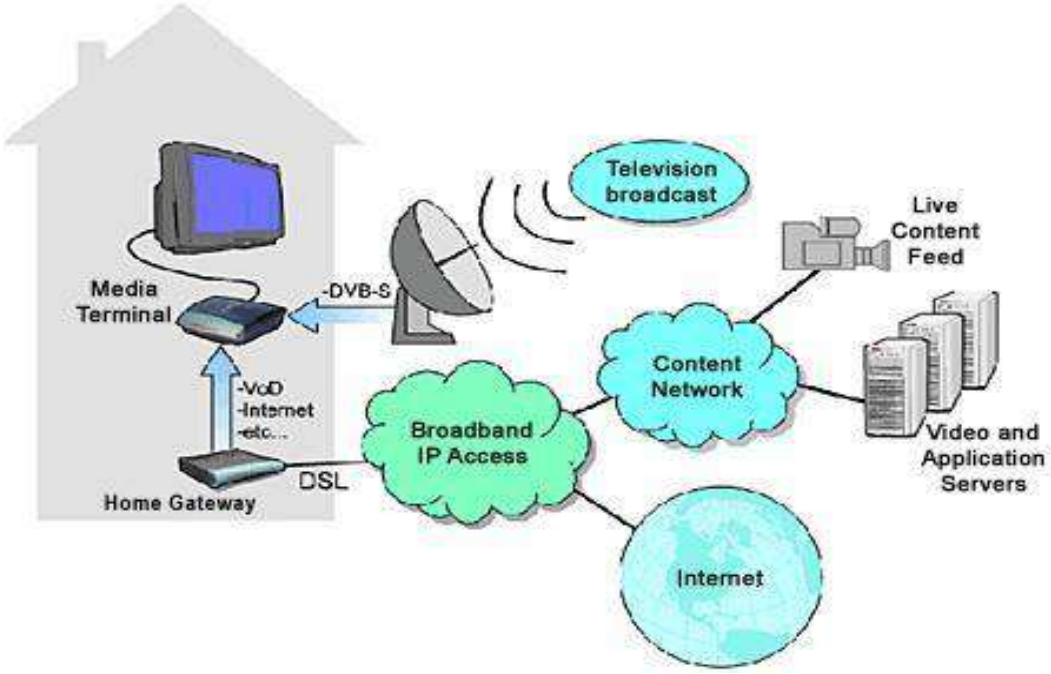


Figure an example of Digital video Broadcast

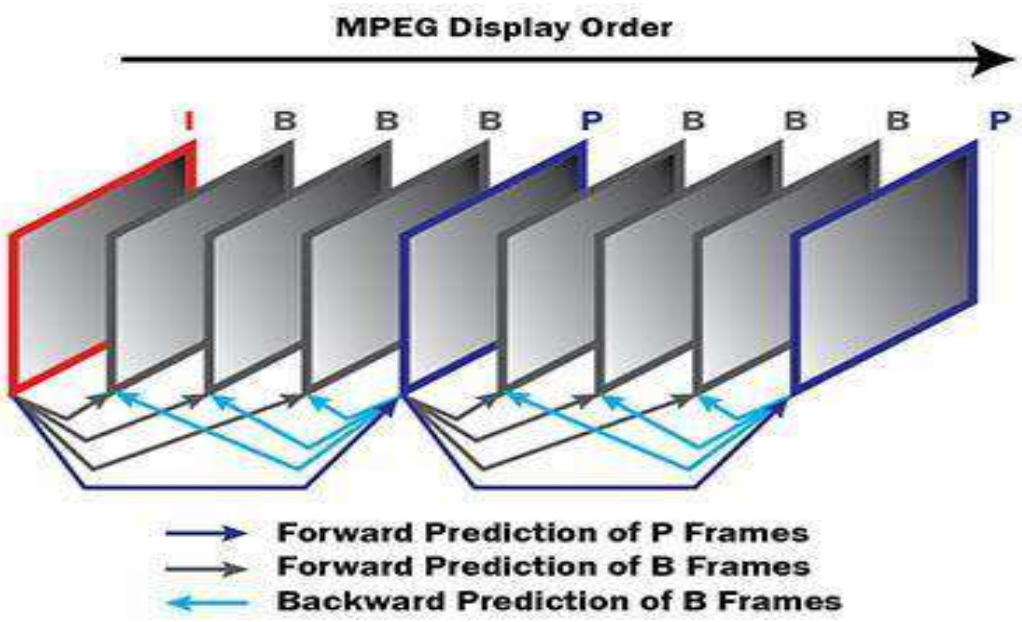


Figure an example of streaming video and compression

SATELLITE COMMUNICATION

UNIT V SATELLITE APPLICATIONS

THEORY

5.1 INTELSAT Series

INTELSAT stands for *International Telecommunications Satellite.* The organization was created in 1964 and currently has over 140 member countries and more than 40 investing entities (see http://www.intelsat.com/ for more details).

In July 2001 INTELSAT became a private company and in May 2002 the company began providing end-to-end solutions through a network of teleports, leased fiber, and *points of presence* (PoPs) around the globe.

Starting with the Early Bird satellite in 1965, a succes- sion of satellites has been launched at intervals of a few years. Figure 1.1 illustrates the evolution of some of the INTELSAT satellites. As the figure shows, the capacity, in terms of number of voice channels, increased dramatically with each succeeding launch, as well as the design lifetime.

These satellites are in *geostationary orbit,* meaning that they appear to be stationary in relation to the earth. At this point it may be noted that geosta- tionary satellites orbit in the earth's equatorial plane and their position is specified by their longitude.

For international traffic, INTELSAT covers three main regions—the *Atlantic Ocean Region* (AOR), the *Indian Ocean Region* (IOR), and the *Pacific Ocean Region* (POR) and what is termed *Intelsat America's Region*.

For the ocean regions the satellites are positioned in geostationary orbit above the particular ocean, where they provide a transoceanic telecommunications route. For example, INTELSAT satellite 905 is positioned at 335.5° east longitude.

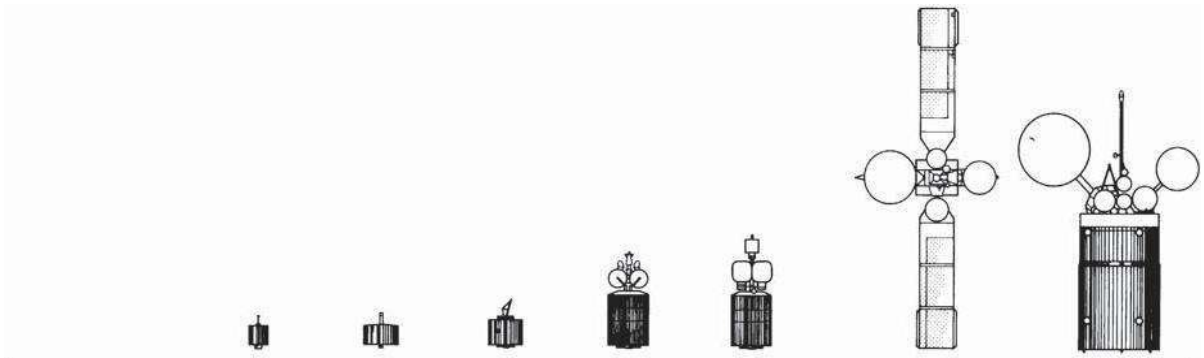
The INTELSAT VII-VII/A series was launched over a period from October 1993 to June 1996. The construction is similar to that for the V and VA/VB series, shown in Fig. in that the VII series has solar sails rather than a cylindrical body.

The VII series was planned for service in the POR and also for some of the less demanding services in the AOR. The antenna beam coverage is appropriate for that of the POR. Figure 1.3 shows the antenna beam footprints for the C -band hemispheric cover- age and zone coverage, as well as the spot beam coverage possible with the Ku-band antennas (Lilly, 1990; Sachdev et al., 1990).

When used in the AOR, the VII series satellite is inverted north for south (Lilly, 1990), minor adjustments then being needed only to optimize the antenna pat- terns for this region. The lifetime of these satellites ranges from 10 to 15 years depending on the launch vehicle.

Recent figures from the INTELSAT Web site give the capacity for the INTELSAT VII as 18,000 two-way telephone circuits and three TV channels; up to 90,000 two-way telephone circuits can be achieved with the use of "digital circuit multiplication."

The INTELSAT VII/A has a capacity of 22,500 two-way telephone circuits and three TV channels; up to 112,500 two-way telephone circuits can be achieved with the use of digital circuit multiplication. As of May 1999, four satellites were in service over the AOR, one in the IOR, and two in the POR.



Designation: Intelsat	I	II	III	IV	IV A	V	V A/V B	VI
Year of first launch	1965	1966	1968	1971	1975	1980	1984/85	1986/87
Prime contractor	Hughes	Hughes	TRW	Hughes	Hughes	Ford Aerospace	Ford Aerospace	Hughes
Width (m)	0.7	1.4	1.4	2.4	2.4	2.0	2.0	3.6
Height (m)	0.6	0.7	1.0	5.3	6.8	6.4	6.4	6.4
Launch vehicles		Thor Delta		Atlas-Centaur		Atlas-Centaur and Ariane	Atlas-Centaur and Ariane	STS and Ariane
Spacecraft mass in transfer orbit (kg)	68	182	293	1385	1489	1946	2140	12,100/3720
Communications payload mass (kg)	13	36	56	185	190	235	280	800
End-of-life (EOL) power of equinox (W)	40	75	134	480	800	1270	1270	2200
Design lifetime (years)	1.5	3	5	7	7	7	7	10
Capacity (number of voice channels)	480	480	2400	8000	12,000	25,000	30,000	80,000
Bandwidth (MHz)	50	130	300	500	800	2137	2480	3520

Figure 5.1 INTELSAT Series

The INTELSAT VIII-VII/A series of satellites was launched over the period February 1997 to June 1998. Satellites in this series have similar capacity as the VII/A series, and the lifetime is 14 to 17 years.

It is standard practice to have a spare satellite in orbit on high-reliability routes (which can carry preemptible traffic) and to have a ground spare in case of launch failure.

Thus the cost for large international schemes can be high; for example, series IX, described later, represents a total investment of approximately $1 billion.

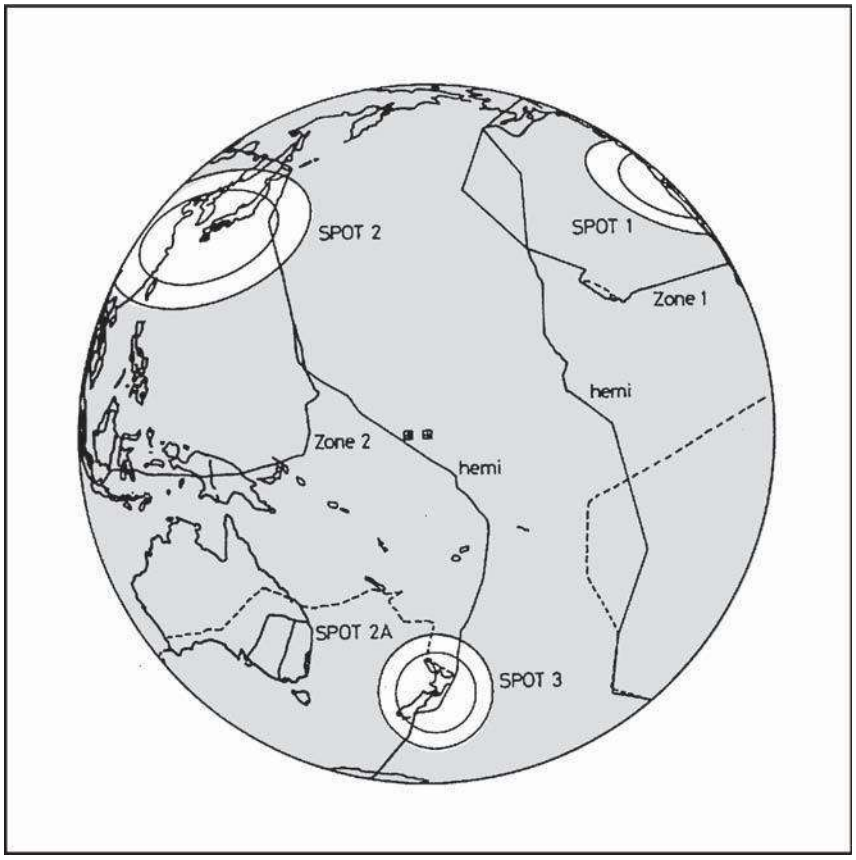


Figure 5.2 Region of glob

5.2 INSAT

INSAT or the ***Indian National Satellite System*** is a series of multipurpose geo-stationary satellites launched by ISRO to satisfy the telecommunications, broadcasting, meteorology, and search and rescue operations.

Commissioned in 1983, INSAT is the largest domestic communication system in the Asia Pacific Region. It is a joint venture of the Department of Space, Department of Telecommunications, India Meteorological Department,

All India Radio and Doordarshan. The overall coordination and management of INSAT system rests with the Secretary-level INSAT Coordination Committee.

INSAT satellites provide transponders in various bands (C, S, Extended C and K_u) to serve the television and communication needs of India. Some of the satellites also have the Very High Resolution Radiometer (VHRR), CCD cameras for metrological imaging.

The satellites also incorporate transponder(s) for receiving distress alert signals for search and rescue missions in the South Asian and Indian Ocean Region, as ISRO is a member of the Cospas-Sarsat programme.

5.2.1 INSAT System

The Indian National Satellite (INSAT) System Was Commissioned With The Launch Of INSAT-1B In August 1983 (INSAT-1A, The First Satellite Was Launched In April 1982 But Could Not Fulfil The Mission).

INSAT System Ushered In A Revolution In India's Television And Radio Broadcasting, Telecommunications And Meteorological Sectors. It Enabled The Rapid Expansion Of TV And Modern Telecommunication Facilities To Even The Remote Areas And Off-Shore Islands.

5.2.2 Satellites In Service

Of The 24 Satellites Launched In The Course Of The INSAT Program, 10 Are Still In Operation.INSAT-2E

It Is The Last Of The Five Satellites In INSAT-2 Series{Prateek }. It Carries Seventeen C-Band And Lower Extended C-Band Transponders Providing Zonal And Global Coverage With An Effective Isotropic Radiated Power (EIRP) Of 36 Dbw.

It Also Carries A Very High Resolution Radiometer (VHRR) With Imaging Capacity In The Visible (0.55-0.75 μm), Thermal Infrared (10.5-12.5 μm) And Water Vapour (5.7-7.1 μm) Channels And Provides 2x2 Km, 8x8 Km And 8x8 Km Ground Resolution Respectively.

INSAT-3A

The Multipurpose Satellite, INSAT-3A, Was Launched By Ariane In April 2003. It Is Located At 93.5 Degree East Longitude. The Payloads On INSAT-3A Are As Follows:

12 Normal C-Band Transponders (9 Channels Provide Expanded Coverage From Middle East To South East Asia With An EIRP Of 38 Dbw, 3 Channels Provide India Coverage With An EIRP Of 36 Dbw And 6 Extended C -Band Transponders Provide India Coverage With An EIRP Of 36 Dbw).

A CCD Camera Provides 1x1 Km Ground Resolution, In The Visible (0.63 - 0.69 μm), Near Infrared (0.77-0.86 μm) And Shortwave Infrared (1.55-1.70 μm) Bands.

INSAT-3D

Launched In July 2013, INSAT-3D Is Positioned At 82 Degree East Longitude. INSAT-3D Payloads Include Imager, Sounder, Data Relay Transponder And Search & Rescue Transponder. All The Transponders Provide Coverage Over Large Part Of The Indian Ocean Region Covering India, Bangladesh, Bhutan,Maldives, Nepal, Seychelles, Sri Lanka And Tanzania For Rendering Distress Alert Services

INSAT-3E

Launched In September 2003, INSAT-3E Is Positioned At 55 Degree East Longitude And Carries 24 Normal C-Band Transponders Provide An Edge Of Coverage EIRP Of 37 Dbw Over India And 12 Extended C-Band Transponders Provide An Edge Of Coverage EIRP Of 38 Dbw Over India.

KALPANA-1

KALPANA-1 Is An Exclusive Meteorological Satellite Launched By PSLV In September 2002. It Carries Very High Resolution Radiometer And DRT Payloads To Provide Meteorological Services. It Is Located At 74 Degree East Longitude. Its First Name Was METSAT. It Was Later Renamed As KALPANA-1 To Commemorate Kalpana Chawla.

Edusat

Configured For Audio-Visual Medium Employing Digital Interactive Classroom Lessons And Multimedia Content, EDUSAT Was Launched By GSLV In September 2004. Its Transponders And Their Ground Coverage Are Specially Configured To Cater To The Educational Requirements.

GSAT-2

Launched By The Second Flight Of GSLV In May 2003, GSAT-2 Is Located At 48 Degree East Longitude And Carries Four Normal C-Band Transponders To Provide 36 Dbw EIRP With India Coverage, Two K_u Band Transponders With 42 Dbw EIRP Over India And An MSS Payload Similar To Those On INSAT-3B And INSAT-3C.

INSAT-4 Series:

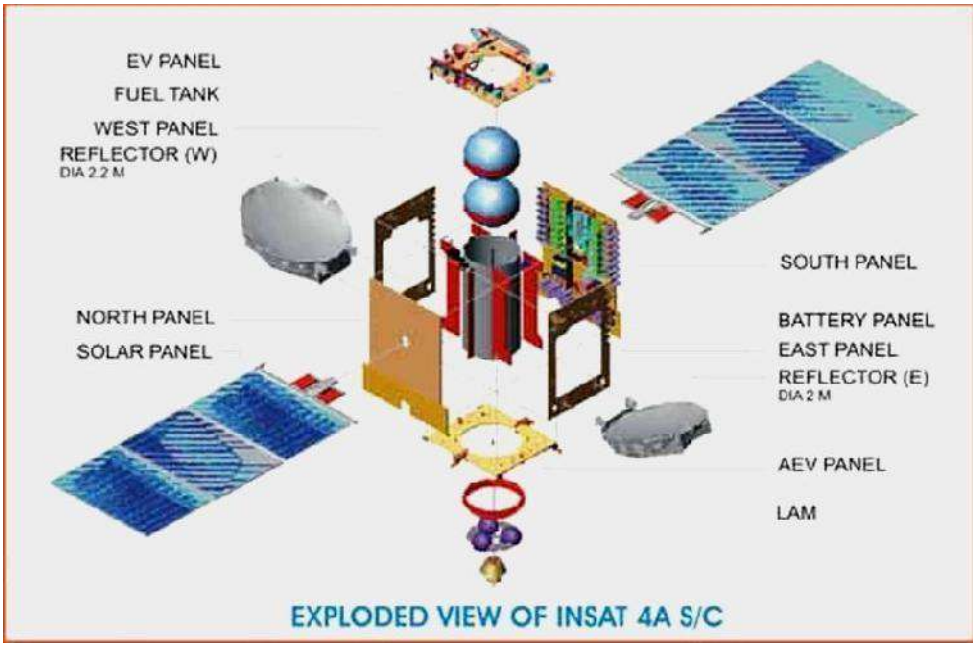


Figure 5.3 INSAT 4A

INSAT-4A is positioned at 83 degree East longitude along with INSAT-2E and INSAT-3B. It carries 12 K_u band 36 MHz bandwidth transponders employing 140 W TWTAs to provide an EIRP of 52 dBW at the edge of coverage polygon with footprint covering Indian main land and 12 C -band 36 MHz bandwidth transponders provide an EIRP of 39 dBW at the edge of coverage with expanded radiation patterns encompassing Indian geographical boun dary, area beyond India in southeast and northwest regions.[8] Tata Sky, a joint venture between the TATA Group and STAR uses INSAT-4A for distributing their DTH service.

- INSAT-4A
- INSAT-4B
- Glitch In INSAT 4B
- China-Stuxnet Connection
- INSAT-4CR
- GSAT-8 / INSAT-4G
- GSAT-12 /GSAT-10

5.3 VSAT

VSAT stands for *very small aperture terminal* system. This is the distinguishing feature of a VSAT system, the earth-station antennas being typically less than 2.4 m in diameter (Rana et al., 1990). The trend is toward even smaller dishes, not more than 1.5 m in diameter (Hughes et al., 1993).

In this sense, the small TVRO terminals for direct broadcast satellites could be labeled as VSATs, but the appellation is usually reserved for private networks, mostly providing two-way communications facilities.

Typical user groups include bank- ing and financial institutions, airline and hotel booking agencies, and large retail stores with geographically dispersed outlets.

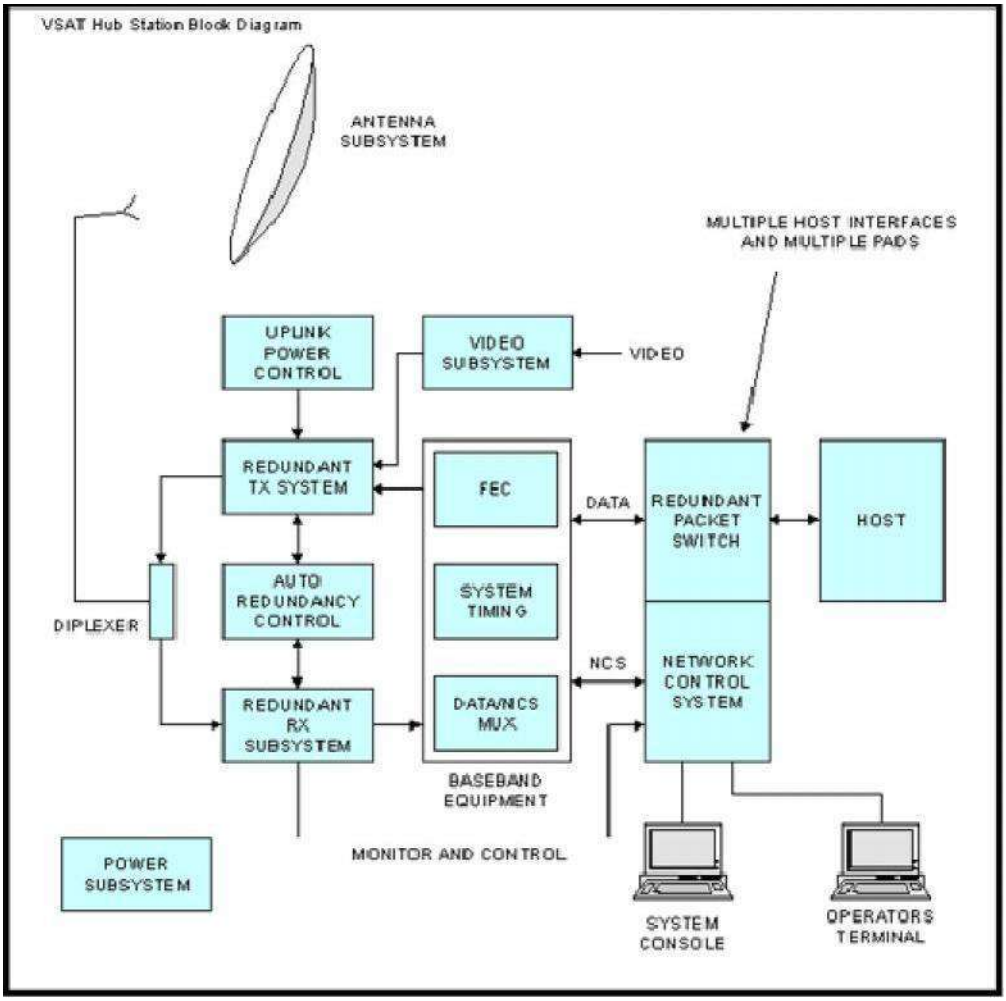


Figure 5.4 VSAT Block Diagrams

5.3.1 VSAT network

The basic structure of a VSAT network consists of a hub station which provides a broadcast facility to all the VSATs in the network and the VSATs themselves which access the satellite in some form of multiple- access mode.

The hub station is operated by the service provider, and it may be shared among a number of users, but of course, each user organ- ization has exclusive access to its own VSAT network.

Time division mul- tiplex is the normal downlink mode of transmission from hub to the VSATs, and the transmission can be broadcast for reception by all the VSATs in a network, or address coding can be used to direct messages to selected VSATs.

A form of *demand assigned multiple access* (DAMA) is employed in some systems in which channel capacity is assigned in response to the fluctuating demands of the VSATs in the network.

Most VSAT systems operate in the Ku band, although there are some C-band systems in existence (Rana et al., 1990).

5.3.2 Applications

- ✓ Supermarket shops (tills, ATM machines, stock sale updates and stock ordering).
- ✓ Chemist shops - Shoppers Drug Mart - Pharmaprix. Broadband direct to the home. e.g. Downloading MP3 audio to audio players.
- ✓ Broadband direct small business, office etc, sharing local use with many PCs.
- ✓ Internet access from on board ship Cruise ships with internet cafes, commercial shipping communications.

5.4 Mobile satellite services

5.4.1 GSM

5.4.1.1 Services and Architecture

If your work involves (or is likely to involve) some form of wireless public communications, you are likely to encounter the GSM standards. Initially developed to support a standardized approach to digital cellular communications in Europe, the "Global System for Mobile Communications" (GSM) protocols are rapidly being adopted to the next generation of wireless telecommunications systems.

In the US, its main competition appears to be the cellular TDMA systems based on the IS-54 standards. Since the GSM systems consist of a wide range of components, standards, and protocols.

The GSM and its companion standard DCS1800 (for the UK, where the 900 MHz frequencies are not available for GSM) have been developed over the last decade to allow cellular communications systems to move beyond the limitations posed by the older analog systems.

Analog system capacities are being stressed with more users that can be effectively supported by the available frequency allocations. Compatibility between types of systems had been limited, if non-existent.

By using digital encoding techniques, more users can share the same frequencies than had been available in the analog systems. As compared to the digital cellular systems in the US (CDMA [IS -95] and TDMA [IS-54]), the GSM market has had impressive success. Estimates of the numbers of telephones run from 7.5 million GSM phones to .5 million IS54 phones to .3 million for IS95.

GSM has gained in acceptance from its initial beginnings in Europe to other parts of the world including Australia, New Zealand, countries in the Middle East and the far east. Beyond its use in cellular frequencies (900 M Hz for GSM, 1800 MHz for DCS1800), portions of the GSM signaling protocols are finding their way into the newly developing PCS and LEO Satellite communications systems.

While the frequencies and link characteristics of these systems differ from the standard GSM air interface, all of these systems must deal with users roaming from one cell (or satellite beam) to another, and bridge services to public communication networks including the Public Switched Telephone Network (PSTN), and public data networks (PDN).

The GSM architecture includes several subsystems

The Mobile Station (MS) -- These digital telephones include vehicle, portable and hand-held terminals. A device called the Subscriber Identity Module (SIM) that is basically a smart -card provides custom information about users such as the services they've subscribed to and their identification in the network

The Base Station Sub-System (BSS) -- The BSS is the collection of devices that support the switching networks radio interface. Major components of the BSS include the Base Transceiver Station (BTS) that consists of the radio modems and antenna equipment.

In OSI terms, the BTS provides the physical interface to the MS where the BSC is responsible for the link layer services to the MS. Logically the transcoding equipment is in the BTS, however, an additional component.

The Network and Switching Sub-System (NSS) -- The NSS provides the switching between the GSM subsystem and external networks along with the databases used for additional subscriber and mobility management.

Major components in the NSS include the Mobile Servi ces Switching Center (MSC), Home and Visiting Location Registers (HLR, VLR). The HLR and VLR databases are interconnected through the telecomm standard Signaling System 7 (SS7) control network.

The Operation Sub-System (OSS) -- The OSS provides the support functions responsible for the management of network maintenance and services. Components of the OSS are responsible for network operation and maintenance, mobile equipment management, and subscription management and charging.

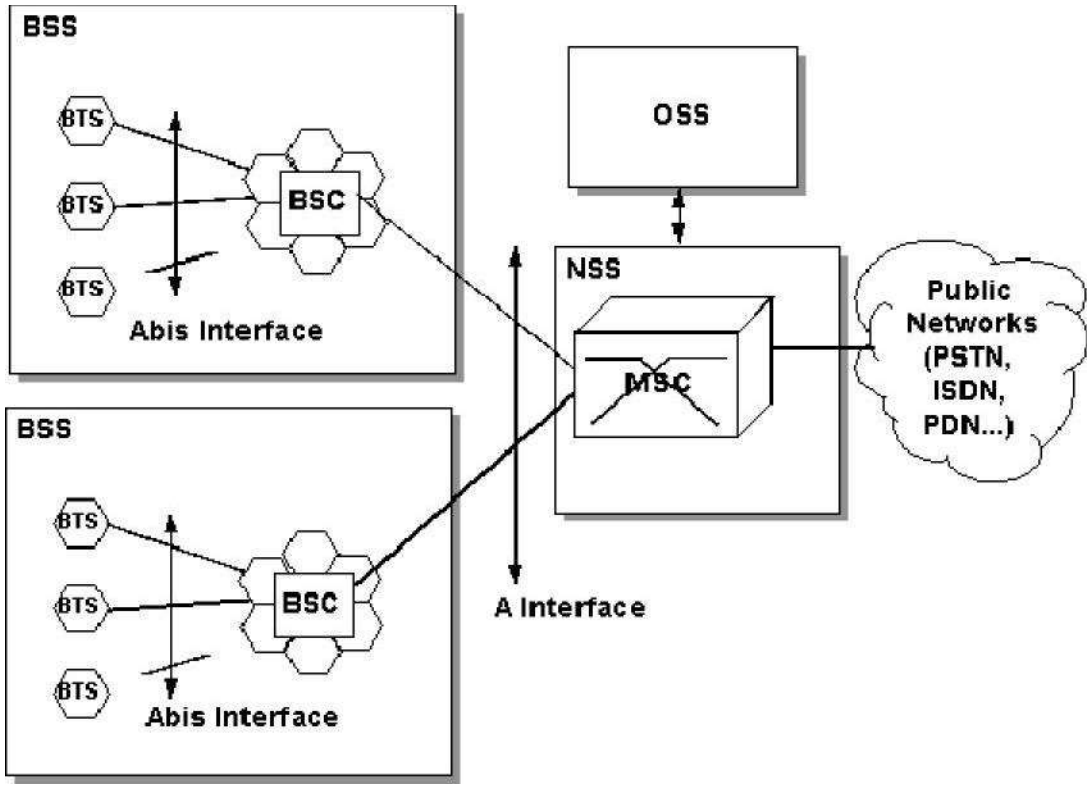


Figure 5.5 GSM Block Diagrams

Several channels are used in the air interface

- ✓ **FCCH** - the frequency correction channel - provides frequency synchronization information in a burst
- ✓ **SCH** - Synchronization Channel - shortly following the FCCH burst (8 bits later), provides a reference to all slots on a given frequency
- ✓ **PAGCH** - Paging and Access Grant Channel used for the transmission of paging information requesting the setup of a call to a MS.
- ✓ **RACH** - Random Access Channel - an inbound channel used by the MS to request connections from the ground network. Since this is used for the first access attempt by users of the network, a random access scheme is used to aid in avoiding collisions.
- ✓ **CBCH -** Cell Broadcast Channel - used for infrequent transmission of broadcasts by the ground network.
- ✓ **BCCH** - Broadcast Control Channel - provides access status information to the MS. The information provided on this channel is used by the MS to determine whether or not to request a transition to a new cell

- ✓ **FACCH -** Fast Associated Control Channel for the control of handovers
- ✓ **TCH/F** - Traffic Channel, Full Rate for speech at 13 kbps or data at 12, 6, or 3.6 kbps
- ✓ **TCH/H** - Traffic Channel, Half Rate for speech at 7 kbps, or data at 6 or 3.6 kbps

5.5 Mobility Management

One of the major features used in all classes of GSM networks (cellular, PCS and Satellite) is the ability to support roaming users. Through the control signaling network, the MSCs interact to locate and connect to users throughout the network.

"Location Registers" are included in the MSC databases to assist in the role of determining how, and whether connections are to be made to roaming users. Each user of a GSM MS is assigned a Home Location Register (HLR) that is used to contain the user's location and subscribed services.

Difficulties facing the operators can include

a. Remote/Rural Areas. To service remote areas, it is often economically unfeasible to provide backhaul facilities (BTS to BSC) via terrestrial lines (fiber/microwave).

b. Time to deploy. Terrestrial build-outs can take years to plan and implement.
c. Areas of 'minor' interest. These can include small isolated centers such as tourist resorts, islands, mines, oil exploration sites, hydro-electric facilities.
d. Temporary Coverage. Special events, even in urban areas, can overload the existing infrastructure.

5.5.1 GSM service security

GSM was designed with a moderate level of service security. GSM uses several cryptographic algorithms for security. The A5/1, A5/2, and A5/3 stream ciphers are used for ensuring over-the-air voice privacy.

GSM uses General Packet Radio Service (GPRS) for data transmissions like browsing the web. The most commonly deployed GPRS ciphers were publicly broken in 2011The researchers revealed flaws in the commonly used GEA/1.

5.5.2 Global Positioning System (GPS)

The Global Positioning System (GPS) is a satellite based navigation system that can be used to locate positions anywhere on earth. Designed and operated by the U.S. Department of Defense, it consists of satellites, control and monitor stations, and receivers. GPS receivers take information transmitted from the satellites and uses triangulation to calculate a user's exact location. GPS is used on incidents in a variety of ways, such as:

- ✓ To determine position locations; for example, you need to radio a helicopter pilot the coordinates of your position location so the pilot can pick you up.
- ✓ To navigate from one location to another; for example, you need to travel from a lookout to the fire perimeter.
- ✓ To create digitized maps; for example, you are assigned to plot the fire perimeter and hot spots.
- ✓ To determine distance between two points or how far you are from another location.

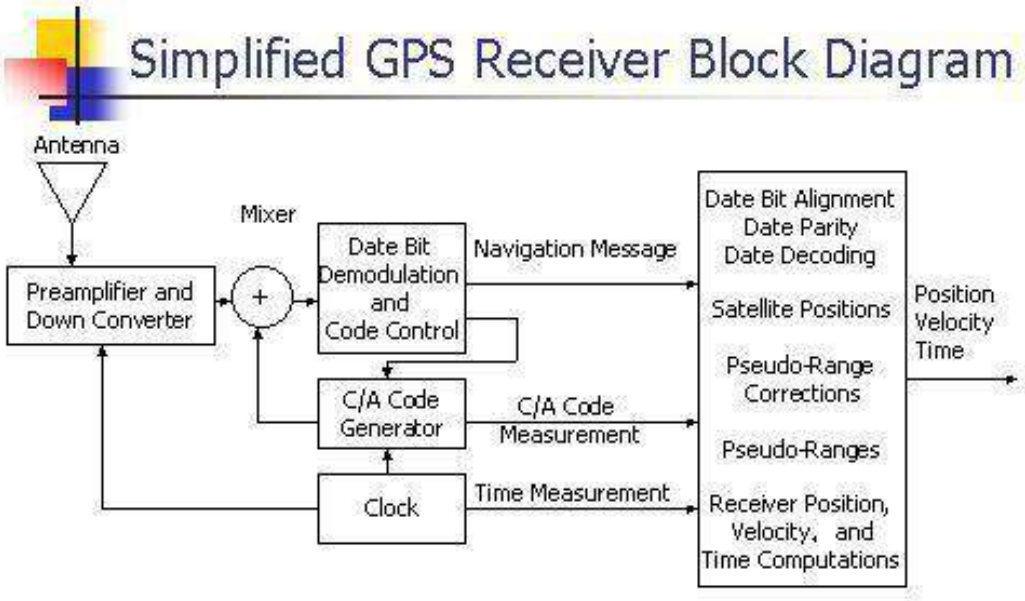


Figure 5.6 GPS Block Diagrams

The purpose of this chapter is to give a general overview of the Global Positioning System, not to teach proficiency in the use of a GPS receiver. To become proficient with a specific GPS receiver, study the owner's manual and practice using the receiver.

The chapter starts with a general introduction on how the global positioning system works. Then it discusses some basics on using a GPS receiver.

Three Segments of GPS:

Space Segment — Satellites orbiting the earth

The space segment consists of 29 satellites circling the earth every 12 hours at 12,000 miles in altitude. This high altitude allows the signals to cover a greater area. The satellites are arranged in their orbits so a GPS receiver on earth can receive a signal from at least four satellites at any given time. Each satellite contains several atomic clocks.

Control Segment — The control and monitoring stations

The control segment tracks the satellites and then provides them with corrected orbital and time information. The control segment consists of five unmanned monitor stations and one Master Control Station. The five unmanned stations monitor GPS satellite signals and then send that information to the Master Control Station where anomalies are corrected and sent back to the GPS satellites through ground antennas.

User Segment — The GPS receivers owned by civilians and military

The user segment consists of the users and their GPS receivers. The number of simultaneous users is limitless.

How GPS Determines a Position

The GPS receiver uses the following information to determine a position.

- ✓ Precise location of satellites

✓When a GPS receiver is first turned on, it downloads orbit information from all the satellites called an almanac. This process, the first time, can take as long as 12 minutes; but once this information is downloaded, it is stored in the receiver's memory for future use.

- ✓ Distance from each satellite

The GPS receiver calculates the distance from each satellite to the receiver by using the distance formula: distance = velocity x time. The receiver already knows the velocity, which is the speed of a radio wave or 186,000 miles per second (the speed of light).

- ✓ Triangulation to determine position

The receiver determines position by using triangulation. When it receives signals from at least three satellites the receiver should be able to calculate its approximate position (a 2D position). The receiver needs at least four or more satellites to calculate a more accurate 3D position.

Using a GPS Receiver

There are several different models and types of GPS receivers. Refer to the owner's manual for your GPS receiver and practice using it to become proficient.

- ✓ When working on an incident with a GPS receiver it is important to:
- ✓ Always have a compass and a map.
- ✓ Have a GPS download cable.
- ✓ Have extra batteries.
- ✓ Know memory capacity of the GPS receiver to prevent loss of data, decrease in accuracy of data,or other problems.
- ✓ Use an external antennae whenever possible, especially under tree canopy, in canyons, or while flying or driving.
- ✓ Set up GPS receiver according to incident or agency standard regulation; coordinate system.
- ✓ Take notes that describe what you are saving in the receiver.

5.6. INMARSAT

Inmarsat-Indian Maritime SATellite is still the sole IMO-mandated provider of satellite communications for the GMDSS.

- Availability for GMDSS is a minimum of 99.9%

Inmarsat has constantly and consistently exceeded this figure & Independently audited by IMSO and reported on to IMO.

Now Inmarsat commercial services use the same satellites and network &Inmarsat A closes at midnight on 31 December 2007 Agreed by IMO – MSC/Circ.1076. Successful closure programme almost concluded Overseen throughout by IMSO.

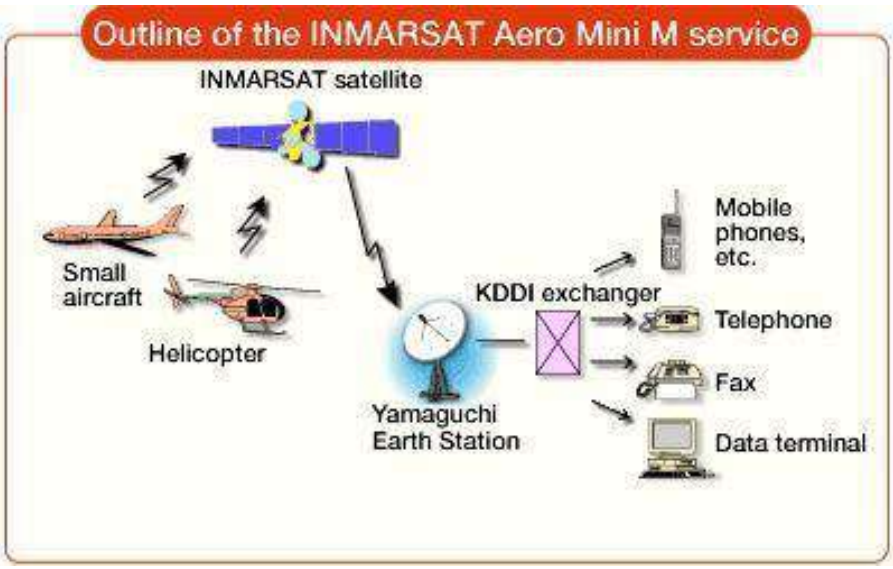


Figure 5.7 INMARSAT Satellite Service

GMDSS services continue to be provided by:

- Inmarsat B, Inmarsat C/mini-C and Inmarsat Fleet F77
- Potential for GMDSS on FleetBroadband being assessed

- The IMO Criteria for the Provision of Mobile Satellite Communications Systems in the Global Maritime Distress and Safety System (GMDSS)

- Amendments were proposed; potentially to make it simpler for other satellite systems to be approved

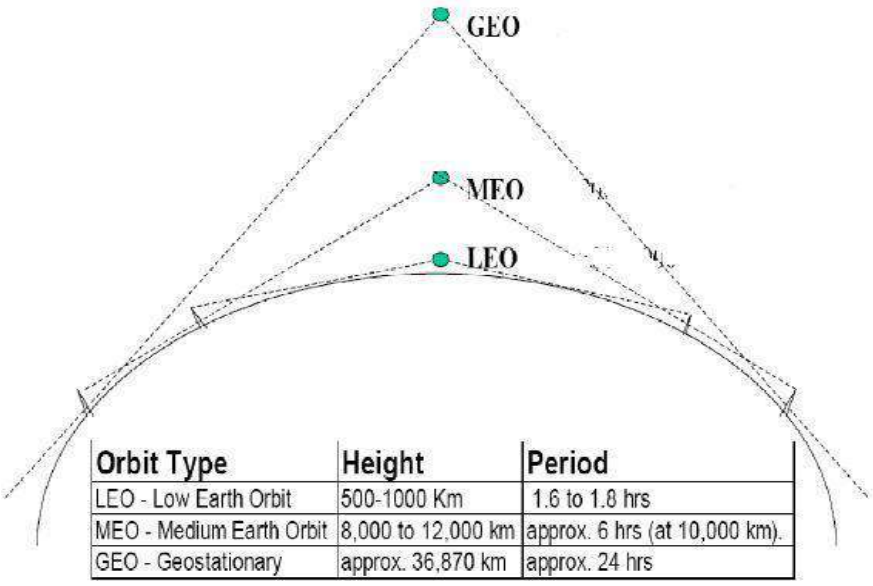
- The original requirements remain and were approved by MSC 83
  - No dilution of standards

- Minor amendments only; replacement Resolution expected to be approved by the IMO 25th Assembly

- Inmarsat remains the sole, approved satcom provider for the GMDSS

5.7 LEO: Low Earth Orbit satellites have a small area of coverage. They are positioned in an orbit approximately 3000km from the surface of the earth

- They complete one orbit every 90 minutes
- The large majority of satellites are in low earth orbit
- The Iridium system utilizes LEO satellites (780km high)
- The satellite in LEO orbit is visible to a point on the earth for a very short time



Orbit Type	Height	Period
LEO - Low Earth Orbit	500-1000 Km	1.6 to 1.8 hrs
MEO - Medium Earth Orbit	8,000 to 12,000 km	approx. 6 hrs (at 10,000 km).
GEO - Geostationary	approx. 36,870 km	approx. 24 hrs

Figure 5.8 LEO, MEO & GEO range

5.8 MEO: *Medium Earth Orbit* satellites have orbital altitudes between 3,000 and 30,000 km.

- They are commonly used used in navigation systems such as GPS

5.9 GEO: *Geosynchronous (Geostationary) Earth Orbit* satellites are positioned over the equator. The orbital altitude is around 30,000-40,000 km

- There is only one geostationary orbit possible around the earth
  - Lying on the earth's equatorial plane.
  - The satellite orbiting at the same speed as the rotational speed of the earth on its axis.
  - They complete one orbit every 24 hours. This causes the satellite to appear stationary with respect to a point on the earth, allowing one satellite to provide continual coverage to a given area on the earth's surface
  - One GEO satellite can cover approximately 1/3 of the world's surface

They are commonly used in communication systems

- Advantages:
  - Simple ground station tracking.
  - Nearly constant range
  - Very small frequency shift

- Disadvantages:
  - Transmission delay of the order of 250 msec.
  - Large free space loss.
  - No polar coverage

- Satellite orbits in terms of the orbital height:
- According to distance from earth:
  - Geosynchronous Earth Orbit (GEO) ,
  - Medium Earth Orbit (MEO),
  - Low Earth Orbit (LEO)

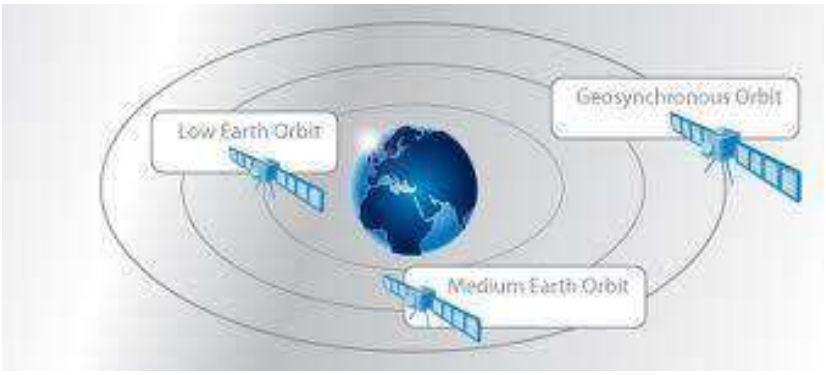


Figure 5.9 LEO, MEO & GEO Orbits

LEO / MEO / GEO / HEO (cont.)

	Name	Number	Panel	No./Panel	altitude	deg.
LEO	STARSYS	24	6	4	1300km	60
	ORBCOMM	24	4	6	785km	45
	GLOBALSTAR	48	8	6	1400km	52
	IRIDIUM	66	6	11	765km	86
MEO	Name	Number	Panel	No./Panel	altitude	deg.
	INMARSAT P	10	2	5	10300km	45
	ODYSEEY	12	3	4	10370km	55
	GPS	24	6	4	20200km	55
	GLONASS	24	3	8	19132km	64.8
HEO	Name	Number	Panel	No./Panel	altitude	deg.
	ELLIPSO	24	4	6	A:7800km P:520km	63.4
	MOLNIYA	4	1	4	A:39863km P:504km	63.4
	ARCHIMEDES	4	4	1	A:39447km P:926km	63.4

9

Figure 5.10 Diff b/w LEO, MEO & GEO Orbits

GEO: 35,786 km above the earth, MEO: 8,000-20,000 km above the earth & LEO: 500-2,000 km above the earth.

5.10 Satellite Navigational System: Benefits

- ❖ Enhanced Safety
- ❖ Increased Capacity
- ❖ Reduced Delays

Advantage

- o Increased Flight Efficiencies
- o Increased Schedule Predictability
- o Environmentally Beneficial Procedures

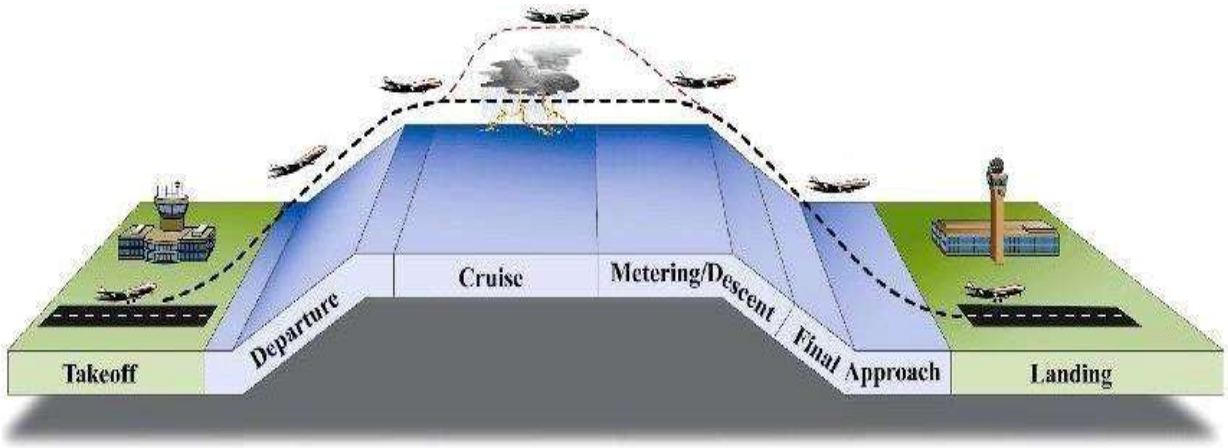


Figure 5.11 LEO, MEO & GEO Orbits

- Using ICAO GNSS Implementation Strategy and ICAO Standards and Recommended Practices
- GPS Aviation Use Approved for Over a Decade
  - Aircraft Based Augmentation Systems (ABAS) – (e.g. RAIM)
- Space Based Augmentation System (SBAS) since 2003
  - Wide Area Augmentation System (WAAS) augmenting GPS

- Development of GNSS Ground Based Augmentation System (GBAS) Continues
  - Local Area Augmentation System (LAAS)
- GNSS is Cornerstone for National Airspace System

5.11 Direct Broadcast satellites (DBS)

Satellites provide *broadcast* transmissions in the fullest sense of the word, because antenna footprints can be made to cover large areas of the earth.

The idea of using satellites to provide direct transmissions into the home has been around for many years, and the services provided are known generally as *direct broadcast satellite* (DBS) services.

Broadcast services include audio, television, and Internet services.

5.11.1 Power Rating and Number of Transponders

From Table 1.4 it will be seen that satellites primarily intended for DBS have a higher [EIRP] than for the other categories, being in the range 51 to 60 dBW. At a *Regional Administrative Radio Council* (RARC) meeting in 1983, the value established for DBS was 57 dBW (Mead,2000). Transponders are rated by the power output of their high-power amplifiers.

Typically, a satellite may carry 32 transponders. If all 32 are in use, each will operate at the lower power rating of 120 W.

The available bandwidth (uplink and downlink) is seen to be 500 MHz. A total number of 32 transponder channels, each of bandwidth 24 MHz, can be accommodated.

The bandwidth is sometimes specified as 27 MHz, but this includes a 3-MHz guard band allowance. Therefore, when calculating bit-rate capacity, the 24 MHz value is used.

The total of 32 transponders requires the use of both *right- hand circular polarization* (RHCP) and *left-hand circular polarization* (LHCP) in order to permit frequency reuse, and guard bands are inserted between channels of a given polarization.

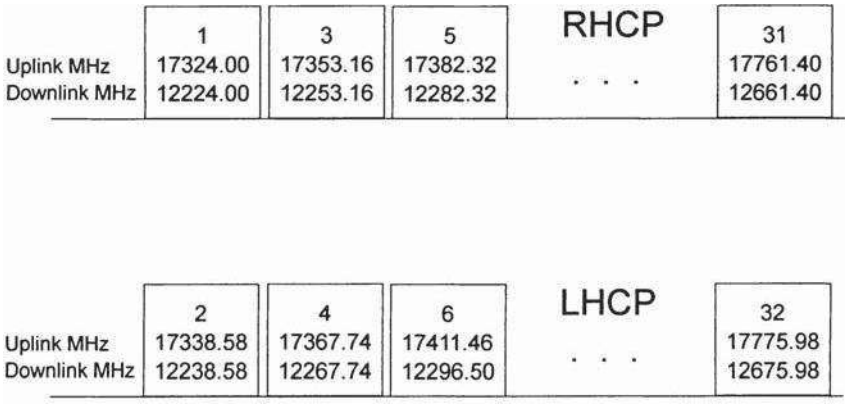


Figure 5.12 DBS Service

5.11.2 Bit Rates for Digital Television

The bit rate for digital television depends very much on the picture format. One way of estimating the uncompressed bit rate is to multiply the number of pixels in a frame by the number of frames per second, and multiply this by the number of bits used to encode each pixel.

5.11.3 MPEG Compression Standards

MPEG is a group within the *International Standards Organization and the International Electrochemical Commission* (ISO/IEC) that undertook the job of defining standards for the transmission and storage of moving pictures and sound.

The MPEG standards currently available are MPEG-1, MPEG-2, MPEG-4, and MPEG-7.

5.12 Direct to home Broadcast (DTH)

DTH stands for Direct-To-Home television. DTH is defined as the reception of satellite programmes with a personal dish in an individual home.

- ✓ DTH Broadcasting to home TV receivers take place in the ku band(12 GHz). This service is known as Direct To Home service.
- ✓ DTH services were first proposed in India in 1996.
- ✓ Finally in 2000, DTH was allowed.
- ✓ The new policy requires all operators to set up earth stations in India

within 12 months of getting a license. DTH licenses in India will cost $2.14 million and will be valid for 10 years.

Working principal of DTH is the satellite communication. Broadcaster modulates the received signal and transmit it to the satellite in KU Band and from satellite one can receive signal by dish and set top box.

5.12.1 DTH Block Diagram

- ✓ A DTH network consists of a broadcasting centre, satellites, encoders, multiplexers, modulators and DTH receivers
- ✓ The encoder converts the audio, video and data signals into the digital format and the multiplexer mixes these signals.

It is used to provide the DTH service in high populated area A Multi Switch is basically a box that contains signal splitters and A/B switches. A outputs of group of DTH LNBs are connected to the A and B inputs of the Multi Switch.

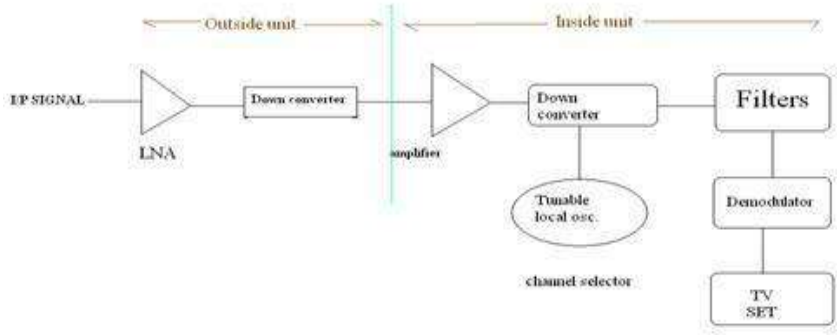


Figure 5.13 DTH Service

5.12.2 Advantage

- ✓ DTH also offers digital quality signals which do not degrade the picture or sound quality.
- ✓ It also offers interactive channels and program guides with customers having the choice to block out programming which they consider undesirable
- ✓ One of the great advantages of the cable industry has been the ability to provide local channels, but this handicap has been overcome by many

DTH providers using other local channels or local feeds.

- ✓ The other advantage of DTH is the availability of satellite broadcast in rural and semi-urban areas where cable is difficult to install.

5.13 Digital audio broadcast (DAB)

DAB Project is an industry-led consortium of over 300 companies

- ✓ The DAB Project was launched on 10th September, 1993
- ✓ In 1995 it was basically finished and became operational
- ✓ There are several sub-standards of the DAB standard
  - ❖ DAB-S (Satellite) – using QPSK – 40 Mb/s
  - ❖ DAB-T (Terrestrial) – using QAM – 50 Mb/s
  - ❖ DAB-C (Cable) – using OFDM – 24 Mb/s
- ✓ These three sub-standards basically differ only in the specifications to the physical representation, modulation, transmission and reception of the signal.
- ✓ The DAB stream consists of a series of fixed length packets which make up a Transport Stream (TS). The packets support 'streams' or 'data sections'.
- ✓ Streams carry higher layer packets derived from an MPEG stream & Data sections are blocks of data carrying signaling and control data.
- ✓ DAB is actually a support mechanism for MPEG.& One MPEG stream needing higher instantaneous data can 'steal' capacity from another with spare capacity.

5.14 World space services

World Space (Nasdaq: WRSP) is the world's only global media and entertainment company positioned to offer a satellite radio experience to consumers in more than 130 countries with five billion people, driving 300 million cars. World Space delivers the latest tunes, trends and information from around the world and around the corner.

World Space subscribers benefit from a unique combination of local programming, original World Space content and content from leading brands

around the globe, including the BBC, CNN, Virgin Radio, NDTV and RFI. World Space's satellites cover two-thirds of the globe with six beams.

Each beam is capable of delivering up to 80 channels of high quality digital audio and multimedia programming directly to World Space Satellite Radios anytime and virtually anywhere in its coverage area. World Space is a pioneer of satellite-based digital radio services (DARS) and was instrumental in the development of the technology infrastructure used today by XM Satellite Radio.

5.15 Business Television (BTV) - Adaptations for Education

Business television (BTV) is the production and distribution, via satellite, of video programs for closed user group audiences. It often has two -way audio interaction component made through a simple telephone line. It is being used by many industries including brokerage firms, pizza houses, car dealers and delivery services.

BTV is an increasingly popular method of information delivery for corporations and institutions. Private networks, account for about 70 percent of all BTV networks. It is estimated that by the mid-1990s BTV has the potential to grow to a $1.6 billion market in North America with more and more Fortune 1,000 companies getting involved. The increase in use of BTV has been dramatic.

Institution updates, news, training, meetings and other events can be broadcast live to multiple locations. The expertise of the best instructors can be delivered to thousands of people without requiring trainers to go to the site. Information can be disseminated to all employees at once, not just a few at a time. Delivery to the workplace at low cost provides the access to training that has been denied lower level employees. It may be the key to re -training America's work force.

Television has been used to deliver training and information within businesses for more than 40 years. Its recent growth began with the introduction of the video cassette in the early 1970s. Even though most programming is produced for video cassette distribution, business is using BTV to provide efficient delivery of specialized programs via satellite.

The advent of smaller receiving stations - called very small aperture terminals (VSATs) has made private communication networks much more economical to operate. BTV has a number of tangible benefits, such as reducing travel, immediate delivery of time-critical messages, and eliminating cassette duplication and distribution hassles.

The programming on BTV networks is extremely cost-effective compared to seminar fees and downtime for travel. It is an excellent way to get solid and current information very fast. Some people prefer to attend seminars and

conferences where they can read, see, hear and ask questions in person. BTV provides yet another piece of the education menu and is another way to provide professional development.

A key advantage is that its format allows viewers to interact with presenters by telephone, enabling viewers to become a part of the program. The satellite effectively places people in the same room, so that sales personnel in the field can learn about new products at the same time.

Speed of transmission may well be the competitive edge which s ome firms need as they introduce new products and services. BTV enables employees in many locations to focus on common problems or issues that might develop into crises without quick communication and resolution.

BTV networks transmit information every business day on a broad range of topics, and provide instructional courses on various products, market trends, selling and motivation. Networks give subscribers the tools to apply the information they have to real world situations.

5.16 GRAMSAT

ISRO has come up with the concept of dedicated GRAMSAT satellites, keeping in mind the urgent need to eradicate illiteracy in the rural belt which is necessary for the all round development of the nation.

This Gramsat satellite is carrying six to eight high powered C -band transponders, which together with video compression techniques can disseminate regional and cultural specific audio-visual programmes of relevance in each of the regional languages through rebroadcast mode on an ordinary TV set.

The high power in C-band has enabled even remote area viewers outside the reach of the TV transmitters to receive programmers of their choice in a direct reception mode with a simple .dish antenna.

The salient features of GRAMSAT projects are:

i. Its communications networks are at the state level connecting the state capital to districts, blocks and enabling a reach to villages.

ii. It is also providing computer connectivity data broadcasting, TV - broadcasting facilities having applications like e- governance, development information, teleconferencing, helping disaster management.

iii. Providing rural-education broadcasting.

However, the Gramsat projects have an appropriate combination of following activities.

(i) Interactive training at district and block levels employing suit able configuration

(ii) Broadcasting services for rural development

(iii) Computer interconnectivity and data exchange services

(iv) Tele-health and tele-medicine services.

5.17 Specialized services

5.17.1Satellite-email services

The addition of Internet Access enables Astrium to act as an Internet Service Provider (ISP) capable of offering Inmarsat users a tailor -made Internet connection.

With Internet services added to our range of terrestrial networks, you will no longer need to subscribe to a third party for Internet access (available for Inmarsat A, B, M, mini-M, Fleet, GAN, Regional BGAN & SWIFT networks).

We treat Internet in the same way as the other terrestrial networks we provide, and thus offer unrestricted access to this service. There is no time - consuming log-on procedure, as users are not required to submit a user-ID or password.

Description of E-mail Service

Astrium's E-Mail service allows Inmarsat users to send and receive e-mail directly through the Internet without accessing a public telephone network.

Features and Benefits

- ✓ No need to configure an e-mail client to access a Astrium e-mail account
- ✓ Service optimized for use with low bandwidth Inmarsat terminals
- ✓ Filter e-mail by previewing the Inbox and deleting any unwanted e-mails prior to downloading
- ✓ No surcharge or monthly subscription fees
- ✓ Service billed according to standard airtime prices for Inmarsat service used

5.17.2 Video Conferencing (medium resolution)

Video conferencing technology can be used to provide the same full, two - way interactivity of satellite broadcast at much lower cost. For Multi -Site meetings, video conferencing uses bridging systems to connect each site to the others.

It is possible to configure a video conference bridge to show all sites at the same time on a projection screen or monitor. Or, as is more typical, a bridge can show just the site from which a person is speaking or making a presentation.

The technology that makes interactive video conferencing possible, compresses video and audio signals, thus creating an image quality lower than that of satellite broadcasts.

5.17.3. Satellite Internet access

Satellite Internet access is Internet access provided through communications satellites. Modern satellite Internet service is typically pro vided to users through geostationary satellites that can offer high data speeds, with newer satellites using Ka band to achieve downstream data speeds up to 50 Mbps.

Satellite Internet generally relies on three primary components: a satellite in geostationary orbit (sometimes referred to as a geosynchronous Earth orbit, or GEO), a number of ground stations known as gateways that relay Internet data to and from the satellite via radio waves (microwave), and a VSAT (very-small-aperture terminal) dish antenna with a transceiver, located at the subscriber's premises.

Other components of a satellite Internet system include a modem at the user end which links the user's network with the transceiver, and a centralized network operations center (NOC) for monitoring the entire system.

APPLICATIONS

Figure example of INSAT -3 satellites

Figure example of Weather forecasting satellite

MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

GE8076 PROFESSIONAL ETHICS IN ENGINEERING

Semester - 08

Notes

MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

Vision

To excel in providing value based education in the field of Electronics and Communication Engineering, keeping in pace with the latest technical developments through commendable research, to raise the intellectual competence to match global standards and to make significant contributions to the society upholding the ethical standards.

Mission

- ✓ To deliver Quality Technical Education, with an equal emphasis on theoretical and practical aspects.
- ✓ To provide state of the art infrastructure for the students and faculty to upgrade their skills and knowledge.
- ✓ To create an open and conducive environment for faculty and students to carry out research and excel in their field of specialization.
- ✓ To focus especially on innovation and development of technologies that is sustainable and inclusive, and thus benefits all sections of the society.
- ✓ To establish a strong Industry Academic Collaboration for teaching and research, that could foster entrepreneurship and innovation in knowledge exchange.
- ✓ To produce quality Engineers who uphold and advance the integrity, honour and dignity of the engineering.

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

1. To provide the students with a strong foundation in the required sciences in order to pursue studies in Electronics and Communication Engineering.
2. To gain adequate knowledge to become good professional in electronic and communication engineering associated industries, higher education and research.
3. To develop attitude in lifelong learning, applying and adapting new ideas and technologies as their field evolves.
4. To prepare students to critically analyze existing literature in an area of specialization and ethically develop innovative and research oriented methodologies to solve the problems identified.
5. To inculcate in the students a professional and ethical attitude and an ability to visualize the engineering issues in a broader social context.

PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO1: Design, develop and analyze electronic systems through application of relevant electronics, mathematics and engineering principles.

PSO2: Design, develop and analyze communication systems through application of fundamentals from communication principles, signal processing, and RF System Design & Electromagnetics.

PSO3: Adapt to emerging electronics and communication technologies and develop innovative solutions for existing and newer problems.

GE 8076 PROFESSIONAL ETHICS IN ENGINEERING L T P C 3 0 0 3

OBJECTIVES:

- To enable the students to create an awareness on Engineering Ethics and Human Values, to instill Moral and Social Values and Loyalty and to appreciate the rights of others.

UNIT I HUMAN VALUES 10

Morals, values and Ethics – Integrity – Work ethic – Service learning – Civic virtue – Respect for others – Living peacefully – Caring – Sharing – Honesty – Courage – Valuing time – Cooperation – Commitment – Empathy – Self confidence – Character – Spirituality – Introduction to Yoga and meditation for professional excellence and stress management.

UNIT II ENGINEERING ETHICS 9

Senses of 'Engineering Ethics' – Variety of moral issues – Types of inquiry – Moral dilemmas – Moral Autonomy – Kohlberg's theory – Gilligan's theory – Consensus and Controversy – Models of professional roles - Theories about right action – Self-interest – Customs and Religion – Uses of Ethical Theories UNIT **III ENGINEERING AS SOCIAL EXPERIMENTATION 9**

Engineering as Experimentation – Engineers as responsible Experimenters – Codes of Ethics – A Balanced Outlook on Law.

UNIT IV SAFETY, RESPONSIBILITIES AND RIGHTS 9

Safety and Risk – Assessment of Safety and Risk – Risk Benefit Analysis and Reducing Risk - Respect for Authority – Collective Bargaining – Confidentiality – Conflicts of Interest – Occupational Crime – Professional Rights – Employee Rights – Intellectual Property Rights (IPR) – Discrimination

UNIT V GLOBAL ISSUES 8

Multinational Corporations – Environmental Ethics – Computer Ethics – Weapons Development – Engineers as Managers – Consulting Engineers – Engineers as Expert Witnesses and Advisors – Moral Leadership –Code of Conduct – Corporate Social Responsibility

TOTAL: 45 PERIODS

OUTCOMES:

- Upon completion of the course, the student should be able to apply ethics in society, discuss the ethical issues related to engineering and realize the responsibilities and rights in the society.

TEXTBOOKS:

1. Mike W. Martin and Roland Schinzinger, "Ethics in Engineering", Tata McGraw Hill, New Delhi, 2003.
2. Govindarajan M, Natarajan S, Senthil Kumar V. S, "Engineering Ethics", Prentice Hall of India, New Delhi, 2004.

REFERENCES:

1. Charles B. Fleddermann, "Engineering Ethics", Pearson Prentice Hall, New Jersey, 2004.
2. Charles E. Harris, Michael S. Pritchard and Michael J. Rabins, "Engineering Ethics – Concepts and Cases", Cengage Learning, 2009
3. John R Boatright, "Ethics and the Conduct of Business", Pearson Education, New Delhi, 2003
4. Edmund G Seebauer and Robert L Barry, "Fundamentals of Ethics for Scientists and Engineers", Oxford University Press, Oxford, 2001
5. Laura P. Hartman and Joe Desjardins, "Business Ethics: Decision Making for Personal Integrity and Social Responsibility" McGraw Hill education, India Pvt. Ltd., New Delhi 2013
6. World Community Service Centre, ' Value Education', Vethathiri publications, Erode, 2011

Web sources:

1. www.onlineethics.org
2. www.nspe.org
3. www.globalethics.org 4. www.ethics.org

UNIT I HUMAN VALUES

Morals, values and Ethics – Integrity – Work ethic – Service learning – Civic virtue – Respect for others – Living peacefully – Caring – Sharing – Honesty – Courage – Valuing time – Cooperation – Commitment – Empathy – Self confidence – Character – Spirituality – Introduction to Yoga and meditation for professional excellence and stress management.

1. MORALS, VALUES AND ETHICS

MORALS

Moral is concerned with the principles of right and wrong behaviour. Morals relate to duty or obligations. It pertains to those actions of right and wrong, virtue and vice. Very often the terms
'morality' and 'ethics' often used interchangeably and closely related, but it is essential to identify the differences between these two with respect to the study of ethical matters. Morality tends to be more general and perspective.

There are some ways to acquire moral beliefs. The following six ways are helpful in acquiring moral beliefs:

- Authority
- Logic
- Sense experience
- Emotion
- Intuition and
- Science

Morality relates to human conduct. Most of the human have acquired and practise moral principles. The whole-heartedly accept them. Moral principles imply that human beings and animals are to be treated with respect and dignity. These principles are, for examples, against stealing others property, against telling lies etc

Morality also guides human beings on aspirations, ideals and values. It guides in understanding human nature, tradition and society. It also determines ones place in society and universe.

If anyone is directed to obey parental authority in ones childhood ages, he/she will generally feel it convenient to accept other authorities in later years. If anyone learns to control the desires of somebody else, he further can substitute other influences for parental authority in later periods. Other ability to respond to any kind of moral guidance is relying upon their ability to control their own actions. Moral philosophers are of the view that moral theories and moral actions are necessary parts to be a human. They are the guiding principles that every citizen should hold.

'Ethics' generally refers only to professional behaviour
'Morality' refers to any aspect of human action very often
'Values' are principles of some one's being god or bad

VALUES

- Value makes actions, characters, traits and objects of anyone good or bad. Evaluating the moral qualities of people or actions and their non-moral characters enviably raises the qualities of people or actions and their non-moral characters enviably raises the qualities of the nature or source of those value.
- A value is defined as a principle that promotes well-being or prevents harm. Values are our guidelines for our success—our paradigm about what is acceptable." Personal values are defined as:

"Emotional beliefs in principles regarded as particularly favourable or important for the individual. Values are individual in nature.

- Values are comprised of personal concepts of responsibility, entitlement and respect
- Values are shaped by personal experience, may change over the span of a lifetime and may be influenced by lessons learned.
- Values may vary according to an individual's cultural, ethnic and/or faith-based background.
- In spite of all the change around you, decide upon what you will never change: your core values.
- Take your time to decide what they are but once you do, do not compromise on them for any reason. *Integrity* is one such value.
- Human values emerge due to two factors:
    1. The impingement of society and its meanings and norms on the fulfillment of the individuals needs or drives.
    2. The introduction of his own awareness, choice and judgement in need fulfillment.
- These two processes are interdependent. Values are essentially social products, and at the same time involve the individuals assumption of certain common goals and purposes of the social milieu that have become a part of him.

TYPES OF VALUES

The five core human values are: (1) Right conduct, (2) Peace, (3) Truth, (4) Love, and (5) Nonviolence.

1. Values related to RIGHT CONDUCT are:

- **SELF-HELP SKILLS:** Care of possessions, diet, hygiene, modesty, posture, self reliance, and tidy appearance
- **SOCIAL SKILLS:** Good behaviour, good manners, good relationships, helpfulness, No wastage, and good environment, and
- **ETHICAL SKILLS:** Code of conduct, courage, dependability, duty, efficiency, ingenuity, initiative, perseverance, punctuality, resourcefulness, respect for all, and responsibility

2. **Values related to PEACE are:**

- Attention, calmness, concentration, contentment, dignity, discipline, equality, equanimity, faithfulness, focus, gratitude, happiness, harmony, humility, inner silence, optimism, patience, reflection, satisfaction, self-acceptance, self-confidence, self-control, self-discipline, self-esteem, self-respect, sense control, tolerance, and understanding

3. **Values related to TRUTH are:**

- Accuracy, curiosity, discernment, fairness, fearlessness, honesty, integrity (unity of thought, word, and deed), intuition, justice, optimism, purity, quest for knowledge, reason, self-analysis, sincerity, sprit of enquiry, synthesis, trust, truthfulness, and determination.

4. **Values related to LOVE are:**

- Acceptance, affection, care, compassion, consideration, dedication, devotion, empathy, forbearance, forgiveness, friendship, generosity, gentleness, humanness, interdependence, kindness, patience, patriotism, reverence, sacrifice, selflessness, service, sharing, sympathy, thoughtfulness, tolerance and trust

5. Values related to NON-VIOLENCE are:

(a)**PSYCHOLOGICAL**: Benevolence, compassion, concern for others, consideration, forbearance, forgiveness, manners, happiness, loyalty, morality, and universal love

(b)**SOCIAL:** Appreciation of other cultures and religions, brotherhood, care of environment, citizenship, equality, harmlessness, national awareness, perseverance, respect for property, and social justice.

PERSEVERANCE

- It is defined as persistence, determination, resolution, tenacity, dedication, commitment, constancy, steadfastness, stamina, endurance and indefatigability. To persevere is described as to continue, carry on, stick at it (in formal), keep going, persist, plug away, (informal), remain, stand firm, stand fast, hold on and hang on. Perseverance builds character.

ACCURACY

- It means freedom from mistake or error; conformity to truth or to a standard or model and exactness. Accuracy is defined as correctness, exactness, authenticity, truth, veracity, closeness to truth (true value) and carefulness. The value of accuracy embraces a large area and has many implications. Engineers are encouraged to demonstrate accuracy in their behavior through the medium of praise and other incentives. Accuracy includes telling the truth, not exaggerating, and taking care over one's work.

DISCERNMENT

- It means discrimination, perception, penetration, and insight. Discernment means the power to see what is not obvious to the average mind. It stresses accuracy, especially in reading character or motives. Discrimination stresses the power to distinguish or select what is true or genuinely excellent. Perception implies quick and often sympathetic discernment, as of shades of feelings. Penetration implies a searching mind that goes beyond what is obvious or superficial.

Universal Values

- Responsibility
- Commitment
- Integrity
- Patriotism

One of the most important characteristics of moral judgements is that they express human values. Not all expressions of values are also moral judgements, but all moral judgements do express the value of human beings. Thus, understanding morality requires investigating what people value and why.

There are three principle types of values which humans can have: preferential values, instrumental values and intrinsic values.

Preference value, the expression of preference is the expression of some value people hod. When people say that they prefer to play sports, they are saying that we value that activity.

Instrumental values are values like ambition, courage, persistence, politeness etc. They are not the end but a mean of achieving terminal values.

Intrinsic value- something which has intrinsic value is valued purely for itself- it isn't used simply as a means to some other end and it isn't simply "preferred" above other possible options.

Evolution of Human Values

The human values evolve because of the following factors:

- The impact of norms of the society on the fulfilment of the individual's needs or desires.
- Developed or modified by one's own awareness, choice, and judgment in fulfilling the needs.
- By the teachings and practice of Preceptors (Gurus) or Savoir or religious leaders.
- Fostered or modified by social leaders, rulers of kingdom, and by law (government).

PERSONAL ETHICS

- Simply put, all individuals are morally autonomous beings with the power and right to choose their values, but it does not follow that all choices and all value systems have an equal claim to be called ethical.
- Actions and beliefs inconsistent with the Six Pillars of Character - trustworthiness, respect, responsibility, fairness, caring and citizenship - are simply not ethical.

PERSONAL ETHICS - Everyday examples

- Software piracy
- Expense account padding
- Copying of homework or tests
- Income taxes
- Borrowing nuts and bolts, office supplies from employer
- Copying of Videos or CD's
- Plagiarism
- Using the copy machine at work

RELIGION AND ETHICS

The "Golden Rule" is a basic tenet in almost all religions: Christian, Hindu, Jewish, Confucian, Buddhist, Muslim.

- "Do unto others as you would have others do unto you."
- "Treat others as you would like them to treat you" (Christian).
- "Hurt not others with that which pains you" (Buddhist)
- "What is hateful to yourself do not do to your fellow men" (Judaism)
- "No man is a true believer unless he desires for his brother that which he desires for himself" (Islam)

MORALITY AND ETHICS

- Concerns the goodness of voluntary human conduct that affects the self or other living things
- Morality (Latin *mores*) usually refers to any aspect of human action
- Ethics (Greek *ethos*) commonly refers only to professional behaviour
- Ethics consist of the application of fundamental moral principles and reflect our dedication to fair treatment of each other, and of society as a whole.
- An individual's own values can result in acceptance or rejection of society's ethical standards because even thoughtfully developed ethical rules can conflict with individual values.

ASPECTS OF ETHICS

There are two aspects to ethics:

- The first involves the ability to discern right from wrong, good from evil and propriety from impropriety.
- The second involves the commitment to do what is right, good and proper. Ethics entails action.

Morality is different from Ethics in the following ways:

S.NO	MORALITY	ETHICS
1	More general and prescriptive based on customs and traditions.	Specific and descriptive. It is a critical reflection on morals
2	More concerned with the results of wrong action, when done.	More concerned with the results of a right action, when not done.
3	Thrust is on judgment and punishment, in the name of God or by laws.	Thrust is on influence, education, training through codes, guidelines, and correction.
4	In case of conflict between the two, morality is given top priority, because the damage is more. It is more common and basic	Less serious, hence second priority only. Less common. But relevant today, because of complex interactions in the modern society
	Example: Character flaw, corruption, extortion, and crime.	Example: Notions or beliefs about manners, tastes, customs, and towards laws.

- As against morals and ethics, laws are norms, formally approved by state, power or national or international political bodies. Breaking the norms is called *crime*, and invites specific punishment.
- Our values associate emotions to our experiences and guide our choices, decisions and actions.
- A person's observations on its environment are filtered through his values to determine whether or not he should expend energy to do something about his experiences.
- A person who values gold and sees a large bag of gold (a positive value) in his path as he walks, will be motivated to reach down and pick it up.
- A person who values his life and knows about venomous snakes will retreat from the sound of a rattle snake (a negative value) from nearby, when he is walking in the desert. Said in another way,
- *"Values are the scales we use to weigh our choices for our actions, whether to move towards or away from something."*Not all values have the same weight or priority.
- **Dr. Abraham Maslow** illustrated this with his hierarchy of human needs. Survival has a higher priority than security, which has a higher priority than social acceptance. Self-esteem can only be addressed to the degree that social acceptance is fulfilled. Similarly, self-actualization can only be pursued to the degree that self-esteem has been satisfied.
- A person's beliefs, values and identity are usually acquired unconsciously based on his personal experience or observations of others' experiences as to what produces desirable or undesirable results in the environment.
- A baby's learning *to walk and talk* is a clear example of identifying with human adults, valuing the act of being able to have the mobility and communication ability of an adult and the belief, based on unconscious observation, that humans can do walk and do talk with each other.
- Physiologists have identified the parts of the human brain that are involved in producing behaviour in accordance with beliefs and values.

- All information collected by human senses is passed through a net-like group of cells, known as the Reticular Activating System (RAS), located near the top of the brain stem. The RAS compares the data received with accepted values, positive and negative (threats), and beliefs stored in memory and determines whether or not immediate action is required. The results of the RAS's comparison are communicated to the 'amygdala' near the mid-brain.
- The 'amygdala' produces neuro-chemicals that cause emotions consistent with the nature of and proportional to the match between environment and values and beliefs. The neuro-chemicals initiate the chemical processes needed for the action to be taken.
- If the emotions produced are strong enough, the perceived information is blocked from reaching the logical, rational and conscious executive centre of the brain, the pre-frontal lobes. In which case, the resulting behaviour will be automatic, not necessarily logical or rational, and completely in accordance with the person's strongest held beliefs, values and/or identity.
- By positive affirmations, one can modify or create new beliefs about a person's identity and/or what is important to him (values). Verbal repetition of statements intended to become new beliefs, and values will result in these being stored for use by the RAS for comparison with the environment being experienced. This is the mechanism how the beliefs or values are modified.

ETHICS

- The word "Ethics" has been derived from Greek word "ethos" which meant "customs". The word "Ethics has several meanings. Accordingly,
    1. It is an activity and area of inquiry. It is the activity of understanding moral values, resolving moral issues and justifying moral judgements.
    2. Ethics is the word that refers to morals, values, and beliefs of the individuals, family or the society. The word has several meanings. Basically it is an activity and process of inquiry. It is different from non-moral problems, when dealing with issues and controversies.
    3. It also refers to a particular set of beliefs, attitudes, and habits of individuals or family or groups concerned with morals. It is used to mean 'morally correct'.
    4. The study on ethics helps to know the people's beliefs, values, and morals, learn the good and bad of them, and practice them to maximize their well-being and happiness. It involves the inquiry on the existing situations, form judgments and resolve the issues.
- In addition, ethics tells us how to live, to respond to issues, through the duties, rights, responsibilities, and obligations.
- In religion, similar principles are included, but the reasoning on procedures is limited.
- The principles and practices of religions have varied from to time to time (history), region (geography, climatic conditions), religion, society, language, caste and creed. But ethics has grown to a large extent beyond the barriers listed above.

Engineering ethics

- It is the activity and discipline, aiming at understanding moral values that have to guide engineering practice, rsolving moral issues and justifying moral judgements in engineering.
- "Engineering Ethics" refers to the set of specific moral problems and issues related to engineering
- "Engineering Ethics" are accepted codes and standards of conduct, approved by various engineering societies.

Ethics can be classified into:

- **"Engineering Ethics"-** Related to engineers, engineering practice and industries
- **"Professional Ethics"** - Related to various professional like Doctors, Lawyers
- **"Business Ethics"** – Related to people involved in business

- **"Work Ethics" –** Related to employers and employees at work place

Mahatma Gandhi, the father of our nation insits the importance of values and ethics when he talks about seven social issues, which are as follows,

1. Wealth without work
2. Pleasure without conscience
3. Knowledge without character
4. Commerce (business) without morality (ethics)
5. Science without humanity
6. Religion without sacrifice
7. Politics without principles

According to Gandhiji, the above said seven social issues can help society, only if ethics is followed. In ethics, the focus is to study and apply the principles and practices, universally.

2. INTEGRITY

What the word "Integrity" means,

- The English word integrity is derived from latin word "integritas" which denotes "the diminished or unimpaired condition of a thing".
- A thing that has integrity is good and it need not be improved further.
- Hence the state of full integrity is a state that ought to be preserved if not preserved earlier.

Any action that leads to a loss of integrity is in need of justification.

Definition:

"Integrity is used to refer to people who act in ways that are consistent with their own code of principles. It is generally a personal choice to uphold oneself to consistently moral and ethical standards".

- It means "soundness of moral character".
- The condition of being free from defects or flaws: durability, firmness, solidity, soundness, stability, strength, wholeness.
- State of being entirely whole: completeness, Unity of thought, open mindedness and honest.

Integrity is one of the core qualities that any professional practitioners should possess. It also refers to honesty and open mindedness either with oneself or others. It generally involves the discovery of truth and its communication. It refers to the capacity to communicate the truth in proper manner so that it enables the client and others to make informed-decisions.

Integrity is one of the self-direction virtues on commitment and on putting understanding to action. It refers to the unity, which is a consistency among human attitudes, emotions and conduct in relation to justified moral values. Thus integrity acts as a link between responsibility in private and pubic life.

With the help of the integrity, the virtues of self-respect and pride in the job can be made possible. It prevents the attitude among the individuals that they are not responsible for their wrong doings in the job. It creates an enthusiasm among the individuals for achieving excellent performance in their job and also makes them to ensure that the job is done well.

Maintaining or practicing integrity needs courage. This courage is obtained when wisdom and integrity join hands. The integrity of the engineers is most essential in the following works:

- Engineering research and testing
- In the use of intellectual property
- Client professional confidentiality
- Expert testimonials and
- Failure to inform the public

Moral integrity is defined as a virtue, which reflects a consistency of one's attitudes, emotions, and conduct in relation to justified moral values.

Various types of integrity:

- ➢ Personal integrity
- ➢ Professional integrity
- ➢ Business integrity
- ➢ Academic integrity
- ➢ Research integrity

1. Personal Integrity:

- Person possessing integrity is honest, trustful and cannot be bribed. A person possessing integrity is trustworthy.

2. Professional Integrity:

- Integrity means moral wholeness a single sense of self across a wide range of circumstances.
- Professional Integrity is essential for maintaining excellence in their professions and for keeping the public's trust.
- For an engineer, professional integrity is the key under difficult situations. An engineer with professional integrity will be ready for any kind of challenge and additional responsibility.

3. Business Integrity:

- Business integrity is the reliability with which the business undertakes its traditions with the various parties.
- The values of honesty and integrity are the foundation of an organization's reputation. Employee working at company need to possess organizational integrity in addition to personal integrity.
- Organizational integrity assures all employees' combined efforts align with organizational values and commitments.

4. Academic Integrity:

- Honesty as an engineer begins with honesty in studying to become an engineer.

5. Research Integrity

- Research should be demonstrated honestly. For example, If an experiment is done, we should think right about it and also we should report what makes the experiment invalid. We need to explain all our results and also explain the things that can be eliminated in the experiment.

An engineer, as a professional, consultant, practicing engineer or manager, should be a person of integrity in discharging his duties and responsibilities to the society.

3. WORK ETHIC

Work ethic is a value based on hard work and diligence. It is also a belief in the moral benefit of work and its ability to enhance character.

Understanding work ethics

The term work ethic is very difficult to explain in just a couple of sentences. However, in simple words, work ethics are standard measures that control all functioning in a professional environment. This means, as an employee you are expected to be honest, sincere and diligent about the work we are assigned. The practices we opt for should be 'clean', acceptable and should abide by concerned laws.

On smaller grounds, you should be understanding, just and true towards other colleagues. As an employee, we should be perceived as hardworking, reliable, honest, diligent and worthy of the monetary compensation we receive.

There are six fundamental social challenges that face modern business. They are the need to achieve ecological balance, the human element in work, improving economic and social productivity, global pressures, demands and needs, balancing ethics and economics in business decision making, and to help in designing social partnerships for resolving societies major problem.

Five characteristics of a good work ethic

In organizations there may be two kinds of employee, some individuals try to get by doing as little work as possible; other possess a dedication that leads them to give it their all every day. People who possess a

strong work ethic embody certain principles that guide their work behaviour, leading them to produce high-quality work consistently.

The following five characters ensure the possession of work ethic in an employee:

a) Reliability
b) Dedication
c) Productivity
d) Co-operation
e) Character

Reliability

Reliability goes hand in hand with a good work ethic. If individuals with a good work ethic say they are going to attend a work function or arrive at a certain time, they do, as they value punctuality.

Individuals with a strong work ethic often want to appear dependable, showing their employers that they are workers to whom they can turn. Because of this, they put effort into portraying and proving this dependability by being reliable and performing consistently.

Dedication

Employees with a good work ethic are dedicated to their jobs and will do anything they can to ensure that they perform well. Often this dedication leads them to change jobs less frequently, as they become committed to the positions in which they work.

They also often put in extra hours beyond what is expected, and truly dedicate themselves to their positions.

Productivity

Since they work at a consistently fast pace, individuals with a good work ethic are often highly productive.

They commonly get large amounts of work done more quickly than others who lack their work ethic, as they don't quit until they've completed the tasks with which they were presented.

This high level of productivity shows the fact that these individuals want to appear to be strong workers. The more productive they are, the more beneficial to the company they appear to those managing them.

Cooperation

Cooperative work can be highly beneficial in the business environment, something that individuals with a strong work ethic know well. Because they recognize the usefulness of cooperative practices such as teamwork they often put an extensive amount of effort into working well with others.

These individuals commonly respect their bosses enough to work with any individuals with whom they are paired in a productive and polite manner, even if they do not enjoy working with the individuals in question.

Character

Individuals with a good work ethic often also possess generally strong character. This means they are self-disciplined, pushing themselves to complete work tasks instead of requiring others to intervene.

They are also often very honest and trustworthy.

To demonstrate their strong character, these workers embody these positives traits daily, likely distinguishing themselves from the rest.

Factors that demonstrate a strong work ethic

A work ethic is a set of moral principles an employee uses in his job. Certain factors come together to create a strong work ethic. To experience and maintain a good as well as strong work ethics in an organization, the following five mandatory factors are to be considered. These are the factors which influence an employee's sincerity, and dedication in the work assigned to him, also strengthens his career and adds in his responsibilities.

A strong work ethic can improve your career-A strong work ethic is vital to a company achieving its goals. Every employee, from the CEO to entry-level workers, must have a good work ethic to keep the company functioning at its peak.

Integrity- stretches to all aspects of an employee's job. An employee with integrity fosters trusting relationships with clients, co-workers and supervisors. Co-workers value the employee's ability to give honest feedback. Client trust the employee's advice. Supervisors trust on the employee's high moral standards and not to steal from the company or create problems.

Sense of Responsibility- A strong sense of responsibility affects how an employee works and the amount of work she does. When the employee feels personally responsible for her job performance, she shows up on time, puts in her best effort and completes projects to the best of her ability.

Emphasis on Quality- Some employees do only the bare minimum, just enough to keep their job intact. Employees with a strong work ethic care about the quality of their work. They do their best to produce great work, not merely churn out what is needed. The employee's commitment to quality improves the company's overall quality.

Discipline- It takes a certain level of commitment to finish our tasks every day. An employee with good discipline stays focussed on his goals and is determined to complete his assignments. These employees show a high level of dedication to the company, always ensuring they do their part.

Sense of Teamwork- Most employees have to work together to meet a company's objectives. An employee with a high sense of teamwork helps a team meet its goals and deliver quality work. These employees respect their peers and help where they can, making collaborations go smoother.

How to increase employee work ethic

A work ethic is typically something ingrained within a person. There are, simply put, lazy people who are impossible to motivate. However, other factors, both economical and psychological, can affect an employee's work ethic. Most people can be encouraged to greater performance, once the right motivating factors are found. This can be a process of trial and error because each individual may have different motivators. However, there are some basic guidelines we can follow to increase employee work ethic.

Step 1: Expect our managers to set a good example.

Make sure they are serving as role models for the rest of our employees.

Step 2: Create a public recognition system.

Rewarding an employee's good work ethics can be a great motivator for other employees who may not be as productive.

Employee of the month competitions and special rewards for those who do their job well may encourage those with a poor work ethic to try harder. Human beings thrive on recognition and feeling appreciated, and these are very powerful motivating factors.

Step 3: Set clear goals and milestones.

In some cases, employees may feel overwhelmed with a project if they are not entirely sure how to complete it or if it looks insurmountable.

Break apart projects into tasks that have clear goals. Set milestones with clear target dates so employees know exactly what we expect of them and how long they have to complete the task.

Step 4: Monitor potential troublemakers.

Almost every office has at least one person who is there for the pay check and not much else. These people can cause dissension among the ranks and bring down not only the morale of the rest of the staff, but also the productivity levels for the company.

Weed through new applicants to make sure they will have job dedication. Monitor current troublemakers, set strict guidelines they must follow or encourage them to seek employment elsewhere.

Step 5: create a monetary award system.

Some employees will be motivated only by the promise of receiving a bonus or a raise if they complete certain tasks and improve their performance. While not all companies may have the resources to give large monetary awards to their employees, even simple gift card challenges and free products can encourage lackadaisical employees.

Negative Work Ethic

Definition

Negative work ethics may be the behaviour of a single individual or something more systematic; regardless of the specifics, identifying the signs is the first step toward correcting it.

Companies like to promote positive work ethics because it needs results in happier and more productive employees, just as it is important to understand a positive work ethic, however, it is equally important to recognize the signs of a negative work ethic.

Negative Influences of Bad Work Ethics

Lack of Productivity

The most obvious sign of a negative work ethic is lack of productivity. Lack of productivity costs the company time and money: essentially paying the employee for doing nothing.

Attendance

A positive work ethic means showing up on time every time, and using sick days for their designated purpose than a vacation by proxy. A negative work ethic, on the other hand, looks to get the most out of the system.

A good worker, for instance, may arrive late every once in a while, but also stays late to make up the time. A bad worker will assume that showing up late is normal, and do so beyond the range of what the company considers acceptable.

Politics

Every company experiences a certain amount of office politics, as different departments compete for different resources enter into otherwise professional relationships.

Someone with a negative work ethic, however, may let office politics consume him: stoking the fire of

disconnect around a perceived rival and worried more about his comparative standing than the well-being of the company as a whole. Such employees might even instigate political crises forcing senior management to spend time and resources calming everyone down rather than getting along with the business at hand.

Esprit de Corps

A good company seeks to foster loyalty among their workers: making them feels like family members as much as employees toiling for a salary. Someone with a negative work ethic, however, fails to engage in office environment. It may be a repeated refusal to participate in company activities such as picnics or mixers.

How to Deal with Bad Work Ethics in Co-workers

One employee's bad work ethic can hamper productivity throughout the workplace. A work ethic is a set of values people have about the benefits and importance of working hard and being productive. Values are subjective, so a co-worker doesn't necessarily have a bad work ethic if his opinions about working aren't in line with ours.

However, we should address a fellow employee's work ethic if he's making it difficult for us to complete our job duties. Address the problem with the employee directly first, but sometimes we have to involve a manager.

Step 1

- Avoid the temptation to wait for a co-worker to figure out that his bad work ethic is affecting us.

- Speak to the co-worker in private, and explain the problem by giving specific examples of how his failure to complete work hampered our ability to get our job done.
- Explain the problem with a teamwork perspective, pointing out how he and others have an important role to fulfil in the workplace.

Step 2

- Find out if co-worker understands how to complete his assigned tasks when we discuss work-ethic problems with him.
- Employees sometimes get duties from managers that they don't have the skills to fulfil, so they avoid those duties.
- Recommend that a co-worker ask a manager for guidance or training on how to complete tasks he doesn't understand. Help him ourself if we can, but don't do his job for him.

Step 3

- Bear in mind that a co-worker might not be getting his job done because he has personal problems that are distracting him.
- Don't feel compelled to take on a co-worker's problems, but we can show understanding by giving him some slack on the job while he sorts out his troubles.
- In such cases, ask him to consider whether taking time off from work would be beneficial in tackling his problems.

Step 4

- Tell our manager about the problems a co-worker's bad work ethic is causing if our other efforts to help him fail.
- Don't make the issue personal when we tell our manager about the matter.
- Present our manager with business-related reasons the co-workers poor work habits are affecting the workplace.
- Consider things such as whether the co-worker's behaviour is creating a backlog of work for us and others.

Combination of mental characteristic and behaviour that distinguishes a person or a group also it is the combination of traits and qualities distinguishing the individual nature of a person or thing. Simply character refers to the distinguishing / unique nature of something.

Character deals with how people think and behave issues such as right and wrong, justice and equity, and other areas of human conduct.

Moral character or character is an evaluation of a particular individual's stable moral qualities. The concept of the character can imply a variety of attributes including the existence or lack of virtues such as empathy, courage, fortitude, honesty, and loyalty, or of good behaviours or habits.

Moral character primarily refers to the assemblage of qualities that distinguish one from another-although on a cultural level, the set of moral behaviours to which a social group adheres can be said to unite and define it culturally as distinct from others. Psychologist Lawrence Pervin defines moral character as "a disposition to express behaviour in consistent patterns of functions across a range of situations".

4. SERVICE LEARNING

- It is an educational method by which participants learn and develop through active participation in service that is conducted in and meets the needs of a community.
- The service learning includes the characteristics of the work, basic requirements, security of the job, and awareness of the procedures, while taking decisions and actions.
- Service learning may be defined as the non-paid activity, in which service is provided on voluntary basis to the public (have-nots in the community), non-profitable institutions, and charitable organizations. It is the service during learning.

- This includes training or study on real life problems and their possible solutions, during the formal learning. i.e., courses of study.
- It is a pedagogy that provided chances to directly interact with local agencies and effect change in the community.
- National youth leadership council defines service learning as a "philosophy, pedagogy and model for community development that is used as an instructional strategy to meet the goals.
- In the industrial scenario, adoption, study, and development of public health or welfare or safety system of a village or school is an example of service learning by the employees.

KEY COMPONENTS

Service-learning combines experimental learning and community service opportunities. It can be distinguished in the following ways:

- **Curricular connections** - Integrating learning into a service project is a key to successful service-learning. Academic ties should be clear and build upon existing disciplinary skills.
- **Student voice** - Beyond being actively engaged in the project itself, students have the opportunity to select, design, implement and evaluate their service activity, encouraging relevancy and sustained interest. In community settings, this is alternatively called youth voice.
- **Student discussion** - students discuss their learning experience during in-class discussions.
- **Reflection** – The balance of reflection and action allows the trainee to be constantly aware of the impact of their work.
- **Community partnership**- Partnership with community agencies are used to identify genuine needs, provides mentorship, and contributes input such as labor and expertise towards completing the project.
- **Authentic community**- local community members or service recipients are involved in determining the significance and depth of the service activities involved.
- **Assessment**- well structured assessment with constructive feedback through reflection provides valuable information regarding service learning which aids sustainability and replication of services.

In 2008, the National Youth Leadership council released service Learning standards for Quality practice which are as follows:

- Meaningful service
- Link to curriculum
- Reflection
- Investigation
- Diversity
- Partnership
- Progress monitoring
- Action
- Demonstration
- Recognition

To distinguish high quality from low quality service learning experiences, Youth service California has

published "seven elements of high quality service learning" as follows:

- Integrated Learning
- High quality service
- Collaboration
- Student voice
- Civic responsibility
- Reflection
- Evaluation

What are the benefits of service learning?

- Identifies and researches local needs or issues
- Combines academic curriculum with service
- Motivates participants to make a difference in their communities
- Opportunities to address community needs and issues
- Develops responsible citizens
- Fosters a sense of caring for others
- First-hand experience with economic, social, cultural, and political contexts and factors
- Greater depth of understanding by connecting course work to the issues and concerns in the community
- Opportunities to contribute to the mission and/or purpose of an agency through service and volunteerism.
- Experience in working with diverse communities
- Understanding how the non-profit, government or educational sector functions

TYPES OF SERVICE LEARNING

i. **Direct service learning:**
Person-to-person, face-to-face projects in which service impacts individuals to receive direct help from students. Eg. Tutoring, working with elders

ii. **Indirect service learning:**
Projects with benefits to a community as opposed to specific individuals. Eg. Food and clothing drives, environmental cleaning.

iii. **Advocacy service learning:**
Working, acting, speaking, writing, teaching, presenting, informing etc. on projects that encourage action or create awareness on issues of public interest. Eg. Disaster preparedness, care for environment, violence and drug prevention, promoting reading, safety of oneself.

iv. **Research service learning:**
Surveys, studies, evaluations, experiments, data gathering, interviewing etc. to find and report information on topics in the public interest. E.g. Water testing, flora and fauna studies.

EFFECT OF SERVICE LEARNING ON ENGINEERING EDUCATION

- Many engineering educators see service learning as the solution to several problems in engineering education today.
- Recently, a change in engineering curriculum is done to focus more on practical aspects of engineering.
- Educators see service learning as a way to both implement a constructivism in engineering education as well as match the teaching styles to the learning styles of engineering students.
- As a result, many engineering schools have begun to integrate service learning into their curricula and there is now a journal dedicated to emphasize the importance of service in engineering.

5. CIVIC VIRTUES

- Civic comes from a Latin word “civitas” which means “civilized” or “living in city”. Now think of the word citizen. Virtue comes from the Latin word “virtus” which means being “moral” or “good”.
- Virtues are positive and preferred values. Virtues are desirable attitudes or character traits, motives and emotions that enable us to be successful and to act in ways that develop our highest potential.
- Virtues are tendencies which include, solving problems through peaceful and constructive means and follow the path of the golden mean between the extremes of ‘excess and deficiency’. They are like habits, once acquired, they become characteristics of a person.

- Civic virtues are the moral duties and rights, as a citizen of the village or the country or an integral part of the society and environment.
- An individual may exhibit civic virtues by voting, volunteering, and organizing welfare groups and meetings. Individually any one knows what is good and what is bad. In spite of this, mostly people act unethically due to the following reasons:
    1. Unawareness
    2. Insensitivity of issues
    3. Selfishness
    4. Faulty reasoning
    5. Pressure
- The duties are:
    1. To pay taxes to the local government and state, in time.
    2. To keep the surroundings clean and green.
    3. Not to pollute the water, land, and air by following hygiene and proper garbage disposal. For example, not to burn wood, tyres, plastic materials, spit in the open, even not to smoke in the open, and not to cause nuisance to the public, are some of the civic (duties) virtues.
    4. To follow the road safety rules.
- On the other hand, the rights are:
    1. To vote the local or state government
    2. To contest in the elections to the local or state government.
    3. To seek a public welfare facility such as a school, hospital or a community hall or transport or communication facility, for the residents.
    4. To establish a green and safe environment, pollution free, corruption free, and to follow ethical principles. People are said to have the right to breathe in fresh air, by not allowing smoking in public.
    5. People have inalienable right to accept or reject a project in their area. One has the right to seek legal remedy, in this respect, through public interest petition.

George Washington embodied the civic virtues as indispensable for a self-governing administration. These virtues are divided into four categories:

1. **Civic Knowledge**
    - Citizens must understand what the Constitution says about how the government is working, and what the government is supposed to do and what not to do.
    - We must understand the basis of our responsibilities as citizens, besides duties and rights.
    - We must be able to recognize when the government or another citizen infringes upon our rights.
    - It implies that the government requires the participation of the enlightened citizens, to serve and survive.
2. **Self-Restraint**
    - For citizens to live in a free society with limited government each citizen must be able to control or restrain himself; otherwise, we would need a police state—that is, a dictatorial government to maintain safety and order.
    - He advocated for morality and declared that happiness is achieved and sustained through virtues and morals.
    - He advocated and demonstrated self-restraint several times in his private and public life, and naturally he was a great leader.
3. **Self-Assertion**
    - Self-assertion means that citizens must be proud of their rights, and have the courage to stand up in public and defend their rights.
    - Sometimes, a government may usurp the very rights that it was created to protect.

- In such cases, it is the right of the people to alter or abolish that government (e.g., voting rights, rights call back).

4. Self-Reliance

- Citizens who cannot provide for themselves will need a large government to take care of them.
- Once citizens become dependent on government for their basic needs, the people are no longer in a position to demand that government act within the confines of the Constitution.
- Self-reliant citizens are free citizens in the sense that they are not dependent on others for their basic needs.
- Only a strong self-reliant citizenry will be able to enjoy fully the blessings of liberty.
-

6. RESPECT FOR OTHERS

"The true measure of a man is how he treats you when others are not looking."

Showing other people respect is a critical part of maintaining important personal relationships. Learning to respect people's efforts, abilities, opinions and quick will help keep us happy and successful in our interpersonal life. Respecting ourself can help us move forward with the confidence to make a habit of respect and share it with the people around us.

Treating people with respect makes a nicer place to live in. it is very easy. All you have to do is, "treat people the way you like others to treat you". Here are a few ideas

- Don't insult people or make fun of them
- When you speak, listen to others
- Value other peoples opinion
- Be considerate of peoples likes and dislikes
- Don't tease or harass people
- It is not good to talk about people behind their back
- Have respect to other peoples feelings
- Don't compel anybody to do something he/she does not want to do.

Respect for other individuals can be shown in many forms as given above, but the following four ways may teach us how to respect others' opinions, also the need of self-respect in the society, in an organization, wherever people being together while working, living, and when they meet publicly.

1. Showing gratitude
2. Respecting others opinions
3. Respecting our enemies
4. Respecting ourself.

1. Showing gratitude:

Thank people for their assistance and their support on a regular basis. It is important to remember all the people who've helped us on your journey.

Show respect by saying thanks. Remember to thank our parents, siblings, co-workers, classmates, friends, teachers, neighbours.

a) Remember to speak politely to everyone:

Compliment the achievements of others. When others are successful, draw attention to it and celebrate their ability and their achievement. Learn to recognize when other people put forth extra effort and achieve something and praise them for it with sincerity.

b) Be sincere:

Always be sincere in our every work, whether it is in our work place or home. Sincere work definitely gets its reward one day. Sincerity is the first step among the steps to success in life.

c) Respect the abilities of others:

Try to recognize when someone is capable of doing something on his or her own and mind our own business to show that person the respect he or she deserves.

2. Respecting others opinion:

a) Be a good listener:

Practice active listening to show people that we have respect for their opinions and ideas. Watch and be quiet when someone else is talking and spend time thinking actively about what they're saying.

b) Ask lots of questions:

To show respect for other people's opinions, question them. Ask open-ended, leading questions that show you're fully engaged with their ideas and that we are listening closely.

c) Learn about the perspectives of others:

Learning to empathize with other people who have very different experiences and perspectives than our own will help us learn to show respect. Be proud of our own opinions and perspectives, but don't assume everyone feels the same way and avoid spotting them in an awkward position.

d) Respectfully disagree:

When we have to dissent, do it calmly and by treating our conversation with tact. Respect the perspective of the other person. Don't insult their opinion or ideas, even if we disagree with them.

3. Respecting your enemies:

a) Don't judge people before we get to know them:

Give people the benefit of the doubt, even people of whom we might have a bad first impression.

b) Decide to like people:

It's too easy to come up with reasons to dislike someone, to disrespect someone, or to dismiss them. Decide to like them, and it'll be much easier to show respect.

c) Worry about our own backyard:

Don't get mixed up in other people's business and create unnecessary enemies.

4. Respecting ourself:

a) Take care of ourself:

To show respect for ourself, try and give ourself the same consideration that you give everyone else.

b) Avoid self-destructive behaviours:

Drinking to excess on a regular basis or habitually self-deprecating ourself will tear us down in mind and body. Try to work actively to build ourself up and surround ourself with encouraging, enlightening, helpful people.

d) Stay healthy:

Make regular visits to the doctor to make sure we are healthy and fit. Exercise regularly and eat well. Start developing easy routines, even walking a few miles a day or doing some light stretches to get in touch with our body and maintain it. Cut out junk foods and eat a variety of nourishing foods.

e) Be ambitious:

Develop plans for ourself and specific steps for carrying them out. Plot an upward trajectory for ourself to keep ourself moving forward in life and staying satisfied. Show respect for ourself by being the best version of ourself we can be.

As per Kant's argument one should never treat people merely as things. However they should be treated as autonomous moral arguments. Capitalism and advance in technology, forces to think of people merely as things. People deserve respect because they have the worth of rational beings inherently, they have the capacity for autonomy i.e., governing lives on the basis of moral principles, and for exercising a good will i.e., the careful effort to do what is right.

7. LIVING PEACEFULLY

➢ The world of peace, purity and prosperity is usually remembered with words such as paradise, heaven and so on. In the world of peace, human beings are like flowers, a country is like a bouquet of flowers and the world is like a garden of flowers.

➢ To live peacefully one should install peace within self. Charity begins at home. Then one can spread peace to family, organization where one works and then to the world including the environment.

➢ Whatever may be the job or profession one should have good environment and working atmosphere to do his job and carry out his responsibilities?

➢ There should not be any tension or over pressure, unnecessary interference or disturbance from others though they are superiors/seniors.

➢ The workers should have peaceful atmosphere at office as well as at home. The rules and regulations must also be to a limited extent in order to make the atmosphere peaceful.

➢ The layout of the business /industries, security of the job and environment with safety are also required for any professionals to live peacefully.

➢ Peace is inner silence filled with power of truth. Peace consists of pure thoughts, pure feelings and wishes. When the energy of thought, word and action is balanced one is said to be in peace.

➢ To lead peaceful life, one requires "Peace of mind". Peace of mind leads to "living peacefully" by all. If one wants to live peacefully, he/she has to follow certain priciples in his/her life. They are:

➢ One has to believe in God

➢ Home has to be made a place of friendliness, refreshment and peace.

➢ One has to be patient and considerate towards others.

➢ We have to work towards removal of social injustice.

➢ Work towards reconciliation between individual, groups and nations, is indeed.

➢ One has to behave in a loving way towards all men and women.

➢ We have to have a caring and loving attitude towards others.

➢ One has to be conscious of his/her daily living

➢ One has to play his/her role against an form of exploitation and oppression

➢ One has to live as simple as possible

➢ We have to serve others in avoiding any form of violence

➢ Nurture

1. Discipline, duty and self regulation
2. Pure thoughts in ones soul
3. Creativity in one's head
4. Beauty in ones heart

➢ GET

1. Good health/Body

➢ ACT

1. Not hurting and torturing others

➢ Helping the needy with head, heart and hands

The following are the factors that promote living peacefully

1. Healthy family situations and labour relations
2. Service to the needy with love and sympathy
3. Secured job and motivational speech.
4. Conductive environment.

8. CARING

Definitions

- ✓ Caring is feeling for other. It is a process which exhibits the interest in the welfare of others with fairness, impartiality and justice in all activities among the employees.
- ✓ It includes showing respect to the feeling of others.
- ✓ Caring is reflected in activities such as friendship, membership in social clubs and professional societies and through various transactions in the family, fraternity, community, country and in international councils.
- ✓ It is a process and product which incorporates supports, sharing and respect. It encompasses the unity of mind, body and spirit of the holistic person with the broader content of ones environment.

As a normal human being when an individual is dealing or moving with neighbours, friends, colleagues, even with their family members, they must have some interests about the welfare of the other persons at least to some extent. These type of caring is essential in the work place too. Caring for others and having interest in the solution of their grievances will definitely bring in a good work environment. This type of adjustment among the workers or between the executives and subordinates in the work spot is also necessary for the successful implementation of the workload assigned. This type of morality of care, leads to concentration rather than impartiality and justice. The individual with justice orientation and caring in any dispute will be interested bothering about the impact on others. But the individual with care orientation will try to identify the best course of action that preserves the interest of all those people involved. Even actions taken by such care-orientated people will have least amount of damage to the relationships among the person.

Caring means, to show a responsible and responsive attitude as a father would do towards his children. There is more gratification in being a "caring person" than in just being a nice person. A caring attitude builds good will which is the best kind of insurance that a person can have. Caring does not cost anything.

Engineers, Professionals, Lawyers, Doctors, Managers should adopt "caring" attitude towards society people, to justify their positions. Caring attitude will boost the efficiency and productivity of a company. Nowadays caring for the environment including the fauna and flora has become a necessity for our survival.

9. SHARING

- ✓ Sharing is a process that describes the transfer of knowledge, experience, commodities and facilities with others.
- ✓ The transfer should be genuine, legal, positive, voluntary and without any expectation in return.
- ✓ Through sharing, experience, expertive, misdom and other benefits reach more people faster.
- ✓ As we all know, happiness is multiplied and suffering is reduced by sharing.
- ✓ Sharing paves the way for peace.
- ✓ Commercially speaking, the profit is maximized by sharing.
- ✓ Technologically the productivity and utilization are maximized by sharing.
- ✓ Code sharing is one of the best examples of sharing which improves productivity.

In the industrial arena, code-sharing in airlines for bookings on air travels and the common Effluent Treatment Plant constructed for small-scale industries in the industrial estates, are some of the examples of sharing. The co-operative societies for producers as well as consumers are typical examples of sharing of the goods, profit and other social benefits.

One story to illustrate the importance of sharing is as follows:

The shouting...the screaming…the fighting. That was the breaking point for me as I poured out my woes to my mother. "How can I get them to *share* as well as we did as kids?" I pleaded Laughter was her reply. "Well, thanks a lot, mom," I said. "I'm sorry," she chuckled, "but you didn't always share." She went

on to explain about the "**Box of Misbehaved Toys**." Every time we fought over a toy, she would quietly take that and put it into the box.

Yes, I did remember that box. I also remember it wasn't always fair since one person may have caused all the commotion. But my mother was consistent. No matter what the reason for the struggle was, the toy disappeared into the box for one week. No questions asked, and no chance of parole.

My siblings and I soon learned that sharing a toy was better than losing it. Often, one person would decide to just wait for a time when no one else was playing with the toy, rather than fight and lose it. It was not a perfect system, but I tried it anyway That box was a shock to my kids and it was close to full, within a few days…..As the weeks progressed, I noticed the box was emptier and the arguing was less. Today, I heard quiet music to my ears as my son said to his sister, "That's OK, you can play with it."

This story illustrates the worthy joy of sharing as compared to the pain of losing.

10. HONESTY

✓ Honesty is the human quality of communicating with a truthful,direct and complete intent.It is related to truth as a value.

✓ It includes both honesty to others and to oneself and about ones own motives and inner reality.Honesty is a virtue and it is exhibited in two aspects namely

a)Truthfullness
b)Trustworthiness

a) **Truthfulness** is to face the responsibilities upon telling the truth. One should keep ones word or promise. By admitting ones mistake committed, it is easy to fix them.
Reliable engineering judgement, maintenance of truth, defending the truth, and communicating the truth only when it does good to others are some of the reflections of truthfulness.

b)**Trustworthiness** is maintain integrity and taking responsibility for personal performance. People abide by law and live by mutual trust they build trust through reliability and authenticity.Rhey admit their own mistakes.

Honesty is mirrored in many ways. The common reflections are

a)Beliefs
b)Communication
c) Decision
d) Actions
e) Intended and unintended results achieved

The value of the engineering services depends on honesty. Unreliable engineering judgement will be the worst. Honesty also refers to the maintenance of truth or not to misuse the truth.

Some of the actions of an engineer that leads to dishonesty are as follows.

1. Lying

Means a person happened to be intentionally with less knowledge or less awareness, communicating wrong or misguided information..

2. Deliberate deception:

With insufficient data or proof, an engineer may judge or decide just to impress customers or employers.

3. Withholding the information:

It means hiding the facts during communication to ones superior or subordinate intentionally.(An engineer during the proposals to his executive ,fails to indicate some of the

negative aspects of the project,even though he is not lying,dishonesty may be considered as a form of omission.)

4. Maintaining confidentiality:

Engineers should not discuss some confidential information without the knowledge of the clients.Mostly such Confidential information may be either,informing to the engineer by client or finding out by the engineer during the process of the work carried out by the client.

Honesty normally includes the activities like –not liking,not stealing,not involving in bribes and kickbacks.It refers to paying respect to the property of others.

- **Honesty in Beliefs**: It denotes intellectual honesty(forming of one's beliefs without self deception).
- **Honesty in Speech**:It refers to the action of not deceiving or not intentionally misleading others. For instance, acts like pretending, manipulating somebody's attention, intentionally lying, misleading and withholding some pertinent information which someone or the client has to know.
- **Honesty in Act**:It means that the individual should not steal,or manipulate accounts,or get bribes and kickbacks.
- **Honesty in discretion**:It means that an employee should not interfere with the decisions of the employer or the client.He should not interfere with the confidential matters.

11. COURAGE

Courage in the tendency to accept and face risks and difficult tasks in rational ways and with self control.courage is classified into three types:

- ➢ Physical courage
- ➢ Social courage
- ➢ Intellectual courage

- In **physical courage**, the thrust is on the adequacy of the physical strength. People with high adrenalin may be prepared to face thrilling challenges and will take excellent decisions.
- The **social courage** involves the changes in the decisions and actions, based on the conviction for or against certain social behaviours. This requires leadership abilities and empathy to motivate the followers, for the social cause.
- The **intellectual courage** is inculcated in people through acquired knowledge and experience.
- One should perform SWOT analysis, SWOT strength, weakness, opportunity and threat. These will help anyone to face the future with courage.
- Opportunities and threat existing and likely to exist in future are also to be studied and measures to be planned. This anticipatory management will help anyone to face the future with courage.
- Facing the criticism, owing responsibility and accepting the mistakes are the expressions of courage.
- For instance a fire-fighter courageously runs into a burning building because he or she is protecting life and property. Part of his courage comes from his duty to his job and community but the rest comes from a courageous instinct that kicks in.
- Ethics is more than just following a set of rules; it is a part of our deeply-held belief system that makes-up the core of our character. It is worth stepping out in courage and making personal sacrifices.

Prof.Sathish Dhawan, chief of ISRO was reported for his courage of owing responsibility when the previous space mission failed, but credited Prof. A.P.J.Abdul Kalam when the subsequent mission succeeded.The courageous people own the following characteristics.

- ✓ Perseverance (sustained hard work)
- ✓ Experimentation (preparedness to face unexpected / unintended results)

✓ Involvement

✓ Commitment

STEPS TO DEVELOP COURAGE:

- Learn to face Reality
- Set a good value system and commit to adopt it.
- Build character
- Practice makes things permanent.

12. VALUING TIME

- Time is rare resource. Once it is spent, it is lost forever. It cannot be stored or removed.
- The history of great reformers, innovators and achievers stressed the importance of time. The proverbs "time and tide wait for nobody" and "procrastination is the chief of time" illustrates this point.
- It refers to making the best use of time as time is always limited.
- Effective time management allows individuals to assign specific time slots to activities as per their importance.
- An anecdote to highlight the value of time is as follows:
  - ➢ To realize the value of one year, ask the student who has failed in the examinations.
  - ➢ To realize the value of one month, ask the mother who has delivered a premature baby.
  - ➢ To realize the value of one week, ask the editor of weekly.
  - ➢ To realize the value of one day, ask the daily wage labors.
  - ➢ To realize the value of one hour, ask the lovers longing to meet.
  - ➢ To realize the value of one minute, ask a person who has missed the train.
  - ➢ To realize the value of one second, ask the person who has survived an accident.
  - ➢ To realize the value of one millisecond, ask the person who has won the bronze medal in Olympics.
  - ➢ To realize the value of one microsecond, ask the NASA team of scientists.
  - ➢ To realize the value of one nanosecond, ask a hardware engineer.
- So, as an engineer, we should realize the value of time.
- Time management includes the following activities:
  - ✓ Effective planning of a task
  - ✓ Setting goals and objectives
  - ✓ Setting deadlines to finish your works
  - ✓ Delegation of responsibilities
  - ✓ Prioritizing activities as per their importance
  - ✓ Spending the right time on the right activity

13. CO-OPERATION

According to the ethical principles, cooperation has to exist in all respects between the superiors and the subordinates, among the workers and between industry and the customers. Lack of co-operation leads to lack of communication, unavoidable delays, and finally may lead to collapse of the design and planning. If proper co-operation is not maintained in business or industries, it leads to lot of frustrations among the employees and finally an entire loss to the society. In such case it is not easy to establish Total Quality Management in the system.

- It is a team spirit present with every individual engaged in engineering.
- Willingness to understand others, think and act together and putting this into practice is cooperation.
- It helps in minimizing the input resources and maximizes the output.

- Cooperation should be developed and maintained at several levels:
  - ➢ Between the employers and employees

- ➢ Between the supervisors and sub ordinates
- ➢ Among the colleagues
- ➢ Between the producers and the suppliers
- ➢ Between it organization and its customers

- The absence of cooperation leads to lack of communication and delay between production and supply. This leads to collapse of the industry over time and an economic loss to the society.
- The various factors affecting successful co operation are
  - ➢ Clash of ego of individuals
  - ➢ Lack of leadership
  - ➢ Conflicts of interests lack of interest

TYPES OF CO-OPERATION

- **Direct Co-operation: I**t implies direct relationship among the individuals. In this type, people do like things together. E.g.: playing together, working together, worship together, ploughing the field together etc.
- **Indirect Co-operation**: Here people do different tasks towards a similar end. E.g.: in a college the principal, lecturers, office assistants, accountants, typist, librarian perform different functions but they make co-operative effort towards a common goal.

A.W.Green has divided co-operation into three types such as:

- **Primary Co-operation:** It is that type of Co-operation in which there is no selfish interest. Every member works for the betterment of all.
- **Secondary Co-operation:** here individual co-operates with others for the achievement of some selfish interests. Each may work in co-operation with others for his own status, power and prestige.
- **Tertiary Co-operation:** different groups make mutual adjustments with each other under certain compelling circumstances. It is purely voluntary in nature. People or groups co-operate with each other according to their sweet will. The attitudes are very opportunistic and selfish.

14. COMMITMENT

- Commitment means alignment to goals and adherence to ethical principles.
- A promise is also a form of commitment by someone to do or not do something.
- An ethical commitment is like a duty or a moral obligation.
- Commitment is the basic requirement for any profession
  For example,
- A design engineer shall exhibit a sense of commitment to make his product as a beneficial contribution to the society.
- Only when the teachers is committed to his job the students will succeed in life and contribute good to the society
- The commitment of top management will naturally lead to committed employees. This is bound to add wealth to oneself one's employer society and the nation at large.

Organizational Commitment and Professional Commitment
Organizational Commitment:

- It is made up of more factors such as faith and acceptance of the organizations set of values and objectives, employee's wish to strive for the organization and a strong will to keep working within it.
- It predicts work variables such as turnover, organizational citizenship behavior and job performance.
- It can be contrasted with other work-related attitudes, such as job satisfaction, defined as an employee's feelings about their job, and organizational identification.

Professional Commitment:

- It signifies an attitude reflecting the strength of the bound between an employee and an organization.
- Teaching is a profession which needs utmost commitment since a teacher not only teaches a student the subjects also he/she train them to behave morally and mopuld them into a perfect individual in a society.
- The quality on teaching depends on the level of teachers involvement in relation to the profession exerted.

Three-component model of Organizational Commitment:

- In 1990, based on observations of several types of organizations. Meyer and Allen develop "The model of three components of organizational commitment"
  - ➢ Affective commitment
  - ➢ Continuity commitment
  - ➢ Normative commitment
- **Affective commitment** is based on the individual's identification with and involvement in the organization.
- It is an emotional commitment, where people that are in a great deal affectively connected to an organization stay within it because they want to.
- **Continuity commitment** is based upon the material and psychological costs involved by ones leaving the organization, people with such kind of commitment remaining within it because they are compelled to do so.
- **Normative commitment** is based upon an ideology or a sense of obligation towards the organization, on the individual's moral belief that it is right and moral to continue within the organization.
- People keep staying within an organization because they think they should. This feeling ob obligation is the result of internalizing the norms exerted on the individual.

Organizational Commitment may be determined by two categories of factors:

- **Individual ones:** where we may include variables of inclinations like professional values, type of personality and demographic variables such as age, gender, educational level, marital status.
- **Organizational factors** such as structure of the job, type of organization, professional experience etc.

15. EMPATHY

- Empathy is social radar. Sensing what others feel about, without their open talk, is the essence of empathy.
- Empathy begins with showing concern, and then obtaining and understanding the feelings of others, from others' point of view.
- It includes the imaginative projection into other's feelings and understanding of other's background such as parentage, physical and mental state, economic situation, and association. This is an essential ingredient for good human relations and transactions.
- Empathy it is distinct from sympathy, pity and emotional contagion.
- To practice 'Empathy', a leader must have or develop in him, the following characteristics:
  - ➢ **Understanding others:** It means sensing others feelings and perspectives, and taking active interest in their welfare.
  - ➢ **Service orientation:** It is anticipation, recognition and meeting the needs of the clients or customers.

➢ **Developing others:** Getting the feedback, acknowledge the strength and then coach the individual by giving correct feedback.

➢ **Leveraging diversity:** This leads to enhanced organizational learning, flexibility, and profitability.

➢ **Political awareness:** It is the ability to read political and social currents in an organization.

- The benefits of empathy include:
  - ➢ Good customer relations (in sales and service, in partnering).
  - ➢ Harmonious labor relations (in manufacturing).
  - ➢ Good vendor-producer relationship (in partnering)
- Through the above three, we can maximize the output and profit, as well as minimizing the loss.
- While dealing with customer complaints, empathy is very effective in realizing the unbiased views of others and in admitting one's own limitations and failures.
- According to Peter Drucker, purpose of the business is not to make a sale, but to make and keep a customer. Empathy assists one in developing courage leading to success.
- **Empathy and Trust** are two major platforms for effective understanding and proper communication.
- **Empathy** is important in developing certain solutions, trials to win and retaining in business and to avoid conflicts.
- **Empathy with trust** is essential for dealing situations with complaints and retaining the customers and ultimately to develop a successful personal and business relationship.

TYPES OF EMPATHY:

- **Affective empathy also called emotional empathy:** the capacity to respond with an appropriate emotion to another's mental states.
- It is subdivided into following scales such as **Empathic concern** which means sympathy and compassion for others in response to their suffering and **personal distress** which is a self-centered feelings of discomfort and anxiety in response to another's suffering.
- **Cognitive empathy**: the capacity to understand another's perspective or mental state.
- It is subdivided into following scales such as **Perspective taking** which is a tendency to spontaneously adopt others psychological perspectives and **Fantasy** which is a tendency to identify with fictional characters.

16. SELF CONFIDENCE

✓ People with self confidence feel that they are equal to others,even in the situations when the others are in positions with better and greater formal power.

✓ The people with self confidence also recognizes the value of building the self confidence of others and normally would not be threatened by doing it so.

✓ Thus self confidence in everyone develops a sense of partnership, respect and accountability and helps the company to get maximum efforts and ideas and guidelines from everyone.

➢ Certainty in one's own capabilities values and goals in self confidence.

➢ These people are usually positive thinking flexible and willing to change. They respect others.

➢ Self confidence in positive attitude, where in the individual has some positive and realistic view of himself.

➢ People with self confidence exhibit courage and unshakable faith in their abilities, whatever may be their positions.

➢ They are not influenced by threats and are prepared to face them for any unexpected consequences.

➢ The people with self confidence have the following characteristics:

- Self-assured standing
- Willing to listen and learn from others

- Frankness to speak the truth
- Respect others efforts

➢ On the contrary, some leaders expose others when failure occurs and own the credit when success comes.

➢ The factors that shape self confidence in a person are
  - Heredity and family environment
  - Friendship
  - Influence of superiors
  - Training in the organization

➢ Self confidence or having belief in our self has been directly connected to an individual's social network, the activities they participate in, and what they hear about themselves from others.

Positive factors of self confidence have been linked to factors such as psychological health, mattering to others, and both body image and physical health.

➢ On the contrary, low confidence level in adolescents has been shown to be an important predictor of unhealthy behaviours and psychological problems such as suicidal ideation later in life.

➢ Self confidence can vary and be served in a variety of dimensions. An individual's self-confidence can vary in different environments, such as home or in school/workplace.

➢ During adolescence confidence level of students is affected by age, race, ethnicity, puberty, health, body, and weight, and body image, involvement in physical activities, gender presentation, and gender identity. Components of one's social and academic life affect their self-confidence.

STEPS TO BUILD OUR SELF-CONFIDENCE:

Step 1: Changing our perspective
- Identify our talents
- Take pride in our good qualities
- Recognize our insecurities and discuss it with our friends
- Bounce back from our mistakes
- Adapt a more positive mind-set
- Stop comparing our self to others

Step 2: Taking actions
- Accept compliments gracefully
- Help others
- Stick to our principles
- Get rid of as many sources of negativity as we can
- Make eye contact when we talk to people
- Put care into our appearance

Step 3: Continuing to build our confidence
- Avoid perfectionism since it paralyses us and keep us away from attaining our goals.
- Always be thankful for what we have
- Address our own addressable flaws
- Spend time with people who make we feel good
- Live in the present moment – Yoga and meditation can help us live in the present moment and to get more in touch with our mind body.

17. CHARACTER

Character is the combination of personal qualities or features that make each person unique. Teachers, parents and community members help children build positive character qualities. For example, the six pillars of character are

✓ Trustworthiness

- ✓ Respect
- ✓ Responsibility
- ✓ Fairness
- ✓ Caring
- ✓ Citizenship.

Ethics is the study of human actions. It deals with issues such as defining, "right and wrong" as well as the grey area in between. Ethics seeks answers to questions like what is "good behaviour" and what should be valued?

Schools often have character education programs that focus on the qualities of character that are honoured by most cultures and traditions. Character education is the development of knowledge, skills, and abilities that encourage children and young adults to make informed and responsible choices.

Ethics are a philosophical reflection of moral beliefs and practices. The Greek and Roman philosophers were particularly interested in discussions related to ethics. Religions and faiths each have their own ethical systems to guide their people . Ethical decision making involves the process of making informed decisions when faced with difficult dilemmas with many alternative solutions.

Characteristics of Ethical people in the Workplace

Of the many characteristics that business looks for in candidates, ethics is one of the most important. Human resources officials commonly seek individuals who possess highly defined ethics, as a strong ethical base improves the likelihood that the worker is a productive. Many of the characteristics associated with an ethical individual are desirable ones that hope to have in their workforce.

Honesty

Ethical workers value honesty and are honest at all cost. This means that they remain honest even when being honest isn't the easiest road to take.

For example, if an ethical employee makes a mistake, he does not lie about the situation in an attempt to make himself seem less culpable. Having an employee who is overtly honest allows management trust the employee more implicitly and reply upon him.

Responsibility

Workers who are take ethical responsibility seriously and do all them can to complete the tasks with which they are charged.

Reliability

When ethical team members say they are going to do something, they follow through. They are reliable at all times and can be trusted to complete projects of great importance.

Goal- Oriented

Ethical individuals are often goal-focused and able to dedicate themselves fully to their job tasks. These individuals often recognize the importance of working to better them and improve the overall success of their company; they are willing to work toward reaching potentially challenging goals.

Job-Focused

Ethical employees remain focused on their jobs at all times, not allowing themselves to become distracted, as doing so pulls them away from the duties of their occupations. These individuals are never found working on a task that is not related to the job in question; as they recognize that their on- the –job is to be spent only doing job related tasks.

18. SPIRITUALITY ("S E NSE OF S E L F")

Spirituality is a process of personal transformation, either in accordance with traditional religious ideals, or, increasingly, oriented on subjective experience and psychological growth independently of any specific religious context.

In a more general sense, it may refer to almost any kind of meaningful activity or blissful experience. It still denotes a process of transformation, but in a context separate from organized religious institutions, termed "spiritual but not religious".

What is spirituality?

Spirituality has many definitions, but at its core spirituality helps to give our lives context. It's not necessarily connected to a specific belief system or even religious worship. Instead, it arises from our connection with ourself and with others, development of our personal value system, and our search for meaning in life.

For many, spirituality takes the form of religious observance, prayer, meditation or a belief in a higher power. For others, it can be found in nature, music, art or a secular community. Spirituality is different for everyone.

Meaning of spirituality

One of a great gift of spiritual knowledge is that it realigns our sense of self to something we may not have even ever imagined was within us. Spirituality says that even if we think we are small and limited, it simply isn't so. We are greater and more powerful than we have ever imagined. A great and divine light exist inside of us. This same light is also in everyone we know and in everyone we will ever know in the future. We may think we are limited to just our physical body and state of affairs-including our gender, race, family, job and status in life-but spirituality comes in and says " There is more than this".

Notice that spirit sounds similar to words like inspire and expire. These are two of the main themes of spiritual journey:

- ✓ ***Inspiring spirituality***: Allowing us to be filled with inspiration, this also translates into love, joy, wisdom, peacefulness and service.
- ✓ ***Expiring spirituality***: Remembering that an inevitable expiration waits to take us away from the very circumstance we may think are so very important right now.

Spirituality differs from religion

Although religion and spirituality are sometimes used interchangeably, they really indicate two different aspects of the human experience. We might say that spirituality is the mystical face of religion.

Spirituality is the wellspring of the divinity that pulsates, dances and flows as the source and essence of every soul. Spirituality relates more to your personal search, to finding greater meaning and purpose in your existence. Some elements of spirituality include the following:

- Looking beyond outer appearance to the deeper significance and soul of everything.
- Love and respect for god.
- Love and respect for ourself.
- Love and respect for everybody

Religion is most often used to describe an organized group or culture that has generally been sparked by the fire of a spiritual or divine soul. Religions usually act with a mission and intention of presenting specific teaching and doctrines while nurturing and propagating a particular way of life.

Spirituality is:

- Beyond all religions yet containing all religions.
- Beyond all science yet containing all science.
- Beyond all philosophy yet containing all philosophy.

Spirituality and stress relief

Some stress relief tools are very tangible: exercising more, eating healthy foods and talking with friends. A less tangible but no less use useful way to find stress relief is through spirituality.

How can spirituality help with stress relief?

Spirituality has many benefits for stress relief and overall mental health.

Feel a sense of purpose. Cultivating your spirituality may help uncover what's most meaningful in your life. By clarifying what's most important, you can focus less on the unimportant things and eliminate stress.

Connect to the world. The more you feel you have a purpose in the world, the less solitary you feel even when we are alone. This can lead to a valuable inner peace during difficult times.

Release control. When we feel part of a greater whole, we realize that we aren't responsible for everything that happens in life. We can share the burden of tough times as well as the joys of life's blessings with those around us.

Lead a healthier life. People who consider themselves spiritual appear to be better able to cope with stress and heal from illness or addiction faster.

Discovering your spirituality

Uncovering your spirituality may take some self-discovery. Here are some questions to ask ourself to discover what experiences and values define us:

- What are our important relationships?
- What do we value most in your life?
- What people give us a sense of community?
- What inspires us gives you hope?
- What brings us joy?
- What are our proudest achievements?

The answers to such questions help us identify the most important people and experiences in our life. With this information, we can focus our search for spirituality on the relationships and activities in life that have helped define us as a person and those that continue to inspire our personal growth.

19. INTRODUCTION TO YOGA AND MEDITATION FOR PERSONAL EXCELLENCE AND STRESS MANAGEMENT

- Yoga and meditation when practiced together strengthen the mind body connection, improving overall fitness and well being.
- Many styles of yoga combine meditation with the physical routines, which use controlled breathing throughout the yoga poses.
- We can meditate without practicing yoga by simply relaxing, clearing our mind and concentrating on controlled breathing. Both yoga and meditation, when used consistently, have proven health benefits.

YOGA-DEFINITION:

Yoga is one of the most popular fitness practices around the world,. yoga is the act of harmonising your body, mind and soul. Yoga is a spiritual practice that has nothing to do with religion. Yoga has been mistaken as a product of Hinduism, just because it took birth in the same region as the religion. Yoga is about complete transformation - physically, psychologically and spiritually. This means, that yoga has the power to transform your mind, body and soul.

History and Origin:

The word yoga is derived from the Sanskrit word yuj, which means 'to join' or 'to unite'. According to Yogic scriptures, practicing yoga results in the union of individual consciousness with universal consciousness, bringing about a perfect harmony between the mind and the body and man and nature. Yoga is now a part of UNESCO's Intangible Cultural Heritage list.

Benefits of practicing Yoga

1. **Increased flexibility**
   - ➢ Yoga poses focus on stretching and lengthening the muscles.
   - ➢ Increased flexibility will help us with daily movements such as lifting and bending.
   - ➢ It reduces body aches and joint pains.

2. **Emotional boost**

- ➢ Both yoga and meditation improve mental focus and provide a general feeling of well being.
- ➢ Yoga makes you breathe slowly, and focus on the present. It shifts the balance from the sympathetic nervous system to the parasympathetic nervous system which is calming and restorative. It lowers breathing and heart rates, reduces blood pressure and increases blood flow to the intestines and reproductive organs.
- ➢ In 2012, a study published in "Alternative Therapies in Health and medicine" found yoga participants happy and peaceful. Meditation provides an emotional boost through deep relaxation and it can be done anywhere. We can give ourself an emotional boost by taking a 10 minute meditation break right at our desk.

3. **Better diet**

Studies suggest that practicing yoga improves fitness and body awareness, leading to better eating habits. This in turn leads to increased self esteem and the desire to take care of our body.

4. **Improved health**

- ➢ Adding yoga or meditation to our life will improve the quality of our life. Improved health means we can participate in more physical activities and just feel better in the things we do daily.
- ➢ The exercises in yoga help you relax and gradually increase blood circulation, especially in your hands and feet. Yoga also pumps more oxygen to your cells, which function better consequently.
- ➢ Yoga also boosts levels of haemoglobin and red blood cells, which carry oxygen to the tissues.

TYPES OF YOGA:

1.SIMHASANA: It requires your face and body to resemble a lion's roar, and anyone with any level of fitness can do this asana. Kneel down on the floor and cross your ankles. Put your palms on the knees and spread them out, pressing the fingers firmly against each knee. Inhale deeply, open your mouth and stick your tongue out, towards the chin. Keep your eyes wide open and contract the muscles in front of the throat. Exhale through your mouth, making a distinct sound.

Benefits:

Tones facial muscles, works as a stress-buster, reduces double chin.

2.TADASANA: It requires one to stand erect like a mountain. This is the starting position for all yoga poses. Stand with your heels slightly apart and let your arms hang beside the torso. Gently lift and spread your toes and the balls of your feet, then place them back softly on the floor. Balance your weight equally on your feet. Lift your ankles and tighten your thigh muscles while rotating them inwards. Elongate your torso as you inhale. Broaden your collarbone and elongate your neck. Your ears, shoulders, hips and ankles should be in one line. Check your alignment by standing against a wall. Breathe easy.

Benefits:Works major muscle groups, increases focus and concentration.

3.VRIKSHASANA: Put your right foot high up on your left thigh. The sole of the foot should be flat and placed firmly. Keep your left leg straight and balance yourself. Inhale and raise your arms over your head, and bring your palms together. Keep your spine straight. Take deep breaths. Exhale slowly, and bring your hands and foot down. Repeat with the other foot.

Benefits: Improves balance, strengthens legs and back.

4.NAUKASANA: Lie with your back on the mat. Keep your feet together and your hands by your side. Take a deep breath. While exhaling, gently lift your chest and feet off the ground, stretching your arms towards your feet. Ensure that your eyes, fingers and toes are in one line. Hold till you feel your ab muscles contract. Then exhale and relax.

Benefits: Tightens abdominal muscles, strengthens shoulders and upper back, provides stability.

5.PASCHIMOTTANASANA: Sit on the floor in Padmasana. Keep the spine erect, and stretch your legs out to your front so that your toes point to the ceiling. Take deep breaths and stretch your hands above your head without bending your elbows. Stretch your spine as much as you can. Breathe out, and bend forward from your thighs. Bring your hands down and try to touch your toes. Hold the position and try pulling your toes backwards till you feel a stretch on your hamstrings. Breathe in, hold the position for 60 to 90 seconds. Exhale and go back to the starting position.
Benefits:Stimulates the centre of the solar plexus, tones the tummy, stretches hamstrings, thighs and hips.

6.PAVANAMUKTASANA: Lie flat on your back, with your arms beside you and legs straight. Inhale as you bend your knees and bring them towards your chest as you exhale. Let your thigh put pressure on the abdomen. Hold the knees properly in place by clasping your hands underneath the thighs. Inhale again, and as you exhale, lift your head, allowing your chin to touch your knees. Hold the position for 60 to 90 seconds, while breathing deeply. Exhale slowly, and release your knees while allowing your head to rest on the floor. Bring your hands onto either side of your body, palms facing the ground. Relax.

Benefits:Reduces gastric problems, triggers fat burning, strengthens back and abdominal muscles, tones muscles in the legs and arms.

7.UTTANPADASANA: Lie down on your back, legs stretched out, and ankles touching each other. Keep your arms close to your body, palms facing down. Take a deep breath. Now breathe out slowly, tilt backwards so that your head touches the floor. Stretch as much as you can. Inhale deeply and raise your legs, making a 45-degree angle with the floor. Hold for 15 to 30 seconds, breathing normally. Exhale deeply, and lift your legs to make a 90-degree angle with the floor. Breathing normally, hold the posture for 30 seconds. Again take a deep breath and slowly bring your legs back to the starting position.
Benefits: Helps reduce fat from lower abdominal region, hips and thighs, cures back pain, improves functioning of reproductive organs, improves blood circulation.

8.KUMBHAKASANA: Start in a pose similar to a push-up with your arms extended under your knees and hands under your shoulders and arms . Breathe in, keep your back and spine straight. Now keeping your hands flat and your fingers spread , pull in your abdominal muscles. Hold for 15 to 30 seconds.
Benefits:Helps reduce fat from thighs, buttocks, shoulders, back and the belly area.

What is Meditation?

Meditation is a practice that has been associated with almost all religions and civilizations across the world. Since it is so closely associated with religion, many people take meditation to be the same thing as praying.

Benefits of Meditation

Meditation has two important benefits:

1. Meditation prevents stress from getting into the system.
2. Meditation releases accumulated stress that is in the system.

Both of these happen simultaneously, leaving one refreshed and joyful.

Physical Benefits of Meditation

With Meditation, the physiology undergoes a change and every cell in the body is filled with more prana (energy). This result in joy, peace, enthusiasm as the level of prana in the body increases.

1. Lowers high blood pressure.
2. Reduces anxiety attacks.
3. Decreases any tension related pain such as tension headaches, ulcers, insomnia, muscle and joint problems.

4. Increases serotonin production that improves mood and behavior.
5. Improves the immune system.
6. Increases the energy level as we gain an inner source of energy.

Mental Benefits of Meditation

Meditation brings the brainwave pattern into an Alpha state so that the mind becomes fresh, delicate and beautiful.

With regular practice of meditation,

1. Anxiety decreases
2. Emotional stability improves
3. Creativity increases
4. Happiness increases
5. Intuition develops
6. Gain clarity and peace of mind
7. Problems become smaller
8. Meditation sharpens the mind by gaining focus and expands through relaxation.
9. A sharp mind without expansion causes tension, anger and frustration.
10. An expanded consciousness without sharpness can lead to lack of action.
11. The balance of a sharp mind and expanded consciousness brings perfection.
12. Meditation makes us aware so that our inner attitude determines our happiness.

Other Benefits of Meditation

1. Emotional steadiness and harmony
   It calms us whenever we feel overwhelmed, unstable or emotionally shut down.
2. Meditation brings harmony to the creation.
3. Consciousness evolves
   - ➢ With the assimilation of meditation into daily life, our consciousness evolves.
   - ➢ As time goes on, we are able to experience the higher and refined states of consciousness.
   - ➢ When our consciousness evolves and expands, the disturbances in our life become negligible. We start living in the moment and let go of the past.
4. Personal Transformation
   - ➢ Meditation can bring about a true personal transformation. The questions that arise in the mind are as follows:
     1. What is the meaning of Life?
     2. What is the purpose of Life?
     3. What is this world?
     4. What is love?
     5. What is knowledge?
   - ➢ As we live through answering them, we will witness that life is transformed to a richer level.

Stress Management

- ➢ Modern life is full of stressful situations, fatigue from long hours and little sleep, allergies, anxiety disorders etc.
- ➢ Regular yoga practice helps to reduce stress responses in our body, according to a study in the 2010 issue of "Psychosomatic Medicine".
- ➢ Meditation is also an effective stress reducer that is used to reduce anxiety, panic disorders and agoraphobia, an anxiety disorder.
- ➢ The rest obtained by meditation is deeper than the rest obtained by sleep. The deeper our rest, the more dynamic our activity would be.

UNIT -2

ENGINEERING ETHICS

Introduction

Engineering ethics is the activity and discipline aimed at

(a) Understanding moral values that ought to guide engineering profession or practice,
(b) Resolving moral issues in engineering, and
(c) Justifying the moral judgments in engineering.

It deals with set of moral problems and issues connected with engineering. Engineering ethics is defined by the codes and standards of conduct endorsed by engineering (professional) societies with respect to the particular set of beliefs, attitudes and habits displayed by the individual or group.

Another important goal of engineering ethics is the discovery of the set of justified moral principles of obligation, rights and ideas that ought to be endorsed by the engineers and apply them to the concrete situations.

Engineering is the largest profession and the decisions and actions of engineers affect all of us in almost all areas of our lives, namely public safety, health, and welfare.

IMPORTANCE OF ENGINEERING ETHICS

Engineering ethics is important both in contributing to safe and useful technological products and in giving meaning to engineers.

The direct aim is to increase one's ability to reason clearly and carefully about moral questions. Improving the ability to reflect carefully on moral issues can be accomplished by improving various practical skills that will help produce autonomous thought about moral issues.

- Moral awareness: Proficiency in recognizing moral problems and issues in engineering.
- Moral Reasoning: Comprehending, Clarrifying and assessing arguments on opposing sides of moral issues.
- Moral imagination: Discerning alternative responses to moral issues and receptivity to creative solutions for practical difficulties.
- Moral Communication: Precision in the use of a common ethical language, a skill needed to express and support ones moral views adequately to others.
- Moral Reasonableness: The willingness and ability to be morally reasonable.
- Respect for persons: Genuine concern for the well being of others as well as oneself.
- Moral hope: Enriched appreciation of the possibilities of using rational dialogue in resolving moral conflicts.
- Integrity: Maintaining Moral integrity and integrating ones professional life and personal convictions.

SCOPE:

The scope of engineering ethics is as follows:

a) Ethics of a workplace which involves the co-workers and employees in an organization.

b) Ethics is related to the product or work which involves the transportation, warehousing and use, besides the safety of the end product and the environment outside the factory.

c) Personal meaning and commitments matter in engineering ethics, along with principles of responsibility that are stated in codes of ethics and are incumbent on all engineers.

d) Promoting responsible conduct is even more important than punishing wrong – doing.

e) Ethical dilemmas arise in engineering, because moral values are myriad and can conflict.

f) Engineering projects are social experiments that generate both new responsibilities and risks and engineers share responsibility for creating benefits, preventing harm and pointing out dangers.

g) Moral values permeate all aspects of technological development, and hence ethics and excellence in engineering go together.

h) Engineering ethics should explore both micro and macro issues, which are often connected.

APPROACH

There are conventionally two approaches in the study of ethics:

1. **Micro ethics**-Which deals with decision and problems of individuals, professionals, and companies.

2. **Macro ethics**-Which deals with the societal problems on a regional/national level. Macro ethics concern more global issues, such as directions in technological development, the laws that should not be passed and the collective responsibilities of groups such as engineering professional societies and consumer groups.

For example: Global issues, collective responsibilities of groups such as professional societies and consumer groups.

1. Sense of Engineering Ethics

In the first sense, engineering ethics consists of the responsibilities and rights that need to be endorsed by those engaged in engineering and also desirable ideals and personal commitments in engineering.

In a second sense, ethics is the study of morality; it studies which actions, goals, principles, policies and laws are morally justified. Engineering ethics is the study of the decisions, policies and values that are morally desirable in engineering practice and research.

These two senses are normative. They refer to justified values and choices to things that are desirable.

There are two different senses (meanings) of engineering ethics, namely

- **Normative sense**
- **Descriptive sense**

The **normative sense** includes the following:

1. Knowing moral values, finding accurate solutions to moral problems and justifying moral judgments in engineering practices.

2. Study of decisions, policies, and values that are morally desirable in the engineering practice and research, and Using codes of ethics and standards and applying them in their transactions by engineers.

The Descriptive sense-refers to what specific individual or group of engineers believe and act, without justifying their beliefs or actions.

2. Variety of Moral Issues

Engineering disasters draw the attention and caution about compromising safety beyond the level of acceptable risk. Examples are as follows,

- Challenger explosion,
- Accident at nuclear plants and Chernobyl and Three Mile Island ,
- Accident at Chemical plant at Bhopal,
- The Exxon Valdez oil spill,
- Natural disasters such as earthquakes in Gujarat and Iran,
- Indiscriminate us of asbestos.

Micro-Ethics: Typical and everyday problems in an engineer's life or an entire engineering office.

Macro-Ethics: Societal problems which are not addressed until they unexpectedly resurface on a regional or national scale.

Major ethical issues:

The decision to report to higher authorities demands careful judgement about the nature of the moral problems and about the consequences of such a decision. Some of the examples are as follows,

- A junior engineer on a building site was given the task of developing plans, which should be checked by a senior. But the plans were not checked and simply stamped and sent for work. The junior engineer reported this to his immediate superior but he dismissed the protest. Where the junior engineer should should go now?
- A civil engineer frequently walked through a building site for which he had no responsibility but noted several problems with safety and with the quality of work. What should he do?
- A senior consultant engineer who followed the correct procedures in the development of major city center building has come to know that there were serious design faults in the foundation of the building. What should he do?
- These few cases demonstrate that the critical issues are involved. The need to report may involve criticism of the actions of colleagues of an organization.
- Whose responsibility is it to report to higher authorities?
- What will be the effects on the professional, his family and career, on the organization, etc.?
- Are there any alternatives to such reporting?

To whom he should report-to his professional body, to government, to senior management, to concerned pressure groups or to the press.

The principal features of such reporting are as follows,

- ✓ The information is conveyed outside approved organizational channels and the person may be under pressure from supervisor not to report it.
- ✓ The information sent to a group may be new to them
- ✓ The information may be a significant moral problem concerning the organization. For example, Criminal behavior, injustice to workers, threats to public safety etc.,

Analysis of issue in ethical problems:

The first step in solving any ethical problem is to completely understand all the issues involved. Once these issues are determined, a solution to the problem becomes apparent.

The issues involved in understanding ethical problems can be split into three categories,

- ✓ Normative
- ✓ Factual
- ✓ Conceptual

The reasons for people including the employers and employees, behaving unethically may be classified into three categories:

1. Resource Crunch:

Due to pressure, through time limits, availability of money or budgetary constraints, and technology decay or obsolescence.
Pressure from the government to complete the project in time (e.g., before the elections), reduction in the budget because of sudden war or natural calamity (e.g., Tsunami) and obsolescence due technology innovation by the competitor led to manipulation and unsafe and unethical execution of projects.
Involving individuals in the development of goals and values and developing policies that allow for individual diversity, dissent, and input to decision –making will prevent unethical results.

2. Opportunity:

(a) Double standards or behavior of the employers towards the employees and the public.

(b) Management projecting their own interests more than that of their employees. Some organizations over-emphasize short-term gains and results at the expense of themselves and others.

(c)Emphasis on results and gains at the expense of the employees,

(d) Management by objectives, without focus on empowerment and improvement of the infrastructure.

3. Attitude:
Poor attitude of the employees set in due to
(a) Low morale of the employees because of dissatisfaction and downsizing,

(b) Absence of grievance redressed mechanism,

(c) Lack of promotion or career development policies or denied promotions,

(d) Lack of transparency,

(e) Absence of recognition and reward system, and

(f) Poor working environment

Giving ethics training for all, recognizing ethical conduct in work place, including ethics in performance appraisal, and encouraging open discussion on ethical issues, are some of the directions to promote positive attitudes among the employees. To get firm and positive effect, ethical standards must be set and adopted by the senior management, with input from all personnel.

3. Types of Inquiries

Engineering ethics combines inquiries into values, meanings, and facts.

- ✓ **Normative inquiries**, which are most central, seek to identify the values that should guide individuals and groups.
- ✓ **Conceptual inquiries** seek to clarify important concepts or ideas, whether the ideas are expressed by single words or by statements and questions.
- ✓ **Factual or descriptive inquiries** seek to provide facts needed for understanding and resolving value issues.

Normative Inquiries

- Engineering ethics involves normative inquiries aimed at identifying and justifying the morally desirable norms or standards that ought to guide individuals or groups.
- Normative questions are about what ought to be and what is good. Some examples of normative questions:

(i) How far does the obligation of engineers to protect public safety extended in given situations?

(ii) When, if ever, should engineers be expected to blow the whistle on dangerous practices of employers for whom they work?

(iii) Whose values ought to be primary in making judgments about acceptable risks in a design for a public transport system: those of management, senior engineers, government, voters, or some combination of these?

(iv) Which particular laws and organizational procedures affecting engineer are morally warranted?

(v) What moral rights should engineers be recognized as having in order to help them fulfill their professional obligations?

For these practical questions, normative inquiries have the theoretical goal of justifying particular moral judgment.

Conceptual Inquiries

Conceptual inquiries are directed toward the clarifying the meaning of concepts, principles, and issues in engineering ethics. For example,

(i) What does “safety” mean and how is it related to idea of “risk”?

(ii) What does it mean when codes of ethics say engineers should “protect the safety, health and welfare of the public”?

(iii) What is a bribe?

(iv) What is a profession and what defines professionals?

Conceptual issues have to do with the meaning or applicability of an idea.

In engineering ethics, this might mean defining what constitutes a bribe as opposed to an acceptable gift, or determining whether certain business information is proprietary.

In the case of bribe, the value of the gift is probably a well-known fact. What is not known is whether accepting it will lead to unfair influence on a business decision. For example, conceptually it must be determined if the gift of tickets to a sporting event by a potential supplier of parts for your projects is meant to influence your decision or just is a nice gesture between friends. Of course, like

factual issues, conceptual issues are not always clear-cut and will often result in controversy.

Once the factual and conceptual issues have been resolved, at least to the extent possible, all that remains is to determine which moral principle is applicable to the situation. Resolution of moral issues is often more obvious. Once the problem is defined, it is usually clear which moral concept applies and the correct decision becomes obvious. In our example of a "gift" offered by a sales representative, once it is determined whether it is simply a gift or is really a bribe, then the appropriate action is obvious. If we determine that it is indeed a bribe, then it cannot ethically be accepted.

Factual Inquiries

Factual inquiries – also called descriptive inquiries- seek to uncover information bearing upon value issues.Where possible, researchers attempt to conduct factual inquiries using proven scientific techniques.

They provide important information about the business realities of contemporary engineering profession, the effectiveness of professional societies in fostering moral conduct, the procedures used in making risk assessments, and psychological an understanding of the background conditions that generate moral problems.

It is also enables us to deals realistically with alternative ways of resolving those problems. Factual inquiries involve what is actually known about a case, i.e., what the facts are. Although this concept seems straightforward, the facts of a particular case are not always clear and may be controversial. An example of facts that are not necessarily clear can be found in the controversy in contemporary society regarding abortion rights. There is great disagreement over the point at which life begins, and at which point a fetus can be legally protected.

In engineering, there are controversies over facts as well. For example, global warming is of great concern to society as we continue to emit greenhouse gases into the atmosphere. Greenhouse gases, such as carbon dioxide, trap heat in atmosphere. This is thought to lead to a generalized warming of the atmosphere as emissions from automobiles and industrial plants increase the carbon dioxide concentration in the atmosphere. This issue is of great importance to engineers since they might be required to design new products, or redesign old ones to comply with stricter environmental standards, if this warming effect indeed proves to be a problem.

However, the global warming process is only barely understood and the need to curtail emission of these gases is a controversial topic. If it were known exactly what the effects of emitting greenhouse gases into the atmosphere would be, the engineer's role in reducing this problem would be clearer.

4. Moral Dilemmas

Ethical dilemmas or moral dilemmas are situation in which moral reasons come into conflict, or in which the applications of moral values are problematic, and it is not immediately obvious what should be done.

Examples

Case study 1:

Generally, Justice is defined as speaking the truth and paying one's debts. Socrates quickly refutes this account by suggesting that it would be wrong to repay certain debts – for example, to return a borrowed weapon to a friend who is not in his right mind. Now there is a conflict between two moral norms:

(i) Repaying one's debts
(ii) Protecting others from harm.

And in this case, Socrates maintains that protecting others from harm is the norm that takes priority.

Case study 2:

When the Sartre's student is informed that his brother had been killed in the German offensive, he wanted to avenge his brother and to fight forces that he regarded as evil.

But the student's mother was living with him, and he was her consolation in life. The student believed that he had conflicting obligations. Sartre describes him as being torn between two kinds of morality; one of the limited scopes but certain efficacy, personal devotion to his mother; the other of much wider scope but uncertain efficacy, attempting to contribute to the defeat of an unjust aggressor.

Concept of Moral Dilemma

What is common to the two well-known cases is conflict. In each case, an agent regards herself as having moral reasons to do each of two actions, but doing both actions is not possible. Ethicist has called situations like these ***moral dilemmas***.

The crucial features of a moral dilemma are these:

(i) The agent is required to do each of two actions;
(ii) The agent can do each of the actions; but the agent cannot do both of the actions.

The agent thus seems condemned to moral failure; no matter what she does, she will do something wrong.

It is more important to protect people from harm than to return a borrowed weapon. And in any case, the borrowed item can be returned later, when the owner no longer poses a threat to others. Thus, in this case we can say that the requirement to protect others from serious harm ***overrides*** the requirement to repay one's debts by returning a borrowed item when it's owner so demands.

When one of the conflicting requirements overrides the other we do not have a **genuine moral dilemma**. In order to have a **genuine moral dilemma**, it must be true that neither of the conflicting requirements be overridden.

<u>Problems</u>

The conflicts in both the case studies arose because there is more than one moral precept. More than one precept sometimes applies to the same situation, and in some of these cases the precepts demand conflicting actions. One obvious solution here would be to arrange the precepts **hierarchically.**

However, there are at least two problems with this obvious solution.

1. First it just does not seem credible to hold that moral rules and principles should be hierarchically ordered. While the requirements to keep one's promises and to prevent harm to others can clearly conflict with each other, it is far from clear that one of these requirements should ***always*** prevail over the other. In case study 1, the obligation to prevent harm is clearly stronger.
2. The ***second problem*** with this easy solution is deeper. Even if it were plausible to arrange moral precepts hierarchically, situations can arise in which the same precept gives rise to conflicting obligations. To illustrate this, let us consider case study 3.

Case study 3:

Sophie and her two children are at a Nazi concentration camp. A guard confronts Sophie and tells her that one of her children will be allowed to live and one will be killed. But it is Sophie who must decide which child will be killed.

Sophie can prevent the death of either of her children, but only by condemning the other to be killed. The guard makes the situation even more excruciating by informing Sophie that if she chooses neither, then both will be killed. With this added factor, Sophie has a morally compelling reason to choose one of her children. But for each child, Sophie has an apparently equally strong reason to save him or her. Thus the same moral precept gives rise to conflicting obligations. These cases are termed as ***symmetrical.***

Dilemmas and Consistency

What is troubling is that theories that allow for dilemmas fail to be ***uniquely action-guiding***. A theory can fail to be uniquely action guiding in either of two ways:

- By not recommending any action in a situation that is moral

Or

- By recommending incompatible actions.

Theories that generate genuine moral dilemmas fail to be uniquely action-guiding.

Moral Residue

In case study-1, suppose the student joins the Free French forces. It is likely that he will experience remorse or guilt for having abandoned his mother.

Suppose if he is staying with his mother and not joined the Free French forces, he also would have appropriately experienced remorse or guilt.

Thus, the agent faces a genuine moral dilemma.

In case study- 3, no matter which of her children Sophie saves, she will experience enormous guilt for the consequences of that choice.

In these cases, proponents of the argument from moral residue must claims that four things are true.

(1) When the agent acts, she experiences remorse or guilt;

(2) That she experiences these emotions is appropriate and called for;

(3) Had the agent acted on the other of the conflicting requirements, she would also have experienced remorse or guilt;

(4) In the latter case, these emotions would have been equally appropriate and called for.

In these situations, then remorse or guilt will be appropriate no matter what the agent does, and these emotions are appropriate only when the agent has done something wrong. Therefore, these situations are genuinely dilemmatic.

Types of Moral Dilemmas

One distinction is between ***epistemic*** conflicts and ***ontological*** conflicts.

***Epistemic* conflicts**: It involves conflicts between two moral requirements and the agent does not know which of conflicting requirements takes precedence in her situation. Everyone concedes that there can be situations where one requirement does take priority over the other with which in conflicts, though at the time action is called for, it is difficult for the agent to tell which requirements prevails.

Ontological* conflicts:** These are conflicts between two moral requirements, and neither is overridden. This is not simply because the agent does not ***know which requirement is stronger-neither is. Genuine moral dilemmas are ontological.

Another distinction is between ***self-imposed*** moral dilemmas and dilemmas imposed on an agent ***by the world,*** as it were.

***Self-imposed moral dilemmas*:** This conflict arises because of the agent's wrongdoing. If an agent made two promises that he knew were in conflict, then, throw his own actions he created a situations in which it is not possible for him to discharge both of his requirements.

Dilemmas imposed on the agent by the world: It does not arise because of the agent's wrongdoing.

E.g. Sophie's choice of selecting one child.

Yet another distinction is between ***obligation dilemmas*** and ***prohibition dilemmas.***

Obligation dilemmas: These are situations in which more than one feasible action is obligatory.

E.g. Sartre's student

Prohibition dilemmas: These are the cases in which all feasible actions are forbidden.

E.g. Sophie's choice

5. Moral Autonomy

Moral autonomy is defined as, decisions and actions exercised on the basis of moral concern for other people and recognition of good moral reasons. Alternatively, moral autonomy means 'self determinant or independent'.

The autonomous people hold moral beliefs and attitudes based on their critical reflections rather than on passive adaption of the conventions of the society or profession. Moral autonomy may also be defined as skill and habit if thinking rationally about the ethical issues, on the basis of moral concern.

Viewing engineering as social experimentation will promote autonomous participation and retain one's professional identity. Periodical performance appraisals, tight-time schedules and fear of foreign competition threatens this autonomy.

It appears that the blue-collar workers with the support of the union can adopt better autonomy than the employed professionals. Some of the instances are as below;

(i) A steel plant worker who refused to dump oil into a river in an unauthorized manner was threatened with dismissal, but his union saw to it that the threat was never carried out.

(ii) In the case of automobile plant inspector who repeatedly want his supervisors about poorly

welded panels that allowed carbon monoxide from the exhaust to leak into the cab. Receiving no satisfactory response from the company, he blew the whistle. The company wanted to fire him, but pressure from the union allowed him to keep his job.

Only recently the legal support has been obtained by the professional societies in exhibiting moral autonomy by professionals in this country as well as in the west.

The engineering skills related to moral autonomy are listed as follows:

1. Proficiency in recognizing moral problems in engineering and ability to distinguish as well as relate them to problems in law, economics, and religion,

2. Skill in comprehending clarifying and critically assessing arguments on different aspects of moral issues.

3. Ability to form consistent and comprehensive view points based on facts,

4. Awareness of alternate response to issues and creative solutions for practical difficulties,

5. Sensitivity to genuine difficulties and subtleties, including willingness to undergo and tolerate some uncertainty while making decisions,

6. Using rational dialogue in resolving moral conflicts and developing tolerance of different perspectives among morally reasonable people, and

7. Maintaining moral integrity

- ***Autonomy*** which is the independence in making decisions and actions is different from authority.
- ***Authority*** provides freedom for action, specified within limits, depending on the situation.
- ***Moral autonomy*** and ***respect for authority*** can coexist. They are not against each other.
- If the authority of the engineer and the moral autonomy of the operator are in conflicts, a consensus is obtained by the two, upon discussion and mutual understanding their limits.

6. Moral Development Theories

1. KOHLBERG THEORY:

The concept of moral autonomy has been discussed with the help of kohlberg's theory.Moral autonomy is based on the psychology of moral development.This theory is based on the pioneering work in the area of moral development put forth by Lawrence Kohlberg. According to this theory,there are three levels of moral development ,which is based on the kinds of reasoning and motivation adopted by individuals with regard to moral questions.

- Pre-conventional,
- Conventional, and
- Post-conventional

In the pre-conventional level, right conduct for an individual regarded as whatever directly benefits oneself.At this level, individuals are motivated by obedience or the desire to avoid punish mentor to satisfy their own needs or by the influence by power on them. All young children exhibit this tendency.

At the conventional level, people respect the law and authority. Rules and norms of one's family or group or society is accepted, as the standard of morality.Individuals in this level want to please or satisfy, and get approval by others and to meet the expectations of the society rather than their self interest (e.g., good boy, good girl). Loyalty is regarded s most important. Many adults do not go beyond this level.

At the post-conventional level, people are called autonomous. They think originally and want to live by universally good principles and welfare of others. They have them do unto you'. They maintain moral integrity, self-interest. They live by principled conscience. They follow the golden rule, 'Do unto others as you would have them do unto you'. They maintain moral integrity, self-respect and respect for others.

Kohlberg believed that individuals could only progress through these stages, one stage at a time. He believed that most of the moral development occurs through social interactions.

Kohlberg's six stages of moral development:

Level	Stage	Social orientation
Pre-Conventional	1	Obedience and punishment
	2	Individualism ,Instrumentalism, and Exchange
Conventional	3	Good boy/Good girl
	4	Law and order
Post-Conventional	5	Social contract
	6	Principled conscience

Moral development Kohlberg's stage theory:

Approximate age range	Stage	Sub stages
Birth to 9	Pre - Conventional	Avoid punishment
		Gain reward
Age 9 to 20	Conventional	Gain Approval and avoid disapproval
		Duty and guilt
Age 20 + and may be never	Post-Conventional	Agreed upon rights
		Personal moral stands

2. GILLIGAN'S THEORY

Carol Gilligan is one of the student of Kohlberg and she found that Kohlberg's theory had a strong male bias.And also the studies were of typically male preoccupation with general rules and rights.

- According to Gilligan's studies, men had a tendency to solve problems by applying abstract moral principles.
- Men were found to resolve moral dilemma by choosing the most important moral rule, overriding rules.
- In contrast, women gave importance to preserve personal relationships with all the people involved.
- The context oriented emphasis on maintaining personal relationships was called the ethics of care, in contrast with the ethics of rules and rights adopted by men.
- Gilligan revised the three levels of moral development of Kohlberg, as stages of growth towards ethics of caring.

Gilligan's stages of cognitive development

Approximate age range	Stage	Goal
Not listed	Pre-Conventional	Goal is individual survival
Transition is from selfishness to responsibility to others		
Not listed	Conventional	Self sacrifice is goodness
Transition is from goodness to truth that she is a person too		
May be never	Post-Conventional	Principle of nonviolence- Do not hurt others or self

The **pre conventional level**, which is same as that of Kohlberg's first one, right conduct, is viewed in a selfish manner solely as what is good for oneself.

The second level called **conventional level**, the importance is on not hurting others, and willing to sacrifice one's own interest and help others. This is the characteristic feature of women.

At the **post conventional level**, a reasoned balance is found between caring about others and pursuing the self interest. The balance one's own need and the needs of others, is aimed while maintaining relationship based on mutual caring. This is achieved by context-oriented reasoning, rather than by hierarchy of rules.

The theories of moral development by Kohlberg and Gilligan **differ** in the following respects.

Kohlberg's theory	**Carole Gilligan's theory**
A .Basic aspects	
Is based on the study on men	Is based on the study of both men and women.
Men give importance to moral rule.	Women always want to keep up the personal relationships with all the persons involved in the situations.
Ethics of roles and rights.	Women give attention to circumstance leading to critical situations rather than rules. (context-oriented and ethics of care)
B .Characteristic feature	
Justice	Reasons
Factual	Emotional
Right or Wrong	Impact on Relationship
Logic Only	Compassion too
Logic and Rule based	Caring and Concern
Less of Caring	More of Caring

Matter of Fact(Practical)	Abstract
Present Focus	Future Focus
Strict Rules	Making Exceptions
Independence	Dependence
Rigid	Human Oriented
Taking a Commanding Role	Shying away from Decision Making
Transactional Approach	Transformational Approach

The difference in these two theories is explained through the well known examples

HEIN'S DILEMMA:

Hein being poor and debtor could not buy the costly medicine for his sick wife, at 10 times the normal cost. Initially he begged the pharmacist to sell at half the price or allow him to pay for it later. Pharmacist refuse to oblige him either way finally he forcibly enter the pharmacy and stole the drug.

According to Kohlberg study men observed that the theft was morally 'wrong' at the conventional level, because the property right was violated.

But men at the post –conventional level, concluded that the theft was 'right', as the life of the human being was in danger.

But women observed that Hein was wrong. They observed that instead of stealing the could have tried other solutions (threatening or payment in installments?) to convince the pharmacist. Gilligan however attributed the decision by women as context-oriented and not on the basis of rules ranked in the order of priority.

7. Consensus and Controversy

Consensus and controversy are factors relevant to moral autonomy.

Consensus-general agreement of opinion .

Controversy- means conflict/disagreement.
The consensus and controversy are playing major roles while considering the moral autonomy. When an individual carry out the moral autonomy, he cannot get the same results and effects as that of the other people.i.e.,the results will be controversy.These kind of unavoidable disagreements require some tolerance among individuals who are autonomous,responsible and reasonable.
The creation of a uniform agreement on ethical values is not the only purpose of teaching ethics of engineering.These views are also expressed by the principles of tolerance.These views can hold good,even if uniform agreement is accomplished by any of studies that would harm the logics of moral autonomy.The similar concept of finding the proper ways and means for promoting tolerance in the practical applications of moral autonomy by engineers should be strictly applied in the evaluated methods of teaching the engineering ethics .

Similarly ,authority should be inducted both in classrooms of engineering .Authority of teachers on students and authority of managers on engineers,would much focus on the values of moral autonomy and ethics.
There are two important relationships between autonomy and authority.

8. Models of Professional Roles

- Promotion of public good is the primary concern of the professional engineers.
- There are several role models to whom the engineers are attracted. These models provoke their thinking, attitudes and actions.

1. **Savior:**

   As a Savior, the engineers save the society from poverty, illiteracy, wastage, inefficiency, ill health, human dignity and lead it to prosperity through technological development and social planning.

   Eg. R.L.Stevenson

2. **Guardian:**

   Engineer knows the direction in which technology should improve and the speed at which it should move.His experience and expertise is certainly needed to contribute to the society and maintain the developments already achieved.

   Eg. Lawrence of Arabia.

3. **Bureaucratic servant:**

   He serves the organization and the employers. The management of an enterprise fixes its goals and assigns the job of problem solving to the engineer, who accepts the challenge and shapes them into concrete achievements.

   Eg. Jamshedji Tata

4. **Social servant:**

   It is one who exhibits social responsibility. The engineer translates the interest and aspirations of the society into a reality.The role of engineers is not only to provide service to others but also their responsibility to the society.An engineer should keep in mind that his real master is the society.Thus engineers act as servants of society.

   Eg. Sir M.Viswesvaraya

5. **Social Enabler and Catalyst:**

   One who changes the society through technology. The engineer must assist the management and the society to understand their needs and make informed decisions on the desirable technological development and minimize the negative effects of technology on people and their living environment. Thus, he shines as a social enabler and a catalyst for further growth.

   Eg. Sri Sundarlal Bahuguna

6. **Game Player:**

   He is neither a servant nor master. An engineer is an assertive player, not a passive player who may carry out his master's voice. He plays a unique role successfully within the organization, enjoying the excitement of the profession and having the satisfaction of surging ahead in a competitive world.

   Eg. Narayanamurthy, Infosys and Dr.Kasthurirangan, ISRO

PROFESSION

Profession-is defined as any occupation/job/vocation that requires advanced expertise, self regulation. It brings high status, socially and economically.

The characteristics of profession are:

- ✓ Advanced expertise
- ✓ Self regulation
- ✓ Public good

Professional-relates to person or any work that a person does on profession, and it requires expertise, self regulation and results in public good. The term professional mean a 'person' as well as a 'status'.

Professionalism- is the status of the professional which implies certain attitudes or typical qualities that are expected to be a professional.

The criteria for achieving and sustaining professional status or professionalism are:

1. Advanced Expertise:

The Expertise includes sophisticated skills and theoretical knowledge in exercising judgment. This means a professional should analyze the problem in specific known area, in an objective manner.

2. Self-Regulation:

One should analyze the problem independent of self-interest and direct to a decision towards the best interest of the clients/customers .autonomous judgment is expected. In such situations, the codes of conduct of professional societies re followed s guidance.

3. Public Good:

One should not be a mere paid employee of an individual or a teaching college or manufacturing origination, to execute whatever the employer wants one to do. The job should be recognized by the public. The concerted efforts in the job should be towards promotion of the welfare, safety, and health of the public.

CHARACTERISTICS:

The characteristics of the 'profession' as distinct from 'non professional' occupation' are listed as follows:

1. Extensive training:

Entry into the profession requires an extensive period of training of intellectual (competence) and moral (integrity) character.

The theoretical base is obtained through formal education, usually in an academic institution.

It may be a Bachelor degree from a college or university or an advanced degree conferred by professional schools.

2. Knowledge and skills:

Knowledge and skills (competence) are necessary for the well-being of the society.

- Knowledge of the physicians protects us from disease and restores health.
- The lawyer's knowledge is useful when we are sued of a crime, or if our business in to be merged or closed or when we buy a property.
- The chartered accountant's knowledge is important for the success of recording financial transactions or when we file the income return.

The knowledge, study, and research of the engineers re required for the safety of the airplane, for the technological advances and for national defense.

3. Monopoly:

The monopoly control is achieved in two ways:

(a) The profession convinces the community that only those who have graduated from the professional school should be allowed to hold the professional title. The profession also gains the control over professional schools by establishing accreditation standards
(b) By persuading the community to have a license, they are liable to pay penalties.

4. Autonomy in workplace:

Professional engaged in private practice have considerable freedom in choosing their clients or patients.

Even the professionals working in large organizations exercise a large degree of impartiality, creativity and discretion (care with decision and communication)in carrying their responsibilities. Besides this, professionals are empowered with certain rights to establish their autonomy.

Accordingly physicians must determine the most appropriate medical treatment for their patients and lawyers must decide on the most successful defense for their clients .The possession of specialized knowledge is thus a powerful defense of professional autonomy.

5. Ethical Standards:

Professional societies promulgate the codes of conduct to regulate the professionals against their abuse or any unethical decisions and actions (impartiality, responsibility) affecting the individuals or grouper the society.

9. Theories about Right action

There are four ethical theories, each differing based on the moral concept involved.

- **Utilitarianism** seeks to produce the most utility, defined as a balance between good and bad consequences of an action, taking into account the consequences for everyone affected.
- **Duty Ethics** contends that there are duties that should be performed regardless of whether these acts lead to the most good.
- **Rights Ethics** emphasizes that we all have moral rights, and any action that violates these rights is ethically unacceptable. Like duty ethics, Rights Ethics also do not take into account, the overall good of the actions.

- **Virtue Ethics** regards those actions as right that manifests good traits (virtues) and regards those actions as bad that display bad traits (vices).

1. Utilitarianism

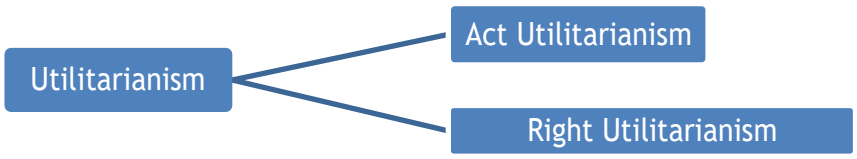
- The emphasis in Utilitarianism is not on maximizing the well being of the individual but rather on maximizing the well being of the society.
- Thus it is somewhat a collectivist approach.
  Eg. Building of Dams
- Dams often lead to great benefit to society by providing stable supplies of drinking water, flood control and recreational approaches.
- However, these benefits come at the expense of people who live in areas that will be flooded by the dam and are required to find new homes.
- Utilitarianism tries to balance the needs of society with the needs of the individual.

Two problems associated with Utilitarianism

- In the example of building a dam, what is best for everyone may be bad for a particular individual or group of individuals.
- Frequently, it is impossible to know exactly what the consequences of an action are. It is often impossible to know exactly what are the possible outcomes, especially when humans are involved as subjects of the experiments.

So, maximizing the benefit to society involves guess work and if the guess goes wrong, it would end up in major risk.

Despite these problems, Utilitarianism is a very valuable tool for ethical problem solving.



- ✓ **Act Utilitarianism** focuses on individual actions rather than on rules.
  The most common rules are "don't steal, be honest and don't harm others" etc.
  J.S.Mill felt that individual actions should be judged based on the most good was a produced and nothing wrong if the rule can also be broken.
- ✓ **Rule Utilitarianism** differs from act utilitarianism as it holds that moral rules are the most important.
  Rule Utilitarians contend that, although adhering to these rules might not always maximize good in a particular situation, overall, adhering to moral rules will ultimately lead to the most good.

Cost – Benefit Analysis :

It is an application of utilitarianism. In Cost – Benefit Analysis, the costs of the project as well as the benefits out of it are assessed. The projects with the highest ratio of benefits to costs alone can be implemented.

2. Duty Ethics

- Duty Ethics was proposed by Immanuel Kant (1724 – 1804).
- The list of duties are
  - Be honest
  - Be fair to others
  - Do not cause suffering to other people.
- These actions should be our duties as they express respect for persons.

3. Rights Ethics

- Right Ethics was formulated by John Locke (1632 – 1704).
- It holds that people have fundamental rights that other people have a duty to respect.
- Duty Ethics and Rights Ethics are just two sides of the same coin.
- In duty ethics, people have duties, an important one of which is to protect the rights of others.
- In right ethics, people have fundamental rights that others have duties to protect.

<u>Two Problems with duty and rights ethics</u>

- ✓ The basic rights of one person may conflict with the basic rights of another group.
  Eg. Building of a dam.
- ✓ If people's land happens to be in the way of a proposed dam, people have rights to own their property. This is sufficient to stop the dam project.
- ✓ Hence rights and duty ethics do not resolve the conflict. In this case, utilitarian approach of trying to determine the most good can be applicable.

Duty and rights ethics always gives importance to the good of a single individual. It does not consider the overall good for society.

4. Virtue ethics:

This emphasizes on the character rather than the rights or duties. The character is the pattern of virtues. The theory advocated by Aristotle, stressed on the tendency to act at proper balance between extremes of conduct, emotion, desire, attitudes to find the golden mean between the extremes of excess and deficiency.The examples shown below illustrate the theory,

Virtue	Excess	Golden mean	Deficient
Truthfulness(governs communication)	Revealing all in violation of tact and confidentiality	Necessary and sufficient to proper person.	Secretive
Courage(face risk)danger,	Roguishness, bold	Firm and humble	Cowardice
Generosity (Giving)	Wasting resources	Give in appropriate measure	Miserly
Friendliness(governs relationship)	Without anger, effusive	Within decent limits	Bad tempered
Green environment	Exploitation	Protection	Neglect
Work and earn	Tiresome work	Balance of work and leisure	Lazy on work

On the other hand, the virtue ethics highlighted on the action aimed at achieving common good and social good such as social justice, promotion of health, creation of useful and safe technological product and services.

Five types of virtues that constitute responsible professionalism are as follows,

- Public spirited virtues,
- Proficiency virtues,
- Team work virtues
- Self governance virtues
- Cardinal virtues

Which theory to use?

In solving ethical problems, we do not have to choose from these theories.

Rather, we can use all of them to analyze a problem from different angles and see what result each of the theories gives us.

This allows us to examine a problem from different perspectives to see what conclusion each one reaches. Frequently, the result will be the same even though the theories are very different.

For e.g. Discharge of waste from chemical plant.

- Consider a chemical plant near a small city that discharges a hazardous waste into the ground water.
- If the city takes its water from wells, significant health problems for the community may result.
- **Rights ethics** indicates that this pollution is unethical since it causes harm to many of the residents.
- **A utilitarian analysis** also comes to the same conclusion since the benefits of the community is of main concern here.
- **Virtue ethics** say that discharging wastes into ground water is irresponsible and harmful to individuals and so should not be done.
- In this case, all the ethical theories lead to the same conclusion.

10. Uses of Ethical theories

Ethical theories have three important uses.

- Resolving moral dilemmas
- Justifying moral obligations
- Relating professional and ordinary morality

(i) Resolving moral dilemmas

- Ethical theories help to resolve moral dilemmas.
- ***Virtue ethics*** resolves dilemma by emphasizing on character and relationships.
  - ✓ ***Utilitarianism*** resolve dilemma by emphasizing on goodness of the public.
  - ✓ ***Duty ethics*** indicates the duties to protect the public.
  - ✓ ***Rights ethics*** emphasizes the rights of the public to be protected.

- Ethical theories provide more precise sense of what kind of information are relevant in resolving moral dilemma.
- The ethical theories offer ways to rank the relevant moral considerations based on their order of importance.
- The ethical theories help us to identify the alternate possibilities to resolve the problem and provide a systematic framework for comparing the alternatives.
- By providing frameworks for moral arguments, the theories strengthen our ability to reach balanced and insightful judgements.

(ii) Justifying moral obligations

Obligations is nothing but the moral responsibilities of an individual.These responsibities are the duties to perform the right acts in a moral way.In connection with function of engineers ,they have the responsibilities or obligations for performing regular inspections,identifying potential benefits and risks of tasks when compared with others.Ethical theories are used to justify the obligations of engineers involved in technological development.

(iii) Relating Professional and Ordinary morality

Safety is the ordinary morality which the engineers also try to achieve as a consequence of their work.Thus ethical theories relate professional and ordinary morality.

11. Self Interest

- Self interest is being good and acceptable to oneself.
- It is very ethical to possess self interest.
- As per **utilitarian theory**, this interest provides respect for others too.
- **Duty ethics** recognizes this aspect as duties to ourselves. Then only one can help others.
- **Right ethicist** stresses our rights to pursue our own good.
- **Virtue ethics** also accepts the importance of self interest.
- In ***ETHICAL EGOISM***, the self is conceived in a highly individualistic manner. It says that every one of us should always promote one's own interest.
- But Self interest should not degenerate into Egoism i.e., maximizing only own good.
- The ethical egoists hold that the society benefits to maximum when
    i) The individuals pursue their personal good
    ii) The individual organizations pursue maximum profit in a competitive enterprise.

    In such pursuits, both individuals and organizations should realize that independence is not the only important value. We are also interdependent.
- Self interest is necessary initially to begin with. Later, it should develop into showing concern for others, in the family as well as society.
- One's self interest should not harm others.
- The principles of "Live and let (others) live" and "reasonably fair competition" are recommended to professionals by the ethicists.

12. Customs

Definitions:

- ✓ Long established habits or behavior of a society.
- ✓ Practice followed by people of a particular group or icon.
- ✓ Habitual Practice of a person
- ✓ Common tradition or usage so long established that it has the force or validity of the law.

Ethical pluralism:

Various cultures in our pluralistic society lead to tolerance for various customs, beliefs and outlooks. In the same way, ethical pluralism also exists. It means that even reasonable people too will not agree on all moral issues.

Ethical relativism:

According to this principle, actions are considered morally right when approved by law or custom. Actions are considered wrong when they violate the laws or customs.Laws appear to be objective ways for judging values. The laws and customs tend to be definite, clear and real, but not always.Eg. Apartheid laws of South Africa violated the human rights of the native Africans.

Ethical relativism assumes that the values are subjective at the cultural level.As per ethical relativism, the actions and laws of Hitler who vowed on Anti- Semitism and killed several million Jews would be accepted as right.

Moral rationalism or moral contextualism:

According to this, the moral judgements must be made in relation to certain factors, which may vary from case to case.

Reasons for accepting ethical relativism:

Customs and law appear to be definite,real and clear cut.

Values are treated as subjective at the cultural level encouraging the virtue to tolerate differences among societies.

It is impossible to have commonly acceptable rules because moral judgments must be made in relation to some factors that may vary from situation to situation.

13. Religion

Religion have played major role in shaping moral views and moral values, over geographical regions.

Christianity has influenced the western countries, Islam in the middle-East countries, Buddhism and Hinduism in Asia and Confucianism in China. Further there is a strong psychological link between the moral and religious beliefs.Religions support moral responsibility. They have set high moral standards. Faith in the religions provides trust and this trust inspires people to be moral. The religions insist on tolerance and moral concern for others.Many professionals who posses religious beliefs are motivated to be morally responsible. Each religion lays stress on certain high moral standards.

E.g. Hinduism holds polytheistic (many gods) view, and virtues of devotion. Christianity beliefs in one deity and emphasizes on virtues of love, faith and hope. Buddhism is non-theistic and focuses on compassion and Islam on one deity. Judaism stresses the virtue of righteousness.

But many religious sections have adapted poor moral standards. Example many religion do not

recognize equal rights for women. The right to worship is denied for some people.

People are killed in the name of or to promote religion. Thus the conflicts exist between the secular and religious people and between one religion and another. Hence religious views have to be morally scrutinized.

2 Marks:

1. What is meant by moral autonomy? (N/D 15)
2. Mention the various types of inquiry. (N/D 15)
3. Define moral dilemma (N/D 16)
4. State Kohlberg's theory (N/D 16)
5. Define engineering ethics (N/D 13)
6. State the importance of ethical theories (A/M 17)
7. State Gilligan's theory (M/J 16)
8. What is meant by consensus (M/J 16)
9. What are the two important versions of utilitarianism? (N/D 17)

16 Marks:

1. Explain the gilligan's theory for moral development (8)
2. What are the different types of model of professional roles? (8)
3. Explain the theory of right ethics and its classification (8)
4. What is meant by self interest? Relate the term with ethical egoism with suitable examples. (8)
5. Discuss in detail about the concept of (i).Moral dilemma (ii). Moral autonomy (16)
6. Discuss in detail the various theories about right action (16)
7. Describe kohlbergs and gilligan's theory on moral autonomy (8)
8. Explain in detail about the senses of engineering ethics. (16)

UNIT-III ENGINEERING AS SOCIAL EXPERIMENTATION

1.ENGINEERING AS EXPERIMENTATION
2.ENGINEERS AS RESPONSIBLE EXPERIMENTERS
3.CODES OF ETHICS
4.A BALANCED OUTLOOK ON LAW

1. ENGINEERING AS EXPERIMENTATION

- Experimentation is commonly recognized as playing as essential role in the design process.
- All products of technology present some potential dangers and thus engineering is an inherently risky activity. In order to underscore this fact and help in exploring its ethical implications, engineering should be viewed as an experimental process.
- It is not an experiment solely in laboratory controlled conditions. It is an experiment on a social scale involving human subjects.
- Before manufacturing a product or providing a project, we make several assumptions and trials, design and redesign and test several times till the product is observed to be functioning satisfactorily. We try different materials and experiments. From the test data obtained we make detailed design and retests. Thus, design as well as engineering is iterative process as illustrated in Fig. 3.1.

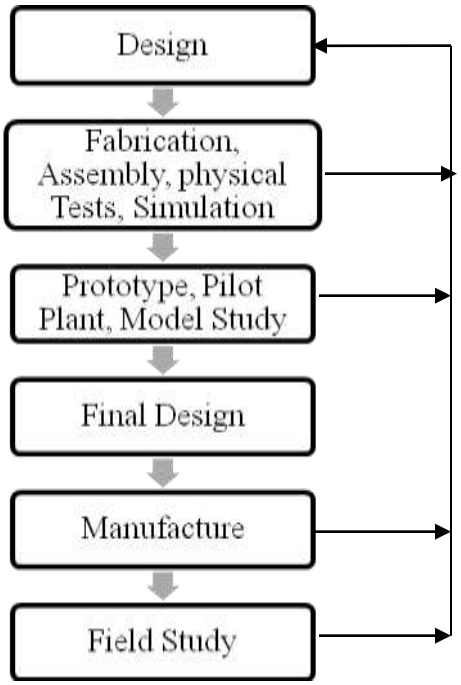


Fig 3.1 Design as an interactive process

- Several redesigns are made upon the feedback information on the performance or failure in the field or in the factory. Besides the tests, each engineering project is modified during execution, based on the periodical feedback on the progress and the lessons from other sources. Hence, the development of a product or a project as a whole may be considered as an experiment.

ENGINEERING PROJECTS VS STANDARD EXPERIMENTS

We shall now compare the two activities, and identify the similarities and contrasts.

A. Similarities

1. ***Partial ignorance*:**
   - The project is usually executed in partial ignorance. Uncertainties exist in the model assumed.
   - There are uncertainties about the nature of the stresses the finished product will encounter. There are uncertainties in the abstract model used for the design calculations. The behavior of materials purchased is uncertain and not constant. They may vary with the suppliers, processed lot, time, and the process used in shaping the materials (e.g., sheet or plate, rod or wire, forged or cast or welded).
   - There may be variations in the grain structure and its resulting failure stress. It is not possible to collect data on all variations. In some cases, extrapolation, interpolation, assumptions of linear behavior over the range of parameters, accelerated testing, simulations, and virtual testing are resorted.
2. ***Uncertainty*:**
   - The final outcomes of projects are also uncertain, as in experiments. Sometimes unintended results, side effects (bye-products), and unsafe operation have also occurred.
   - Unexpected risks, such as undue seepage in a storage dam, leakage of nuclear radiation from an atomic power plant, presence of pesticides in food or soft drink bottle, an new irrigation canal spreading water-borne diseases, and an unsuspecting hair dryer causing lung cancer on the user from the asbestos gasket used in the product have been reported.
3. ***Continuous monitoring*:**
   - Monitoring continually the progress and gaining new knowledge are needed before, during, and after execution of project as in the case of experimentation.
   - The performance is to be monitored even during the use (or wrong use!) of the product by the end user/beneficiary.
4. ***Learning from the past*:**
   - Usually engineers learn from their own earlier design and operating results, as well as from those of other engineers, but unfortunately that is not always the case.
   - Lack of established channels of communication, misplaced pride in not asking for information, embarrassment at failure or fear of litigation and plain neglect often impede the flow of such information and lead to many repetitions of past mistakes.
   - Here are a few examples:
   - The Titanic lacked a sufficient number of lifeboats decades after most of the passengers and crew on the steamship Arctic had perished because of the same problem.
   - “Complete lacks of protection against impact by shipping caused Sweden’s worst ever bridge collapse on Friday as a result of which eight people were killed.” Thus reported the New Civil Engineer on January 24, 1980. Engineers now recommend the use of floating concrete bumpers that can deflect ships, but that recommendation is rarely heeded as seen by the 1993 collapse of the Bayou Cannot Bridge that cost 43 passengers of the Sunset Limited their lives.
   - Valves are notorious for being among the least reliable components of hydraulic systems. It was a pressure relief valve, and a lack of definitive information regarding its open or

shut state, which contributed to the nuclear reactor accident at Three Mile Island on March 28, 1979. Similar malfunctions had occurred with identical valves on nuclear reactors at other locations. The required reports had been filed with Babcock and Wilcox, the reactor's manufacturer, but no attention had been given to them .

- NASA uses the metric system while Lockheed Martin uses the English system when building a satellite Courtesy of NASA

  Cost of the lost orbiter: $125 million

  Inflation-adjusted: $165.6 million

  In 1999 a team of Lockheed Martin engineers used the English system of measurement, while the rest of the team used the metric system for a Mars orbiter. The use of two different measurement systems prevented the spacecraft's navigation coordinates from being transferred from a spacecraft team in Denver to a lab in California.

  The orbiter was then lost in space, and NASA was out $125 million. These examples illustrate why it is not enough for engineers to rely on handbooks and computer programs without knowing the limits of the tables and algorithms underlying their favorite tools. They do well to visit shop floors and construction sites to learn from workers and testers how well the customers' wishes were met.

  The art of back-of-the-envelope calculations to obtain ballpark values with which to quickly check lengthy and complicated computational procedures must not be lost. Engineering demands practitioners who remain alert and well informed at every stage of a project's history and who exchange ideas freely with colleagues in related departments.

B. Contrasts

The scientific experiments in the laboratory and the engineering experiments in the field exhibit several contrasts as listed below:

1. ***Experimental control*:**
   - In standard experiments, members for study are selected into two groups namely A and B at random.
   - Group A are given special treatment. The group B is given no treatment and is called the 'controlled group'. But they are placed in the same environment as the other group A. This process is called the *experimental control*.
   - This practice is adopted in the field of medicine. In engineering, this does not happen, except when the project is confined to laboratory experiments. This is because it is the clients or consumers who choose the product, exercise the control.
   - It is not possible to make a random selection of participants from various groups. In engineering, through random sampling, the survey is made from among the users, to assess the results on the product. .
2. ***Humane touch*:**
   - Engineering experiments involve human souls, their needs, views, expectations, and creative use as in case of social experimentation.
   - This point of view is not agreed by many of the engineers. But now the quality engineers and managers have fully realized this humane aspect.
3. ***Informed consent*:**
   - Engineering experimentation is viewed as Societal Experiment since the subject and the beneficiary are human beings.

- In this respect, it is similar to medical experimentation on human beings. In the case of medical practice, moral and legal rights have been recognized while planning for experimentation.
- Informed consent is practiced in medical experimentation. Such a practice is not there in scientific laboratory experiments. Informed consent has two basic elements:

***Knowledge*:** The subject should be given all relevant information needed to make the decision to participate.

***Voluntariness*:** Subject should take part without force, fraud or deception. Respect for rights of minorities to dissent and compensation for harmful effect are assumed here.

For a valid consent, the following conditions are to be fulfilled:

- Consent must be voluntary
- All relevant information shall be presented/stated in a clearly understandable form

Consenter shall be capable of processing the information and make rational decisions.

4. The subject's consent may be offered in proxy by a group that represents many subjects of like-interests Informed consent when bringing an engineering product to market, implies letting the customer know the following:

(*a*) The knowledge about the product

(*b*) Risks and benefits of using the product and

(*c*) All relevant information on the product, such as how to use and how not to use (do's and don'ts).

The relevant factual information implies, that the engineers are obliged to obtain and assess all the available information related to the fulfillment of one's moral obligations (i.e., wrong or immoral use of a product one designs), including the intended and unintended impacts of the product, on the society.

Still there exists a possibility of a large gap of understanding between the experimenter and the subjects (public). Sometimes, the managements have not been willing to disseminate the full information about the project or product beyond the legal requirements, because of the fear of potential competitions and likely exposure to potential litigation.

People object to *involuntary risks* wherein the affected individual is neither a direct participant nor a decision maker. In short, we prefer to be the subjects of our own experiments rather than those of somebody else. If it is an asbestos plant or nuclear plant to be approved, affected parties expect their consent to be obtained. But they are ready to accept *voluntary* risks as in the case of stunts and amazing races.

In case of Koodangulam power project as well as the Sethusamudram Canal Project, Tamil Nadu, several citizen groups including Fishermen Forums have responded. The Central government was able contain many harsh apprehensions and protracted legal and political battles, by providing all relevant information.

5. ***Knowledge gained*:**

- Scientific experiments are conducted to gain new knowledge, while "engineering projects are experiments that are not necessarily designed to produce very much knowledge".
- Engineering experiments at the most help us to,

(*a*) Verify the adequacy of the design,

(*b*) To check the stability of the design parameters, and

(*c*) Prepare for the unexpected outcomes, in the actual field environments.

From the models tested in the laboratory to the pilot plant tested in the field, there are differences in performance as well as other outcomes.

2.ENGINEERS AS RESPONSIBLE EXPERIMENTERS

- Engineers are the main technical enablers or facilitators; they are far from being the sole experimenters. Their responsibility is shared with management, the public, and others. Yet their expertise places them in a unique position to monitor projects, to identify risks, and to provide clients and the public with the information needed to make reasonable decisions.
- From the perspective of engineering as social experimentation, four features characterize what it means to be a responsible person while acting as an engineer: a conscientious commitment to live by moral values, a comprehensive perspective, autonomy, and accountability.
- Although the engineers facilitate experiments, they are not alone in the field. Their responsibility is shared with the organizations, people, government, and others. No doubt the engineers share a greater responsibility while monitoring the projects, identifying the risks, and informing the clients and the public with facts. Based on this, they can take decisions to participate or protest or promote.
- The engineer, as an experimenter, owe several responsibilities to the society, namely,
  1. A conscientious commitment to live by moral values.
  2. A comprehensive perspective on relevant information. It includes constant awareness of the progress of the experiment and readiness to monitor the side effects, if any.
  3. Unrestricted free-personal involvement in all steps of the project/product development (autonomy).
  4. be accountable for the results of the project (accountability).

a) Conscientiousness(Awareness):

- Conscientious moral commitment means:
  (*a*) Being sensitive to full range of moral values and responsibilities relevant to the prevailing situation and
  (*b*) The willingness to develop the skill and put efforts needed to reach the best balance possible among those considerations. In short, engineers must possess open eyes, open ears, and an open mind (i.e., moral vision, moral listening, and moral reasoning).
- This makes the engineers as social experimenters, respect foremost the safety and health of the affected, while they seek to enrich their knowledge, rush for the profit, follow the rules, or care for only the beneficiary. The human rights of the participant should be protected through voluntary and informed consent.

b) **Comprehensive Perspective**

- Conscientiousness is blind without relevant factual information. Hence showing moral concern involves a commitment to obtain and properly assess all available information that is pertinent to meeting moral obligations.
- This means, as a first step, fully grasping the context of one's work, which makes it count as an activity having a moral import. For example, in designing a heat exchanger, if I ignore the fact that it will be used in the manufacture of a potent, illegal hallucinogen, I am showing a lack of moral concern.
- One should not ignore his conscience, if the product or project that he is involved will result in damaging the nervous system of the people (or even the enemy, in case of weapon development).

- A product has a built-in obsolete or redundant component to boost sales with a false claim. In possessing of the perspective of factual information, the engineer should exhibit a moral concern and not agree for this design. Sometimes, the guilt is transferred to the government or the competitors.
- Some organizations think that they will let the government find the fault or let the fraudulent competitor be caught first. Finally, a full-scale environmental or social impact study of the product or project by individual engineers is useful but not possible, in practice.

c) Moral Autonomy

- Moral autonomy is defined as, decisions and actions exercised on the basis of moral concern for other people and recognition of good moral reasons. Alternatively, moral autonomy means 'self determinant or independent'.
- The autonomous people hold moral beliefs and attitudes based on their critical reflection rather than on passive adoption of the conventions of the society or profession.
- Moral autonomy may also be defined as a skill and habit of thinking rationally about the ethical issues, on the basis of moral concern.
- Viewing engineering as social experimentation, and anticipating unknown consequences should promote an attitude of questioning about the adequacy of the existing economic and safety standards. This proves a greater sense of personal involvement in one's work.
- If management views profitability is more important than consistent quality and retention of the customers that discourage the moral autonomy, engineers are compelled to seek the support from their professional societies and outside organizations for moral support.

d) **Accountability**

- The term Accountability means:

  1. The capacity to understand and act on moral reasons

  2. Willingness to submit one's actions to moral scrutiny and be responsive to the assessment of others.
- It includes being answerable for meeting specific obligations, i.e., liable to justify (or give reasonable excuses) the decisions, actions or means, and outcomes (sometimes unexpected), when required by the stakeholders or by law.
- The tug-of-war between of causal influence by the employer and moral responsibility of the employee is quite common in professions. In the engineering practice, the problems are:

  a) The fragmentation of work in a project inevitably makes the final products lie away from the immediate work place, and lessens the personal responsibility of the employee.

  b) Further the responsibilities diffuse into various hierarchies and to various people. Nobody gets the real feel of personal responsibility.

  c) Often projects are executed one after another. An employee is more interested in adherence of tight schedules rather than giving personal care for the current project.

  d) More litigation is to be faced by the engineers (as in the case of medical practitioners). This makes them wary of showing moral concerns beyond what is prescribed by the institutions. In spite of all these shortcomings, engineers are expected to face the risk and show up personal responsibility as the profession demands.

3.CODE OF ETHICS:

- Ethics is generally the discipline or field of study dealing with moral duty or obligation. This gives rise to set of governing principles or values which in turn are used to judge the particular conduct or behaviour.
- Code of ethics exhibits rights, duties and obligations of the members of a profession and a professional society and it have to be adopted by engineering societies as well as engineers.
- Code of ethics provides a framework for ethical judgement for a professional. It serves as a starting point for ethical decision making. A code expresses the circumstances to ethical conduct shared by the members of a profession.
- Code is based on broad principles of truth, honesty and trustworthiness, respect for human life and welfare competence and accountability.
- As a consequence, code of professional ethics is more than a minimum standard of conduct, rather it is a set of principles which should guide professionals in their daily work.

MODEL CODE OF ETHICS:

To keep up these basic codes of ethics professional engineers shall ,

- Offer services,advice on or undertaking assignments only in areas of their competence and practice in a careful and diligent manner.
- Act as faithful agents of their clients or employers.
- Keep themselves informed in order to maintain their competence,strive to advance the body of knowledge within which they practice and provide oppurtunities for the professional development of their subordinates.
- Conduct themselves with fairness and good faith toward clients , colleagues and others.
- Be aware of and ensure that clients and employers are made aware of societal and environmental consequences .
- Report to their association and /or appropriate agencies any illegal or unethical engineering decisions or practices by engineers or others.

ROLES OF CODE OF ETHICS:

The codes exhibit following essential roles:

- INSPIRATION AND GUIDANCE:

1) The codes express the collective commitment of the profession to ethical conduct and public good and thus inspire the individuals.

2) They identify primary responsibilities and provide statements and guidelines on interpretation for the professional and professional societies.

3) The engineering societies like AAES-American Association of Engineering Societies,NSPE-National Society of Professional Engineers,IEEE-Institute of Electrical and Electronics Engineering,AICTE-All India Council for Technical Education etc.,have published codes of ethics.

4) Most of the technological companies like Microsoft,Texas Instruments etc., have set up their own codes in order to take care of ethical problems while dealing with vendors and clients.

- SUPPORT TO ENGINEERS:

1. The codes gives positive support to professionals for taking stands on moral issues.Further they serve as potential legal support to discharge professional obligations.

➢ DETERRENCE AND DISCIPLINE:

1.Code of ethics assist in formally analyzing unethical practices in a profession.

2. If the result of the analysis are found to be correct ,the professional engaged in such a practice can be suspended or his membership can be terminated.

➢ EDUCATION AND MUTUAL UNDERSTANDING:

1) Ethical codes can be taught in the educational institutions , or circulate among the public and officials for imparting the importance of moral issues and values.

2) They also serve for the mutual exchange of views and ideas between public and professionals with regard to the professional commitments and responsibilities.

➢ PUBLIC IMAGE ABOUT PROFESSION:

Having codes will boost up the image amongst the public. Codes can also motivate the engineers to serve the public in an effective manner.

➢ PROMOTING BUSINESS INTERESTS:

Codes help to moralise the various dealings involved in business in order to get benefits and to attain objectives.

➢ ADVANTAGES OF CODE OF ETHICS:

1) Set out the ideals and responsibilities of the profession.
2) Enhance the profile of the profession.
3) Motivate and inspire practitioners.
4) Improve quality and consistency.
5) Raise awareness and consciousness of issues.

➢ LIMITATIONS OF CODE OF ETHICS:

1) Whether such code is desirable or feasible.

2) Whether ethical values are universally or culturally relativistic.

4.BALANCED OUTLOOK ON LAW:

- The balanced outlook on law signifies the need of laws and regulations in directing engineering practices.
- With the help of ethical conduct ,a balance is developed between individual needs and desires against collective need and desires.And the ethical conduct can be only applied with the help of laws.
- Engineers have to play an effective role in enhancing or changing enforceable rules of engineering as well as in enforcing them.And the codes must be enforced with the help of laws.

CASE STUDY:

BABYLONS BUILDING CODE:

This code was developed by Hammurabhi, King of Babylon in the period of 1758.The builders of his time and all the builders were forced to adhere the codes of law.Following are the essential features of law.

1) Householder dies due to poor quality construction; builder will be given death sentence.

2) Builders son has to be given death sentence,if the son of the house owner was killed in destruction of the house.

3) If the householders slave dies for the above mentioned reason,builder have to give one slave to the householder.

4) If the property damaged or destroyed,builder should replace or reconstruct the house.

The aspects of Babylons building code had a terrible impact on moral duties and responsibilities of earlier builders.

UNITED STATES STEAMBOAT CODE:

- In united states alone between the years 1816 and 1847,more than 2500 people were killed and 2000 people were injured due to the explosion of boilers in steamboats.
- Because of this in this year,a law was passed in order to make the provisions for periodical inspections of safety of boilers and engines.
- But it is clear that this law was not effective due to the corruptions of the inspectors and also their inadequate training regarding the safety checking.
- In 1865,the law enforcement had the serious setback by the sinking of riverboat due to boiler explosion.
- Then,American society of mechanical engineers(ASME) formulated the standard procedures for the regulation of the production of steam boilers and their application in various fields.

INDUSTRIAL STANDARDS:

- Standards are designed and recommended by Organisations, Trades and professional association for industrial applications.
- Standards are technical specifications that lay down characteristics of a product such as

1) Size
2) Quality
3) Performance
4) Safety

➢ **The following points indicate some of the objectives and the functional role of standards**.

1) Public, clients and industrial producers are jointed benefits by standards.

2) With the help of designs, names, brands etc.standards maintain a steady and balanced competition in industries.

3) Standards focus on the measure of quality products.

PROBLEMS WITH LAW IN ENGINEERING:

➢ Increasing demand for enforcing legal restrictions in the field of engineering.eg.Safety measures

➢ Minimal compliance.

➢ It is very difficult to update the laws in order to match with the rapidly changing technology.

➢ Many laws remain useless in the form of non-laws.

➢ Some professionals also have to refer the standard with the readymade specifications as a substituition for original thought.

PROPER ROLE OF LAW IN ENGINEERING

➢ If laws are effectively enforced, society will get benefits. Law gives a powerful support to those who desire to act ethically in their activities.

➢ For eg, Rain water harvesting,poliodrops etc.

➢ All rules governing engineering practices should be formulated in such a way to assist engineering professional to undertake responsible social experimentation.

➢ If the experimentation is strong,the rules should not try to cover all the possible outcomes of experiments and they should not force the engineer to follow a rigid course of action.

➢ The Rules and regulations should be broad in nature, but at the same time they should not be written firmly in such a way the engineer to be forced to be accountable for his decisions.

UNIT IV SAFETY, RESPONSIBILITIES AND RIGHTS

Safety and Risk – Assessment of Safety and Risk – Risk Benefit Analysis and Reducing Risk – Respect for Authority – Collective Bargaining – Confidentiality – Conflicts of Interest – Occupational Crime – Professional Rights – Employee Rights – Intellectual Property Rights (IPR) – Discrimination

1.Safety and Risk

ENGINEER'S RESPONSIBILITY FOR SAFETY

INTRODUCTION:

Unarguably the first and most important duty of an engineer is to ensure the safety of the people who use the product designed by him. It could be seen from all of the code of ethics of the professional engineering societies that safety is of paramount importance to the engineer. The engineers have a responsibility to the society to produce products that are safe. Engineering necessarily involves risk. It is impossible to design any products to be completely risk free. The relationship of risk to safety is mostly very close.

In this chapter we shall explore the following questions:

- What kind of responsibility should an engineer hold to ensure the safety, health and welfare of the public?
- How can the products be designed by the engineers to minimize the risk to the user?
- How much risk is appropriate in an engineering design?
- How to assess the safety and risk? And
- How to do risk benefit analyses?

Safety and Its Concepts:

What is meant by safety?

Safety means the state of being safe. Safe means protected from danger and harm. The term safety is always difficult to describe completely. What may be safe for one person may not be safe for another person. It is because different persons have different perceptions about what is safe. For example, a shaving razor in the hands of a child is never safe as it can be in the hands of an adult. The American heritage dictionary defines safety as freedom from damage, injury or risk. Absolute safety that satisfies all individuals or group under all conditions is neither attainable nor affordable.

Safety Defined

The initial version of William W. Lawrence's definition for safety is as follows:

"A thing is safe if its risk is judged to be acceptable". It means a thing is safe for a person if the perceived risk is less. Similarly, a thing is unsafe if the perceived risk is high. Drawbacks of the Lawrence's Initial Version of Definition:

Some of the drawbacks are

a) ***Under-estimation of risks***: An unsafe product may be considered to be safe because of all faulty view and misjudgment of the person. Example, buying improperly designed coil type water heater which eventually ends up with a severe electric shock.

In the above example, the judgment about the product has failed which is against the[2] Lawrence definition.

b) ***Over-estimation of risk:*** A product whose risks are comparatively less may be considered unsafe because of over safety concern of a person. Example, thinking that adding chlorine in drinking water will kill a lot of people in this case according to Lawrence definition the water became unsafe the

moment we judged the risks is unacceptable for us. But our common concept of safety says chlorinated drinking water is safer. This again contrasts the Lawrence definition.

c) ***No-estimation of risk:*** For the person who does not judge bout risk the product may be either safe or unsafe. Example, Purchasing a LPG gasoline fuel driven car without judging anything about its safety.

So in order to overcome the above said contradictions, Lawrence proposed a modified version of definition for safety.

A Modified Lawrence Definition of Safety

Definition***:"A thing is safe with respect to a given person or group at a given time, if its risks were fully known. If those risks would be judged acceptable, in light of settled value principles".***

In the modified Lawrence definition the term 'things' represent not only products, but also service, processes, etc. Therefore the definition can be extended to the design, finance, international affairs, etc.

CONCEPT OF RELATIVE SAFETY

Safety is expressed frequently in terms of degree and comparisons. We often use words such as fairly 'safe'. The relative safety expresses the safety of a thing in comparison with safety of similar things.

Example: stating that airplane travel is safer than car travel and car travel is safer than travel in a bike.

ENGINEERS AND SAFETY

Criteria to ensure safe design.

It is universally accepted that safety should be an integral part of any engineering design. In order to ensure the safe design, the following criteria should be met:

1) A design should comply with legal standard for product safety and other applicable loss.
2) An acceptable design should meet be standard of single code accepted engineering practice.
3) Alternative designs that are potentially safer should be explore
4) Finally, the designed product should be tested using prototype to determine:

Whether the products meets the specifications, and whether the product is safety use.

Designing for safety

(Incorporating safety into the engineering design process)

Step1: Define the problem. It includes the issues of safety in the product definition and specification.

Step2: Generate multiple alternate design solutions.

Step3: Analyze each design solution. It evaluates pros and cons of each solution.

Step4: Test the solutions.

Step5: Select the best solution.

Step6: Implement the chosen solution.

During step 1-6, a safety and risk criteria should be given paramount importance than other issues.

RISK AND ITS CONCEPT

What is meant by risk?

A risk is the potential that something unwanted and harmful may occur.

The American heritage dictionary discusses the possibility of suffering harm or loss. Generally the term 'risk' is synonymously used adverse effects are harm. The term 'harm' may be defined as an invention or limitation of a person's freedom or well-being.

The three most important types of well-being are physical well-being psychological well- being, and economical well-being.

Effect of risk it includes dangers of bodily harm, economic loss and environmental degradation.

Cause of risk risks or harms are caused by delayed job completion, faulty product or systems, and economically or environmentally injurious solutions to technological problems.

RISK DEFINED

William w. Lawrence has defined risk as ***"a compound measure of the probability and magnitude of adverse effect.***

Mathematically,

Risk=probability of harm *magnitude or consequence of the harm

In simple words the risk is the product of the likelihood and the magnitude of the harm.

A relatively slight harm having more probability of occurring might constitute a greater risk than a relatively large harm having lesser probability of occurring.

Natural Hazards and Disaster

A natural hazard such as floods, earthquake, droughts, volcanoes etc greatly threatens and damages the long lifelines of human populations.

A disaster is a serious disruptive event coincides with state of insufficient preparation.

In recent years, engineering and technologies have greatly reduced some of the ill effects of natural hazards and disasters.

Thus engineers should be aware of the ethical and professional issues regarding risk.

Factors Influencing Risk

Since the concept of risk is subjective in nature, it depends on many factors. They are:

1. Voluntary Vs Involuntary risk

If the person knowingly takes any risk, then he feels it safe. If the same risk is forced to him, then he feels it unsafe. In simple terms, the voluntary risks are considered as safe and the involuntary risks are considered as unsafe.

2. Short-term Vs Long-term consequences

A thing, which causes a short-lived illness or disability, seems safer than a thing that will result in permanent disability. An activity for which there is a risk of getting a fractured leg will appear much less risky than an activity with a risk of a spinal fracture since a broken leg will be painful and disabling for a few months, but full recovery is in the norm. Spinal fractures, can lead to permanent disability.

3. Delayed Vs Immediate Risk

An activity whose harm is delayed for many years will seem much less risky than something with an immediate effect. For example, for several years now, Americans have been warned about the adverse long-term health effects of a high fat diet. This type of diet can lead to chronic heart problems or stroke later in life. Yet may be ignored these warnings and are unconcerned about a risk that is so far in future. These same people might find an activity such as sky diving unacceptably risky since an accident will cause immediate injury or death.

Something that one person feels safe may seem very unsafe to someone else. This creates some confusion for the engineer who has to decide whether a project is safe enough to be pursued. It is up to the engineer and company management to use their professional judgment to determine whether a project can be safely implemented.

4. Expected Probability

A relative slight harm having more probability of occurring(say, 50: 50 chance) seems to be a greater/unacceptable risk than a relatively a severe harm having lesser probability of occurring(say, 1 in 1,00,000).

5. Reversible Effects

Something will seem less risky if the bad effects are ultimately reversible. This concept is similar to short term risk Vs Long term risk.

6. Threshold Levels for Risk

Something that is risky only at fairly high exposures will seem safer than something with a uniform exposure to risk.For example the probability of being in an automobile accident is the same regardless of how often you drive.you can reduce the likelihood of being in an accident by driving less often. From the above discussion, it is understood that something is unsafe or risky to one person may seem very safe to someone else. This creates the great challenge for the engineers to decide on the optimal safety level.

ACCEPTABILITY OF RISK

What is meant by Acceptable Risk?

According to D. Rowe, "A risk is acceptable when those affected are generally no longer apprehensive about it". Apprehensiveness mainly depends on how the risk is perceived by the people.

Elements of Risk Perception

(Factors influencing the perception of risk)

The risk perception is influenced by the factors such as:

1. Whether the risk is assumed voluntarily;
2. The effect of knowledge on how the probabilities of harm are perceived;
3. Job –related or other pressure that cause people to be aware of risks;
4. Whether the effects of a risky activity or situation are immediately noticeable; and
5. Whether the potential victims are identifiable beforehand.

We shall discuss these elements of risk perception, in detail, in the following sections.

1. **Voluntarism and control**

Voluntary risk: if people take risk knowingly, then their involvement of risk is known as voluntary risk. Many people considered safer if they knowingly take on the risk. Also the people believe that they have 'full control' over their actions.

Examples for voluntary risk

1. Buying a flat/house near a chemical plant that emits low level of a toxic waste into the air, because the property values are very low.
2. Participating in a potentially adventurous sport such as motorcycle racing, skiing, boxing, hang-gliding, bungee jumping, etc without much safety guards.

Controlled risk

If the risk taken is within the control limit, which can be controlled by any means, then the risk is known as controlled risk.

Examples for controlled risk:

In practice, all the dangerous sports such as motorcycle racing, skiing, hang-gliding, bungee jumping,

horseback riding, boxing etc are carried out under the assumed control of the participants. They use all safety guard to keep the risk under control.

2. Effect of information on Risk Assessments

The information about a harm/danger should be presented in a systematic and appropriate manner. Because the manner in which the required information for decision-making is presented has a great influence on how risks are perceived.

Many case studies and experiments have proved that the manner in which information about a danger is presented can lead to undesirable and wrong perceptions about danger.

The threshold limit of individuals for information varies from person to person.Some would be comfortable only when they have information of deeper depth and quality, while others may be comfortable with minimal information.

Many experiments have drawn the following two conclusions:

1. Options perceived as yielding company gains will tend to be preferred over those from which gains are perceived as risky or only probable.
2. People tend to be more willing to take risks in order to avoid perceived company losses than they are to win only possible gains.

3. Job-related Risks

The exposure of risk depends on the person's job and his work place.

The nature of the job and the working environment will determine the risk level of a person. For example, people working in the coalmines, oil mines shipyards, chemical plants, nuclear power plants, etc have more probability of being exposed to the high risk.

Because of high competition for survival, the employees don't have any options other than undertaking high-risk jobs.

Unions, and occupational and safety regulations should regulate and enforce the employers, to facilitate the standard working environment.

Most importantly, engineers who design and the workers suggestions/complaints regarding their workplace.

4. Magnitude and proximity

Our reaction to risk is affected by the magnification and the personal identification or relationship we have the victims.

For instance, we feel very bad if none of our close relative or friends are subjected to great harm by some accident than if it might affect 20 strangers.

Thus the magnitude of risk and the proximity with the victims greatly influences the degree of reaction to the risk.

LESSONS FOR THE ENGINEERS

Engineers have the challenge to face/overcome the following two different public conceptions of safety.

1. **Positive or optimistic attitude**: Some people assume that things that are familiar, that have not hurt them before and over which they some control, present no real risks.

2. **Negative or pessimistic attitude**: some people feel feared, when an accident kills or harms in large numbers, or affects their relations, they consider those risks as high risks. Therefore, while designing a thing engineers should recognize and consider such widely held perceptions of risks along with their routine technical design issues. So engineers should recognize this as pt of their work.

Engineers should also understand that it is not wise to proceed under an assumption that education will quickly change the people's under-estimation or over-estimation of the risk. The continuous, proper

information about dangers and other issues of risk are necessary to educate the people to have right attitude and perception about the risk.

The risk communication and risk management efforts should be structured as a two way process.

RISK BENEFIT VALUE FUNCTION

The risk and benefits are based on the perception of probable gain and probable loss. A typical risk-benefit value function illustrated.

The risk benefits value function drops sharply on the loss portion that it raises on the gain portion.

One can see that the threshold is added on both gain and loss sides of the function.

The threshold on the loss side is to account for human habit of ignoring smaller risks in order to avoid anxiety overload. It means that no effect is spent to overcome the loss sides of the function.

The threshold on the gain side is to account for normal human inertia. It means that the inherent character of people may delay the process of seeking their own gain.

TYPES OF ACCIDENTS

Engineers should have the knowledge about different nature of accidents so that to try to prevent them. The three important types of accidents are procedural, engineered, and systemic accidents.

1. Procedural Accidents

Procedural accidents are the result of someone making a bad choice or not following established standard procedures.

Example: Road accidents because of not following the rules, in the airline industry, procedural accidents are frequently labeled as “pilot error”. These are accidents caused by misreading, flying when the weather should have dictated, or failure to follow regulations and procedures. In the airline industry, this type of error is not restricted to the pilot; it can also be committed by air traffic controllers and maintenance personnel.

2. Engineered Accidents

Engineered accidents are caused by errors in the design.

Because of minor design error, the device may not perform as expected or the device may not perform well under all circumstances.

Engineers should have the knowledge and experience to anticipate all possible engineered failures during the design stage itself.

Example: Minor casting defects in aircraft turbine blades may cause failure of the system.

3. Systemic Accidents

Systemic accidents are difficult to understand and difficult to control.

They are characterized of very complex technologies and the complex organization that are required to operate them.

Example: Failure of use space shuttle, in which 7 astronauts including Indian born astronaut Chawla were killed.

ASSESSMENT OF SAFETY AND RISK

Relationship between safety, risk and cost

It is always a great challenge to engineers to balance quality and safety against cost. In general, engineer’s tendency is to design and produce high-quality products, but business managers tend to keep the cost down.

Therefore it is necessary to understand the relationship between safety, risk, costs and price. The relationship between them is depicted.

Why both low-Risk and High-Risk Products are Costly?

- A product cost may have two elements:

1. Primary cost of product, and
2. Secondary cost of product.

- The primary cost of product (p) includes production cost and cost of safety measures involved.
- The secondary cost of product(s) includes costs associated with warranty expenses, los of customer good will, litigation, possible downtime in the manufacturing process, etc.

		Safety	
		High	**Low**
Risk	**High**	High safety and High risk High cost , High price Example: Nuclear plant, aircraft, missiles.	Low safety and High risk Low cost, High price Example: automobiles.
	Low	High safety and low risk High cost, medium price Examples: Electrical products, safety values.	Low safety and Low risk low cost , Low price Examples: Electronic goods, computers.

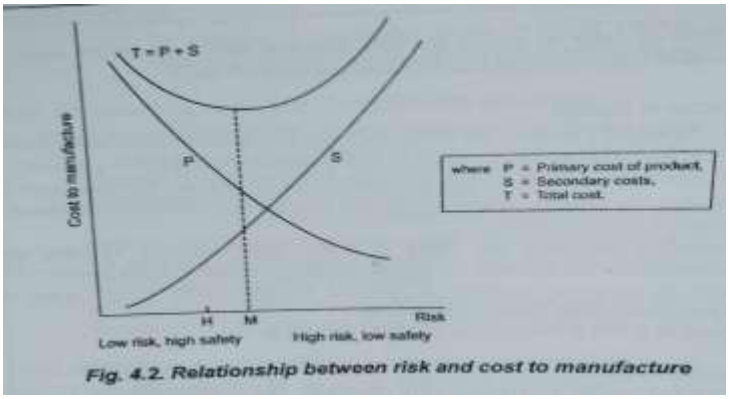


Fig. 4.2. Relationship between risk and cost to manufacture

Figure illustrates Relationship between risk and cost to manufacture.

- A stress on low risk and high safety leads to high primary costs and lower secondary costs, and vice versa.
- The total cost is the sum of primary and secondary costs of product.
- The total cost is minimum at point M, where the slopes of the primary and secondary cost curves are equal in magnitude but opposite in direction.

Therefore it is evident that both low-risk and products are costly. The optimal total cost (M) occurs in between low-risk and high-risk region.

DETERMINATION OF RISK

In order to determine the risk, one should have knowledge about the following criterions.

1. Knowledge of Risk

To assess a risk, an engineer must first identify it. To identify a risk, an engineer must first know the information about the safety of standard products.

Though past experience and historical data provide good information about the safety of standard products, still it is insufficient to completely assess the risk of a product.

The past experience and historical data are inadequate to access the risk, because of the following reasons:

The information is not freely shared among firms, and

There are always new applications of old technology that makes the available information less useful.

Therefore in order to access the risk, engineers and firms should share the information and knowledge about the safety of freely.

2. Uncertainties in Design

While designing a product, the design engineer must deal with many uncertainties. Many of the risks can be expressed as probabilities and as educated guesses.

The uncertainties are in the form of application of the product, materials used for producing the product, changing economic conditions, unfavorable environment conditions, temperature, etc.

Traditionally engineers use 'factor of safety' while designing to cop with uncertainties about materials and actual operating conditions of the product. The factor of safety is proposed to account for unpredictably high loads or unaccountably weak construction material.

A product is said to be safe if its capability exceeds its duty.

3. Testing for safety

Once the product is designed, both prototype and finished devices must be thoroughly tested.

The testing is not just to determine whether the product meets the specification. It should also involve testing to see if the product is safe.

It is essential that in any engineering design, all safety systems be tested to ensure that they work as planned

Different Approaches for Testing

The various testing approaches other than destructive testing are given below:

1. **Scenario Analysis:** This test starts from a given event, and then studies the different consequences that might develop from it.
2. **Failure modes and effects Analysis:** This approach systematically examines the failure modes of each component without focusing on causes or relationships among the elements of a complex system.
3. **Fault-Free analysis:** This approach proposes a system failure and then traces the events back to possible causes at component level.
4. **Event-free Analysis:** This is the reverse of the fault –free analysis. This analysis is very useful in identifying a potentially hazardous situation in the plant.

Out of these several techniques, the fault-free analysis is most effective, because it explains the possible approaches for proper functioning and safety of a complex system.

RISK-BENEFIT ANALYSIS

What is risk-benefit analysis?

Risk benefit analysis is a technique, similar to cost-benefit analysis, used to analyze the risk in a project and to determine whether the project should be carried out or not.

Risk-benefit analysis answers the following questions:

- What are the benefits of the product/project?
- Is the project/product worth the risks connected with its use? and
- Do benefits outweigh the risks?

It is understood that everyone is ready to certain levels of the risk as long as the product/project/activity promises sufficient benefit or gain.

In risk-benefit analysis, the risks and befits of a product/project are assigned money values, and the most favorable ratio between risks and benefits is determined.

Conceptual Difficulties in Risk-Benefit Analysis

Risk benefit analysis is a very difficult process, because of the following reasons:

- In risk benefit analysis, both risk and benefits are very difficult to quantify, because both lie in the future. That is both risk and benefits are associated with uncertainties.
- It should be noticed that who takes the risk and who enjoys the benefits? Therefore it is important to ensure that those who have taken the risks are the beneficiaries of it.
- It is mostly difficult to expressed both risk and benefits in a common set of units . For example , when the risks can be expressed and measured in one set of units and benefits in another then very difficult to do risk-benefits analysis. In this case, risk- benefit analysis is used to judge the relative merits of different designs.

Ethical Implications on Risk Benefit Analysis

While performing the risk-benefit analysis, one should keep in mind the following ethical questions:

a) Under what conditions, someone in society is entitled to impose a risk on someone else on behalf of a supposed benefit to others?

b) How can we consider the worst-case scenarios of person exposed to maximum risks while they are also obtaining only minimum benefits? Are their rights violated? Are they provided safer alternatives?

PERSONAL RISK

If sufficient risk is given to a person, then he can be able to decide whether to participate in risky activity or not.Many experiments have concluded that individuals are more willing to face voluntary risk than involuntary risks, even when the voluntary risks are more harmful than the involuntary ones.Personal risks are difficult to assess especially if they are involuntary personal risks.

Examples for personal risks:

- A person living near a chemical plant voluntarily or involuntarily; and
- A person working in a nuclear power plant or oil refinery plant.

The qualification in assessing personal safety and risk is very difficult to estimate. While assessing the personal risk, one should consider the following ethical question:

- How to access the money value of an individual's life?

- On what basis, the compensation for the risk can be decided?
- Is the compensation for a risk by an amount based on the exposure tolerance of the average person justifiable?
- What will be the compensation if the tolerance level of the person is below or above the average tolerance level?

In order to minimize the above difficulties in assessing personal risks, the analysis employ all the available quantitative measures such as

- Making judgments on the basis of the amount of life insurance taken out by an individual; and
- Assessing a hazardous job by looking at the increased wages a worker demands to carry out the task.

Public Risk and Public Acceptance

Risks and benefits to the public at large can be more easily determined than the personal risks and benefits. Because individual difference tend to even out as large numbers of people are considered.

Assessment studies relating to technological safety can be conducted in a better manner for public risk than for personal risk, as statistical parameters take on greater significance.

In this regard, American's National Highway Traffic safety Administration (NHTSA) has emphasized the following two points:

1. A value human life can be estimated based on loss of future income and other costs associated with an accident.

2. An estimation of quantifiable losses in social welfare is not based on the maximum expenditure allocated to save a life.

Accounting Publicly For Benefits and Risks

Public accountability for risk has been affected by the following problems:

- An expert or even group of experts cannot be expected to know everything. Hence the public processes suffer from incomplete engineering knowledge.
- The uncertainty produced by scientists and regulators also infects the risk regulation. In others words, a refusal to face the hard questions created by lack of knowledge affects the risk regulation.
- Since the conception of risk vary depending on how the facts are presented, therefore special caution should be given when stating probabilities of rare events.

Becoming A Responsible Engineer Regarding Risk:

- The engineer can provide background material to prove the faulty positions.
- Engineers should actively participate in the debates related t safety and risk.
- Engineers should always insist on meaningful numbers and figures when assessing safety and risk.
- Engineers should also recognize the previously mentioned difficulties with measuring risks and benefits in absolute terms.
- Engineers should not be influenced by any influential lobby or trade organization.
- Engineers need to be sensitive to various quantitative value judgments related with human and ethical values.
- Engineers should be aware at the legal liabilities regarding risk.

REDUCING RISK

- As we know, it is impossible to design and manufacture anything to be completely risk free. However it is the responsibility of the engineers to explore all the possible ways to reduce the risk under the given financial and time constraints.
- **Risk management defined:** Risk management is defined as the eradication or minimization of the adverse effects of the pure risks to which n organization is exposed.
- **Elements of a risk management programme** : According to the recent health and safety legislation, the three important elements of a risk management programme are:
- **Risk identification**
- **Risk evaluation**
- **Risk control**

1. Risk Identification

- Risk can be identified by various techniques such as physical inspection, safety audit, job-safety analysis, management and worker discussions, and historical data analysis.

2. Risk Evaluation

- Risk can be measured on the basis of economic, social or legal considerations.
- Economic and social considerations include financial aspects, uninsured cist of accidents, insurance premium, overall effect on the probability, and possible loss of production.
- Legal considerations include possible constraint from compliance with health and safety legislation, code of practice, guidance notes and accepted standards, fire prevention, pollution and product liability.

3. Risk Control

- **Risk control consists of four areas**: risk avoidance, risk retention, risk transfer and risk reduction.
- **Risk avoidance**: It refers to the conscious decision by the management to avoid completely a particular risk by discontinuing the operation producing the risk.
- **Risk retention**: It refers to the legal assignment of the cost of certain potential losses from one party to another.
- **Risk reduction**: It refers to the reduction or elimination of all aspect of accidental loss that lead to a wastage of n organization's assets.

Faulty Assumptions and Their Realities about Safety

There are many misconceptions about safety, some of the popular fault assumptions and their realities about safety are given below:

1. **Assumption**: The principle causes of all accidents are operator error and negligence.

Reality: Accidents are caused by dangerous conditions, which can be corrected.

2. **Assumption**: Producing a safe product always increases the costs.

Reality: If safety is incorporated in the design stage itself, the initial cost will be reduced. If there are any later design changes, it will increase the product cost.

3. **Assumption**: we learn about safety after a product has been completed and tested.

Reality: If safety is not built into the original design, people can be hurt during the testing stage. Unwillingness to change a design means compromising on safety.

4. **Assumption**: Warnings about harms are sufficient. Insurance coverage is cheaper than planning for safety.
Reality: warnings can provide minimal protection against harmful events. Insurance rates are very high.

CASE STUDY OF THE THREE MILE ISLAND

Abstract

The accident at the three miles island unit-2(TMI-2) nuclear power plant in Pennsylvania on March 28, 1979 was one of the most serious in the history of the U.S. nuclear industry. It not only brought to light the hazards associated with nuclear power, but also force the industry to take a closer look at the operating procedures used at the time. What makes the TMI-2 accident such an interesting case study is the series of events, which led up to the partial meltdown of the reactor core. It was a combination of human error, insufficient training, bad operating procedures and unforeseen equipment failure that culminated in a nuclear accident that could have easily been prevented.

CASE STUDY OF THE CHERNOBYL DISASTER

Abstract

The April 1986 disaster at the Chernobyl nuclear power plant in the Ukraine was the product of a flawed soviet reactor design coupled with serious mistakes made by the plant operators in the context of a system where training was minimal. It was a direct consequence of cold war isolation and the resulting lack of any safety culture.

The accident destroyed the Chernobyl-4 reactor and killed 30 people, including 28 from radiation exposure.A further 209 on site were treated for actuate radiation poisoning and among these, 134 cases were confirmed. Nobody off-site suffered from acute radiation effects.However the large areas of Belarus, Ukraine, and Russia and beyond were contaminated in varying degrees.

The Chernobyl disaster was a unique event and the only accident in the history of commercial nuclear power where radiation-related fatalities occurred.

RESPECT FOR AUTHORITY

WHAT IS MEANT BY AUTHORITY?

Authority is the right to make decision, the right to direct the work of other, and the right to give orders.

It is a crucial factor in organization since engineers and employees must be authorized to carry out the jobs assigned to them.

SOURCES OF AUTHORITY

Authority derives from several sources. They are the person's position or rank, and personal attributes such as Knowledge and expertise

1. INSTITUTIONAL AUTHORITY

DEFINITION: Institutional Authority can be defined as the institutional right given to a person to exercise power based on the resources of institution.

It is Authority given by institution to the qualified individual in order to meet their institutional objectives.

This Authority is exercised by the making policy decisions, allocating resources, issuing orders, carrying out the actions, giving recommendations, etc.,

LIMITS OF INSTITUTIONAL AUTHORITY:

In a company, the institutional Authority is given by the owners or stockholders of the company .In practice, sometimes the owners of the company delegate the institutional authority to ineffective and incompetent individuals. Those individuals may be unable to exercise their authority effectively in order to meet the company's objective.

2. EXPERT AUTHORITY:

Apart from institutional authority, there is an authority based on the knowledge and expertise.

Expert authority is the possession of special knowledge, skill or competence to perform some task or to give sound advice.

It is proved that the leaders with the expertise can more effectively guide and motivate others 15 then the conventional leaders. This concept is referred as 'authority of leadership'

In today's organization setup, the staff engineers, advisors, and consultants are given expert authority, while institutional authority is assigned to the line managers.

AUTHORITY Vs POWER

The differences between institutional authority and power are as follows.

S.NO	AUTHORITY	POWER
1.	It is a legal right of a superior which compel his subordinates to perform certain acts	It is the ability of the person to influence the others to perform an act. It may not have legal sanction.
2	It is delegated to an individual by his superior	It is earned by an individual through his own effort
3	It lies in the position held and the authority changes with change in position.	It rests the individual. Even when the position has changed, his power remains with him.
4	It is mostly well defined and finite	It is undefined and infinite.

MORALLY JUSTIFIED AUTHORITY

The institutional authority assigned to employee may ensure in achieving the institutional objectives. But those institutional rights and duties should necessarily be morally justified institutional rights and duties.

The institutional authority is set to be morally justified, only when:

- The goals of the institution are morally permissible or morally desirable, and

- The way of implementation should not violate basic moral duties.

Therefore engineers should have moral obligations to perform only morally permissible institutional duties, when they accept employment.

Accepting Authority

Employees accept/ recognize their employers' authority by accepting the guidance and obeying the directives issued by the employer. Rarely employees disobey an order on moral grounds.

According to Herbert Simon, "a subordinate is said to accept authority whenever he permits his behavior to be guided by the decision of a superior, without independently examining the merits of that decision".

Simon also noted that all employees have limits on "zone of acceptance" in which they are willing to accept authority.

Generally employees are not interested to make an issue of every incident of questionable morality, because of fear of losing their job/position.

Therefore the 'zone of acceptance' can be used as a measure of the lack of individual moral integrity.

Paramount obligations

The codes of ethics of the professional societies state that an engineer's paramount obligation is to protect public health, safety and welfare rather than the obligations of loyalty and faithful service to employers.

As professionals, engineers have obligations to accept their employers' institutional authority. But this does not mean that they have to obey obligations blindly. Therefore the basic moral task of engineers is to be aware of their obligations to obey employers on the one hand and to protect and serve the public and clients on the other hand.

Engineers must weigh their obligations to the public, their employers, their colleagues and others when conflicts between such obligations rise.

COLLECTIVE BARGAINING

What is meant by collective bargaining?

- International Labor Organization (ILO) has defined collective bargaining as "negotiation about working conditions and terms of employment between an employer and one or more representative employee's with a view to reaching agreement".
- The process is collective in the sense that the issue relating to terms and conditions of employment are solved by representatives of employees and employers rather than individuals.
- The term bargaining refers to evolving an agreement using methods like negotiation, discussion, exchange of facts and ideas rather than confrontation.

Process of collective bargaining:

The process of collective bargaining can be summarized in the following three steps:

Step – 1: Presenting the character of demands by the union on behalf of the constituent elements.

Step – 2: Negotiations at the bargaining table.

Step – 3: Reacting an agreement.

Unionism and Professionalism:

- Collective bargaining assumes 'unionism'. Legally, any organization employing more than 20 employees could have a union. In organizations, more than one union is also permitted.
- The employers form unions to safeguard the interests of employees and to prevent exploitation of employee.
- Many professional managers have argued that the ethical aspects of professionalism in engineering are inconsistent with union ideology and practice.
- According to John Kemper, the unionism and professionalism are conflicting with each other. Professionalism offers paramount importance to the interests of society and of employer. But unions, also known as collective bargaining agents, consider the economic interests of the members ahead of the interests of their employer.
- Also, a number of professional societies have emphasized that loyalty to employers and the public is not possible with any form of collective bargaining.
- Even many professional societies indirectly instruct the engineers that they should not become shall not actively participate in strikes, picket lines, or other collective coercive action.
- Thus professional societies oppose unionization because of the issue of conflicting loyalties and on the grounds that is unprofessional.
- In a nutshell, the general view is that it is impossible for the engineer to belong to a union and at the same time to maintain the standards of his profession.

CONCLUSION:

For the above discussions, the following conclusion can be made:

1. We can absorb whether collective bargaining and its tactics are ethical or unethical, only on the basic of the given situation.
2. Though unions often have misused their power and irresponsibly disregarded the public good, the formation of engineering unions should not be considered always unprofessional.
3. The moral assessment of unions is complex. Many morally relevant fact and factors should be considered while judging about any union.

ARGUMENTS OVER UNIONS

There are two arguments in favor of and against unions.

1. ARGUMENTS IN FAVOR OF UNIONS

- Unions play a vital role in achieving high salaries and improved standard of living of employees.
- Unions give employees a greater sense of participation in organization decision making.
- Unions ensure job security and protection against arbitrary treatment to the employees.
- Unions have the ability to resist any orders from the employers to perform unethical acts.
- Unions maintain stability by providing and effective grievance procedure for employee complaints.
- Unions can act as counter force to radial political movements that exploit the employees.

2. ARGUMENTS AGAINST UNIONS

- Unions shatter the economy of a country by playing DISTORTING influence on efficient users of labor.
- Unions remove person to person negotiations between employers and employees. Thus an individual is not given much important in the process of collective bargaining.
- Unions encourage unrest and strained relations between employees and employee.
- Unions encourage the unhealthy concept of job promotion, salary hike, etc on the basis of seniority.
- Unions prevent employer from rewarding individuals for their personal achievements.

EXTERNAL RESPONSIBILITIES

- External responsibilities refer to the responsibilities of the engineers to the outside world.
- The responsibilities to the outside world include :
    1. Confidentiality
    2. Conflict of interest
    3. Occupational crimes.

CONFIDENTIALITY

What is confidentiality?

- It is widely accepted that the engineers have an obligation to keep certain information of the employer/client secret or confidential.
    - Just as with lawyers and medical physicians, engineers also require the confidentiality principle in their profession. For example, lawyers, doctors, and counselors keep information of their clients/patients confidential. In the same way engineers have an obligation to keep proprietary information of their employer or client conflidential.
- Confidentiality is highly emphasized in most engineering codes of ethics. For example, the NSPE states that engineers shall not reveal facts, data or information obtained in a professional capacity without the prior consent of the employer or client authorized by law or this code.

Confidential Information:

- Confidential information is information deemed desirable to keep secret.
- According to the codes of ethics of ABET, Engineers shall treat confidential information coming to them in the course of their assignment as confidential".
- The most commonly considered criterion on the confidential information is as follows: Confidential information is any information that the employer or client would like to keep secret in order to compete effectively against business rivals.
- **In general** confidential covers both sensitive information given by the employer or client and information gained by the professional in work paid by the employer orclient.

TERMS RELATED TO CONFIDENTIAL INFORMATION

1. Privileged Information

- IT refers information that is available only on the basis of special privilege. That is, information available to an employee who is working on a special assignment.
- It includes information that has not yet become to public or known within an organization.
- This term is often used as a synonym for confidential information.

2. Proprietary Information

- It is the information that is owned by a company.
- It refers to a new knowledge established within the organization that can be legally protected from use by others.
- This term is often used as synonym for 'property' and 'ownership'.

3. Trade secrets

- A trade secret can be any type of information that has not public and which an employer has taken steps to keep secret.
- These trade secrets may be about designs, technical processes, plant facilities, quality control systems, business plans, marketing strategies and so on.
- Trade secrets are given limited legal protection against employee or contractor abuse. In the sense, an employer can sue employees or contractor for leaking trade secrets or even for planning to do so.

4. Patents

- **Patents legally** protect specific product from being manufactured and sold by competitors without the permissions of the patent holder.
- **Patents vs. trade secrets:** A patent holder has legally protected monopoly power. But in case of trade secrets, the legal protection is limited to keeping relationships of confidentiality and trust.

Why must engineering be kept confidential?

Many information such as privileged information, proprietary information, and trade secrets are very important for a company to compete in the market. If such information is leaked to competitors, then the competitors may gain competitive edge and may capture the market. Therefore it is in the company's best interest to keep such information confidential as much as possible.

What type of information should be kept confidential?

Some of the type information that should be kept confidential is:

- Information about the unreleased products.
- Test results and data about the products.
- Design or formulas for products
- Data should technical process.
- Organization of plant facilities.
- Quality control procedures.
- Business information such as number of employees working on project, the suppliers' list, marketing strategies, production costs, and production yields.

Justification and limits of confidentiality:

The obligation of confidentiality can be justified at two levels:

First Level: It focuses on three moral considerations- respect for autonomy, respect for promises, and regard for public well-being.

Second Level: It focus on the major ethical theories. It includes justification of confidentiality by right ethicists, duty ethicists and utilitarian.

We shall discuss the above considerations in the following sections.

1. First level of justification of the confidentiality obligation

(a) Respect for autonomy:

- It refers to respect the autonomy of freedom and self-determination of individuals and companies in order to recognize their legitimate control over some information.Without the legitimate control, the individual and companies cannot maintain their privacy and protect their self-interest.

21

(b) Respect for promises:

- It refer to respect the promise (in the form of signing contracts) made by employees to the employer.
- It is the duty of the employee to respect the promises made to the employer.

(c)Respect for public well-being:

- There are public benefits in recognizing confidentiality, relationship within professional context.

- The economic benefits of competitiveness to the public can be promoted only when companies maintain some degree of confidentially concerning their product s.

2. Second level of justification of the confidentiality obligation:

(a)Justification by right ethicists:

- The right ethicists justify employees' obligation of confidentiality by appealing to basic human rights.
- These ethicists argue that the rights of employers, to establish what information should be treated as confidential, should be limited by other basic human rights.
- For instance, no employer should be given a right to safeguard proprietary information by preventing engineers from whistle blowing in cases where those leaked information would save human lives and their rights.

(b) Justification by duty ethicists:

- The duty ethicists insist on the basic duties of both employers and employees to maintain the trust and to commit themselves to an employment agreement they have made.
- They also emphasize that nobody should abuse the property of others.

(c) Justification of utilitarian's:

- **View of rule-utilitarian's**: Rule-utilitarian's view rules governing confidentiality as justified to the extent that such rules protect the most good for the greatest number of people.
- **View of act- utilitarian's**: Act- utilitarian's focus on each situation when an employer decides on some information to be confidential information.

Changing jobs and confidentiality:

- The obligation of protecting confidential information is not over when employees change jobs.
- Legally, an engineer is expected to keep information confidential even after the employee has moved to a new employer.

Management policies for maintaining confidentiality:

Some general management policies for maintaining confidentiality are as follows:

Approach 1 : T o use employment contracts that place special restrictions on future employment. This type of agreement the right of individuals to proceed their careers freely.

Approach 2: To use an employment contact that offers positive in exchange for the restrictions it places on future employment.

Approach 3: To offer an employee a special post-employment annual consulting fee for several year on the condition that he will not work for a direct competitor during that period.

Approach 4: To tighten the security control on the internal flow of information by restricting access to trade secrets. This may create an unhealthy working atmosphere of distrust.

Approach 5: To have unwritten and informal agreements among competing companies not to hire another's more important employees.

However, a better feasible solution is that the employers have to create a sense of professional responsibilities among their staffs that beyond merely following the employment agreements.

CONFLICTS OF INTEREST

What is a conflict of interest?

- In general, conflicts of interest means an individual has two or more desires that all interests cannot be satisfied given the circumstances.
- Professional conflicts of interest are situations where professionals have an interest, if pursued, could keep them from meeting one of their obligations to their employers. Examples:
    - An employee working in a company depositing a substantial investment in a competitors company.
    - An employee working in a company serving as a consultant for a competitors company.

Difference between general conflicts of interest and professional conflicts of interest

- ✓ In general conflicts of interest, satisfying all desires/interests of a person cannot be possible because of physical or economical or other problems. By contrast, the professional conflicts of interest cannot be pursued only because of moral or ethical problems.

Types of conflicts of interest

The three important types of conflicts of interest are:

1. Actual conflicts of interest.
2. Potential conflicts of interest and
3. Apparent conflicts of interest.

1. Actual conflicts of interest

- ✓ The actual conflicts of interest arises when an employee compromise objective engineering judgment.
- ✓ It refers to the loss of objectively in decision-making and inability to faithfully discharge professional duties to employer.
- ✓ Example: A mechanical engineer working in the purchase department of an automobile industry might have his personal influence while offering the contract for supply of raw materials to a vendor .In this case pursing his financial interest with the vendor might lead him not to objectively and faithfully discharge his professional duties to his industry.
- ✓ Thus actual conflicts of interest can corrupt professional judgement.

2. Potential Conflicts of interest

- ✓ The potential conflicts of interest may corrupt professional judgment in the future, if not in the present.
- ✓ Although potential conflicts of interest may not harm the interest of the employer initially, there is a threat that potential conflicts of interest will become actual conflicts of interest at later stage.
- ✓ Example: An engineer becoming a friend with a supplier for his company. In this case,

the engineer may not have conflicts of interest initially. However, in future he may favour his friend, as in the case of actual conflicts of interest.

3. Apparent Conflicts of Interest

- ✓ There are situations in which there is the appearance of a conflict of interest. This type is referred as apparent conflicts of interest.

- ✓ Apparent conflicts of interest actually not corrupting the professional judgment. However it decreases the confidence of the employer and the public in the objectively and trustworthiness of professional services. Thus it harms both the profession and the public.
- ✓ Example: Consider a situation, where a design engineer is based on a percentage of the cost of the design and there is no incentive for him to reduce the costs down. In this context, it may appear that the engineer will

make the design more expensive in order to earn more commission for himself. This appearance of conflict of interest may cause the distrust on the engineer's ability to perform his professional duties.

What Do The Engineering Codes of Ethics Say About Conflicts Of Interest?

Some of engineering codes that address conflicts of interest are given below:

1. Fundamental canons of the NSPE code says:

- ✓ "Engineers shall not be influenced in their professional duties by conflicting interests".
- ✓ "Engineers shall not accept financial or other considerations, including free engineer designs, from material or equipment suppliers for specifying their product".
- ✓ 'Engineers shall not accept commissions or allowances, directly or indirectly, from contractors with work for which the Engineer is responsible".

2. Fundamental canons of ethics of ABET says:

- ✓ "Engineers shall not solicit nor accept gratuities, directly or indirectly, from contractors, their agents, or other parties dealing with their clients or employers in connection with work for which they are responsible".

Conflicts of Interest and Accepting Gifts/Bribes

- ✓ Mostly engineers find themselves in actual, potential, or apparent conflicts of interest are those involving accepting gifts.
- ✓ **What is a bribe?**
  - A bribe is something such as money or a favor, offered or given to someone in a position or trust in order to induce him to act dishonestly.
  - It is something offered to influence or persuade.
- ✓ **What are the ethical reasons for not tolerating bribery?**

Bribes are illegal and immoral because of the following three reasons:

1. Bribery corrupts free market economy system and it is anticompetitive.

2. Bribery corrupts justice and public policy by allowing rich people to make all the rules. In today's business,It is understood that only large and powerful companies will survive, since they are capable of providing bribes.

3. Bribery treats people as commodities that can be bought and sold. This practice degrades the human beings and corrupts both the buyer and the seller.

- ✓ **What is meant by the term 'kickbacks'?**
  - Kickbacks are another form of bribing.
  - Prearranged payments made by contractors to companies or their representatives in exchange for contracts actually granted are called 'kickbacks'.
- ✓ **When is a gift a bribe?**
  (What are the difference between gift and bribe?)
  - Gifts are not bribe as long as they are gratuities of smaller amounts. But bribes are illegal and immoral because they are worth of substantial amounts.
  - Gift may play a legitimate role in the normal conduct of business whereas a bribe influences the judgment.
  - In olden days the following thumb rule was applied: "A gift is a bribe if one can't eat, drink or smoke it in a day".
  - Today a more appropriate thumb rule says:
    "If you think that your offer of a particular gift would have grave or merely embarrassing consequences for your company if made public, then the gift should be considered as a bribe".

What is moonlighting? Does it create conflicts of interest?

- The term moonlighting is used when employee of a company works for another company during his spare time'
- Moonlighting creates conflicts of interest only in special circumstances, such as working for competitors, suppliers, or customers.

Different ways to avoid Conflicts of interest:

It is understood that all conflicts of interests, whether actual, potential, or apparent should be avoided in all possible ways. Some of the effective ways to avoid conflicts of interests are as follows:

- To follow the guidance of company policy.
- In the absence of company policy, one can go for a second opinion from a coworker or manager.
- In the absence of above two options, It is better to examine one's own motives and use ethical problem solving techniques.
- Finally, one follows the statements in the professional code of ethic. Some of the codes have given clear statements to identify whether the given situation is a conflict of interests or not.

What Are Occupational Crimes?

- ✓ *Occupational crimes are illegal acts committed through a person's lawful employment.*
- ✓ It is the secretive violation of laws regarding work activities.
- ✓ When professionals or office workers commit the occupational crimes, it is referred as *'White Collar Crime'*.
- ✓ Most of the occupational crimes are special instances of conflicts of interest. These crimes are motivated by personal greed, corporate ambition, misguided company loyalty, and many other motives.
- ✓ Even crimes that are aimed at promoting the interest of one's employer rather than oneself are also considered as occupational crimes.
- ✓ Occupational crimes impinge on various aspects such as professionalism, loyalty, conflicts of interest, and confidentiality.

Examples of Occupational Crimes

Three cases of occupational crimes that are commonly observed are:

1. Price fixing.
2. Endangering lives and
3. Industrial espionage i.e. industrial spying.

1. Occupational Crimes of Price Fixing

- ✓ While fixing a price for any commodity/product/service, sometimes all competitors come together and jointly set the prices to be charged. These are called as *Pricing cartels.*
- ✓ The above price fixation is unfair and unethical practice. This leads to restraint the free trade and open competition. Thus the above kind of price fixing is an example of occupational crime.
- ✓ ***Case Illustration:*** In 1983, in American state of Washington, six large electrical contractors along with eight company presidents and vice presidents were indicted on charges of fixing bids (contracts) for building public power plants. This is evident instance of occupational crime.
- ✓ In order to avoid the above kind of occupational crimes, the laws are enforced which forbids companies from jointly fixing prices.

2. Endangering Lives

- ✓ Endangering the lives of employees is another kind of occupational crime.
- ✓ Some companies employ workers without disclosing them the harmful health effects and safety hazards about the working environment and the product to be manufactured. In due course of time, workers are exposed to very serious health problems. In this case, the employers are guilty of involved in an occupational crime.
- ✓ ***Case Illustration:*** Manville Corporation, the largest producer of asbestos in U.S, knew that asbestos dust was harmful for their employees' health. It could cause a lung disease named 'asbestosis' and an incurable cancer named 'mesothelioma'. The company kept this

information secret from the employees and the public.
During 1940-1979, over 27 million workers were exposed to asbestos and more than 1,00,000 workers have died. Many victims and their families have successfully filed civil suits to claim damages.

- ✓ The above shocking case study is the typical illustration of an occupational crime committed by the Manville Corporation.

3. Industrial Espionage

- ✓ Industrial espionage means industrial spying. *Espionage refers secret gathering of information in order to influence relationships between two entities.*
- ✓ Keeping information secret is a right. But acquisition of others' secret to one's advantage is espionage. The espionage is one of the most unethical and lawless activities.
- ✓ The vital information are secretly gathered/theft through espionage agents (also called spies).
- ✓ Industrial or corporate espionage is the theft of trade secrets of economic gains. The trade secret may be any of the intellectual properties such as designs, prototypes, formulas, software codes, passwords, manufacturing processes, marketing plans, supplier/contractor details, etc.
- ✓ From the above discussion, it is clear that the industrial espionage is also a typical occupational crime existing in our society.

RIGHTS OF ENGINEERING

PROFESSIONAL RIGHTS AND EMPLOYEE RIGHTS

Engineers not only have many responsibilities, they also have rights to go along with these responsibilities. Always rights, duties and responsibilities go together. In fact, rights and responsibilities go together. In fact, rights and responsibilities are two sides of the same coin.

In most popular definition of right states, "A right is a valid claim to something and against someone which is recognized by the principles of an enlightened conscience".

TYPES OF RIGHTS:

The concept of rights can be categorized into the following three types:

1. Human rights
2. Employee rights
   - (a) Contractual rights,
   - (b) Non-contractual rights.
3. Professional rights

1. Human Rights

- ✓ Human rights are the rights possessed by virtue of being people or moral agents.
- ✓ The fundamental human rights adopted by the united Nation's international bills of human rights are listed
  - (1)Rights to life
  - (2)Rights to liberty
  - (3)Rights to security of person

(4)Rights not to be held in slavery
(5)Rights not to be tortured or subjected to inhuman or degrading punishment (6)Rights to recognition before the law
(7)Rights to impartial trail and protection from arbitrary (8)Rights to freedom to movement
(9) Rights to marriage
(10) Rights not to marry without free consent
(11)Rights to property ownership
(12)Rights to freedom of thought
(13)Rights to peaceful assembly and participation in government.
(14)Rights to social security and work
(15) Rights to education
(16) Rights to participate in and from trade unions
(17)Rights to nondiscrimination

(18)Rights to a minimal standard of living

- ✓ Thus engineer, as human being, have human rights to live and freely practice their legitimate interests.

2. Employee Rights

- ✓ Employee rights are the rights that apply or refer to the status or position of employee

Types of employee rights:

- ✓ (1)Contractual employee rights.
- ✓ (2)Non-contractual employee rights.

(1) Contractual Employee Rights

- ✓ These employee rights are institution rights that arise only due to specific agreements in employee contact.
- ✓ Examples: The contractual employee rights include
    - Rights to receive a salary of a certain amount
    - Rights to receive other company benefits such as bonuses, salary increment, etc.

(2) Non-Contractual Employee Rights

- ✓ These are Rights existing even if not formally recognized in the specific contracts of company policies
- ✓ Examples: The Non-contractual employee rights include
    - Right to choose outside activities;
    - Right to Privacy and employer confidentiality;
    - Right to due process from employer;
    - Right to nondiscrimination and absence of sexual harassment at the workplace

(3) Professional Rights

- ✓ Professional rights are the rights possessed by virtue of being professionals having special moral responsibilities.
- ✓ Examples: The professional rights include
    - Right to exercise one's professional judgment on the basis of his conscience.
    - Rights to refuse to involve in unethical activities
    - Rights to warn the public about harms and dangers
    - Rights to express one's professional judgment, include his rights to disagree
    - Rights to fair recognition and remuneration for professional services.

VARIOUS ASPECTS OF PROFESSIONAL RIGHTS

Rights of professional Conscience:

- ✓ It is one of the most fundamental Rights of engineers.
- ✓ The rights of professional conscience refer to the moral rights to exercise responsible professional judgment in discharging one's professional responsibilities.
- ✓ In simple words, it is the rights to do what everyone agree obligatory for the professional engineer to do.
- ✓ The Rights of conscience is a 'negative' right. because its place on obligation on other people not to interface with exercise
- ✓ In order to exercise the rights of professional conscience, engineers require special resources and support from others. So, this right is also considered as a 'positive' right, as it is placing on the other people an obligation to do more merely not interfering.

Specific Rights:

- ✓ The right of professional conscience is most general professional rights. It consists of many other specific rights. Two important specific rights are:
    (1).Right of conscientious refusal, and
    (2).Right of recognition.

1. Rights of conscientious refusal:

- ✓ The right of conscientious refusal is the right to engage in unethical behavior.
- ✓ According to this rights, no employer can force or pressure an employee to do something that the employee considers unethical unacceptable.
- ✓ The rights of conscientious refusal arises two situations:
    1. Where there is widely shared agreement in the profession regarding ethical and unethical acts, and
    2. Where there is a possibility for disagreement among the people over unethical acts.
- ✓ Thus the engineers should have a moral right to refuse in participating unethical activities. The

example of unethical activities are forging documents, lying, giving or taking bribes , selling the company secrets to others, information theft, etc.

- ✓ Also the employers should have a moral right not to force or pressurize the employee to participate any unethical activities. Employers should not use any revenge techniques against the employees in this regard.

2. Right to Recognition:

- ✓ The right to recognition refers to the engineer's right to professional recognition for their work and accomplishments.
- ✓ The recognition/reward may be of any one of the following types :

1. **Extrinsic Rewards:** These are related to monetary remunerations such as increased salaries, commissions, cash bonus, gain sharing, etc.

2. **Intrinsic Rewards**: These are related to non-monetary remunerations such as acknowledging achievements by issuing appreciations letters, certificates and oral praises, etc.

FOUNDATION OF PROFESSIONAL RIGHTS:

The basic professional rights discussed so far, can justified by the ethical theories as given below.

1. Rights Ethics:

- ✓ Rights ethics emphasizes that should all have human moral rights, and any action that violates these rights is unethical.
- ✓ Thus rights ethicists justify the basic right of professional conscience by referring to the human moral rights.

2. Duty Ethics:

- ✓ Duty ethics emphasizes that is duties should be performed, without considering much about moral rights.
- ✓ For example, if an individual has a right to do something, it is only because others have duties or obligations to support him to do so. In this view, the basic professional rights is justified by reference to others duties to support or not interfere with the work-related exercise of professional s conscience.

3. Utilitarianism:

- ✓ Utilitarian theory argues that the greatest good is promoted by allowing engineers to practice their obligations.
- ✓ Thus the utilitarianism justifies the right of professional conscience referring to the basic goal of producing the most good for the greatest number of people.

WHISTLE BLOWING:What does whistle blowing mean?

✓ Whistle blowing is the act by an employee of informing the public or higher management of unethical or illegal behavior by an employer or supervisor.

✓ It is the act of reporting on unethical conduct within an organization to someone outside of the organization in an effort to discourage the organization from continuing the activity.

✓ According to the codes of ethics of the professional engineering societies, engineers have the professional rights to disclose wrong doing within their organization and expect to take appropriate actions. Thus in a way, whistle blowing is also one of the professional rights of engineers.

✓ On the other hand, the employers/companies view whistle blowing as a bad exercise. Because they feel that whistle blowing can lead to distrust, disharmony, and an inability of employees to work together.

✓ **Example**: Journalists and media persons blow the whistle on politicians to bring out their corruption by publishing articles or informing regulatory authorities.

Whistle Blowing Defined:

✓ **General Definition:** Whistle blowing is alerting relevant persons to some moral or legal corruption, where relevant persons are those in a position to act in response, if only by registering protest.

Subjective Definition: Whistle blowing occurs when an employee or former employee conveys information about a significant moral problem outside approved organization channels to someone in a position to take action on the problem.

The definition has four main parts.

1. Disclosure: Information is intentionally conveyed outside approved organizational channels or in situations where the person conveying it is under pressure from supervisors or others not to do so.

2. Topic: The information concerns what the person believes is a significant moral problems for the organization. Examples of significant problems are serious threats to public or employee safety and well being, criminal behavior, unethical policies or practices and injustices to workers within the organization.

3. **Agent:** The Person disclosing the information is an employee or former employee.

4. **Recipient:** The information is conveyed to a person in an organization that is in a position to act on the problem. The desired response or action might consist in remedying the problem or merely alerting affected parties.

Types of Whistle Blowing:

The four type of whistle blowing are given below:

1. **Internal Whistle Blowing**: Internal Whistle blowing occurs when the information is conveyed to someone within the organization.
2. **External Whistle Blowing**: External Whistle blowing occurs when the information is passed outside the organization.
3. **Open Whistle Blowing:** Open whistle blowing also known as **acknowledged whistle blowing** occurs when the persons openly reveal their identity as they convey the information.
4. **Anonymous Whistle Blowing**: Anonymous whistle blowing occurs when the persons who is

blowing the whistle refuses to reveal his name when making allegations.

Moral Guidelines:

1. When is the whistle-blowing morally permissible?

Richard Degeorge has provided a set of criteria that must be satisfied before whistle blowing can be morally justified. Degeorge Believes that whistle blowing is morally permissible when the following three criteria are met:

1. If the harm that will be done by the product to the public is serious and considerable.

2. If the employees report their concern to their superiors.

3. After not getting satisfaction from immediate superiors ,regular channels within the organization have been used to reach up to the highest levels of management

4. There is reasonable hope that whistle blowing can prevent or remedy the harm.

2. When is the whistle-blowing morally obligatory?

Degeorge believes that whistle blowing is morally obligatory when the following two criteria is met:

1. If the employee has documented evidence that would convince a responsible impartial observer that his view of the situation is correct and the company policy is wrong; and

2. If the employee has strong evidence that making the information policy that public willin fact prevent the threatened the serious harm.

PREVENTING WHISTLE BLOWING

In order to solve the whistle blowing a problem within a company, any one of the following four methods can be used:

1. The company should create a strong ethics culture. There should be clear commitment to the ethical behavior from both employers and employees.

2. The organization should remove rigid channels of communication. Instead of they should encourage free and open communication system within the organization,

3. The companies should be Create ethical review committee with real freedom to investigate the complains and make independent recognitions to top management.

4. There should be willingness on the part of the management to admit mistakes, if necessary. This attribute will set an atmosphere for employee's ethical behavior.

EMPLOYEE RIGHTS:

What are the Employee Rights?

- ✓ As discussed in section employee rights are any rights, moral or legal, that involve the status or being of an employee.
- ✓ In fact the professional rights are also employee rights. For example the professional rights to express dissent about company policies are also evidently employee rights.
- ✓ Employee rights also include fundamental human rights relevant to the employment situation.
- ✓ As stated already, employee rights are of two types:
  - Contractual employee rights and
  - Non-Contractual employee rights.

Right To Outside Activities:

- ✓ As per the basic human rights all employees have the right to practice outside activities of their own interest without any interference from employers.
- ✓ However, the rights of employees to practice outside the activities and should not violate the duties and responsibilities to their jobs.
- ✓ Also the employers have the right to take action when outside activities create a conflict of interest,
- ✓ In addition employees have no right to damage their employer's interests even during non-working hours.

Rights To privacy:

- ✓ The right to personal privacy means the right to have a private life off the job.
- ✓ In other words, the right to privacy is also limited by the legitimate exercise of Employer's rights. The employers can obtain and use information of employees for their effective management of the company. However this personal information should not be given to outsiders.

Rights to Due Process:

- ✓ The rights to due process means right to fair procedures safeguarding protecting the exercise of other rights.
- ✓ This righty also extends to fair procedures in firing, demotion, and other disciplinary actions.
- ✓ In order to implement the right of due process, the following two general procedures can be used:
  1. Written explanations, specifying the reasons, should be given to employees who are penalized in any ways.
  2. An appeals procedure should be established so that an employee can appeal against their penalties if he believes his rights have been violated.

INTELLECTUAL PROPERTY RIGHTS (IPR):

What is an intellectual property?

1. Intellectual property is a property that results from mental labor.
2. The intellectual property is originating mainly from the activities of the human intellect.
3. Intellectual property is the information and original expression that derives its original value from creative ideas with a commercial value.
4. In the legal sense, intellectual property is a patterned invention, a trade secret or copyrighted material.
5. Like other properties intellectual property is also an asset which can be bought or sold licensed and exchanged.

What Are The Intellectual Property Rights?

1. Intellectual property is a class of property originated from the activities of the human intellect.

Any property movable or immovable is legally protected to prevent it from being stolen.

2. These rights are governed by the law on IPR of the country which grants such rights.

3. IPR allows the people to independently own their innovations and creativity, which is similar to

legal protection of any other properties.

4. The IPRs safeguard and encourage the innovations in the various sectors of the benefit of

this society.

Elements of Intellectual Property Rights

The WTO has established seven elements of IPRS, which were agreed by TRIPS, they are:

1. Patents.
2. Industrial Designs
3. Trade Marks
4. Copy Rights
5. Trade Secrets
6. Design of integrated circuits
7. Geographical Indications.

1. PATENTS:

- ✓ Patents are the legal rights approved for new inventions involving scientific and technical knowledge.
- ✓ Patent means an official document giving the holder the sole right to make use or sell an invention and preventing others from copying it.
- ✓ To be patient the invention must be useful, original, new, unusual and hardly noticeable.
- ✓ As invention may be a product, method, apparatus, design, composition of matters etc But one

cannot patient a way of doing business or anything that occurs in nature.

- ✓ The validity period of most of the patents are 20 years from the date of the filling However for the design patents such as new design for a product the patent validity is 14 years.

2. INDUSTRIAL DESIGNS:

- ✓ It is the right to safeguard one's industrial designs.
- ✓ As stated by TRIPS, a design is an idea or conception as to the features to the shape configuration pattern, ornament of composition of lines or colors applied to any article, two or more dimensional or both any industrial process or means which is the finished article appeals and is judged to the solely by the eye or product.

3. TRADEMARKS:

- ✓ Trademarks is a visual symbol in the form of words, phrases, sound or symbol associated with goods or services.
- ✓ Trademark means a registered design or name used to identify a company's goods.
- ✓ It is used to indicate the public the origin of manufacture of the goods affixed with that mark.
- ✓ Examples: Pepsi is a registered trademark in soft drinks; Thomson in electronic goods: and nestle in food products.

4. COPYRIGHTS:

- ✓ Copyright means the legal right held for a certain number of years to print publish sell broadcast perform film or record an original work or any part of it.
- ✓ The copyrights product the expression of the idea, not the idea themselves.
- ✓ The copyright expires fifty years after the death of the author.
- ✓ Example: Poems, paintings, script of movies and computer programs.

5. TRADE SECRETS:

- ✓ Trade secret means a device or techniques used by a company in manufacturing its products etc and keep secret from other companies or the general public.
- ✓ Trade secret such as formulas, patterns, methods, and data compilations are kept secret in order to gain a competitive advantage over the competitor.
- ✓ Though the trade secret cannot be registered like the other intellectual properties thefts of trade secrets are legally considered as crime.
- ✓ Examples: The formula of Fanta soft drink and the formula for making drugs.

6. DESIGN OF INTEGRATED CIRCUITS:

- ✓ It is the right granted to the inventor to prevent anybody making use of the design of integrated circuits, semiconductor devices, and other electronic devices.
- ✓ Example: Invention of a new microprocessor chip.

7. GEOGRAPHICAL INDICATIONS:

- ✓ Geographical indications identify goods as originating in the territory of a country, a origin or a locality in that territory where a specific quality reputations or other characteristics of the goods is the essentially attributed to their geographical origin.

- ✓ The legal rights secured for an intellectual property under appropriate under legislation can be enforced only within the boundaries of the country which grants such rights.
- ✓ Examples: Tirunelveli halwa, Dindugal locks, sivakasi crackers, kancheepuram sarees.

BENEFITS OF INTELLECTUAL PROPERTY RIGHTS:

The benefits of implementing the IPRS are given below:

- ✓ IPRS promote incentives for the inventions industrial and economical developments of a country.
- ✓ IPRS provide incentives for the inventions and ensure adequate returns on commercialization of the invention.
- ✓ IPRs are useful in identifying unprotected areas to avoid violation.
- ✓ IPRs provide use the invention for the public purpose.

DISCRIMINATION:

What is the discrimination?

1. Discrimination is the unequal treatment of an individual intentionally or unintentionally.
2. Discrimination refers to the treating the people unfairly because of one's sex, race, skin color, age or religious outlook.
3. Discrimination based on these aspects of the biological makeup and basic convection is disgraceful.
4. Discrimination violates the fundamental human rights of fair and equal treatment humans.
5. Discrimination defined: Discrimination is a morally unjustified treatment of people on arbitrary or irrelevant grounds.

Preferential Treatments:

1. Preferential treatment mean giving an advantage to a member of a group that is the past was denied equal treatment in particular women and minorities.

2. The preferential treatment s are also referred as reverse preferential treatments, as it reverses the historical order of preferences.

TWO KINDS OF PREFERENTIAL TREATMENT:

1. Week Preferential Treatment: In involves giving an advantage to members of traditionally discriminated-against groups over equally qualified applicants who are members of other group.

2. Strong Preferential Treatment: In involves giving preference to minority applicants women over better qualified applicants from other groups.

ARGUMENTS OVER PREFERENTIAL TREATMENT:

1. Arguments Favoring Preferential Treatment:

- ✓ A rights ethics who favor preferential treatment emphasizes on the principle of compensatory justice. According to them past violations of rights must be compensated. So preference should be given on the basis of the membership in a group that has been disadvantaged in the past.
- ✓ The utilitarianism who favors preferential treatment argues that the women and minorities should be integrated into the economic and social mainstream.

2. Arguments against Preferential Treatment:

- ✓ It can be argued that preferential treatment is a straightforward violation of others people rights to equal opportunity.
- ✓ It is also argued that there is the economic harm that results from a policy of not consistently recruiting the best practice is the best qualified person.
- ✓ The reverse discrimination is unfair in the present the similar to the unfair discrimination made against the disadvantaged groups in the past.

Thus in the present scenario it is very important to find a way to balance these for and against arguments over preferential arguments over preferential treatments in order to achieve the social integration.

SEXUAL HARASSMENT

- ✓ Sexual harassment is a particular undesirable, objectionable from the sex discrimination
- ✓ Definition of sexual harassment.

1.”sexual harassment is any sexual oriented practice that endangers a woman’s job that undermines her job performance and threatens her economic livelihood”.

2.”sexual harassment is the unwanted imposition of the sexual requirements in the context of a relationship of unequal power”.

- ✓ Sexual harassment can be sexual harness as it violates the basic human rights to pursue one’s work free from the pressures, fears, penalties, and insults.

PREPARED BY VERIFIED BY APPROVED BY

UNIT – V GLOBAL ISSUES

Syllabus:

Multinational Corporations – Environmental Ethics – Computer Ethics – Weapons Development – Engineers as Managers – Consulting Engineers – Engineers as Expert Witnesses and Advisors – Moral Leadership –Code of Conduct – Corporate Social Responsibility

MULTINATIONAL CORPORATIONS

A multinational corporation (MNC), also called a transnational corporation (TNC), or multinational enterprise (MNE), is a corporation or an enterprise that manages production or delivers services in more than one country. It can also be referred to as an international corporation. The International Labour Organization (ILO) has defined[citation needed] an MNC as a corporation that has its management headquarters in one country, known as the home country, and operates in several other countries, known as host countries.

The Dutch East India Company was the first multinational corporation in the world and the first company to issue stock. It was also arguably the world's first megacorporation, possessing quasi-governmental powers, including the ability to wage war, negotiate treaties, coin money, and establish colonies.

The first modern multinational corporation is generally thought to be the East India Company. Many corporations have offices, branches or manufacturing plants in different countries from where their original and main headquarters is located.

Some multinational corporations are very big, with budgets that exceed some nations' GDPs. Multinational corporations can have a powerful influence in local economies, and even the world economy, and play an important role in international relations and globalization

Multinational corporations have played an important role in globalization. Countries and sometimes subnational regions must compete against one another for the establishment of

MNC facilities, and the subsequent tax revenue, employment, and economic activity. To compete, countries and regional political districts sometimes offer incentives to MNCs such as tax breaks, pledges of governmental assistance or improved infrastructure, or lax environmental and labor standards enforcement. This process of becoming more attractive to foreign investment can be characterized as a race to the bottom, a push towards greater autonomy for corporate bodies, or both.

However, some scholars for instance the Columbia economist Jagdish Bhagwati, have argued that multinationals are engaged in a 'race to the top.' While multinationals certainly regard a low tax burden or low labor costs as an element of comparative advantage, there is no evidence to suggest that MNCs deliberately avail themselves of lax environmental regulation or poor labour standards. As Bhagwati has pointed out, MNC profits are tied to operational efficiency, which includes a high degree of standardisation. Thus, MNCs are likely to tailor production processes in all of their operations in conformity to those jurisdictions where they operate (which will almost always include one or more of the US, Japan or EU) that has the most rigorous standards. As for labor costs, while MNCs clearly pay workers in, e.g. Vietnam, much less than they would in the US (though it is worth noting that higher American productivity—linked to technology—means that any comparison is tricky, since in America the same company would probably hire far fewer people and automate whatever process they performed in Vietnam with manual labour), it is also the case that they tend to pay a premium of between 10% and 100% on local labor rates.[10] Finally, depending on the nature of the MNC, investment in any country reflects a desire for a long-term return. Costs associated with establishing plant, training workers, etc., can be very high; once established in a jurisdiction, therefore, many MNCs are quite vulnerable to predatory practices such as, e.g., expropriation, sudden contract renegotiation, the arbitrary withdrawal or compulsory purchase of unnecessary 'licenses,' etc. Thus, both the negotiating power of MNCs and the supposed 'race to the bottom' may be overstated, while the substantial benefits that MNCs bring (tax revenues aside) are often understated

Market withdrawal

Because of their size, multinationals can have a significant impact on government policy, primarily through the threat of market withdrawal. For example, in an effort to reduce health care costs, some countries have tried to force pharmaceutical companies to license their patented drugs to local competitors for a very low fee, thereby artificially lowering the price. When faced with that threat, multinational pharmaceutical firms have simply withdrawn from the market, which often leads to limited availability of advanced drugs. In these cases, governments have been forced to back down from their efforts. Similar corporate and government confrontations have occurred when governments tried to force MNCs to make their intellectual property public in an effort to gain technology for local entrepreneurs. When companies are faced with the option of losing a core competitive technological advantage or withdrawing from a national market, they may choose the latter. This withdrawal often causes governments to change policy. Countries that have been the most successful in this type of confrontation with multinational corporations are large countries such as United States and Brazil[citation needed], which have viable indigenous market competitors.

Lobbying

Multinational corporate lobbying is directed at a range of business concerns, from tariff structures to environmental regulations. There is no unified multinational perspective on any of these issues. Companies that have invested heavily in pollution control mechanisms may lobby for very tough environmental standards in an effort to force non-compliant competitors into a weaker position. Corporations lobby tariffs to restrict competition of foreign industries. For every tariff category that one multinational wants to have reduced, there is another multinational that wants the tariff raised. Even within the U.S. auto industry, the fraction of a company's imported components will vary, so some firms favor tighter import restrictions, while others favor looser ones. Says Ely Oliveira, Manager Director of the MCT/IR: This is very serious and is very hard and takes a lot of work for the owner.pk

Multinational corporations such as Wal-mart and McDonald's benefit from government zoning laws, to create barriers to entry.Many industries such as General Electric and Boeing lobby the government to receive subsidies to preserve their monopoly.

Patents

Any multinational corporations hold patents to prevent competitors from arising. For example, Adidas holds patents on shoe designs, Siemens A.G. holds many patents on equipment and infrastructure and Microsoft benefits from software patents. The pharmaceutical companies lobby international agreements to enforce patent laws on others.

Government power

In addition to efforts by multinational corporations to affect governments, there is much government action intended to affect corporate behavior. The threat of nationalization (forcing a company to sell its local assets to the government or to other local nationals) or changes in local business laws and regulations can limit a multinational's power. These issues become of increasing importance because of the emergence of MNCs in developing countries.

Micro-multinationals

Enabled by Internet based communication tools, a new breed of multinational companies is growing in numbers.(Copeland, Michael V. (2006-06-29). "How startups go global".
http://money.cnn.com/2006/06/28/magazines/business2/startupsgoglobal.biz2/index.htm.
Retrieved 2010-05-13.) These multinationals start operating in different countries from the very early stages. These companies are being called micro-multinationals. (Varian, Hal R. (2005-08-25). "*Technology* Levels the Business Playing Field".
The New York Times. http://www.nytimes.com/2005/08/25/business/25scene.html. Retrieved 2010-05-13.) What differentiates micro-multinationals from the large MNCs is the fact that they are small businesses. Some of these micro-multinationals, particularly software development companies, have been hiring employees in multiple countries from the beginning of the Internet era. But more and more micro-multinationals are actively starting to market their products and services in various countries. Internet tools like Google, Yahoo, MSN, Ebay and Amazon make it easier for the micro-multinationals to reach potential customers in other countries.

Service sector micro-multinationals, like Facebook, Alibaba etc. started as dispersed virtual businesses with employees, clients and resources located in various countries. Their

rapid growth is a direct result of being able to use the internet, cheaper telephony and lower traveling costs to create unique business opportunities.

Low cost SaaS (Software As A Service) suites make it easier for these companies to operate without a physical office.

Hal Varian, Chief Economist at Google and a professor of information economics at U.C. Berkeley, said in April 2010, "Immigration today, thanks to the Web, means something very different than it used to mean. There's no longer a brain drain but brain circulation. People now doing startups understand what opportunities are available to them around the world and work to harness it from a distance rather than move people from one place to another."

ENVIRONMENTAL ETHICS

Environmental ethics believes in the ethical relationship between human beings and the natural environment. Human beings are a part of the society and so are the other living beings. When we talk about the philosophical principle that guides our life, we often ignore the fact that even plants and animals are a part of our lives. They are an integral part of the environment and hence have a right to be considered a part of the human life. On these lines, it is clear that they should also be associated with our guiding principles as well as our moral and ethical values.

What is Environmental Ethics?

We are cutting down forests for making our homes. We are continuing with an excessive consumption of natural resources. Their excessive use is resulting in their depletion, risking the life of our future generations. Is this ethical? This is the issue that environmental ethics takes up. Scientists like Rachel Carson and the environmentalists who led philosophers to consider the philosophical aspect of environmental problems, pioneered in the development of environmental ethics as a branch of environmental philosophy.

The Earth Day celebration of 1970 was also one of the factors, which led to the development of environmental ethics as a separate field of study. This field received impetus when it was first discussed in the academic journals in North America and Canada. Around

the same time, this field also emerged in Australia and Norway. Today, environmental ethics is one of the major concerns of mankind.

When industrial processes lead to destruction of resources, is it not the industry's responsibility to restore the depleted resources? Moreover, can a restored environment make up for the originally natural one? Mining processes hamper the ecology of certain areas; they may result in the disruption of plant and animal life in those areas. Slash and burn techniques are used for clearing the land for agriculture.

Most of the human activities lead to environmental pollution. The overly increasing human population is increasing the human demand for resources like food and shelter. As the population is exceeding the carrying capacity of our planet, natural environments are being used for human inhabitation.

Thus human beings are disturbing the balance in the nature. The harm we, as human beings, are causing to the nature, is coming back to us by resulting in a polluted environment. The depletion of natural resources is endangering our future generations. The imbalance in nature that we have caused is going to disrupt our life as well. But environmental ethics brings about the fact that all the life forms on Earth have a right to live. By destroying the nature, we are depriving these life forms of their right to live. We are going against the true ethical and moral values by disturbing the balance in nature. We are being unethical in treating the plant and animal life forms, which coexist in society.

Human beings have certain duties towards their fellow beings. On similar lines, we have a set of duties towards our environment. Environmental ethics says that we should base our behavior on a set of ethical values that guide our approach towards the other living beings in nature.

Environmental ethics is about including the rights of non-human animals in our ethical and moral values. Even if the human race is considered the primary concern of society, animals and plants are in no way less important. They have a right to get their fair share of existence.

We, the human beings, along with the other forms of life make up our society. We all are a art of the food chain and thus closel associated with each other. We, to ether form

our environment. The conservation of natural resources is not only the need of the day but also our prime duty.

COMPUTER ETHICS

Ethics is a set of moral principles that govern the behavior of a group or individual. Therefore, computer ethics is set of moral principles that regulate the use of computers. Some common issues of computer ethics include intellectual property rights (such as copyrighted electronic content), privacy concerns, and how computers affect society. For example, while it is easy to duplicate copyrighted electronic (or digital) content, computer ethics would suggest that it is wrong to do so without the author's approval. And while it may be possible to access someone's personal information on a computer system, computer ethics would advise that such an action is unethical.

As technology advances, computers continue to have a greater impact on society. Therefore, computer ethics promotes the discussion of how much influence computers should have in areas such as artificial intelligence and human communication. As the world of computers evolves, computer ethics continues to create ethical standards that address new issues raised by new technologies.

WEAPONS DEVELOPMENT

A weapon is an instrument used for the purpose of causing harm or damage to people, animals or structures. Weapons are used in hunting, attack, self-defense, or defense in combat and range from simple implements like clubs and spears to complicated modern machines such as intercontinental ballistic missiles. One who possesses or carries a weapon is said to be armed.

In a broader context weapons include anything used to gain an advantage over an adversary or to place them at a disadvantage. Examples include the use of sieges, tactics, and psychological weapons which reduce the morale of an enemy

Classification By user

- what person or unit uses the weapon

- Personal weapons (or small arms) - designed to be used by a single person.
- Hunting weapon - primarily for hunting game animals for food or sport
- Infantry support weapons - larger than personal weapons, requiring two or more to operate correctly.
- Fortification weapons - mounted in a permanent installation, or used primarily within a fortification.
- Mountain weapons - for use by mountain forces or those operating in difficult terrain.
- Vehicle weapons - to be mounted on any type of military vehicle.
- Railway weapons - designed to be mounted on railway cars, including armored trains.
- Aircraft weapons - carried on and used by some type of aircraft, helicopter, or other aerial vehicle.
- Naval weapons - mounted on ships and submarines.
- Space weapons - are designed to be used in or launched from space.

By function

- the construction of the weapon and principle of operation

- Antimatter weapons (theoretical) would combine matter and antimatter to cause a powerful explosion.
- Archery weapons operate by using a tensioned string to launch a projectile.
- Artillery are capable of launching heavy projectiles over long distances.
- Biological weapons spread biological agents, causing disease or infection.
- Chemical weapons, poisoning and causing reactions.
- Energy weapons rely on concentrating forms of energy to attack, such as lasers or sonic attack.
- Explosive weapons use a physical explosion to create blast concussion or spread shrapnel.

- Firearms use a chemical charge to launch projectiles.
- Improvised weapons are common objects, reused as weapons.

- Incendiary weapons cause damage by fire.
- Non-lethal weapons are designed to subdue without killing.
- Magnetic weapons use magnetic fields to propel projectiles, or to focus particle

- Melee weapons operate as physical extensions of the user's body and directly impact their target.
- Missiles are rockets which are guided to their target after launch. (Also a general term for projectile weapons).
- Nuclear weapons use radioactive materials to create nuclear fission and/or nuclear fusion detonations.
- Primitive weapons make little or no use of technological or industrial elements.
- Ranged weapons (unlike Mêlée weapons), target a distant object or person.
- Rockets use chemical propellant to accelerate a projectile
- Suicide weapons exploit the willingness of their operator to not survive the attack.
- Trojan weapons appear on face value to be gifts, though the intent is to in some way to harm the recipient.

By target

- the type of target the weapon is designed to attack

- Anti-aircraft weapons target missiles and aerial vehicles in flight.
- Anti-fortification weapons are designed to target enemy installations.
- Anti-personnel weapons are designed to attack people, either individually or in numbers.
- Anti-radiation weapons target sources of electronic radiation, particularly radar emitters.
- Anti-satellite weapons target orbiting satellites.
- Anti-ship weapons target ships and vessels on water.
- Anti-submarine weapons target submarines and other underwater targets.

- Anti-tank weapons are designed to defeat armored targets.
- Area denial weapons target territory, making it unsafe or unsuitable for enemy use or travel.
- Hunting weapons are civilian weapons used to hunt animals.
- Infantry support weapons are designed to attack various threats to infantry units

CONSULTING ENGINEERS

Consultants are individuals who typically work for themselves but may also be associated with a consulting firm. They, for a fee, gives advice or provides a service in a field of specialized knowledge or training. Most consultants carry their own life and health insurance, pay their own taxes, most have their own tools and equipment. The consultant can work alone or with the client's staff.

Consultants can play a multi-faceted role. They can, for example function as advisors, fixers, bosses, generalists, stabilizers, listeners, advisors, specialists, catalysts, managers or quasi-employees. The actual work that consultants perform for one company to another may vary greatly, i.e. tax account to office decoration. However, the typical underlying reasons that a consultant is hired are universal. A problem exists and the owner or manager of the company has decided to seek the help of an expert.

Bringing in an expert can save time, effort and money. It has been estimated that approximately 3/4 of all companies call upon consultants at one time or another. Many companies claim that they receive a higher return for their invested dollars by using consultants for specific tasks.

Most companies have experienced the problem of needing short-term technical expertise. Perhaps the company's existing staff is already working to capacity. In many cases, the engineering skills required for a project can be satisfied with a full time employee. When they can not fully justify bringing someone on board full time, their answer is to hire a consultant. By doing so, the businessman solves his immediate problem without permanently increasing his payroll and payroll taxes.

Consultants can be hired when the company may not have anyone on staff capable of solving the specific problem. At such times, a costly learning curve on the part of the engineering staff is associated with the project. One example is using a consultant as a viable alternative during the development stages of new products. Hiring a consultant with experience in a given area can then cut days, weeks or even months off a project schedule. In addition, he can help the staff avoid mistakes they may otherwise make. When the project reaches a certain point, the permanent staff can then take over.

Consultants can deal directly with owners and upper management. In this role, consultants can provide an objective third-party view point. Critical objectives can then be identified and advise given in confidence.

Consultants are a viable alternative in assisting in feasibility studies or in proposal preparation.

Perhaps the manager cannot justify shifting the duties of existing staff members.

Another time that consultants become useful is when a company is just starting a business. The development of the company's new product can be begun by the consultant while a full time permanent technical staff member is being hired.

Finding the right consultant can be difficult. Managers can rely on referrals from their friends or hire the consultant who happens to call at the right time. Once the decisions is made to hire a consultant, the need is immediate and one may not have the time to shop for a consultant. As a part of planning ahead, it is wise to meet various consultants on an informal basis before the need to hire one arises. Then when the time comes, you will know exactly who to call for you have already established an informal relationship

ETHICS IN ASCE

To preserve the high ethical standards of the civil engineering profession, the Society's ethics program includes:

- Edict

  The Society maintains a Code of Ethics.

- Enforcement

  The Society enforces the Code by investigating potential violations of the Code and taking disciplinary action if warranted.

- Education

  The Society endeavors to educate its members and the public on ethics issues.

IEEE code of Ethics

1. to accept responsibility in making decisions consistent with the safety, health and welfare of the public, and to disclose promptly factors that might endanger the public or the environment;
2. to avoid real or perceived conflicts of interest whenever possible, and to disclose them to affected parties when they do exist;
3. to be honest and realistic in stating claims or estimates based on available data;
4. to reject bribery in all its forms;
5. to improve the understanding of technology, its appropriate application, and potential consequences;
6. to maintain and improve our technical competence and to undertake technological tasks for others only if qualified by training or experience, or after full disclosure of pertinent limitations;
7. to seek, accept, and offer honest criticism of technical work, to acknowledge and correct errors, and to credit properly the contributions of others;

8. to treat fairly all persons regardless of such factors as race, religion, gender, disability, age, or national origin;

9. to avoid injuring others, their property, reputation, or employment by false or malicious action;

10. to assist colleagues and co-workers in their professional development and to support them in following this code of ethics

Ethics in Indian Institute of Materials and Management

- To consider first, the TOTAL interest to one's organization in all transactions without impairing the dignity and responsibility to one's office;

- To buy without prejudice, seeking to obtain the maximum ultimate value for each Rupee of Expenditure;

- To subscribe and work for honesty and truth in buying and selling, to denounce all forms and manifestations of commercial bribery and to eschew anti-social practices;

- To accord a prompt and courteous reception so far as conditions will permit, to all who call upon a legitimate business mission;

To respect one's obligations and those of one's organization, consistent with good business practice

Ethics in Institute of Engineers

1.1 Engineers serve all members of the community in enhancing their welfare, health and safety by a creative process utilising the engineers' knowledge, expertise and experience.

1.2 Pursuant to the avowed objectives of The Institution of Engineers (India) as enshrined in the presents of the Royal Charter granted to the Institution, the Council of the Institution prescribed a set of "Professional Conduct Rules" in the year 1944 replacing the same with the "Code of Ethics for Corporate Members" in the year 1954 which was revised in the year 1997.

1.3 In view of globalisation, concern for the environment and the concept of sustainable development, it has been felt that the prevailing "Code of Ethics for Corporate Members" needs review and revision in letter and spirit. The engineering organisations world over have updated their Code of Ethics.

1.4 The Council of the Institution vested with the authority in terms of the Present 2(j) of the Royal Charter adopted at its 626th meeting held on 21.12.2003 at Lucknow the "Code of Ethics for Corporate Members" as provided hereinafter.

1.5 The Code of Ethics is based on broad principles of truth, honesty, justice, trustworthiness, respect and safeguard of human life and welfare, competence and accountability which constitute the moral values every Corporate Member of the Institution must recognize, uphold and abide by.

1.6 This "Code of Ethics for Corporate Members" shall be in force till the same is revised by a decision of the Council of the Institution.

CODE OF ETHICS FOR Institute of Engineers

1.0 Preamble

1.1 The Corporate Members of The Institution of Engineers (India) are committed to promote and practice the profession of engineering for the common good of the community bearing in mind the following concerns :

1.1.1 Concern for ethical standard;

1.1.2 Concern for social justice, social order and human rights;

1.1.3 Concern for protection of the environment;

1.1.4 Concern for sustainable development;

1.1.5 Public safety and tranquility.

2.0 The Tenets of the Code of Ethics

2.1 A Corporate Member shall utilise his knowledge and expertise for the welfare, health and safety of the community without any discrimination for sectional or private interests.

2.2 A Corporate Member shall maintain the honour, integrity and dignity in all his professional actions to be worthy of the trust of the community and the profession.

2.3 A Corporate Member shall act only in the domains of his competence and with diligence, care, sincerity and honesty.

2.4 A Corporate Member shall apply his knowledge and expertise in the interest of his employer or the clients for whom he shall work without compromising with other obligations to these Tenets.

2.5 A Corporate Member shall not falsify or misrepresent his own or his associates' qualifications, experience, etc.

2.6 A Corporate Member, wherever necessary and relevant, shall take all reasonable steps to inform himself, his employer or clients, of the environmental, economic, social and other possible consequences, which may arise out of his actions.

2.7 A Corporate Member shall maintain utmost honesty and fairness in making statements or giving witness and shall do so on the basis of adequate knowledge.

2.8 A Corporate Member shall not directly or indirectly injure the professional reputation of another member.

2.9 A Corporate Member shall reject any kind of offer that may involve unfair practice or may cause avoidable damage to the ecosystem.

2.10 A Corporate Member shall be concerned about and shall act in the best of his abilities

for maintenance of sustainability of the process of development.

2.11 A Corporate Member shall not act in any manner which may injure the reputation of the Institution or which may cause any damage to the Institution financially or otherwise.

3.0 General Guidance

The Tenets of the Code of Ethics are based on the recognition that –

3.1 A common tie exists among the humanity and that The Institution of Engineers (India) derives its value from the people, so that the actions of its Corporate Members should indicate the member's highest regard for equality of opportunity, social justice and fairness;

3.2 The Corporate Members of the Institution hold a privileged position in the community so as to make it a necessity for their not using the position for personal and sectional interests.

4.0 And, as such, a Corporate Member –

4.1 should keep his employer or client fully informed on all matters in respect of his assignment which are likely to lead to a conflict of interest or when, in his judgement, a project will not be viable on the basis of commercial, technical, environmental or any other risks;

4.2 should maintain confidentiality of any information with utmost sincerity unless expressly permitted to disclose such information or unless such permission, if withheld, may adversely affect the welfare, health and safety of the community;

4.3 should neither solicit nor accept financial or other considerations from anyone related to a project or assignment of which he is in the charge;

4.4 should neither pay nor offer direct or indirect inducements to secure work;

4.5 should compete on the basis of merit alone;

4.6 should refrain from inducing a client to breach a contract entered into with another duly appointed engineer;

4.7 should, if asked by the employer or a client, to review the work of another person or organisation, discuss the review with the other person or organisation to arrive at a balanced opinion;

4.8 should make statements or give evidence before a tribunal or a court of law in an objective and accurate manner and express any opinion on the basis of adequate knowledge and competence; and

4.9 should reveal the existence of any interest – pecuniary or otherwise – which may affect the judgement while giving an evidence or making a statement.

5.0 Any decision of the Council as per provisions of the relevant Bye-Laws of the Institution shall be final and binding on all Corporate Members

ASME Code of Ethics of Engineers

ASME requires ethical practice by each of its members and has adopted the following Code of Ethics of Engineers as referenced in the ASME Constitution, Article C2.1.1.

CODE OF ETHICS OF ENGINEERS

The Fundamental Principles

Engineers uphold and advance the integrity, honor and dignity of the engineering profession by:

I. Using their knowledge and skill for the enhancement of human welfare;

Being honest and impartial, and serving with fidelity the public, their employers and clients; and

III. Striving to increase the competence and prestige of the engineering profession.

The Fundamental Canons

1. Engineers shall hold paramount the safety, health and welfare of the public in the performance of their professional duties.

2. Engineers shall perform services only in the areas of their competence.

3. Engineers shall continue their professional development throughout their careers and shall provide opportunities for the professional and ethical development of those engineers under their supervision.

4. Engineers shall act in professional matters for each employer or client as faithful agents or trustees, and shall avoid conflicts of interest or the appearance of conflicts of interest.

5. Engineers shall build their professional reputation on the merit of their services and shall not compete unfairly with others.

6. Engineers shall associate only with reputable persons or organizations.

7. Engineers shall issue public statements only in an objective and truthful manner.

8. Engineers shall consider environmental impact in the performance of their professional duties.

The ASME criteria for interpretation of the Canons are guidelines and represent the objectives toward which members of the engineering profession should strive. They are principles which an engineer can reference in specific situations. In addition, they provide interpretive guidance to the ASME Board on Professional Practice and Ethics on the Code of Ethics of Engineers.

1. Engineers shall hold paramount the safety, health and welfare of the public in the performance of their professional duties.

a. Engineers shall recognize that the lives, safety, health and welfare of the c are dependent upon engineering judgments, decisions and practices incorporated into structures, machines, products, processes and devices.

b. Engineers shall not approve or seal plans and/or specifications that are not of a design safe to the public health and welfare and in conformity with accepted engineering standards.

c. Whenever the Engineers' professional judgments are over ruled under circumstances where the safety, health, and welfare of the public are endangered, the Engineers shall inform their clients and/or employers of the possible consequences.

(1) Engineers shall endeavor to provide data such as published standards, test codes, and quality control procedures that will enable the users to understand safe use during life expectancy associated with the designs, products, or systems for which they are responsible.

(2) Engineers shall conduct reviews of the safety and reliability of the designs, products, or systems for which they are responsible before giving their approval to the plans for the design.

.Whenever Engineers observe conditions, directly related to their employment, which they believe will endanger public safety or health, they shall inform the proper authority of the situation.

d. If engineers have knowledge of or reason to believe that another person or firm may be in violation of any of the provisions of these Canons, they shall present such information to the proper authority in writing and shall cooperate with the proper authority in furnishing such further information or assistance as may be required.

2. Engineers shall perform services only in areas of their competence.

a. Engineers shall undertake to perform engineering assignments only when qualified by education and/or experience in the specific technical field of engineering involved.

b. Engineers may accept an assignment requiring education and/or experience outside of their own fields of competence, but their services shall be restricted to other phases of the project in which they are qualified. All other phases of such project shall be performed by qualified associates, consultants, or employees.

3. Engineers shall continue their professional development throughout their careers, and should provide opportunities for the professional and ethical development of those engineers under their supervision.

4. Engineers shall act in professional matters for each employer or client as faithful agents or trustees,and shall avoid conflict of interest or the appearance of conflicts of interest

a. Engineers shall avoid all known conflicts of interest with their employers or clients and shall promptly inform their employers or clients of any business association, interests, or circumstances which could influence their judgment or the quality of their services.

b. Engineers shall not undertake any assignments which would knowingly create a potential conflict of interest between themselves and their clients or their employers.

c. Engineers shall not accept compensation, financial or otherwise, from more than one party for services on the same project, or for services pertaining to the same project, unless the circumstances are fully disclosed to, and agreed to, by all interested parties.

d. Engineers shall not solicit or accept financial or other valuable considerations, for

specifying products or material or equipment suppliers, without disclosure to their clients or employers.

e. Engineers shall not solicit or accept gratuities, directly or indirectly, from contractors, their agents, or other parties dealing with their clients or employers in connection with work for which they are responsible. Where official public policy or employers' policies tolerate acceptance of modest gratuities or gifts, engineers shall avoid a conflict of interest by

complying with appropriate policies and shall avoid the appearance of a conflict of interest.

f. When in public service as members, advisors, or employees of a governmental body or department, Engineers shall not participate in considerations or actions with respect to services provided by them or their organization(s) in private or product engineering practice.

g. Engineers shall not solicit an engineering contract from a governmental body or other entity on which a principal, officer, or employee of their organization serves as a member without disclosing their relationship and removing themselves from any activity of the body which concerns their organization.

h. Engineers working on codes, standards or governmental sanctioned rules and specifications shall exercise careful judgment in their determinations to ensure a balanced viewpoint, and avoid a conflict of interest.

i. When, as a result of their studies, Engineers believe a project(s) will not be successful, they

shall so advise their employer or client.

j. Engineers shall treat information coming to them in the course of their assignments as confidential, and shall not use such information as a means of making personal profit if such action is adverse to the interests of their clients, their employers or the public.

(1) They will not disclose confidential information concerning the business affairs or technical processes of any present or former employer or client or bidder under evaluation, without his consent, unless required by law or court order.

(2) They shall not reveal confidential information or finding of any commission or board of which they are members unless required by law or court order

Designs supplied to Engineers by clients shall not be duplicated by the Engineers for others

k. Engineers shall act with fairness and justice to all parties when administering a construction (or other) contract.

l. Before undertaking work for others in which Engineers may make improvements, plans, designs, inventions, or other records which may justify seeking copyrights, patents, or proprietary

rights, Engineers shall enter into positive agreements regarding the rights of respective parties.

m. Engineers shall admit their own errors when proven wrong and refrain from distorting or altering the facts to justify their mistakes or decisions.

n. Engineers shall not accept professional employment or assignments outside of their regular work without the knowledge of their employers.

o. Engineers shall not attempt to attract an employee from other employers or from the market place by false or misleading representations.

5. Engineers shall build their professional reputation on the merit of their services and shall not compete unfairly with others.

a. Engineers shall negotiate contracts for professional services on the basis of demonstrated

competence and qualifications for the type of professional service required.

b. Engineers shall not request, propose, or accept professional commissions on a contingent basis if, under the circumstances, their professional judgments may be compromised.

c. Engineers shall not falsify or permit misrepresentation of their, or their associates, academic or professional qualification. They shall not misrepresent or exaggerate their degrees of responsibility in or for the subject matter of prior assignments. Brochures or other presentations used to solicit personal employment shall not misrepresent pertinent facts concerning employers, employees, associates, joint venturers, or their accomplishments.

d. Engineers shall prepare articles for the lay or technical press which are only factual. Technical Communications for publication (theses, articles, papers, reports,etc.) which are based on research involving more than one individual (including students and supervising faculty, industrial supervisor/researcher or other co-workers) must recognize all significant contributors. Plagiarism, the act of substantially using another's ideas or written materials without due credit, is unethical. (See Appendix.)

e. Engineers shall not maliciously or falsely, directly or indirectly, injure the professional reputation, prospects, practice or employment of another engineer, nor shall they indiscriminately criticize another's work.

f. Engineers shall not use equipment, supplies, laboratory or office facilities of their employers to carry on outside private practice without consent.

6. Engineers shall associate only with reputable persons or organizations.

a. Engineers shall not knowingly associate with or permit the use of their names or firm names in business ventures by any person or firm which they know, or have reason to believe, are engaging in business or professional practices of a fraudulent or dishonest nature.

b. Engineers shall not use association with non-engineers, corporations, or partnerships to disguise unethical acts.

7. Engineers shall issue public statements only in an objective and truthful manner.

a. Engineers shall endeavor to extend public knowledge, and to prevent misunderstandings of the achievements of engineering.

b. Engineers shall be completely objective and truthful in all professional reports, statements or testimony. They shall include all relevant and pertinent information in such reports, statements or testimony.

c. Engineers, when serving as expert or technical witnesses before any court, commission, or other tribunal, shall express an engineering opinion only when it is founded on their adequate knowledge of the facts in issue, their background of technical competence in the subject matter, and their belief in the accuracy and propriety of their testimony.

d. Engineers shall issue no statements, criticisms, or arguments on engineering matters which are inspired or paid for by an interested party, or parties, unless they preface their comments by identifying themselves, by disclosing the identities of the party or parties on whose behalf they are speaking, and by revealing the existence of any financial interest they may have in matters under discussion.

e. Engineers shall be truthful in explaining their work and merit, and shall avoid any act tending to promote their own interest at the expense of the integrity and honor of the profession or another individual.

8. Engineers shall consider environmental impact in the performance of their professional duties.

a. Engineers shall concern themselves with the impact of their plans and designs on the environment. When the impact is a clear threat to health or safety of the public, then the guidelines for this Canon revert to those of Canon 1.

9. Engineers accepting membership in The American Society of Mechanical Engineers by this action agree to abide by this Society Policy on Ethics and procedures for its implementation.

Moral Leadership

Moral Leadership brings together in one comprehensive volume essays from leading scholars in law, leadership, psychology, political science, and ethics to provide practical, theoretical policy guidance. The authors explore key questions about moral leadership such as:

- How do leaders form, sustain, and transmit moral commitments?
- Under what conditions are those processes most effective?
- What is the impact of ethics officers, codes, training programs, and similar initiatives?
- How do standards and practices vary across context and culture?
- What can we do at the individual, organizational, and societal level to foster moral leadership?

ENGINEERS AS EXPERT WITNESS AND ADVISORS

Engineering expert witnesses are highly credentialed mechanical, safety & civil, geotechnical, chemical and electrical engineers specializing in the areas of design, construction & structural engineering, failure analysis, human factors, occupational safety, metallurgy and more. They provide litigation support through review and evaluation of distressed structures for land slide and erosion cases; performance of forensic studies on hydraulics, power plants, pipelines, boiler systems, traffic, automotive, electrical fire involving electrical systems of machinery; site research and inspection, laboratory testings,

report writing, depositions and court testimony.

Engineers shall endeavor to extend public knowledge, and to prevent misunderstandings of the achievements of engineering.

b. Engineers shall be completely objective and truthful in all professional reports, statements or testimony. They shall include all relevant and pertinent information in such reports, statements or testimony.

c. Engineers, when serving as expert or technical witnesses before any court, commission, or other tribunal, shall express an engineering opinion only when it is founded on their adequate knowledge of the facts in issue, their background of technical competence in the subject matter, and their belief in the accuracy and propriety of their testimony.